Machine Learning Basics

Data Intelligence and Learning (DIAL) Lab
Prof. Jongwuk Lee

Supervised Learning Process

Key Ingredients in ML Models

- **Training data** $\mathcal{D} = \{(x^{(i)}, y^{(i)}) : 1 \le i \le n\}$
- **Machine learning model** $f(x; \mathbf{w})$
- **Error function (or loss function)** $E(\mathbf{w})$
- **An optimization algorithm**
- **Test data**

Goal of Supervised Learning

$$y = f(x)$$

- **Training**: Given a training dataset, estimate the function $f(x)$ that minimizes the error on the training dataset.

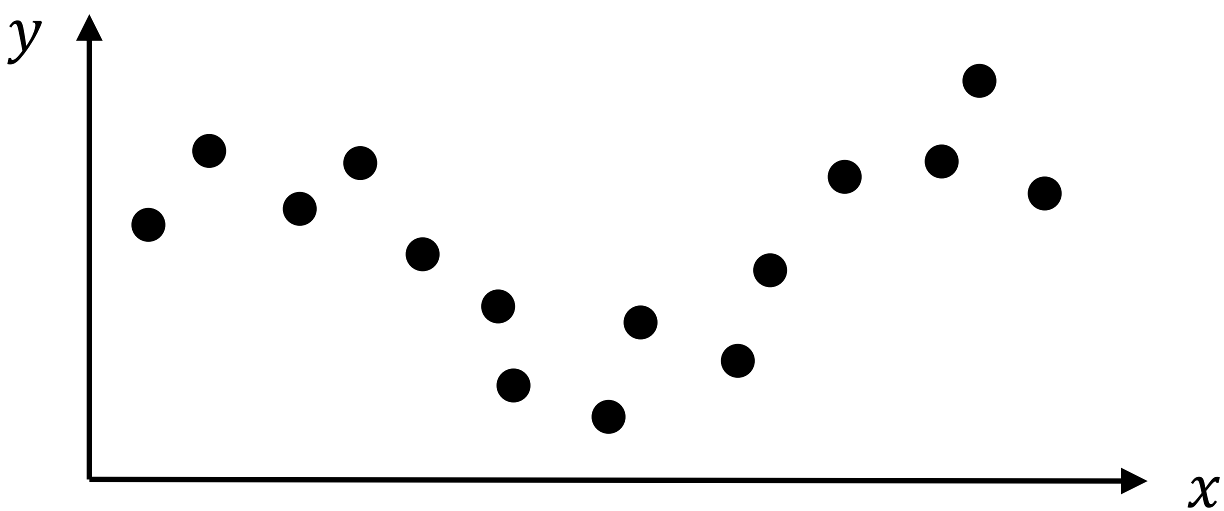
- **Testing**: Apply $f(x)$ to an unseen test example x_{new} and return the predicted value $\hat{y} = f(x_{new})$.

Goal of Supervised Learning

➢ **Given training data $\mathcal{D} = \{(x^{(i)}, y^{(i)}) : 1 \leq i \leq n\}$, we want to find the best function that describes the relationship between x and y.**

- Let $\mathbf{w}$ denote **a parameter vector** for the **function** (or **model**) f.
- Let $\boldsymbol{f}(x; \mathbf{w})$ denote explicit representation with parameters.

$$y \approx f(x; \mathbf{w})$$



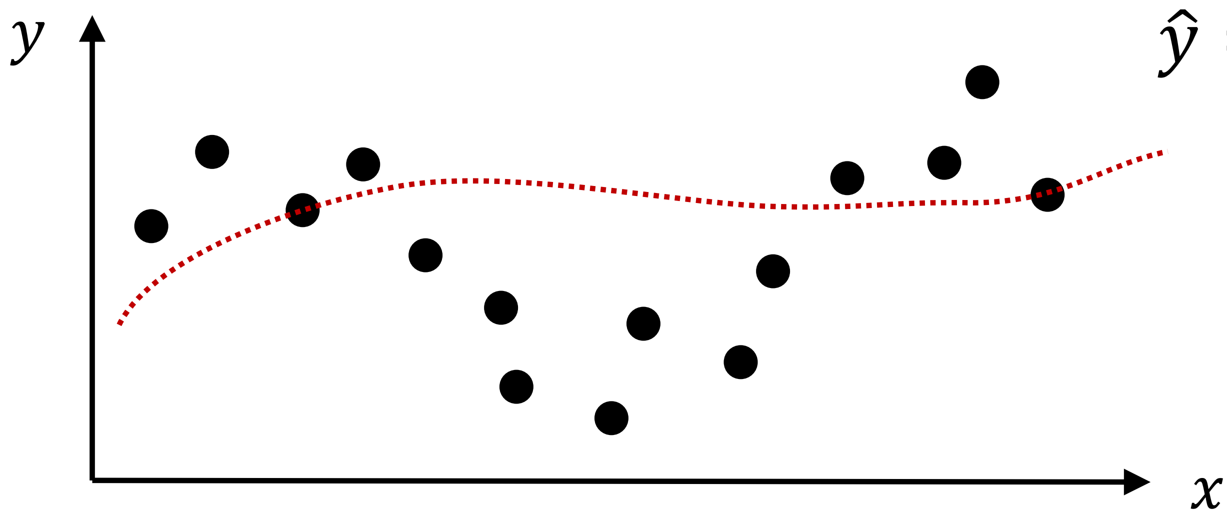
Training: Optimizing Parameters

- **Finding the parameter that best describes a given data**

$$f(x; 1.0,1.0,1.0)$$

Given x, let $\hat{y}$ denote a predicted value for x by $f(x;\mathbf{w})$.

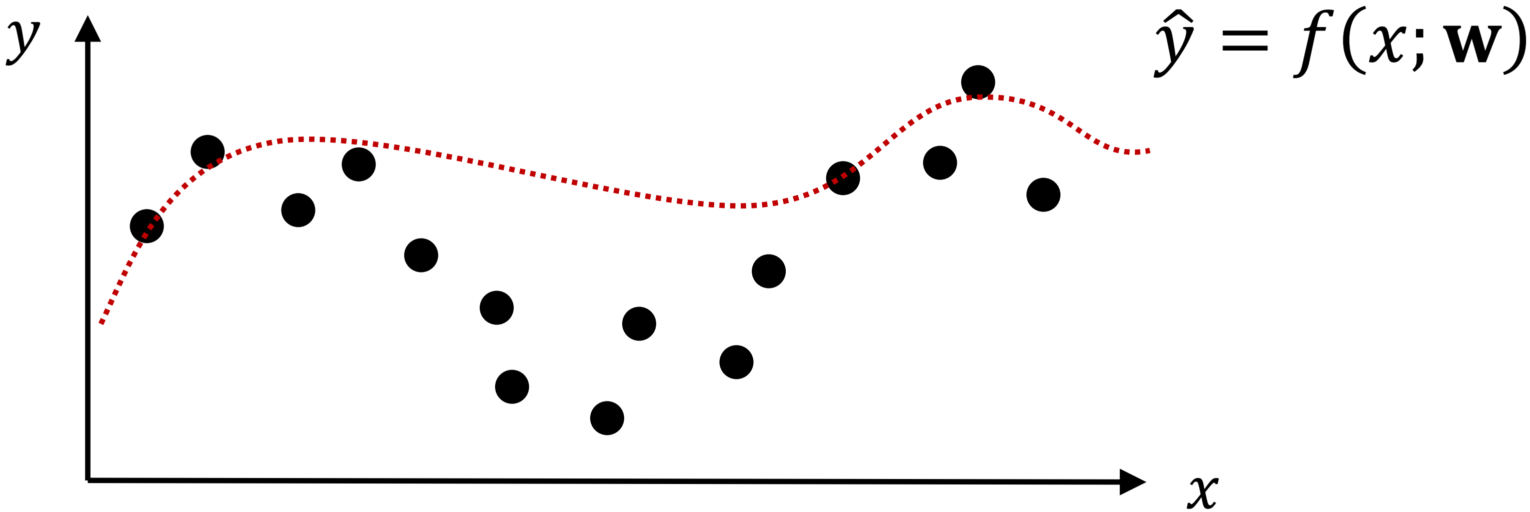
$$\hat{y} = f(x;\mathbf{w})$$



Training: Optimizing Parameters

- Finding the parameter that best describes a given data

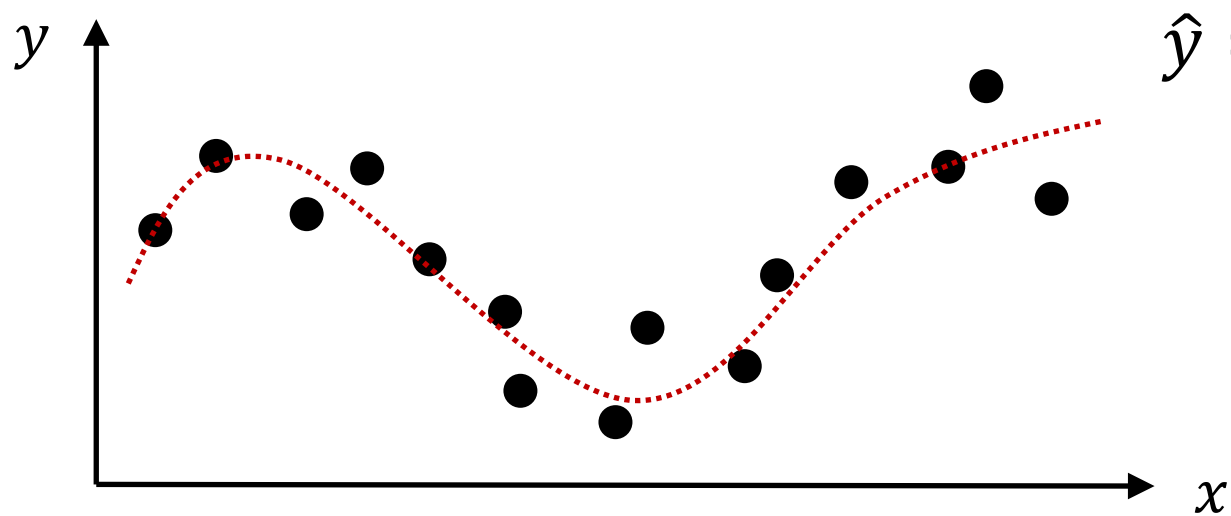
$$f(x; 1.5, 1.0, 1.5)$$



Training: Optimizing Parameters

- **Finding the parameter that best describes a given data**

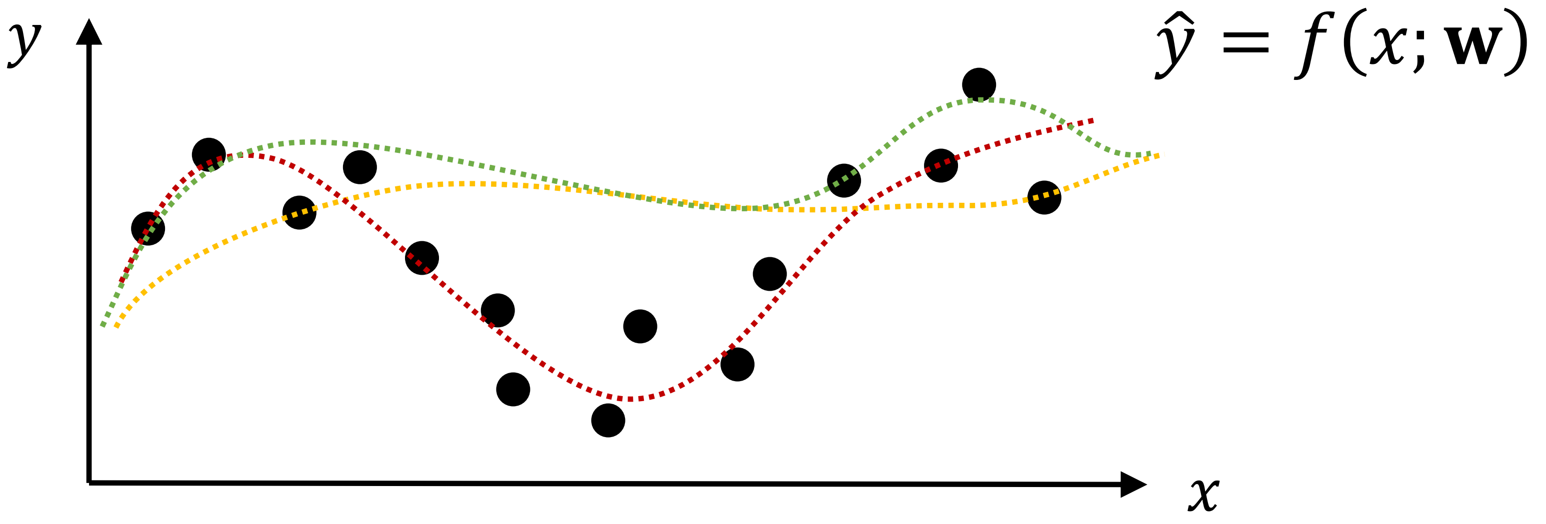
$$f(x; 1.7, -1.2, 1.5)$$



$$\hat{y} = f(x; \mathbf{w})$$

Training: Optimizing Parameters

- How to search an optimal parameter for a given data?
- It is converted to the **optimization problem**.

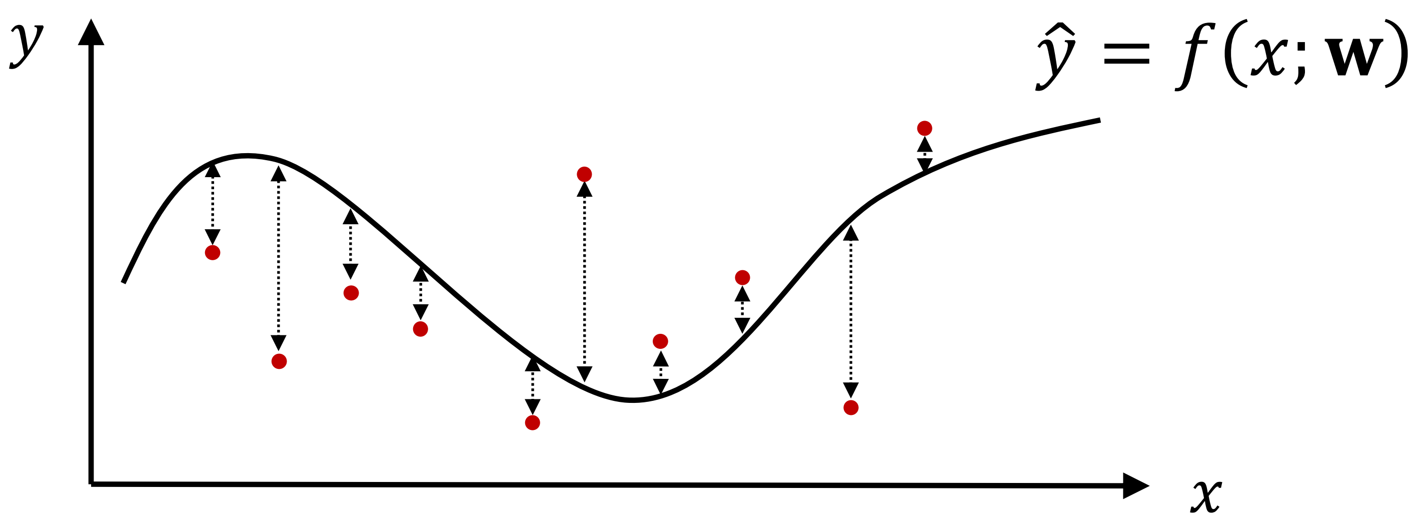


Error Function (or Loss Function)

- To find an **optimal parameter**, we **minimize the error function** between $f(x;\mathbf{w})$ and y.

$$Error(\mathbf{w}) = \frac{1}{n} \sum_{(x,y)\,\in\,\mathcal{D}} \left(y - f(x;\mathbf{w})\right)^2$$

Let $\mathcal{D}$ be a training dataset.



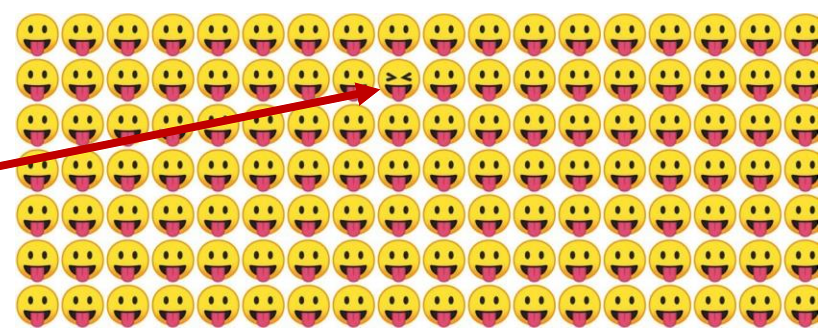
How to Find an Optimal Parameter?

- The error depends on $\mathbf{w}$.

  → **Let's find $\mathbf{w}$ which minimizes the error function.**

$$Error(\mathbf{w}) = \frac{1}{n} \sum_{(x,y) \in \mathcal{D}} \left(y - f(x; \mathbf{w})\right)^2$$

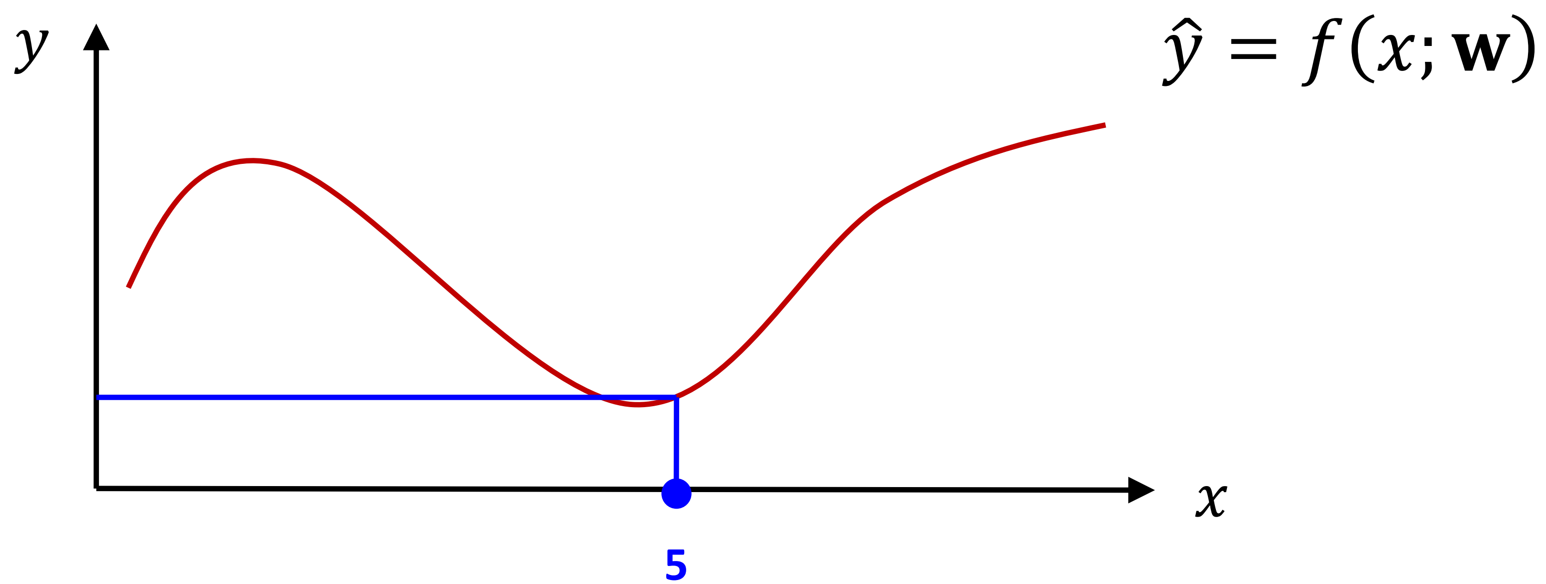
Optimal parameter



Testing: Predicting y from new x

- **Predicting y from x using the trained model**

$$3 = f(5; 1.7, -1.2, 1.5)$$



Simple Optimization Examples

Solving the Optimization Problem

- How to solve the optimization problem with d parameters?

Finding $w_1, w_2, \ldots, w_d$ that minimize the error function:

$$E(w_1, w_2, \ldots, w_d) = \frac{1}{n} \sum_{(\mathbf{x},y) \in \mathcal{D}} \left(y - f(\mathbf{x}; w_1, w_2, \ldots, w_d)\right)^2$$

- Let's first solve an easy problem.

Finding w that minimizes the following error function:

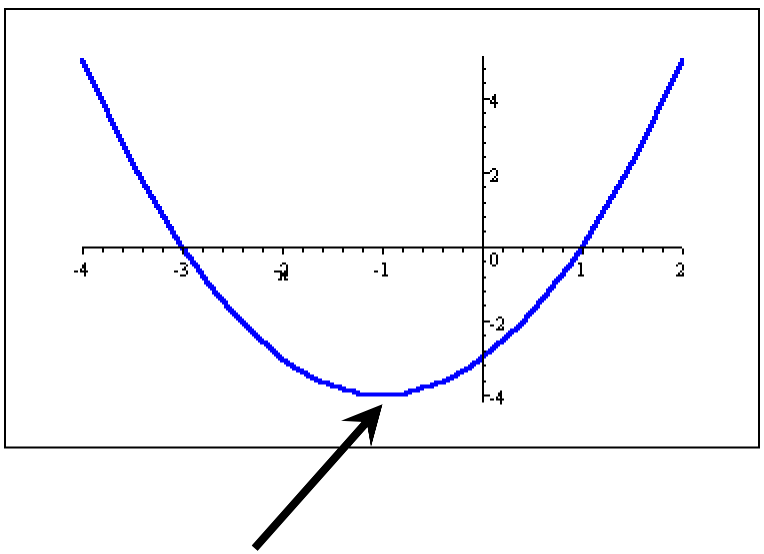
$$E(w) = w^2 + 2w + 3$$

Optimization with One Variable

- **Find w that minimizes the following function.**

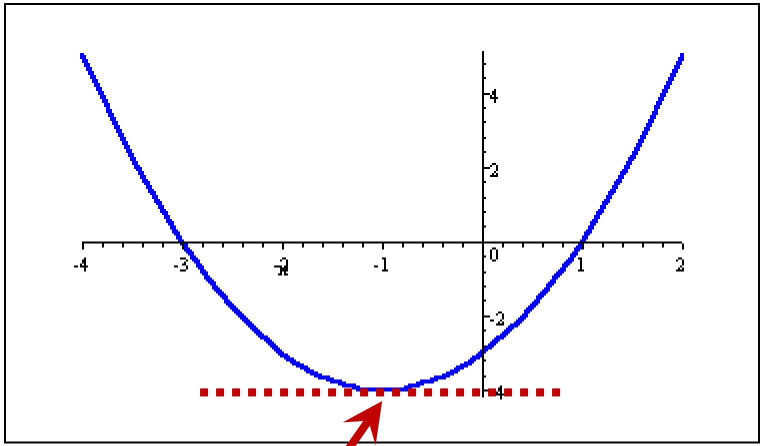
$$f(w) = w^2 + 2w - 3$$

- **How??**



Optimization with One Variable

- **We can find the minimum value.**
  - The minimum value occurs where **the slope of the curve is 0**.
  - **The first derivative of the function = slope of the curve**
  - Set the first derivative to 0 and solve it for x.

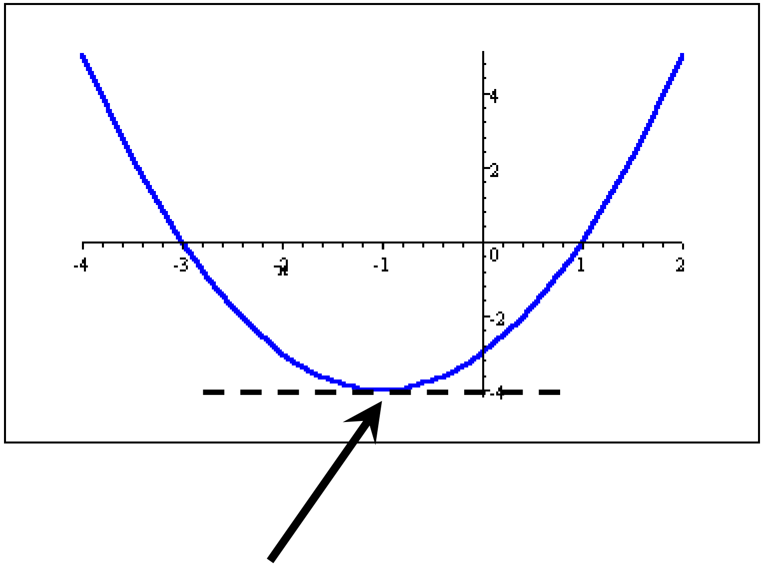


$$\frac{df(w)}{dw} = 0$$

Optimization with One Variable

$$
\begin{aligned}
f(w) &= w^2 + 2w - 3 \\
df(w)/dw &= 2w + 2 \\
2w + 2 &= 0 \\
w &= -1
\end{aligned}
$$

is value of w where $f(w)$ is minimum.

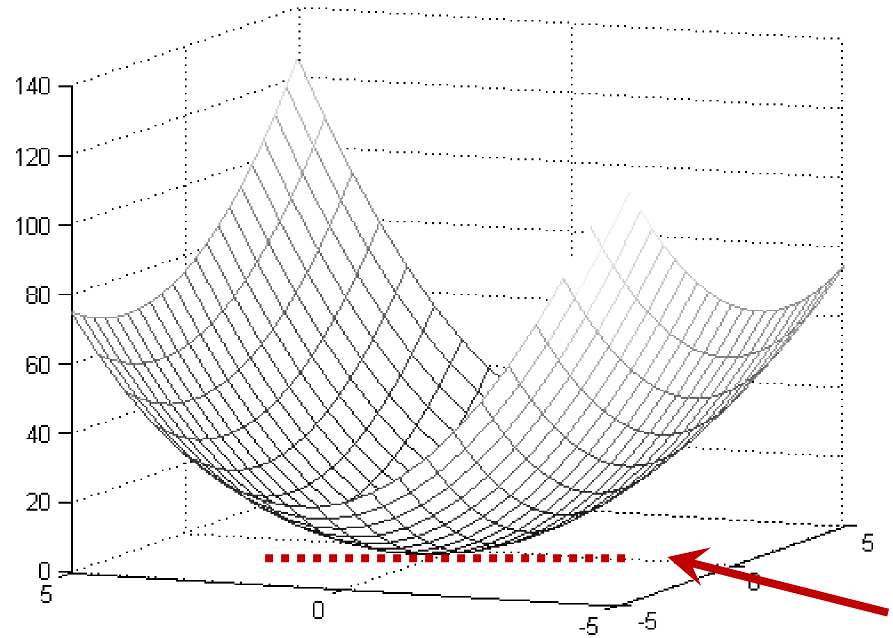


Optimization with Two Variables

- **Quadratic function with two variables**

$$f(w_1, w_2) = w_1^2 + w_1 w_2 - 2w_1 - w_2^2$$

- **$f(\mathbf{w})$ is the minimum, where the derivative of $f(\mathbf{w})$ is zero in all directions.**



$$\frac{df(\mathbf{w})}{d\mathbf{w}} = 0$$

Optimization with Two Variables

➢ **It is still simple enough that we can find the minimum directly.**

$$f(w_1, w_2) = w_1^2 + w_1 w_2 - 2w_1 - w_2^2$$

$$\nabla f(w_1, w_2) = [2w_1 + w_2 - 2, \quad w_1 - 2w_2]$$

- Set both elements of the derivative to 0.
- Give two linear equations in two variables.
- Solve for $w_1, w_{2.}$

$$2w_1 + w_2 = 2\,, \;\; w_1 - 2w_2 = 0$$

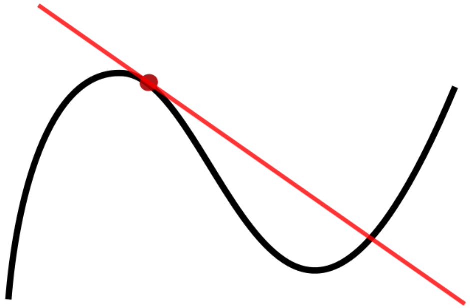
$$\Rightarrow \qquad w_1 = 4/5, w_2 = 2/5$$

Math: Derivative Calculation

Review: Scalar Derivative

- **Derivative means the slope of the tangent line.**

y	a	x^n	e^x	$\log x$
$\frac{dy}{dx}$	0	nx^{n-1}	e^x	$\frac{1}{x}$



Note: a is not a function of x.

y	$u+v$	uv	$y=f(u)$, $u=g(x)$
$\frac{dy}{dx}$	$\frac{du}{dx}+\frac{dv}{dx}$	$\frac{du}{dx}v+\frac{dv}{dx}u$	$\frac{dy}{du}\frac{du}{dx}$

It is called a chain rule.

Gradients

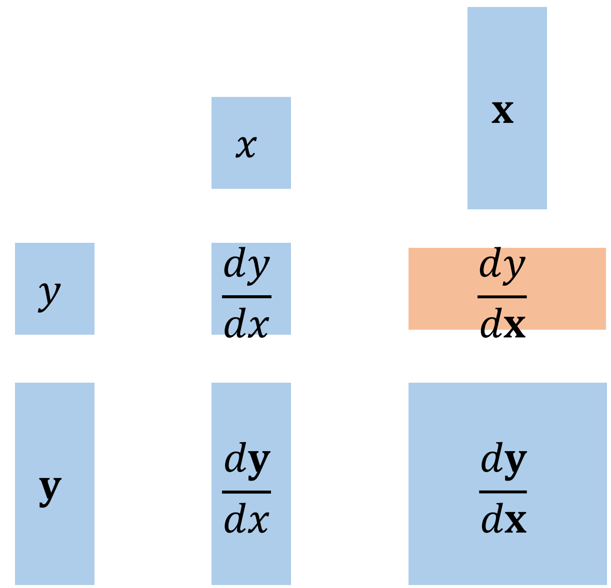
- **Generalize derivatives into vectors.**

		Scalar	Vector
		x	$\mathbf{x}$ $(n, 1)$
Scalar	y	$\frac{dy}{dx}$	$\frac{dy}{d\mathbf{x}}$ $(1, n)$
Vector	$\mathbf{y}$ $(m, 1)$	$\frac{d\mathbf{y}}{dx}$ $(m, 1)$	$\frac{d\mathbf{y}}{d\mathbf{x}}$ (m, n)

Bold lower-case font indicates a vector.

Derivative $\partial y / \partial \mathbf{x}$

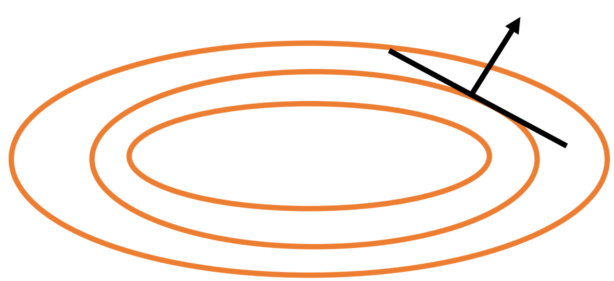
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \frac{\partial y}{\partial \mathbf{x}} = \left[\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \dots, \frac{\partial y}{\partial x_n}\right]$$



- **Example:** $\boldsymbol{f(x_1, x_2) = x_1^2 + 2x_2^2}$

$$\frac{\partial f}{\partial \mathbf{x}} = [2x_1, 4x_2]$$

Given (1, 1), the direction (2, 4) is the gradient, perpendicular to the contour line.



Example: Derivative $\partial y / \partial \mathbf{x}$

y	a	au	$\|\mathbf{x}\|^2$
$\dfrac{dy}{d\mathbf{x}}$	$\mathbf{0}^T$	$a\dfrac{du}{d\mathbf{x}}$	$2\mathbf{x}^{\mathrm{T}}$

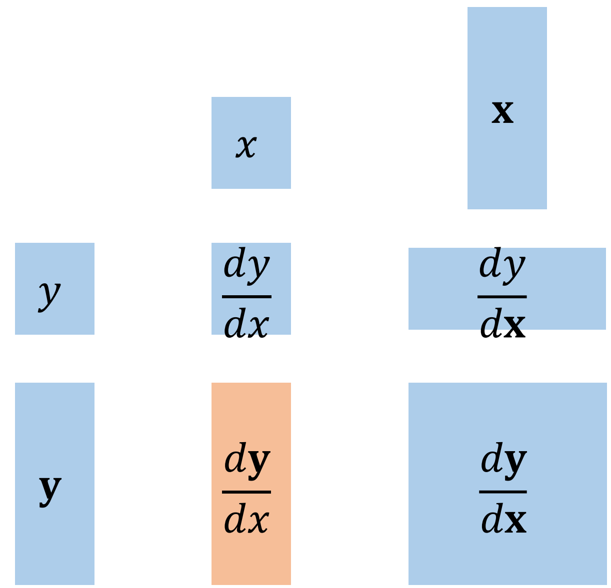
a is not a function of $\mathbf{x}$

$\mathbf{0}$ is zero vector.

y	$u+v$	uv	$\langle \mathbf{u}, \mathbf{v} \rangle$
$\dfrac{\partial y}{\partial \mathbf{x}}$	$\dfrac{\partial u}{\partial \mathbf{x}} + \dfrac{\partial v}{\partial \mathbf{x}}$	$\dfrac{\partial u}{\partial \mathbf{x}} v + \dfrac{\partial v}{\partial \mathbf{x}} u$	$\mathbf{u}^{\mathrm{T}} \dfrac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^{\mathrm{T}} \dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}$

Derivative $\partial \mathbf{y}/\partial x$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \qquad \frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$



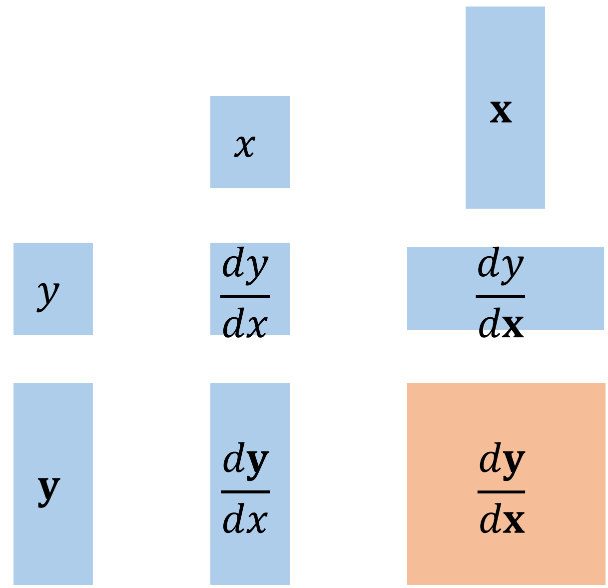
- **While $\partial y/\partial \mathbf{x}$ is a row vector, $\partial \mathbf{y}/\partial x$ is a column vector.**
  - It is called numerator-layout notation.

Derivative $\partial \mathbf{y} / \partial \mathbf{x}$

- **It is the generalized derivative representation.**

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \mathbf{x}} \\ \frac{\partial y_2}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial y_m}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1}, \frac{\partial y_1}{\partial x_2}, \dots, \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1}, \frac{\partial y_2}{\partial x_2}, \dots, \frac{\partial y_2}{\partial x_n} \\ \vdots \\ \frac{\partial y_m}{\partial x_1}, \frac{\partial y_m}{\partial x_2}, \dots, \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$



Example: Derivative $\partial \mathbf{y} / \partial \mathbf{x}$

Bold uppercase font indicates a matrix.

$\mathbf{y}$	$\mathbf{a}$	$\mathbf{x}$	$\mathbf{Ax}$	$\mathbf{x}^T\mathbf{A}$
$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	$\mathbf{0}$	$\mathbf{I}$	$\mathbf{A}$	$\mathbf{A}^T$

$\mathbf{x} \in \mathbb{R}^{\boldsymbol{n}}, \mathbf{y} \in \mathbb{R}^{\boldsymbol{m}}$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{n}}$$

$\mathbf{a}$ is $\mathbf{A}$ are not a function of $\mathbf{x}$.

$\mathbf{0}$ and $\mathbf{I}$ are matrices.

$\mathbf{y}$	$a\mathbf{u}$	$\mathbf{u} + \mathbf{v}$
$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$

Generalize to Matrices

		Scalar	**Vector**	**Matrix**
		x $(1,)$	$\mathbf{x}$ $(n,1)$	$\mathbf{X}$ (n,k)
Scalar	y $(1,)$	$\frac{dy}{dx}$ $(1,)$	$\frac{dy}{d\mathbf{x}}$ $(1,n)$	$\frac{dy}{d\mathbf{X}}$ (k,n)
Vector	$\mathbf{y}$ $(m,1)$	$\frac{d\mathbf{y}}{dx}$ $(m,1)$	$\frac{d\mathbf{y}}{d\mathbf{x}}$ (m,n)	$\frac{d\mathbf{y}}{d\mathbf{X}}$ (m,k,n)
Matrix	$\mathbf{Y}$ (m,l)	$\frac{d\mathbf{Y}}{dx}$ (m,l)	$\frac{d\mathbf{Y}}{d\mathbf{x}}$ (m,l,n)	$\frac{d\mathbf{Y}}{d\mathbf{X}}$ (m,l,k,n)

What is the Chain Rule?

- **If $y = f(u)$ and $u = g(x)$ are differentiable functions, then $y = f\big(g(x)\big)$ is a differentiable function of x.**

- **The chain rule is**

Multiply by the derivative of the inside

Derivative of u in terms of x

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = f'\big(g(x)\big)g'(x)$$

Derivative of y in terms of u

Keep the inside and take the derivative of the outside

Example: Chain Rule

- **Given** $y = (3x^2 - 5x + 2)^4$

- **Use a substitution,** u **= "the inside function."**

$$u = 3x^2 - 5x + 2 \quad \Rightarrow \quad y = u^4$$

- **Break up functions using the chain rule.**

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = (4u^3)(6x - 5) = 4(x^2 - 5x + 2)^3(6x - 5)$$

Example: Chain Rule

- **Given** $y = \log 3x^2$

- **Use a substitution,** u **= "the inside function."**

$$u = 3x^2 \quad \Rightarrow \quad y = \log u$$

- **Break up functions using the chain rule.**

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = \left(\frac{1}{u}\right)6x = \left(\frac{1}{3x^2}\right)6x = \frac{2}{x}$$

Example: Chain Rule

- **Given** $y = \log(2x-1)^3$

- **Use a substitution,** u **= "the inside function."**

$$\boxed{u = v^3 = (2x-1)^3} \Rightarrow \boxed{y = \log u}$$

$$v = (2x-1)$$

- **Break up functions using the chain rule.**

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dv}\frac{dv}{dx} = \left(\frac{1}{u}\right)3v^2(2) = \left(\frac{1}{(2x-1)^3}\right)6(2x-1)^2 = 6\left(\frac{1}{2x-1}\right)$$

Generalize to Vectors

- **Chain rule for scalars**

$$y = f\big(g(x)\big) \quad \Rightarrow \quad \frac{\partial y}{\partial x} = \frac{\partial y}{\partial u}\frac{\partial u}{\partial x}$$

- **Generalize to vectors straightforwardly.**

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial y}{\partial u}\frac{\partial u}{\partial \mathbf{x}} \qquad \frac{\partial y}{\partial \mathbf{x}} = \frac{\partial y}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}} \qquad \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

$(1, n) \quad (1,) \quad (1, n) \qquad (1, n) \quad (1, k) \quad (k, n) \qquad (m, n) \quad (m, k) \quad (k, n)$

Example: Generalize to Vectors

- **Assume** $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n, \;\; y \in \mathbb{R}$

$$z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$$

- **Compute** $\dfrac{\partial z}{\partial \mathbf{w}}$

- **Decompose**

$$a = \langle \mathbf{x}, \mathbf{w} \rangle$$

$$b = a - y$$

$$z = b^2$$

$$\frac{\partial z}{\partial \mathbf{w}} = \frac{\partial z}{\partial b}\frac{\partial b}{\partial a}\frac{\partial a}{\partial \mathbf{w}}$$

$$= 2b \cdot 1 \cdot \mathbf{x}^{\mathrm{T}}$$

$$= 2(\langle \mathbf{x}, \mathbf{w} \rangle - y) \cdot \mathbf{x}^{\mathrm{T}}$$

Example: Generalize to Vectors

- Assume $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathbf{w} \in \mathbb{R}^{n}$, $\mathbf{y} \in \mathbb{R}^{m}$

$$\mathbf{z} = (\mathbf{X}\mathbf{w} - \mathbf{y})^2$$

- Compute $\dfrac{\partial z}{\partial \mathbf{w}}$

- Decompose

$$\mathbf{a} = \mathbf{X}\mathbf{w}$$

$$\mathbf{b} = \mathbf{a} - \mathbf{y}$$

$$\mathbf{z} = ||\mathbf{b}||^2$$

$$\frac{\partial \mathbf{z}}{\partial \mathbf{w}} = \frac{\partial z}{\partial \mathbf{b}} \frac{\partial \mathbf{b}}{\partial \mathbf{a}} \frac{\partial \mathbf{a}}{\partial \mathbf{w}}$$

$$= 2\mathbf{b}^{\mathrm{T}} \cdot \mathbf{I} \cdot \mathbf{X}$$

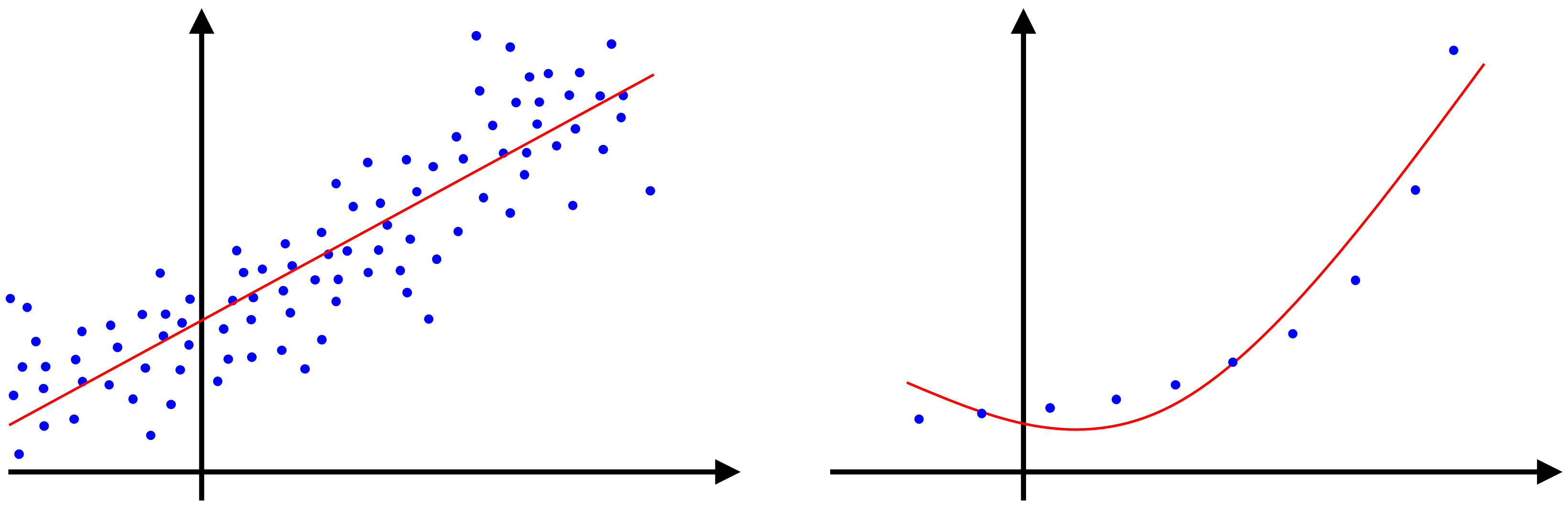
$$= 2(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}} \cdot \mathbf{X}$$

Example: Simple Linear Regression

Regression Models

- Modeling the relationship between **input feature vector $\mathbf{x}$** and a **label y**

- Predicting and forecasting the **continuous value**



Linear Model Formulation

- Given d-dimensional input $\mathbf{x} = [x_1, x_2, \dots, x_d]^{\mathrm{T}}$,

- The linear model has a d-dimensional weight and a bias

$$\mathbf{w} = [w_1, w_2, \dots, w_d]^{\mathrm{T}} \qquad b$$

- The output is a weighted sum of the inputs

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b$$

- Vectorized version

$$y = \langle \mathbf{w}, \mathbf{x} \rangle + b = \mathbf{w}^{\mathrm{T}} \mathbf{x} + b$$

Training Data

- **Data samples to fit model parameters**
  - The more the better

- **Assume that we collect n samples (or instances, examples).**

$$\mathbf{X} = [\,\mathbf{x}_1, \dots, \mathbf{x}_n]^{\mathrm{T}}$$

$$y = [y_1, \dots, y_n]^{\mathrm{T}}$$

$\mathbf{x}_i \in \mathbb{R}^d$

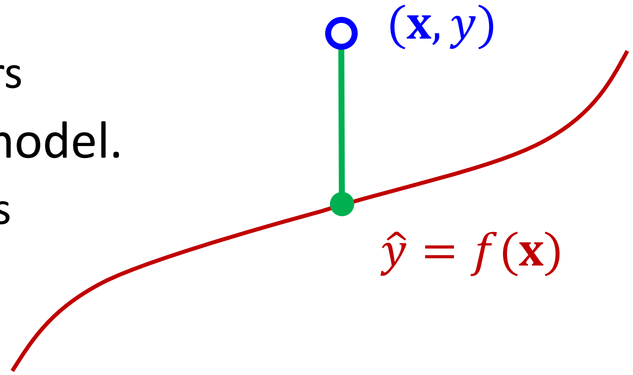
$y_i \in \mathbb{R}$

  - Each sample is represented by a d-dimensional vector.
  - Each label is scalar.

$$\mathbf{x} = [x_1, x_2, \dots, x_d]^{\mathrm{T}}$$

Measuring Errors

- **Compare the true value vs. the estimated value.**
  - Let y be the **true value**.
    - E.g., the actual sale price for used cars
  - Let $\hat{y}$ be the **estimated value** by our model.
    - E.g., the estimated price for used cars



- **Formulate the error function to minimize the difference between the true value and the estimated value.**

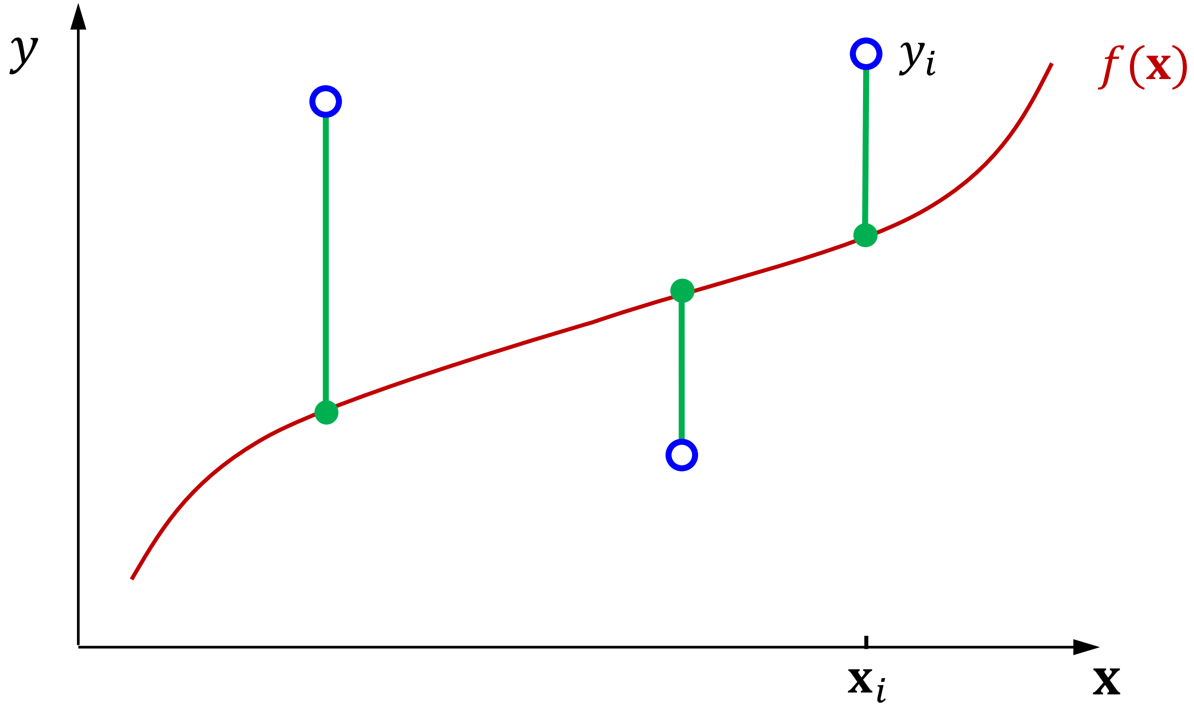
- **E.g., squared loss** $(y - \hat{y})^2$

Error Function (or Loss Function)

- Minimize the **squared residual** between the **actual value** and the **predicted value.**

Mean squared error (MSE) function

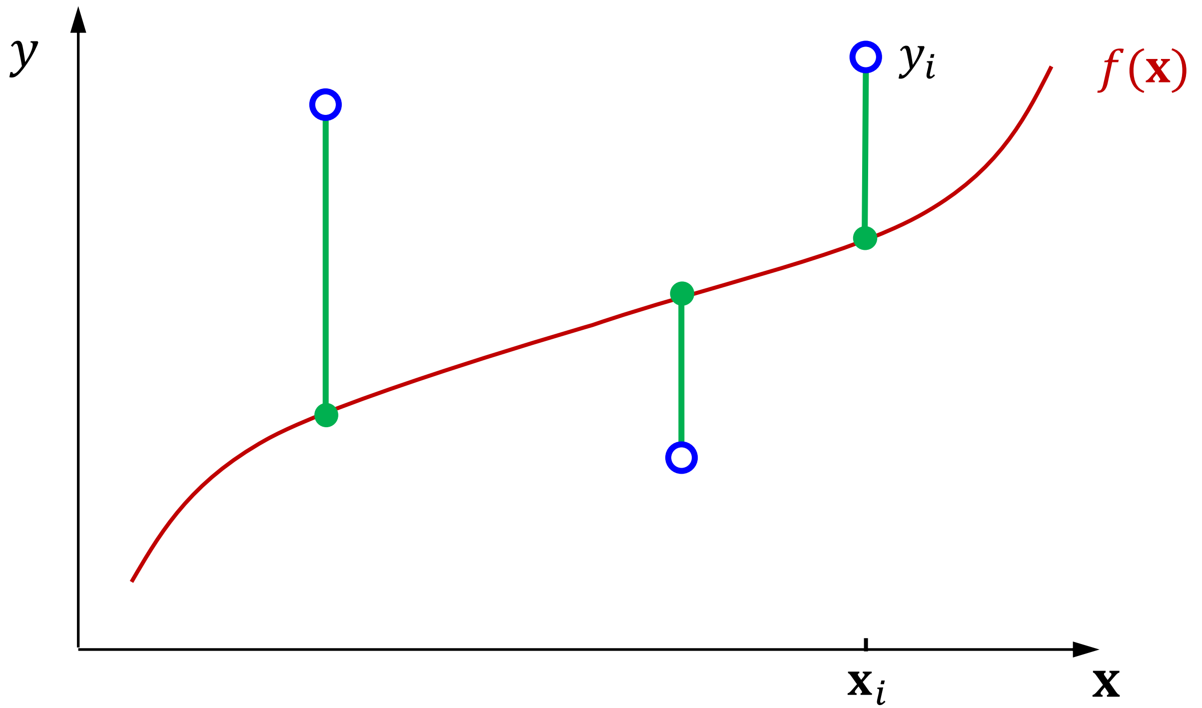
$$E(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(\mathbf{x}_i;\mathbf{w})\right)^2$$



Optimization Problem

- Find **an optimal parameter $\mathbf{w}^*$** minimizing the total errors for all training samples.

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\, E(\mathbf{w})$$

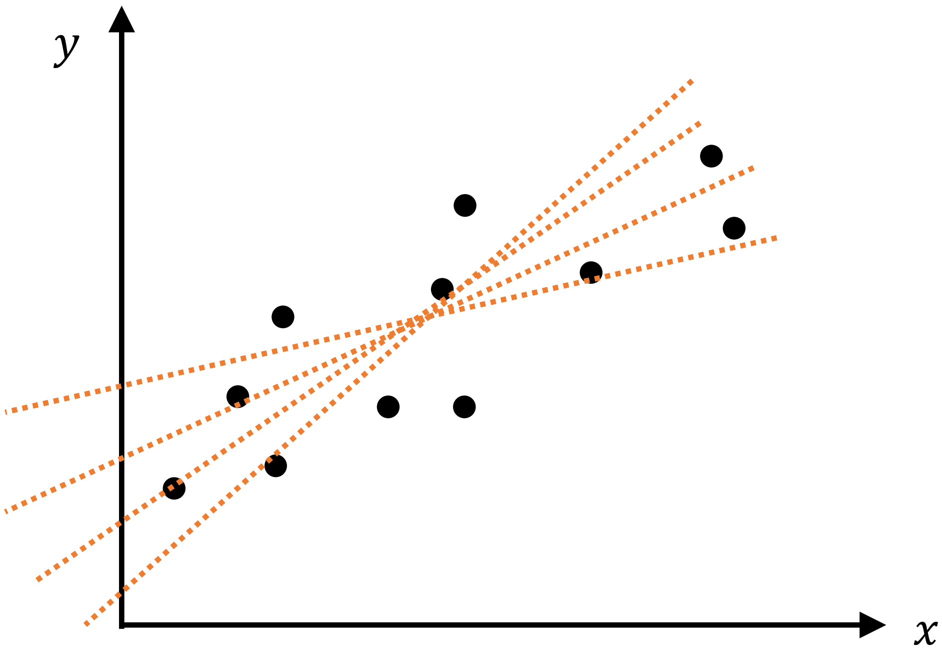


Example: Simple Linear Regression

- Given training data $\mathcal{D} = \{(x^{(i)}, y^{(i)}): 1 \le i \le n\}$,
- We want to train a **linear function**.

$$f(x; w_0, w_1) = w_1 x + w_0$$

w_0: bias, intercept

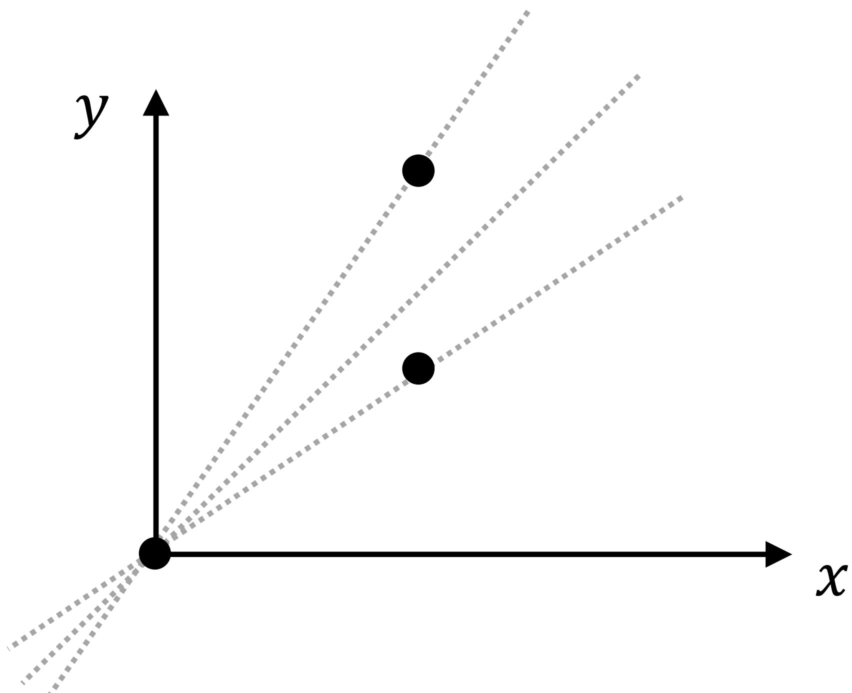


Which line is the best?

Simple Linear Regression Model

- Finding a **linear model** that fits a training dataset

$$f(x; w_0, w_1) = w_1 x + w_0$$

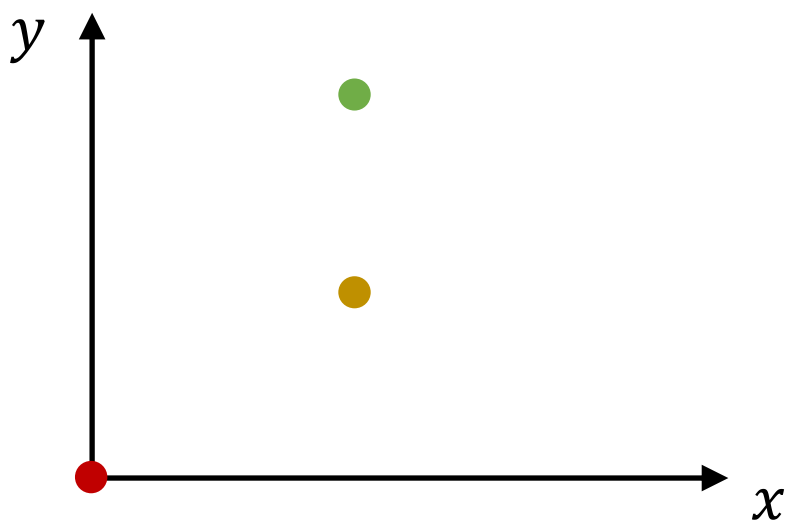


$$Data = \{(0.0, 0.0), (1.0, 1.0), (1.0, 2.0)\}$$

Solving the Optimization Problem

- Finding $\mathbf{w} = (w_0, w_1)$ that minimizes an error function

$$f(x; w_0, w_1) = w_1 x + w_0$$



$$f(0.0; w_0, w_1) \approx 0.0$$

$$f(1.0; w_0, w_1) \approx 1.0$$

$$f(1.0; w_0, w_1) \approx 2.0$$

$$Data = \{(0.0, 0.0), (1.0, 1.0), (1.0, 2.0)\}$$

Solving the Optimization Problem

- Finding $\mathbf{w} = (w_0, w_1)$ that minimizes an error function

$$f(x; w_0, w_1) = w_1 x + w_0$$

$$f(0.0; w_0, w_1) \approx 0.0$$

$$f(1.0; w_0, w_1) \approx 1.0$$

$$f(1.0; w_0, w_1) \approx 2.0$$

$$\Rightarrow$$

$$[f(0.0; w_0, w_1) - 0.0]^2$$

$$[f(1.0; w_0, w_1) - 1.0]^2$$

$$[f(1.0; w_0, w_1) - 2.0]^2$$

Solving the Optimization Problem

- **Finding $\mathbf{w} = (w_0, w_1)$ that minimizes $E(w_0, w_1)$**

$$f(x; w_0, w_1) = w_1 x + w_0$$

$$f(0.0; w_0, w_1) \approx 0.0$$

$$f(1.0; w_0, w_1) \approx 1.0$$

$$f(1.0; w_0, w_1) \approx 2.0$$

$$\Rightarrow \quad E(w_0, w_1) = \sum_{(x,y)\in\mathcal{D}} \big(y - f(x; w_0, w_1)\big)^2$$

Solving the Optimization Problem

- **Finding $\mathbf{w} = (w_0, w_1)$ that minimizes $E(w_0, w_1)$**
  - For simplicity, we use sum instead of mean.

$$E(w_0, w_1) = \sum_{(x,y)\in\mathcal{D}} \left(y - f(x; w_0, w_1)\right)^2$$

$$f(x; w_0, w_1) = w_1 x + w_0$$

$$Data = \{(0.0, 0.0), (1.0, 1.0), (1.0, 2.0)\}$$

$$E(w_0, w_1) = \left(0.0 - f(0.0; w_0, w_1)\right)^2 + \left(1.0 - f(1.0; w_0, w_1)\right)^2 + \left(2.0 - f(1.0; w_0, w_1)\right)^2$$

$$E(w_0, w_1) = (0.0 - w_0)^2 + (1.0 - (w_0 + w_1))^2 + (2.0 - (w_0 + w_1))^2$$

$$E(w_0, w_1) = 2w_1^2 + 3w_0^2 - 6w_1 - 6w_0 + 4w_1 w_0 + 5$$

Solving the Optimization Problem

- **Finding $\mathbf{w} = (w_0, w_1)$ that minimizes $E(w_0, w_1)$**

$$E(w_0, w_1) = (0.0 - w_0)^2 + (1.0 - (w_0 + w_1))^2 + (2.0 - (w_0 + w_1))^2$$

$$E(w_0, w_1) = 2w_1^2 + 3w_0^2 - 6w_1 - 6w_0 + 4w_1 w_0 + 5$$

$$\frac{\partial E}{\partial w_1} = 4w_1 + 4w_0 - 6$$

$$\frac{\partial E}{\partial w_0} = 4w_1 + 6w_0 - 6$$

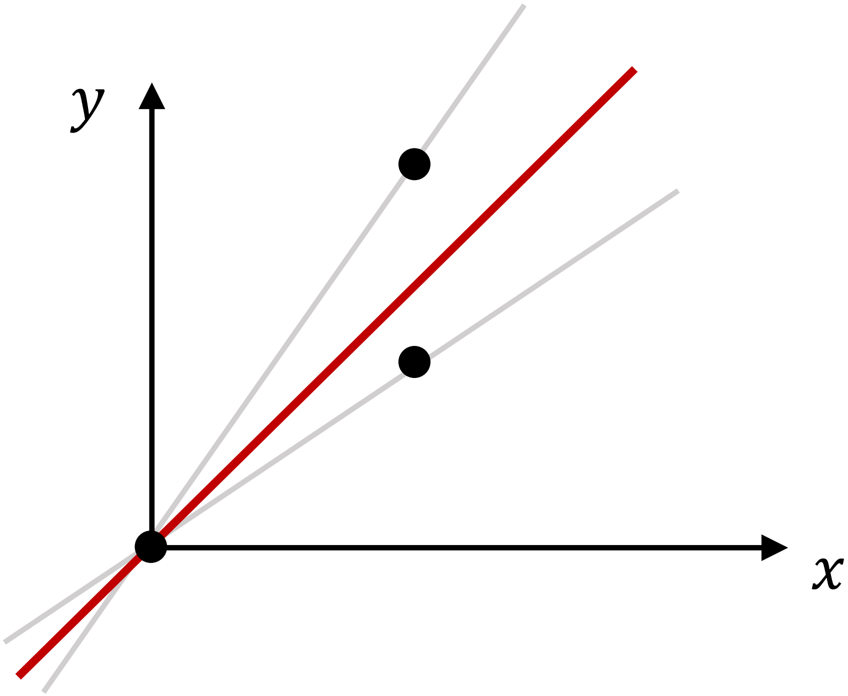
$$\Rightarrow \quad 4w_1 + 4w_0 - 6 = 0 \qquad 4w_1 + 6w_0 - 6 = 0$$

$$\Rightarrow \quad w_0 = 0.0 \qquad w_1 = 1.5$$

Solution of the Linear Model

- **Finally, we have learned a linear model that fits a given training dataset.**

$$f(x; w_0, w_1) = 1.5x + 0.0$$



$$Data = \{(0.0, 0.0), (1.0, 1.0), (1.0, 2.0)\}$$

Q&A