梯度下降

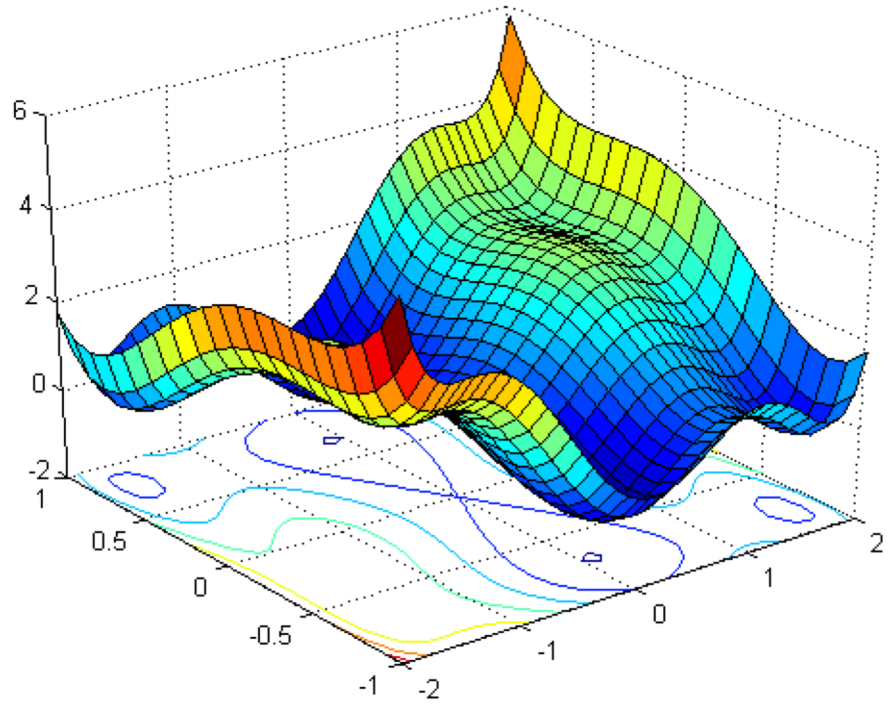
Gradient Descent

郝　奇
南京大学 天文与空间科学学院

Application of Machine Learning in Astronomy

机器学习在天文中的应用

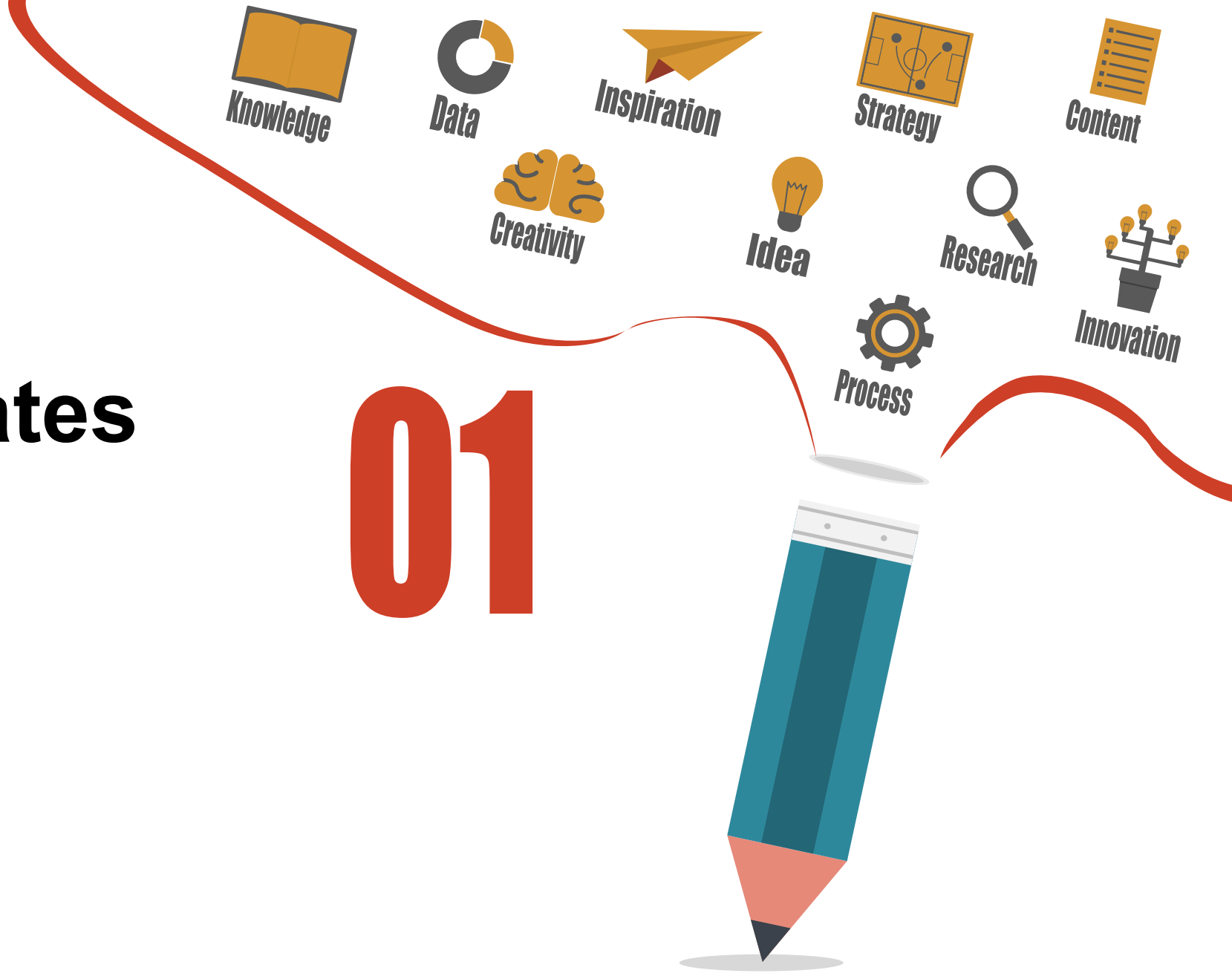
Contents



Cite: Hung-yi Lee, Machine Learning

Learning rates

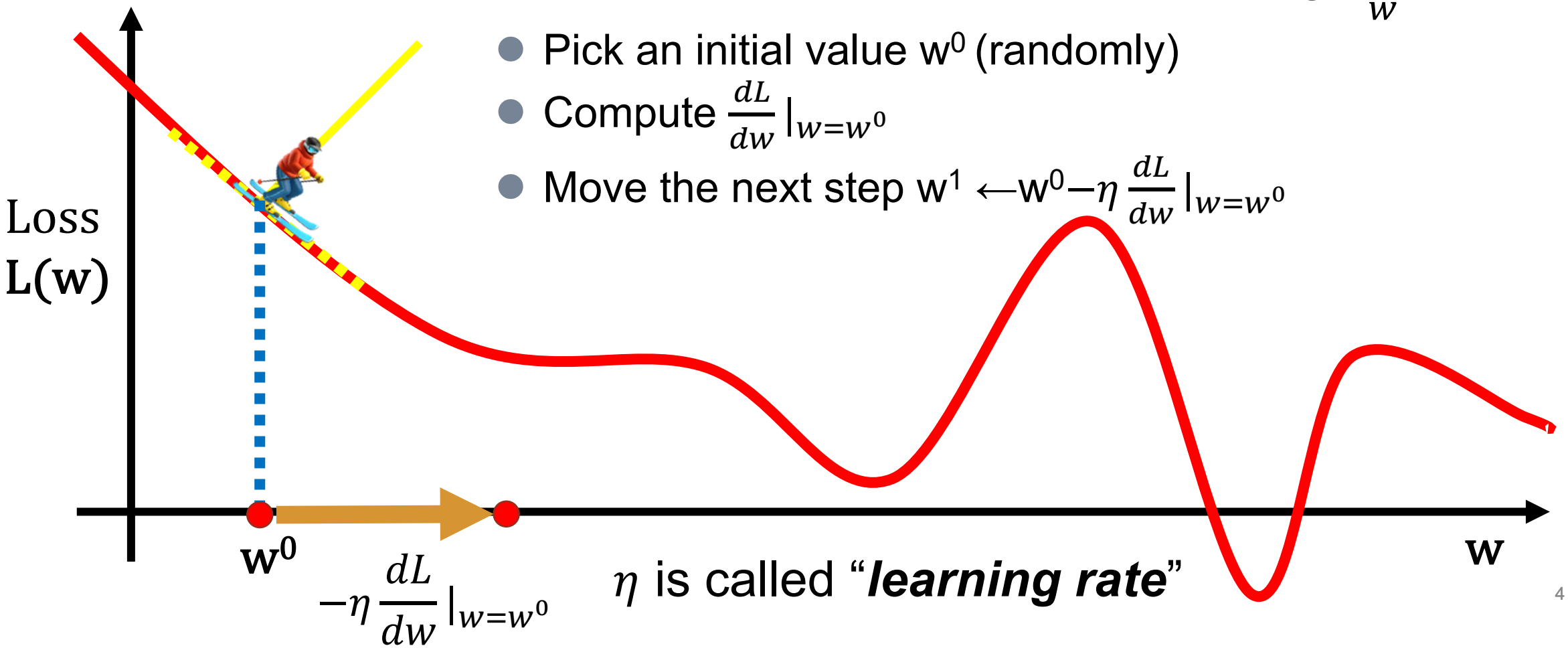
01



1. Learning rates

Previous Lecture

Consider loss function L(w) with one parameter w: $w^* = arg \min_w L(w)$



- Pick an initial value w^0 (randomly)
- Compute $\frac{dL}{dw}|_{w=w^0}$
- Move the next step $w^1 \leftarrow w^0 - \eta \frac{dL}{dw}|_{w=w^0}$

η is called "***learning rate***"

1. Learning rates

$L(w)$: loss function

w: parameters

Previous Lecture

$$w^* = arg\min_w L(w)$$

Suppose that w has two variables $\{w_1, w_2\}$

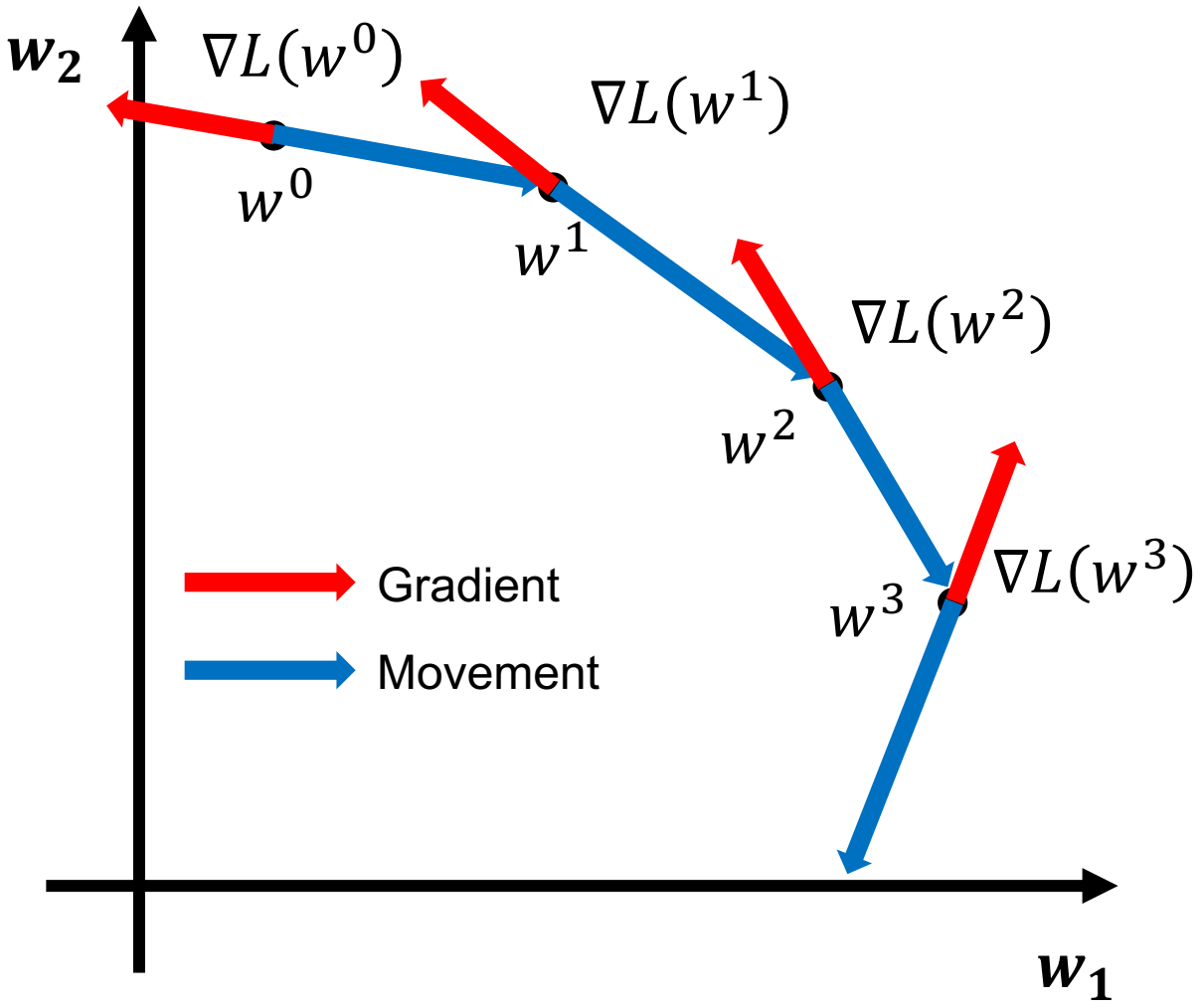
$$\nabla L(w) = \begin{bmatrix} \partial L(w_1^0)/\partial w_1 \\ \partial L(w_2^0)/\partial w_2 \end{bmatrix}$$

Randomly start at $w^0 = \begin{bmatrix} w_1^0 \\ w_2^0 \end{bmatrix}$

$$\begin{bmatrix} w_1^1 \\ w_2^1 \end{bmatrix} = \begin{bmatrix} w_1^0 \\ w_2^0 \end{bmatrix} - \eta \begin{bmatrix} \partial L(w_1^0)/\partial w_1 \\ \partial L(w_2^0)/\partial w_2 \end{bmatrix} \Rightarrow w^1 = w^0 - \eta \nabla L(w^0)$$

$$\begin{bmatrix} w_1^2 \\ w_2^2 \end{bmatrix} = \begin{bmatrix} w_1^1 \\ w_2^1 \end{bmatrix} - \eta \begin{bmatrix} \partial L(w_1^1)/\partial w_1 \\ \partial L(w_2^1)/\partial w_2 \end{bmatrix} \Rightarrow w^2 = w^1 - \eta \nabla L(w^1)$$

1. Learning rates



Start at position w^0

Calculate gradient at w^0: $\nabla L(w^0)$

Move to $w^1 = w^0 - \eta \nabla L(w^0)$

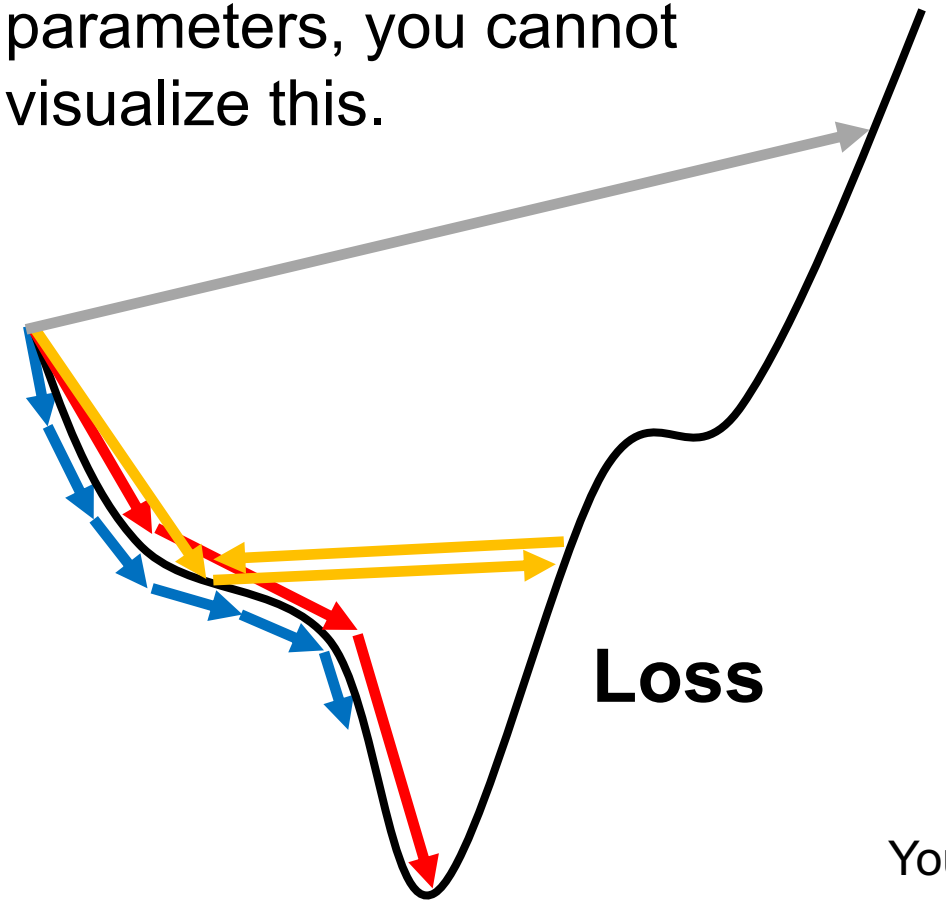
Calculate gradient at w^1: $\nabla L(w^1)$

Move to $w^2 = w^1 - \eta \nabla L(w^1)$

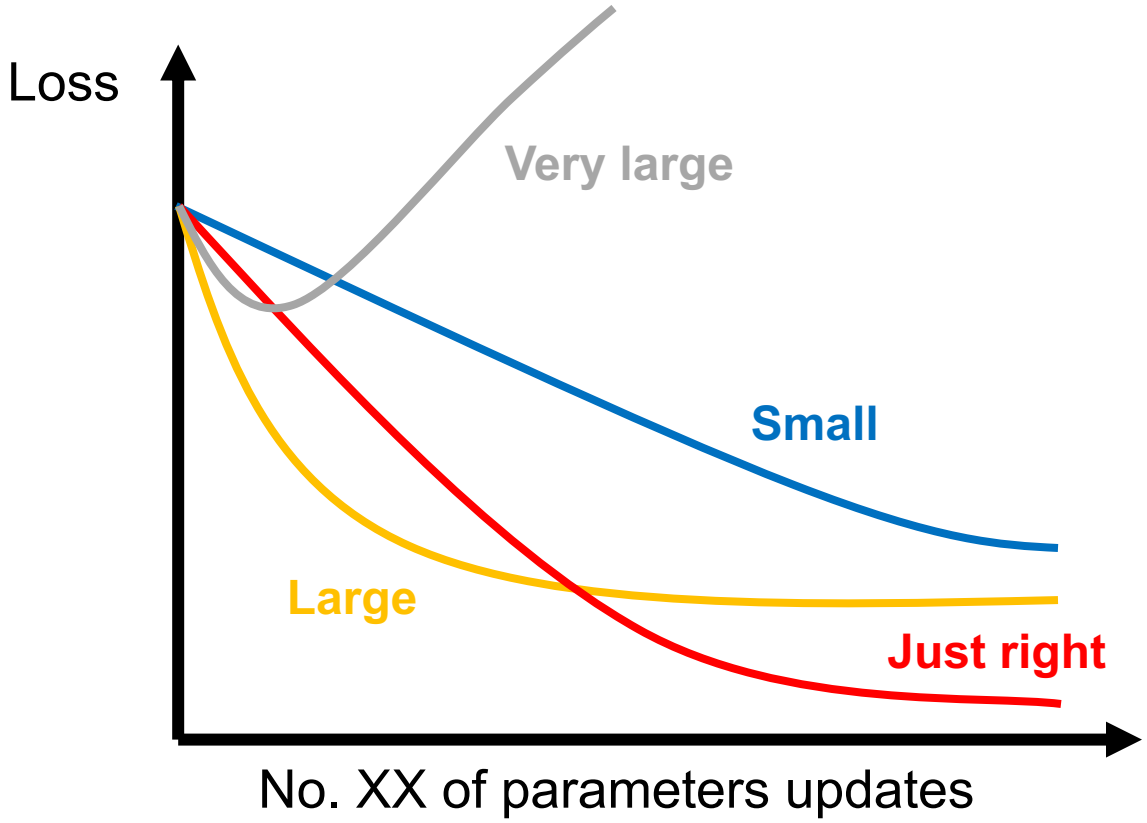
⋮

1. Learning rates

If there are more than three parameters, you cannot visualize this.



$$w^i = w^{i-1} - \eta \nabla L\left(w^{i-1}\right)$$



You can always visualize the **Loss V.S. parameters updates** !

Is there any adaptive learning rate?

1. Learning rates

- **Popular & simple idea:** Reduce the learning rate by some factors every few epochs.
  - At the beginning, we far from the destination, so we use larger learning rate;
  - After several epochs, we are close to the destination, so we reduce the learning rate;
  - E.g. $\eta^t = \eta/\sqrt{t+1}$
- **Learning rate cannot be one-size-fits-all.**
  - Giving different parameters different learning rates.

1. Learning rates

Adagrad

Divide the learning rate of each parameter by the **root mean square of its previous derivatives**

$$\eta^t = \frac{\eta}{\sqrt{t+1}}$$

Vanilla Gradient descent

$$w^{t+1} = w^t - \eta^t g^t$$

$$g^t = \frac{\partial L(w^t)}{\partial w}$$

Adagrad

$$w^{t+1} = w^t - \frac{\eta^t}{\sigma^t} g^t$$

σ^t is the **root mean square** of the previous derivatives of parameter w.

1. Learning rates

Adagrad

$$w^1 = w^0 - \frac{\eta^0}{\sigma^0} g^0 \qquad \sigma^0 = \sqrt{(g^0)^2}$$

$$w^2 = w^1 - \frac{\eta^1}{\sigma^1} g^1 \qquad \sigma^1 = \sqrt{\frac{1}{2}[(g^0)^2 + (g^1)^2]}$$

$$w^3 = w^2 - \frac{\eta^2}{\sigma^2} g^2 \qquad \sigma^2 = \sqrt{\frac{1}{2}[(g^0)^2 + (g^1)^2 + (g^2)^2]}$$

$$\vdots$$

$$w^{t+1} = w^t - \frac{\eta^t}{\sigma^t} g^t \qquad \sigma^t = \sqrt{\frac{1}{t+1}\sum_{i=0}^{t}(g^i)^2}$$

σ^t is the **root mean square** of the previous derivatives of parameter w.

1. Learning rates

Adagrad

Divide the learning rate of each parameter by the **root mean square of its previous derivatives**

$$w^{t+1} = w^t - \frac{\eta^t}{\sigma^t} g^t$$

$$\eta^t = \frac{\eta}{\sqrt{t+1}}$$

$$\sigma^t = \sqrt{\frac{1}{t+1}\sum_{i=0}^{t}(g^i)^2}$$

$$w^{t+1} = w^t - \frac{\eta}{\sqrt{\sum_{i=0}^{t}(g^i)^2}} g^t$$

1. Learning rates

Adagrad

$$g^t = \frac{\partial L(w^t)}{\partial w}$$

Vanilla Gradient descent

$$w^{t+1} = w^t - \eta^t g^t$$

Adagrad

$$w^{t+1} = w^t - \frac{\eta}{\sqrt{\sum_{i=0}^{t}(g^i)^2}} g^t$$

Larger gradient, larger step

Larger gradient, larger step

Contradiction

Larger gradient, smaller step

Larger gradient, smaller step

1. Learning rates

Adagrad

Extremely large

g^0	g^1	g^2	g^3	g^4	g^5	……
0.0001	0.0001	0.0003	0.0002	0.0001	0.1	……

g^0	g^1	g^2	g^3	g^4	g^5	……
10.5	39.2	23.1	50.4	22.8	0.1	……

Extremely small

$$w^{t+1} = w^t - \frac{\eta}{\sqrt{\sum_{i=0}^{t}(g^i)^2}} g^t$$

Creates a contrast effect

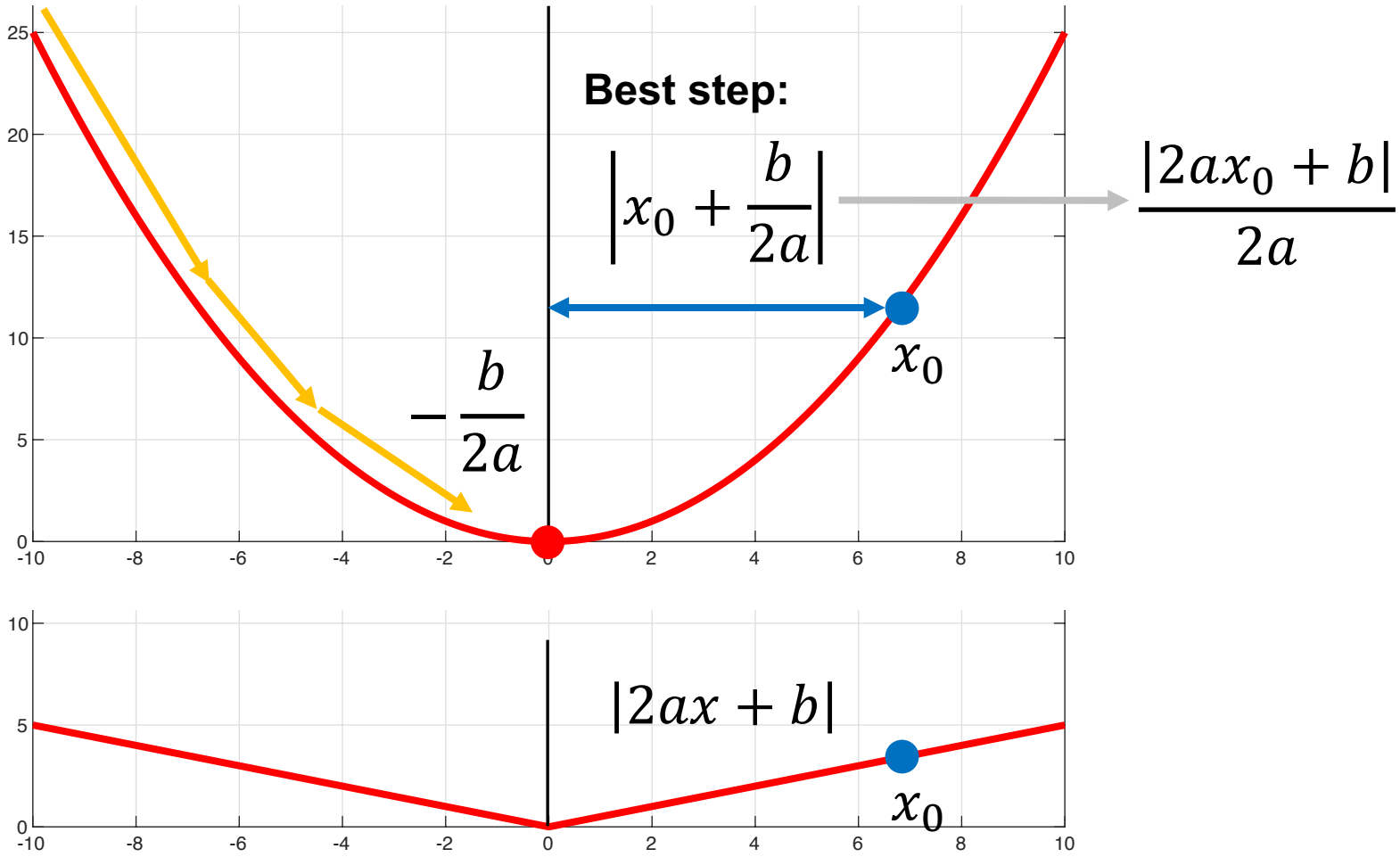
1. Learning rates

Adagrad

Larger 1st order derivative means far from the minima

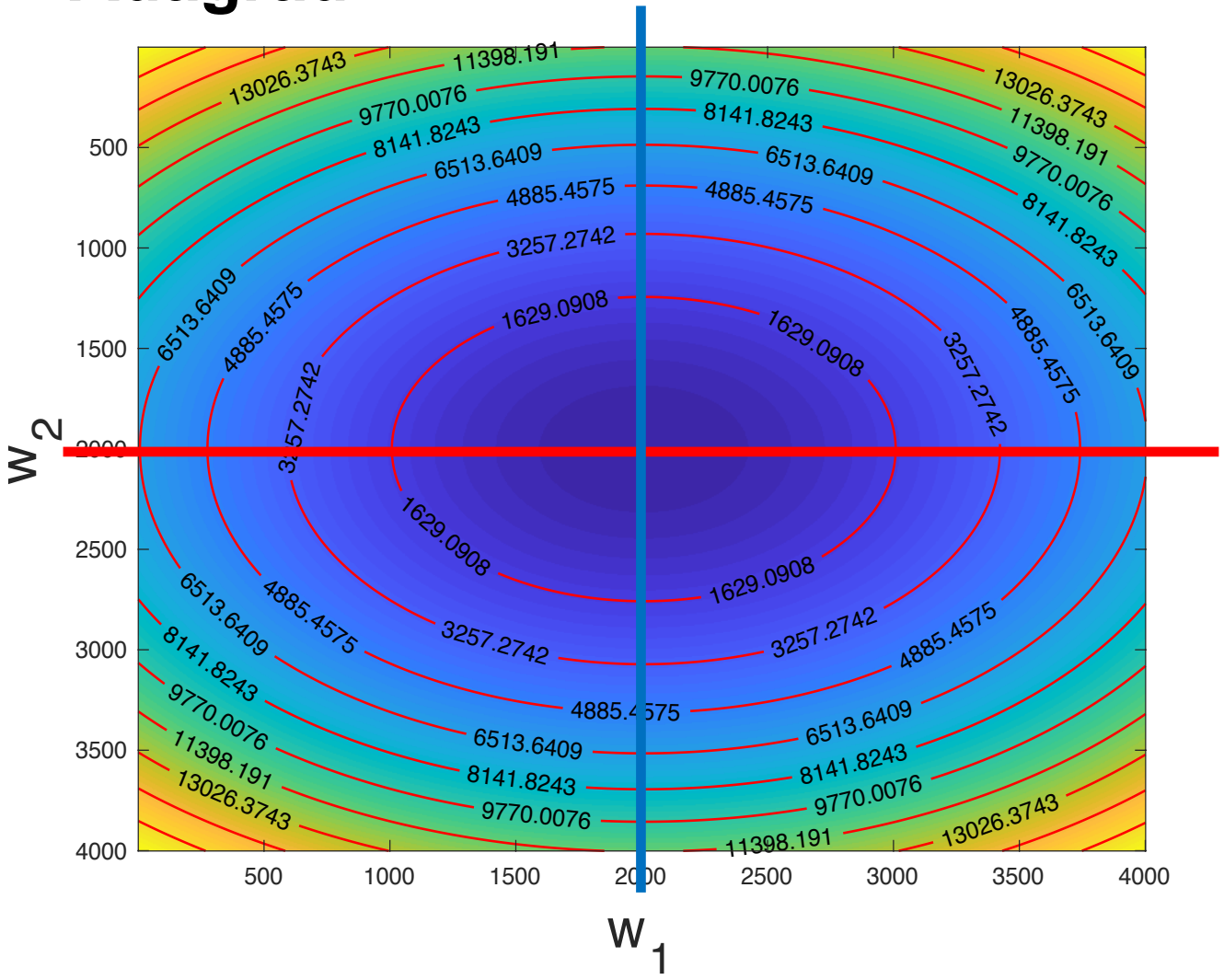
$$y = ax^2 + bx + c$$

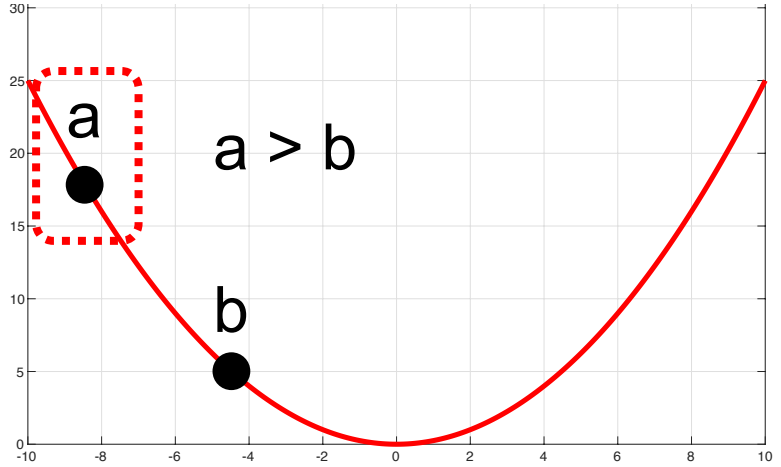
$$\left|\frac{\partial y}{\partial x}\right| = |2ax + b|$$

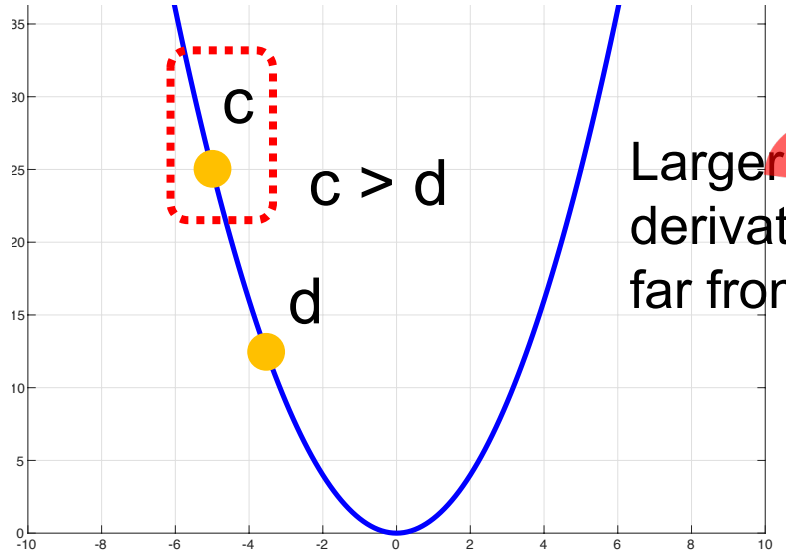


1. Learning rates

Adagrad





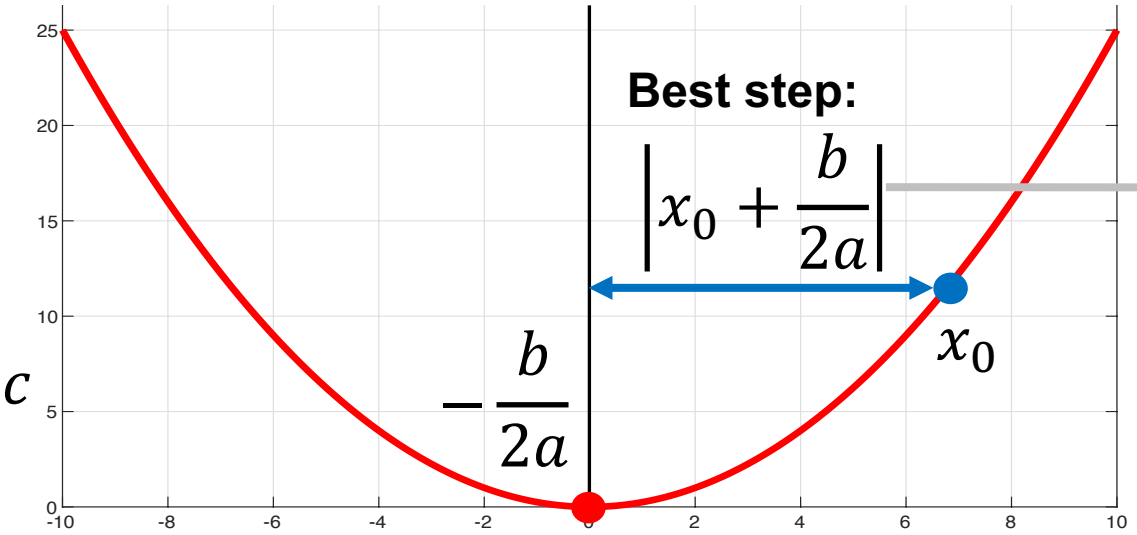


Larger 1st order derivative means far from the minima

1. Learning rates

Adagrad

$$y = ax^2 + bx + c$$



$$\frac{|2ax_0 + b|}{2a}$$

$$\left|\frac{\partial y}{\partial x}\right| = |2ax + b|$$

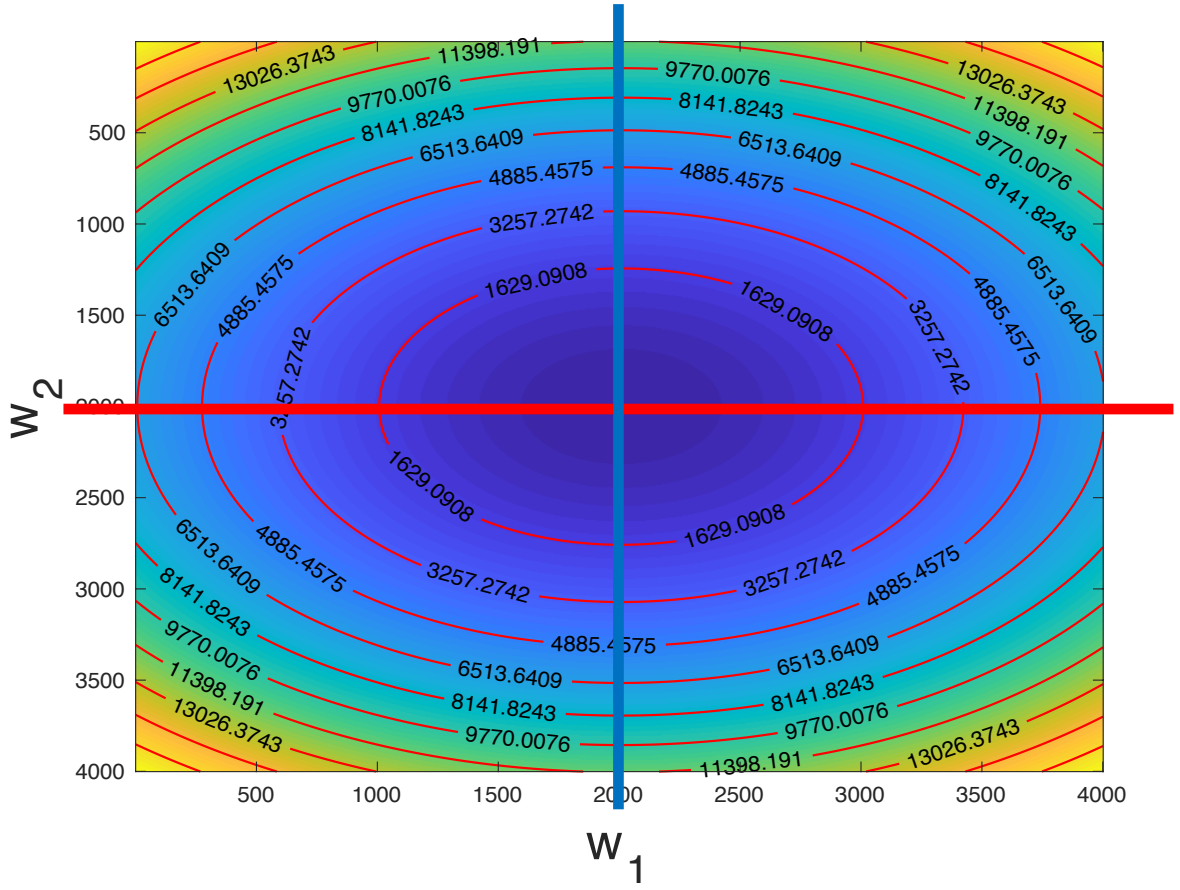
$$\left|\frac{\partial^2 y}{\partial x^2}\right| = 2a$$

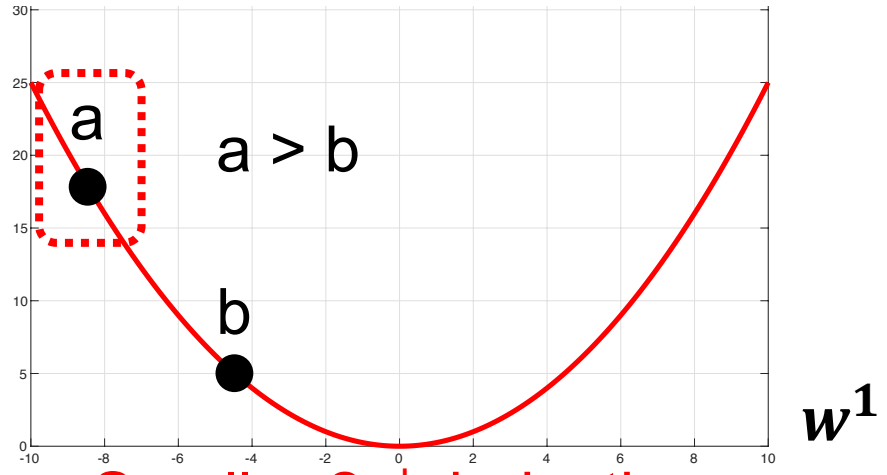
The best step is $\frac{\text{|First derivative|}}{\text{Second derivative}}$

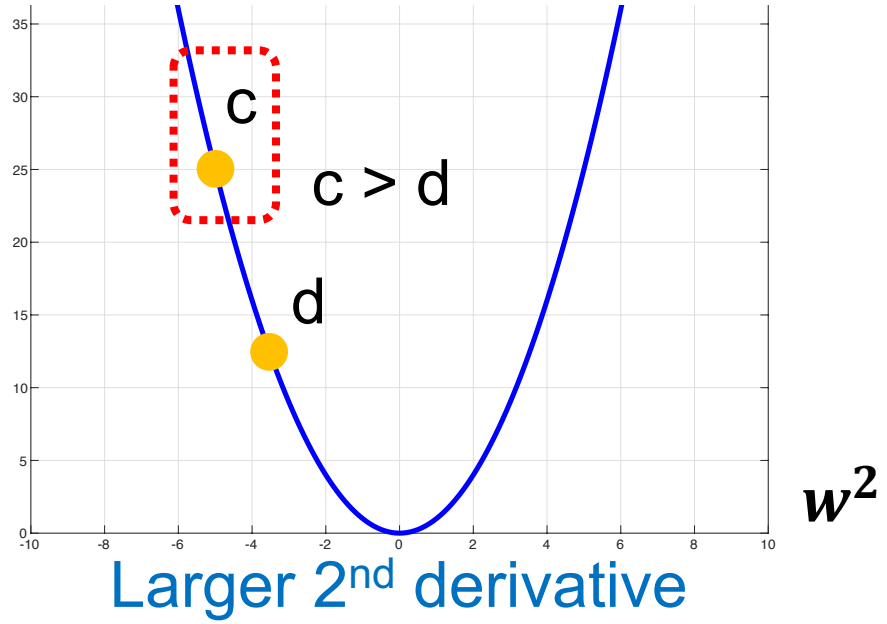
1. Learning rates

Adagrad

The best step is $\frac{|\text{First derivative}|}{\text{Second derivative}}$







Larger 1st order derivative means far from the minima
-- only suitable for one parameter condition

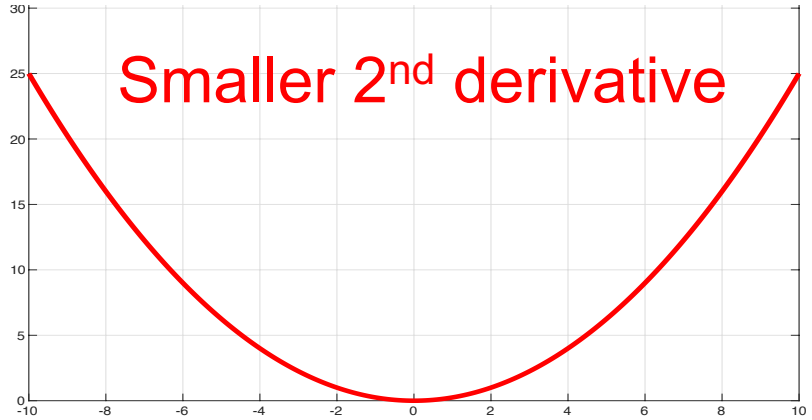
1. Learning rates

Adagrad

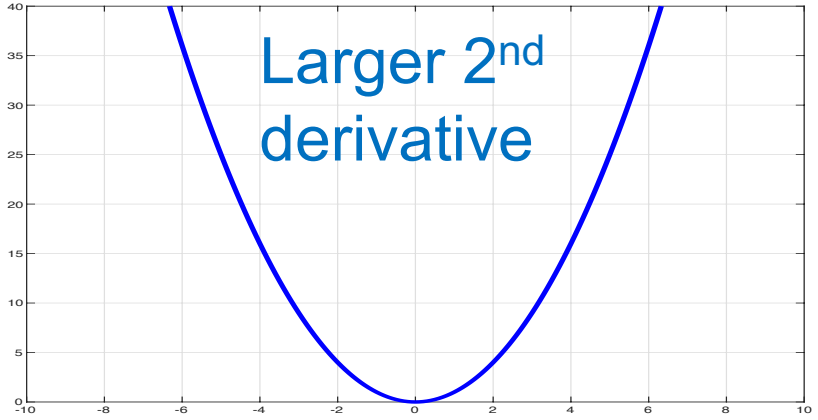
$$w^{t+1} = w^t - \frac{\eta}{\sqrt{\sum_{i=0}^{t}(g^i)^2}} g^t$$

|First derivative| / Second derivative

?

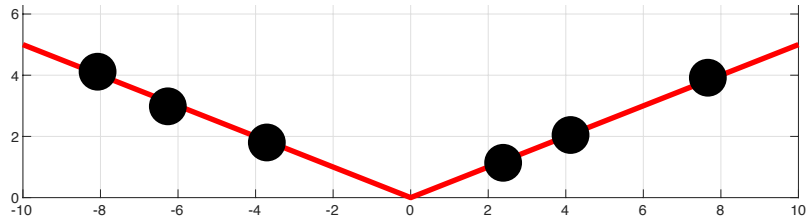


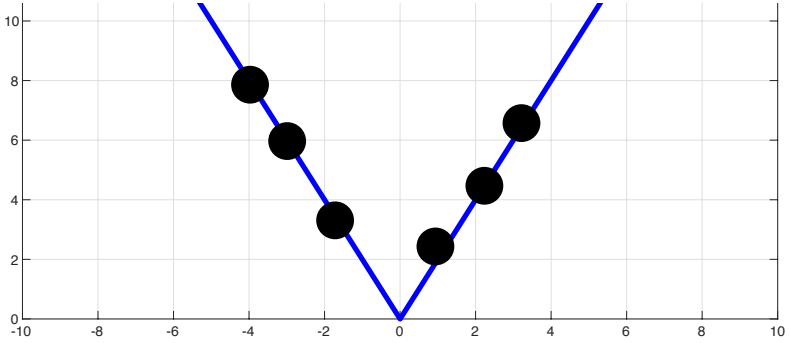
w^1



w^2

$\sqrt{(First\ derivative)^2}$





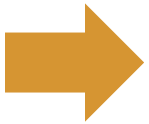
1. Learning rates

Feature Scaling

Previous Lecture

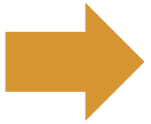
Mathematical transformations -- Scaling

Spine length
50.08
59.09
103.15
38.06
170.25
47.07
74.11
103.15
42.06



Spine length
50.08
59.09
103.15
38.06
170.25
47.07
74.11
103.15
42.06

/ max(spine length)



Spine length
0.294
0.347
0.606
0.224
1.000
0.276
0.435
0.606
0.247

Convert variables to comparable scales

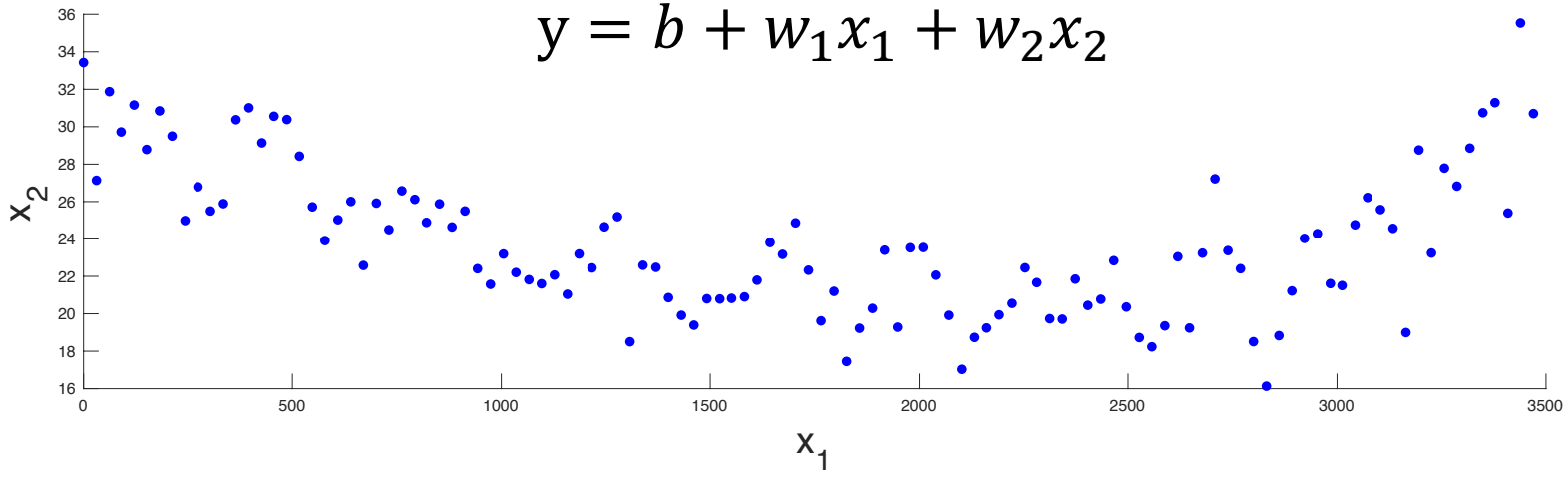
1. Learning rates

Feature Scaling

$$y = b + w_1 x_1 + w_2 x_2$$

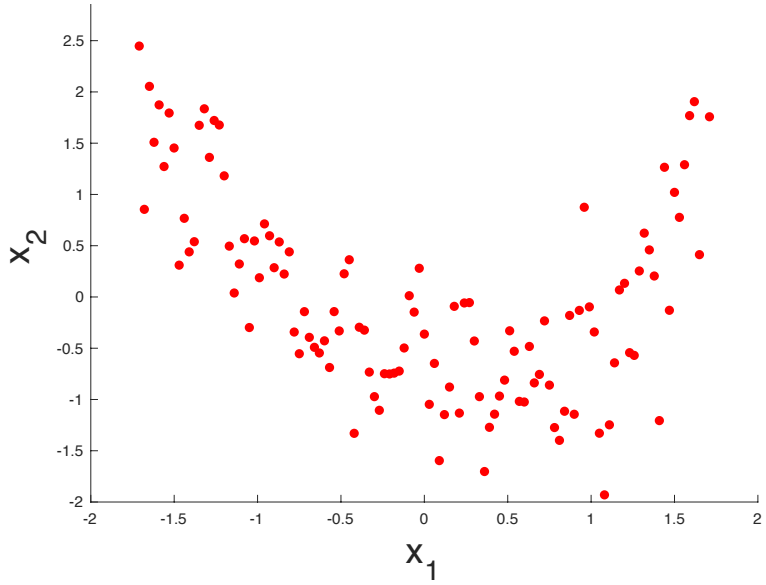
$x_1 \in [0, 3500]$

$x_2 \in [16, 36]$



$x_1 \in [-2, 2.5]$

$x_2 \in [-2, 2]$

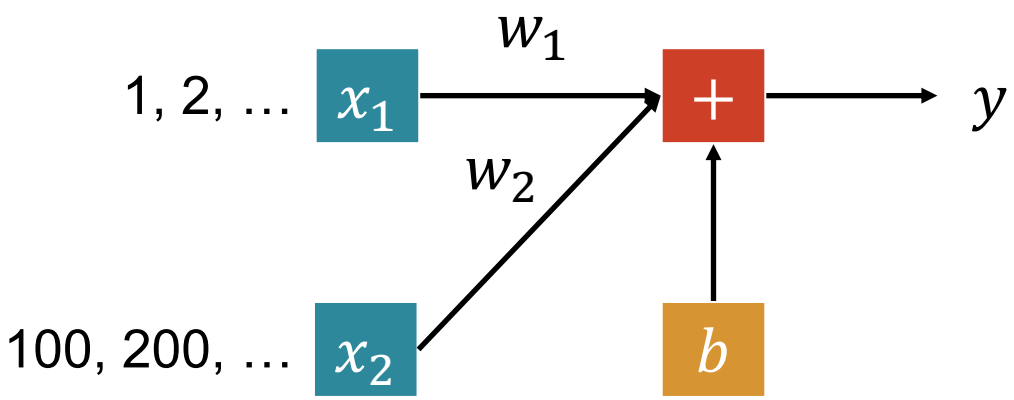


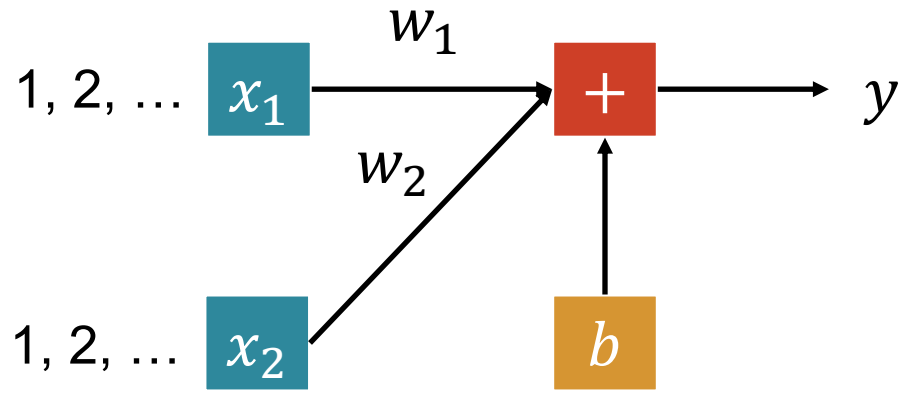
Make different features have the same scaling

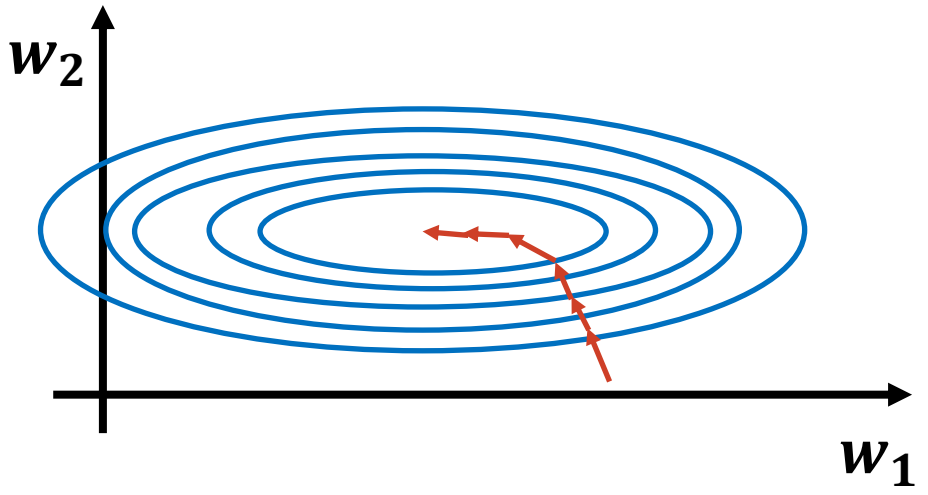
1. Learning rates

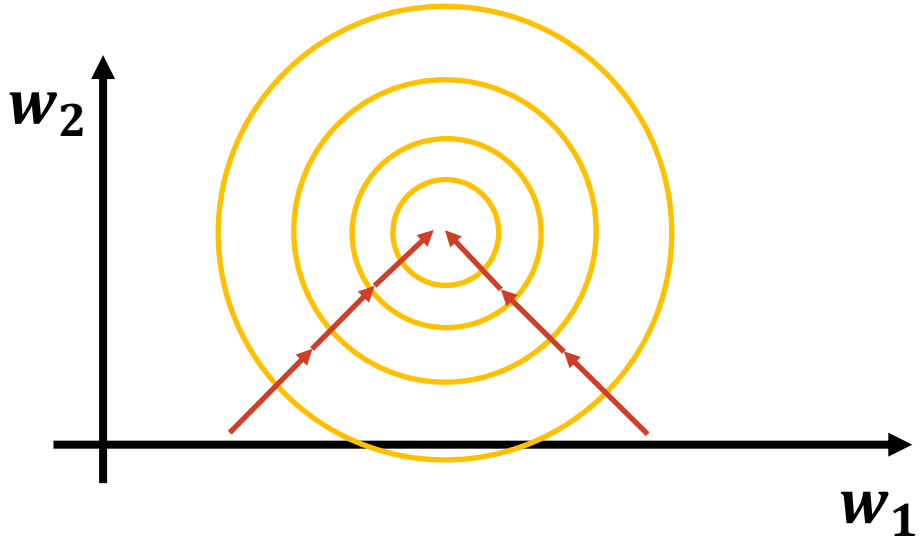
Feature Scaling

$$y = b + w_1 x_1 + w_2 x_2$$



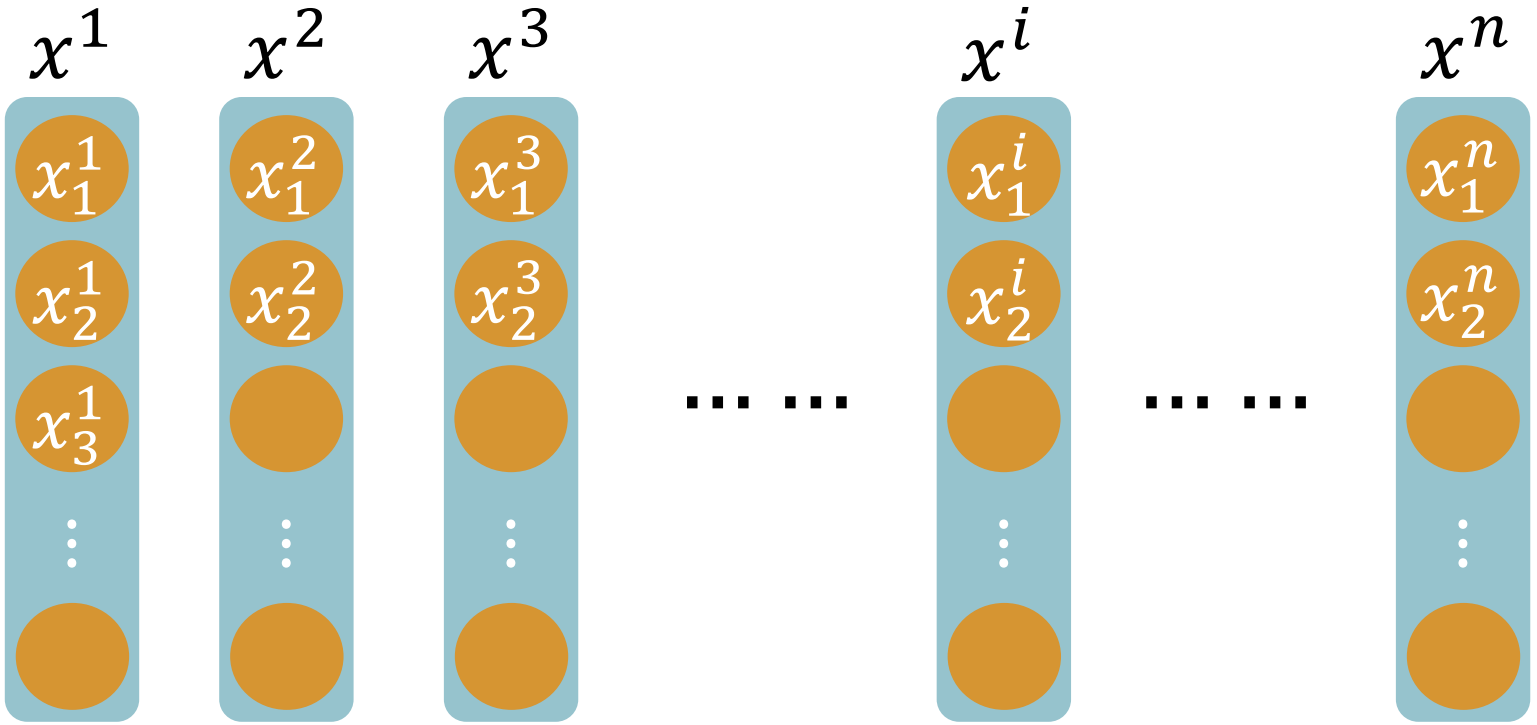






1. Learning rates

Feature Scaling



For each dimension j:

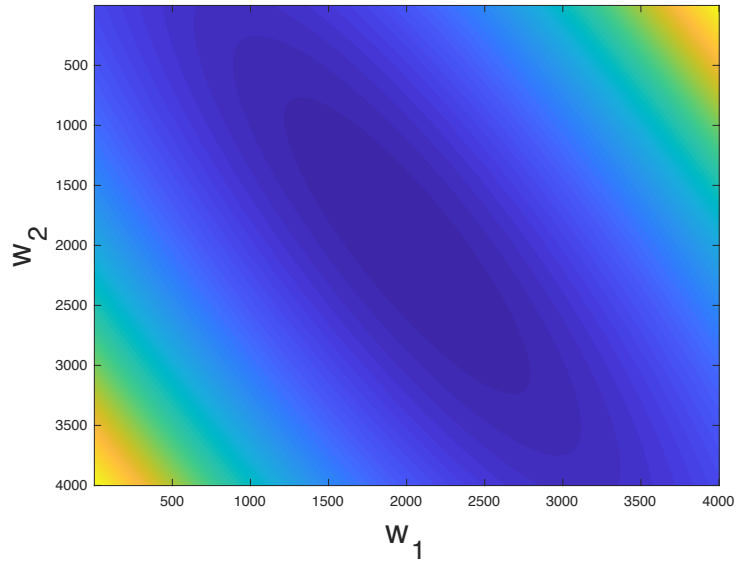
Mean: m_j

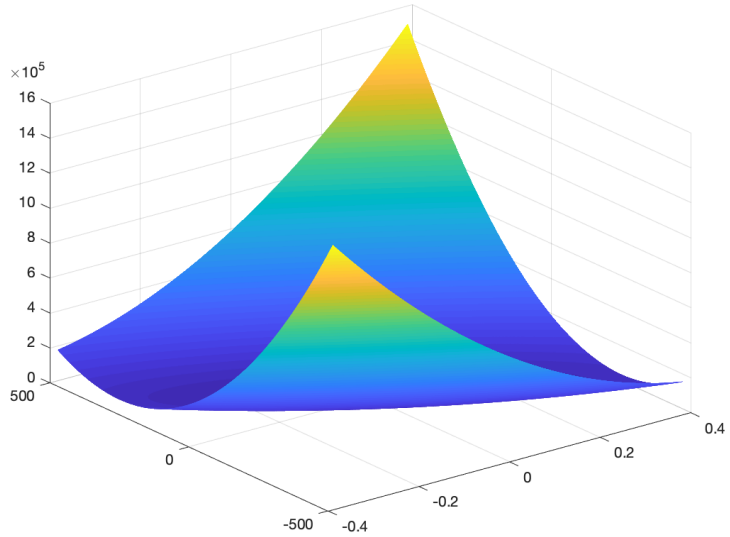
Standard deviation: σ_j

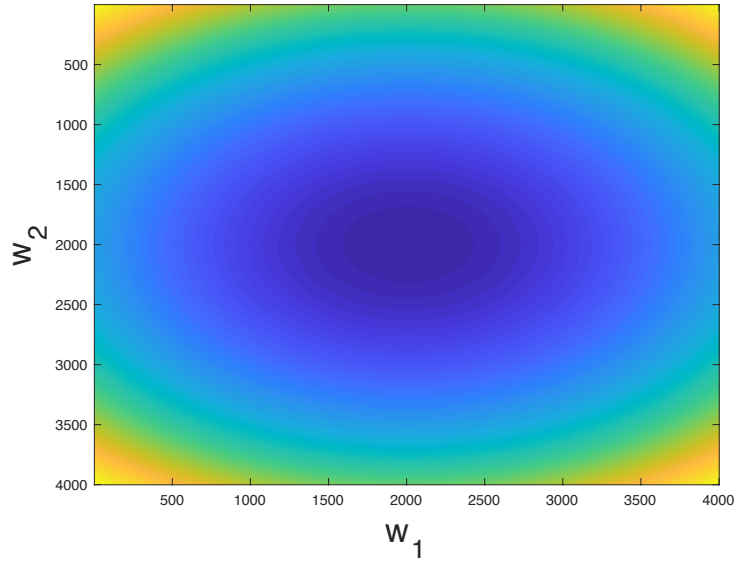
$$x_j^i = \frac{x_j^i - m_j}{\sigma_j}$$

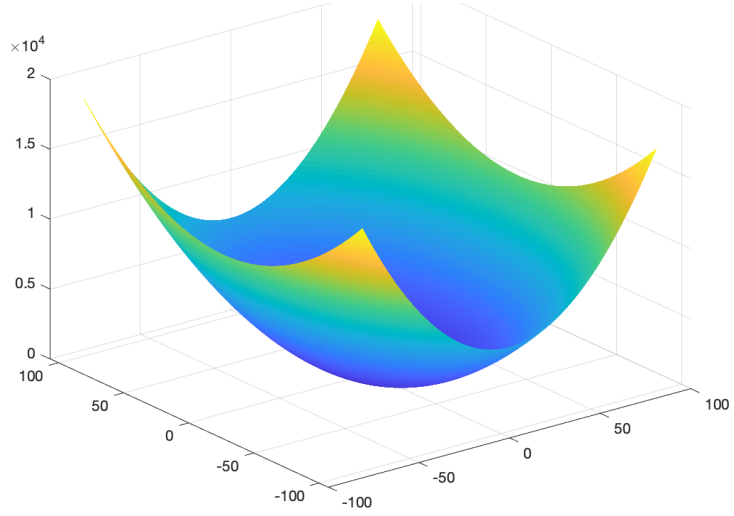
The means of all dimensions are 0
The variances are all 1

1. Learning rates









Stochastic gradient descent

02

2. Stochastic gradient descent

$$\mathrm{L}(f) = \sum_n \left(\hat{y}^n - f(x_i^n)\right)^2$$

Gradient descent **Loss is the summation over all training samples**

$$\mathrm{L} = \sum_n \left(\hat{y}^n - \left(b + \sum w_i x_i^n\right)\right)^2 \qquad w^i = w^{i-1} - \eta \nabla L\left(w^{i-1}\right)$$

Stochastic gradient descent **Loss is current one sample**

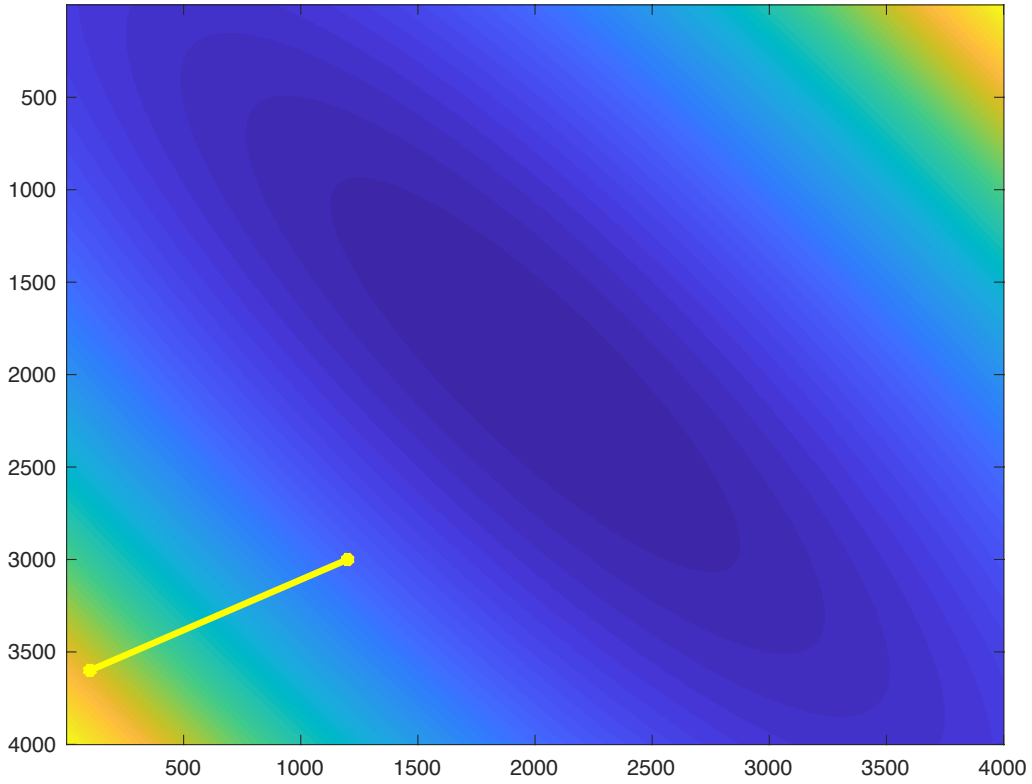
Pick an example x^n

$$\mathrm{L}^n = \left(\hat{y}^n - \left(b + \sum w_i x_i^n\right)\right)^2 \qquad w^i = w^{i-1} - \eta \nabla \mathrm{L}^n\left(w^{i-1}\right)$$

2. Stochastic gradient descent

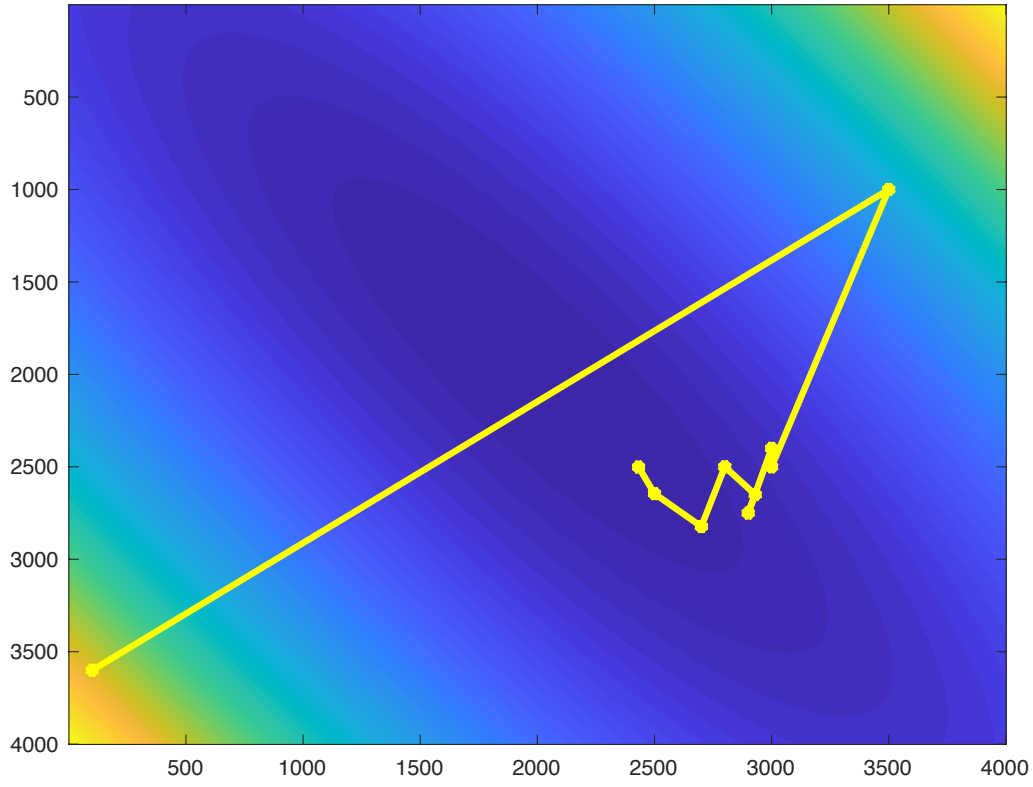
Gradient descent

Update after seeing all samples



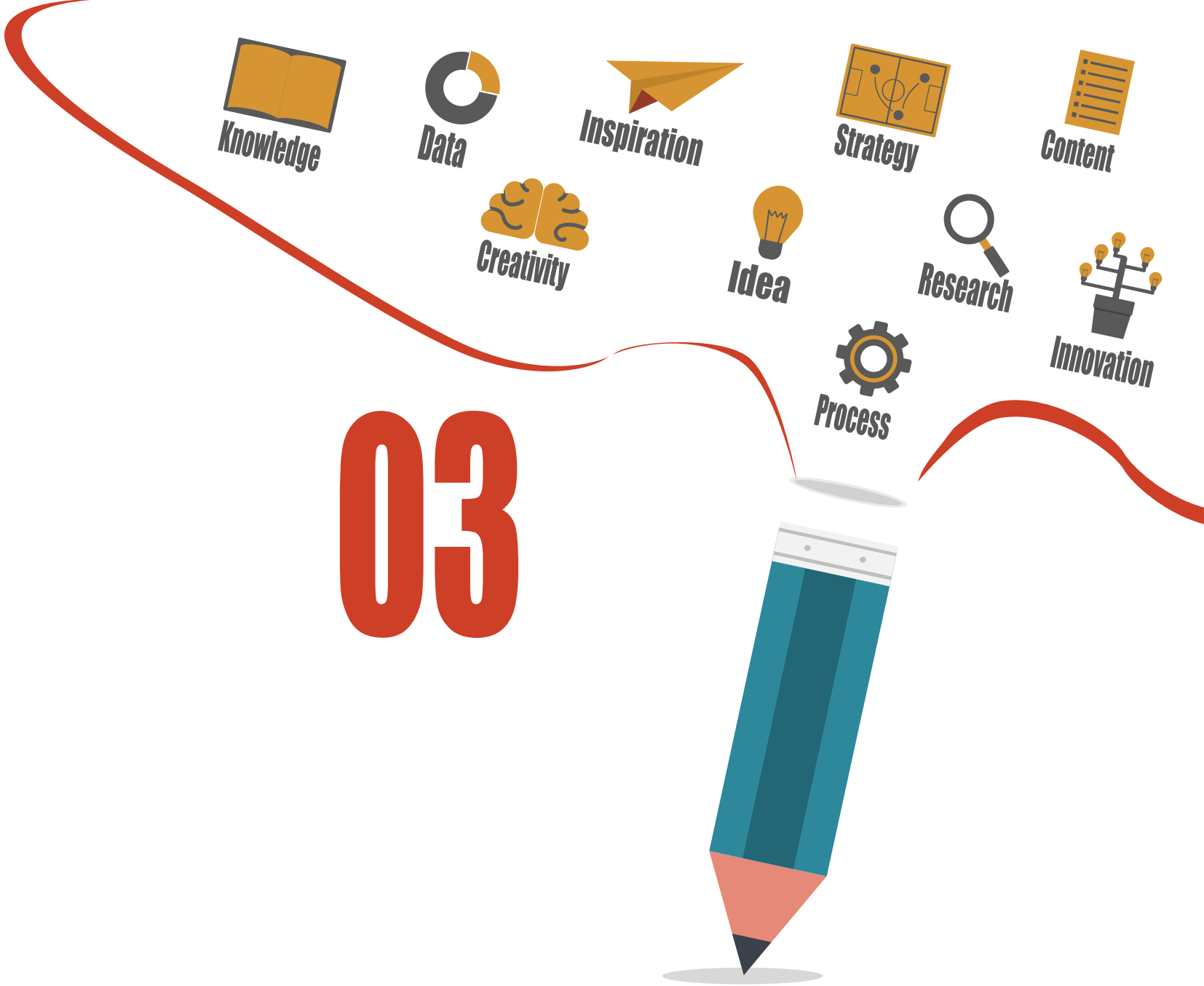
Stochastic gradient descent

Update for each sample;
If there are 10 samples, 10 times faster.



Theory and Limitation

03

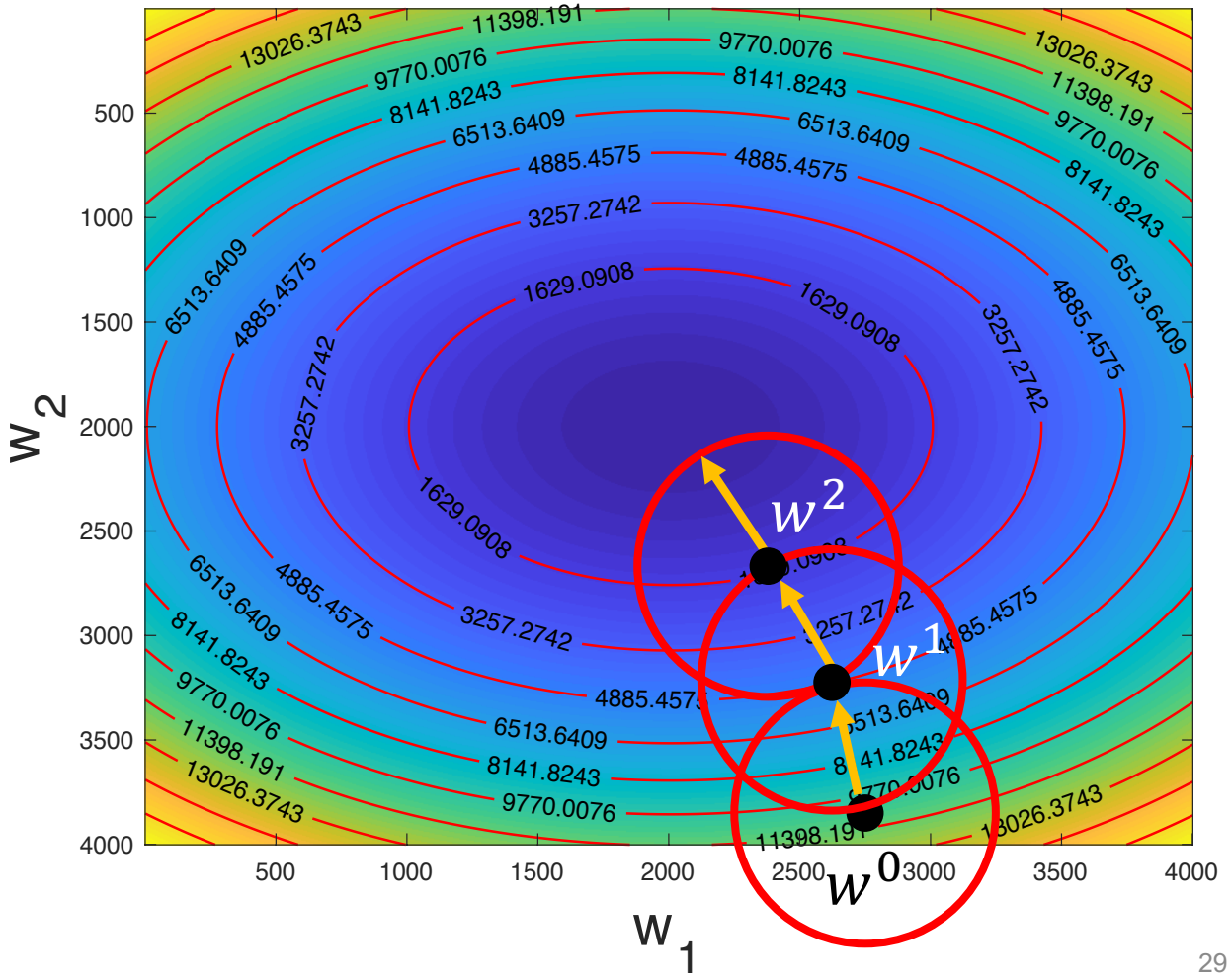


3. Theory and Limitation

Formal Derivation

Suppose that w has two variables $\{w_1, w_2\}$

Given a point, we can easily find the point with the smallest value nearby.



3. Theory and Limitation

Taylor Series

Taylor series: Let $h(x)$ be any function infinitely differentiable around $x = x_0$.

$$h(x) = \sum_{k=0}^{\infty} \frac{h^{(k)}(x_0)}{k!}(x - x_0)^k$$

$$= h(x_0) + \frac{h'(x_0)}{1!}(x - x_0) + \frac{h''(x_0)}{2!}(x - x_0)^2 + \frac{h^3(x_0)}{3!}(x - x_0)^3 + \cdots$$

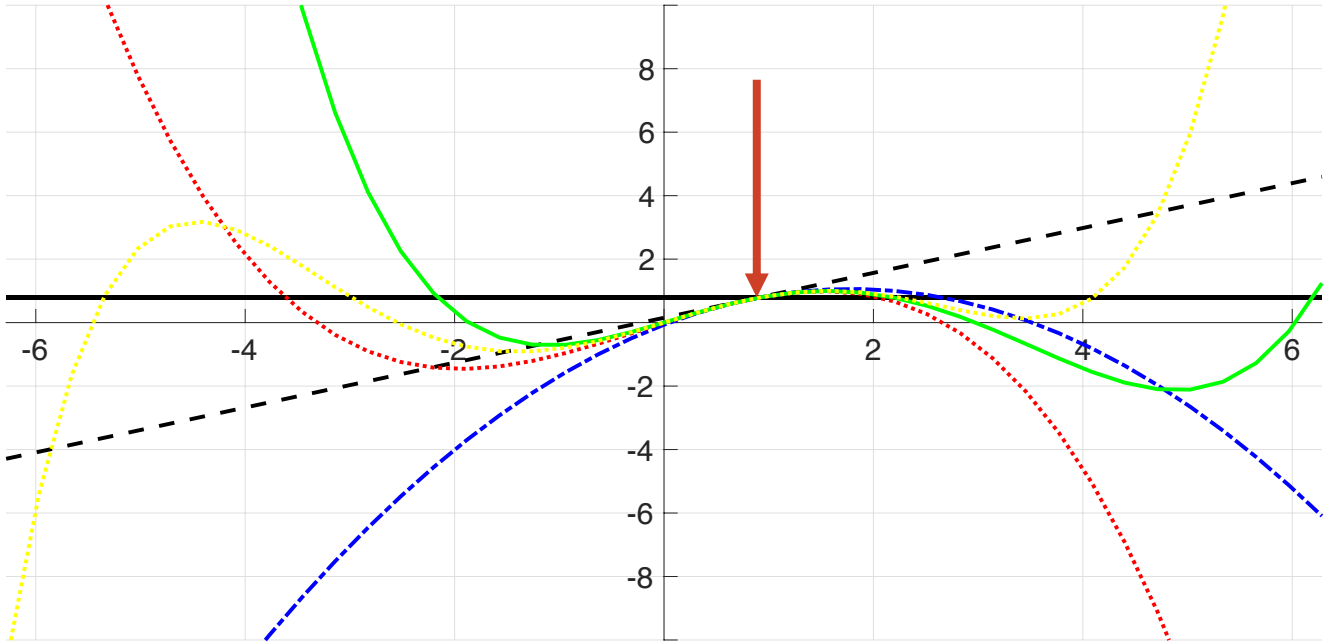
When x is close to x_0 then: $h(x) = h(x_0) + h'(x_0)(x - x_0)$

3. Theory and Limitation

Taylor Series

Let $h(x) = sin(x)$ around $x_0 = \pi/4$.

$$sin(x) = \frac{1}{\sqrt{2}} + \frac{x-\frac{\pi}{4}}{\sqrt{2}} - \frac{\left(x-\frac{\pi}{4}\right)^2}{2\sqrt{2}} - \frac{\left(x-\frac{\pi}{4}\right)^3}{6\sqrt{2}} + \frac{\left(x-\frac{\pi}{4}\right)^4}{24\sqrt{2}} + \frac{\left(x-\frac{\pi}{4}\right)^5}{120\sqrt{2}} - \frac{\left(x-\frac{\pi}{4}\right)^6}{720\sqrt{2}} - \frac{\left(x-\frac{\pi}{4}\right)^7}{5040\sqrt{2}} + \frac{\left(x-\frac{\pi}{4}\right)^8}{40320\sqrt{2}} + \frac{\left(x-\frac{\pi}{4}\right)^9}{362880\sqrt{2}} + \cdots\cdots$$



3. Theory and Limitation

Taylor Series (Multivariable)

$$h(x,y) = h(x_0,y_0) + \frac{\partial h(x_0,y_0)}{\partial x}(x-x_0) + \frac{\partial h(x_0,y_0)}{\partial y}(y-y_0)$$
$$+\text{something related to } (x-x_0)^2 \text{ and } (y-y_0)^2 + \cdots$$

When x and y are close to x_0 and y_0 then:

$$h(x,y) \approx h(x_0,y_0) + \frac{\partial h(x_0,y_0)}{\partial x}(x-x_0) + \frac{\partial h(x_0,y_0)}{\partial y}(y-y_0)$$

3. Theory and Limitation

Formal Derivation

Based on Taylor series:

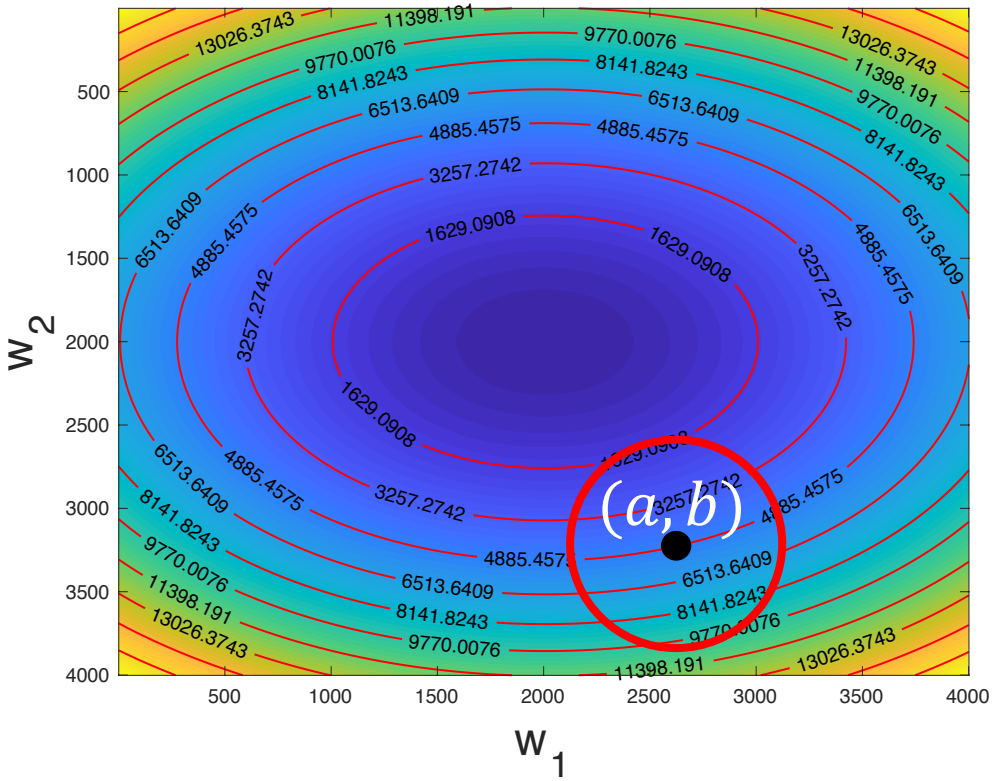
If the red circle is ***small enough***, then in the red circle:

$$L(w) \approx L(a,b) + \frac{\partial L(a,b)}{\partial w_1}(w_1 - a) + \frac{\partial L(a,b)}{\partial w_2}(w_2 - b)$$

$$s = L(a,b)$$

$$u = \frac{\partial L(a,b)}{\partial w_1}, v = \frac{\partial L(a,b)}{\partial w_2}$$

$$L(w) \approx s + u(w_1 - a) + v(w_2 - b)$$



3. Theory and Limitation

Formal Derivation

Based on Taylor series:

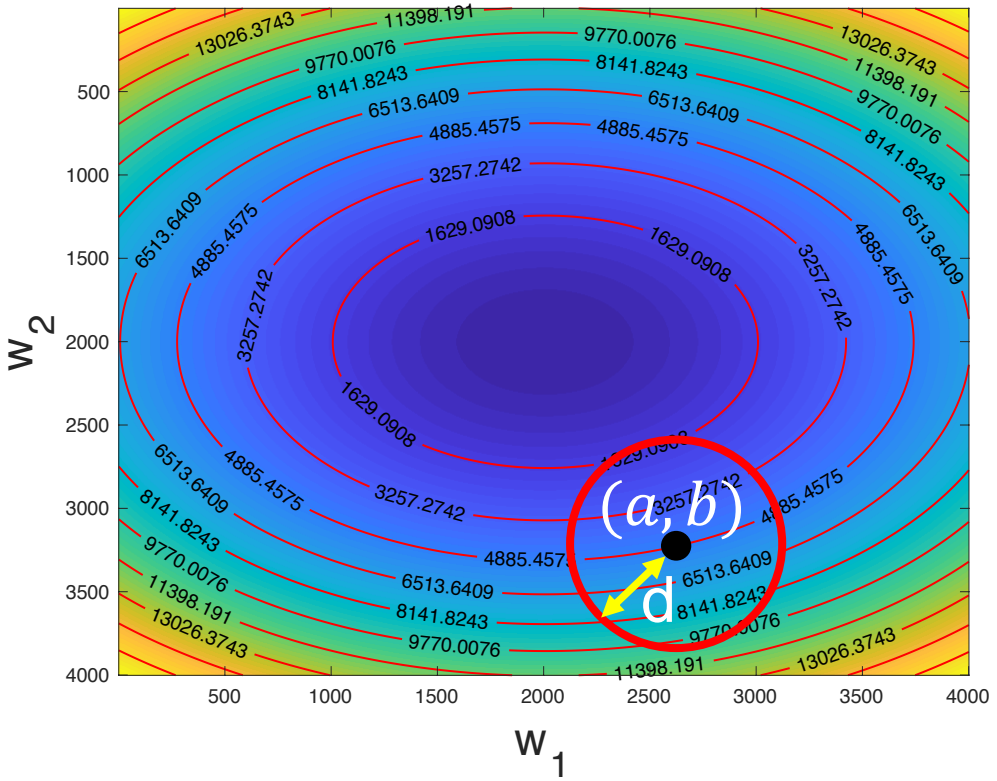
If the red circle is ***small enough***, then in the red circle:

$$L(w) \approx s + u(w_1 - a) + v(w_2 - b)$$

$$\text{where } s = L(a,b), u = \frac{\partial L(a,b)}{\partial w_1}, v = \frac{\partial L(a,b)}{\partial w_2}$$

Find w_1 and w_2 in the red circle **minimizing** $\boldsymbol{L(w)}$

$$(w_1 - a)^2 + (w_2 - b)^2 \leq d^2$$



3. Theory and Limitation

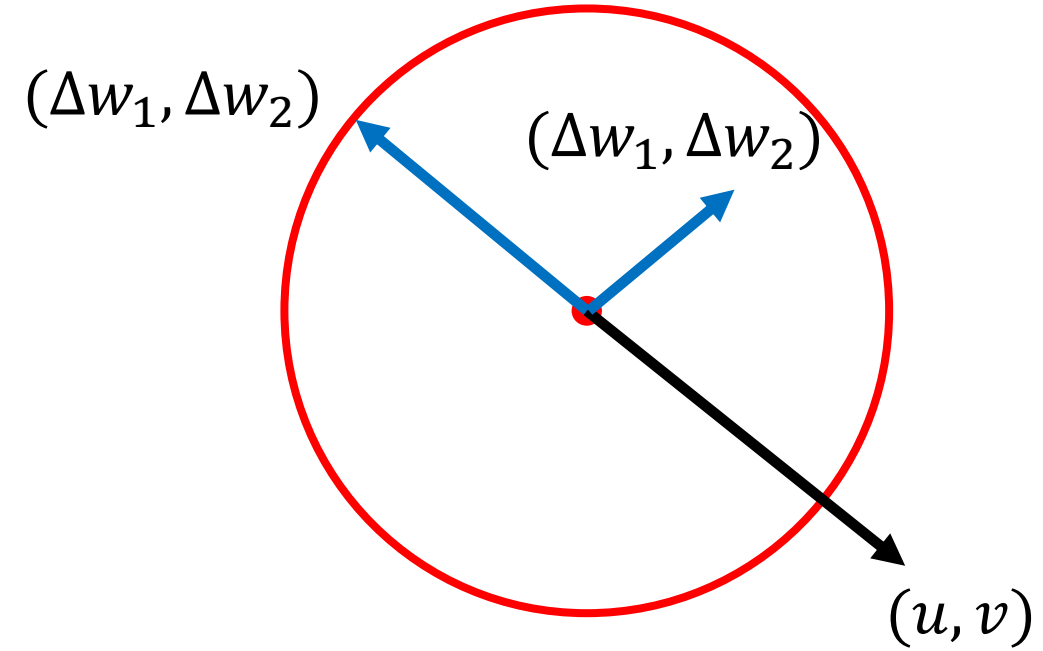
Gradient descent (two variables)

Based on Taylor series:

If the red circle is ***small enough***, then in the red circle:

$$L(w) \approx s + u(w_1 - a) + v(w_2 - b)$$

Δw_1 Δw_2



Find w_1 and w_2 in the red circle **minimizing $L(w)$**

$$(\Delta w_1)^2 + (\Delta w_2)^2 \leq d^2$$

To minimizing $L(w)$

$$\begin{bmatrix} \Delta w_1 \\ \Delta w_2 \end{bmatrix} = -\eta \begin{bmatrix} u \\ v \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} - \eta \begin{bmatrix} u \\ v \end{bmatrix}$$

3. Theory and Limitation

Gradient descent (two variables)

Based on Taylor series:

If the red circle is ***small enough***, then in the red circle:

$$L(w) \approx s + u(w_1 - a) + v(w_2 - b)$$

Constants

where $s = L(a,b), u = \frac{\partial L(a,b)}{\partial w_1}, v = \frac{\partial L(a,b)}{\partial w_2}$

Find w_1 and w_2 yielding the smallest value of $L(w)$ in the circle

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} - \eta \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} - \eta \begin{bmatrix} \frac{\partial L(a,b)}{\partial w_1} \\ \frac{\partial L(a,b)}{\partial w_2} \end{bmatrix}$$

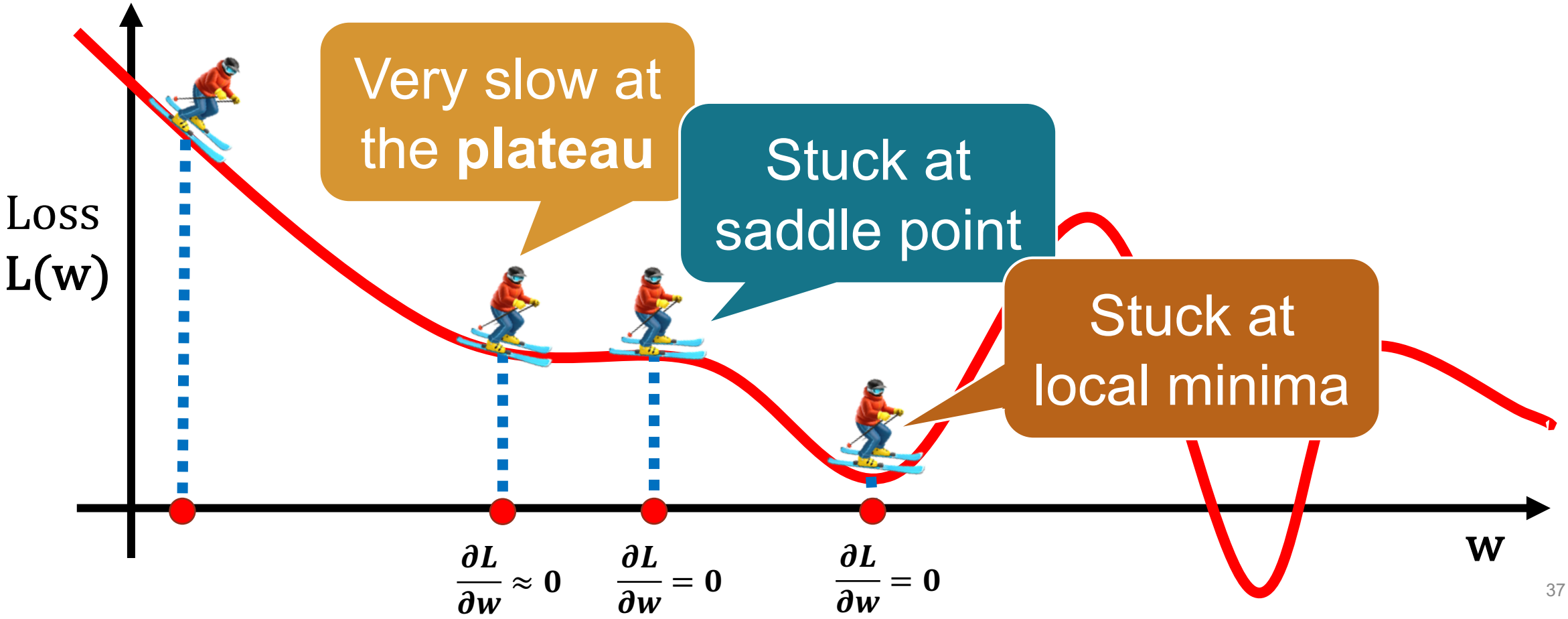
This is the gradient descent.

ONLY satisfied if the red circle (learning rate) is small enough !

3. Theory and Limitation

Limitation of Gradient descent

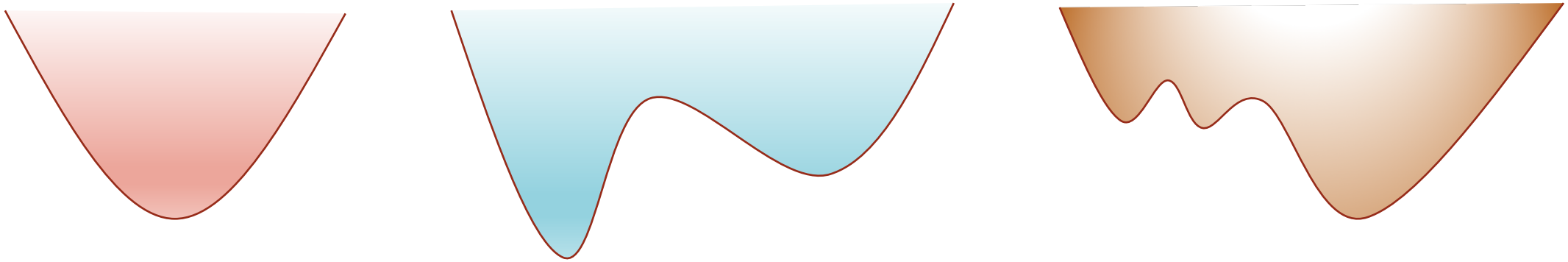
Previous Lecture



3. Theory and Limitation

Convex Functions

- Is finding a w with $\nabla L(w) = 0$ good enough?
  - Yes. But only for convex functions.
- A function is **convex** if the area above the function is a convex set.
  - All values between any two points above function stay above function.
- All w with $\nabla L(w) = 0$ for convex functions are global minima ?



Convex Functions

How do we know if a function is convex?

3. Theory and Limitation

Convex Functions

Some useful tricks for showing a function is convex:

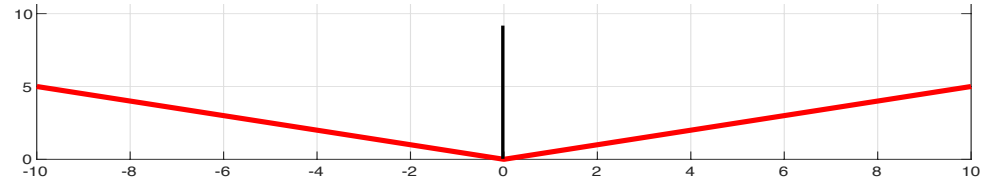
- 1-variable, **twice-differentiable function is convex if $L''(w) > 0$** for all w.

$$y = ax^2 + bx + c$$

$a > 0$

$$\frac{\partial y}{\partial x} = 2ax + b$$

$$\frac{\partial^2 y}{\partial x^2} = 2a$$



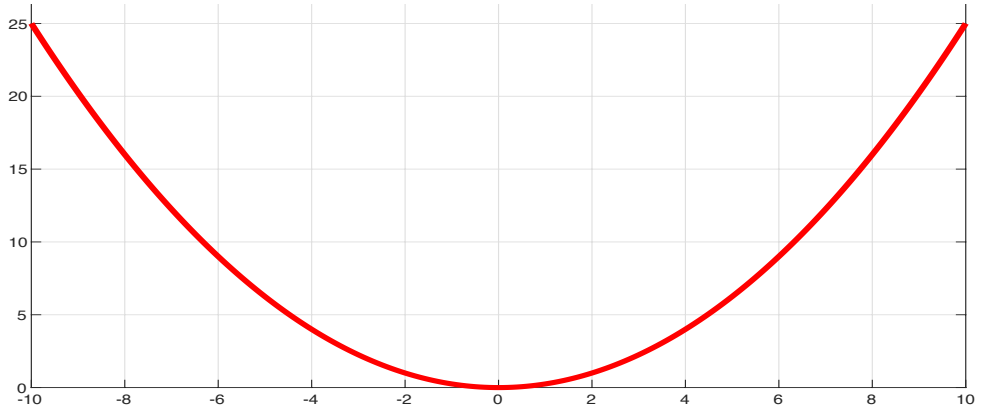
3. Theory and Limitation

Convex Functions

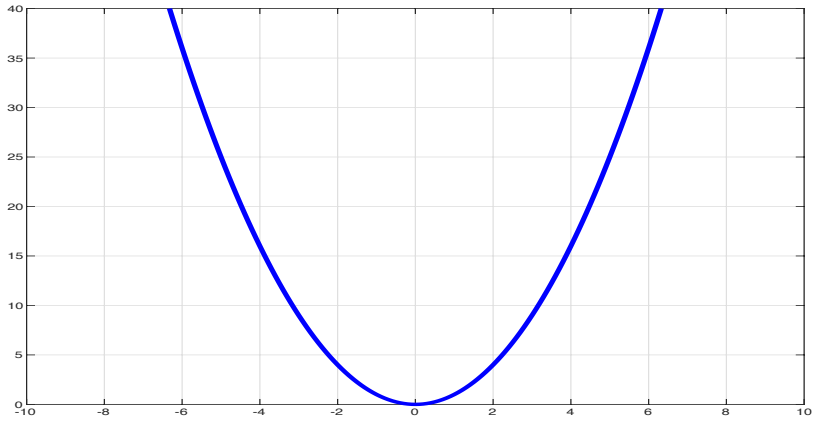
Some useful tricks for showing a function is convex:

- 1-variable, **twice-differentiable function is convex if $L''(w) > 0$** for all w.
- A convex function **multiplied by non-negative constant** is convex.

$$y = ax^2 + bx + c$$



$$y = k * (ax^2 + bx + c), k > 0$$

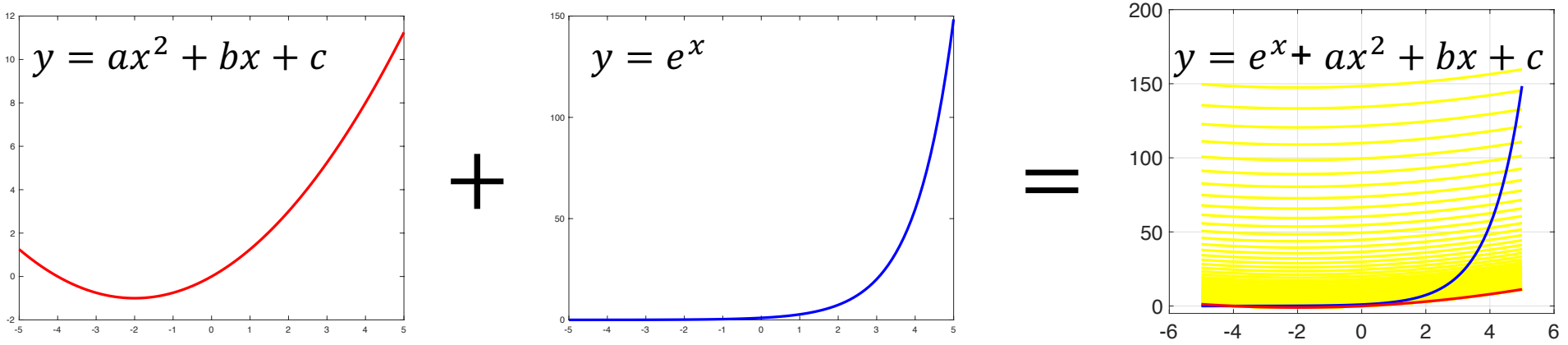


3. Theory and Limitation

Convex Functions

Some useful tricks for showing a function is convex:

- 1-variable, **twice-differentiable function is convex if** $\boldsymbol{L''(w) > 0}$ for all w.
- A convex function **multiplied by non-negative constant** is convex.
- **Norms** and **squared norms** are convex. E.g. $||w_i||$, $||w_i||^2$.
- The **sum of convex functions** is a convex function.



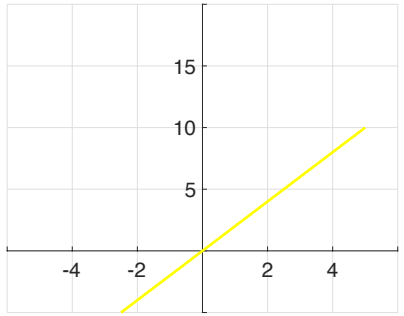
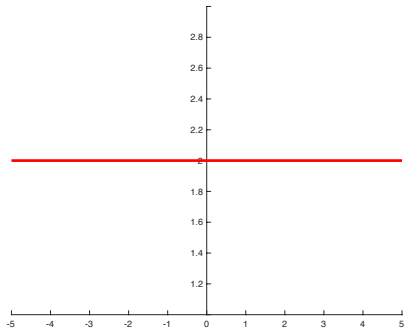
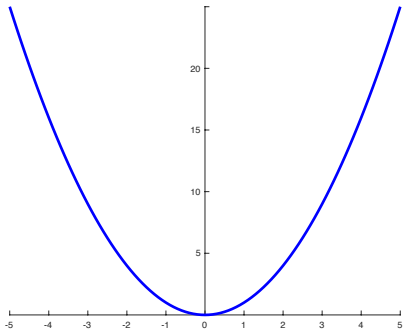
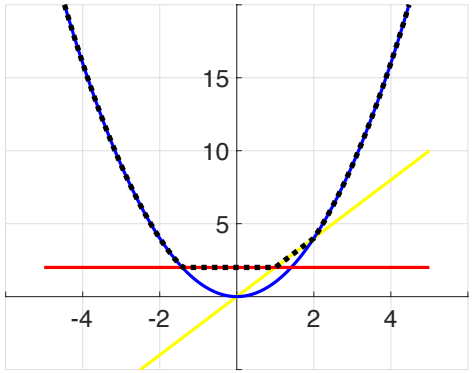
3. Theory and Limitation

Convex Functions

Previous Lecture

Some useful tricks for showing a function is convex:

- 1-variable, **twice-differentiable function is convex if** $\boldsymbol{L''(w) > 0}$ for all w.
- A convex function **multiplied by non-negative constant** is convex.
- **Norms** and **squared norms** are convex. E.g. $||w_i||$, $||w_i||^2$.
- The **sum of convex functions** is a convex function.
- The **max of convex functions** is a convex function. E.g. $f(x) = max\{x^2, 2x, 2\}$.

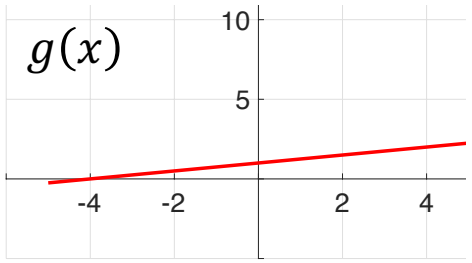
$\boldsymbol{max\{}$    $\boldsymbol{\}} =$ 

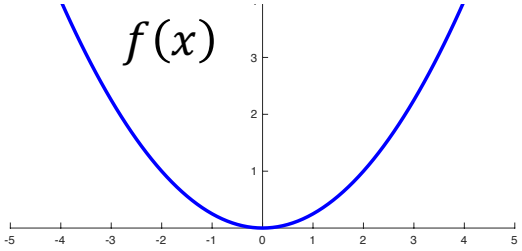
3. Theory and Limitation

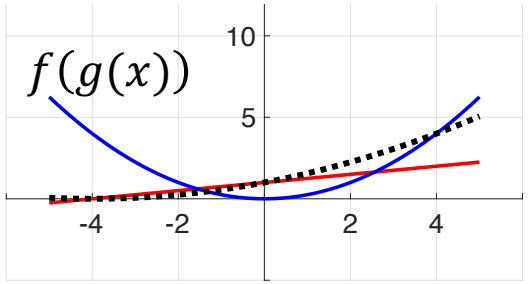
Convex Functions

Some useful tricks for showing a function is convex:

- 1-variable, **twice-differentiable function is convex if** $\boldsymbol{L''(w) > 0}$ for all w.
- A convex function **multiplied by non-negative constant** is convex.
- **Norms** and **squared norms** are convex. E.g. $||w_i||$, $||w_i||^2$.
- The **sum of convex functions** is a convex function.
- The **max of convex functions** is a convex function. E.g. $f(x) = max\{x^2, 2x, 2\}$.
- **Composition of a convex function and a linear function** is convex. E.g. $f(g(x)) = (ax + b)^2$







3. Theory and Limitation

Convex Functions

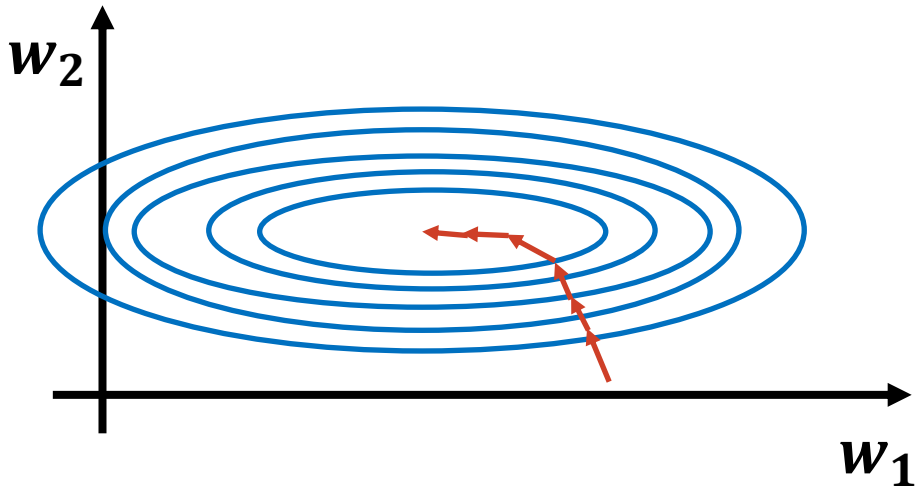
Some useful tricks for showing a function is convex:

- 1-variable, **twice-differentiable function is convex if** $\boldsymbol{L''(w) > 0}$ for all w.
- A convex function **multiplied by non-negative constant** is convex.
- **Norms** and **squared norms** are convex. E.g. $||w_i||$, $||w_i||^2$.
- The **sum of convex functions** is a convex function.
- The **max of convex functions** is a convex function. E.g. $f(x) = max\{x^2, 2x, 2\}$.
- **Composition of a convex function and a linear function** is convex. E.g. $f(g(x)) = (ax + b)^2$
- **Do not true that multiplication of convex functions is convex.**
  - **E.g.** $f(x) = x$, $g(x) = x^2$, $f(x) * g(x) = x^3$.

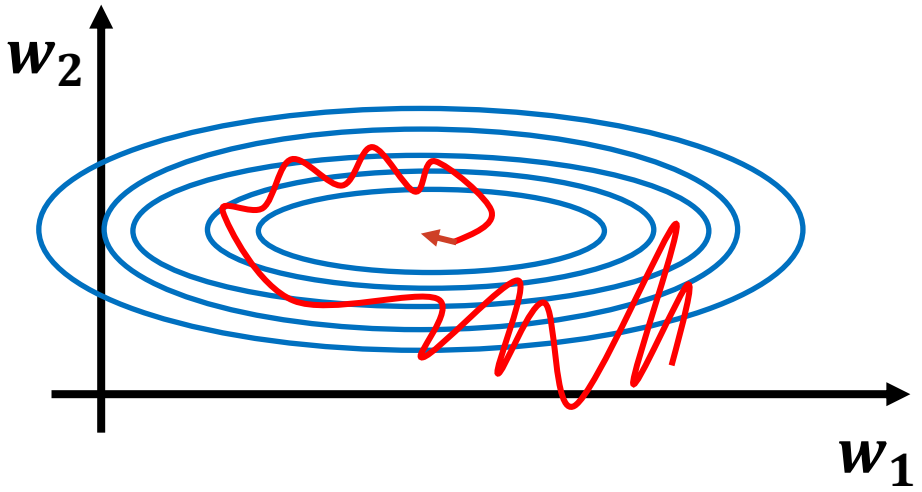
3. Theory and Limitation

Limitation of Stochastic Gradient descent

Gradient descent



Stochastic gradient descent

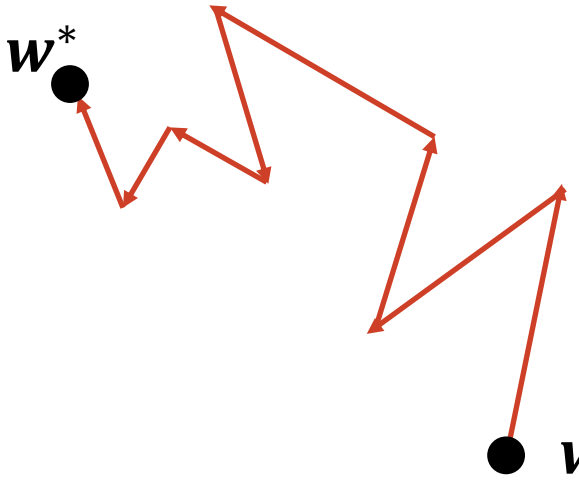


Gradient of random sample might point in the wrong direction.

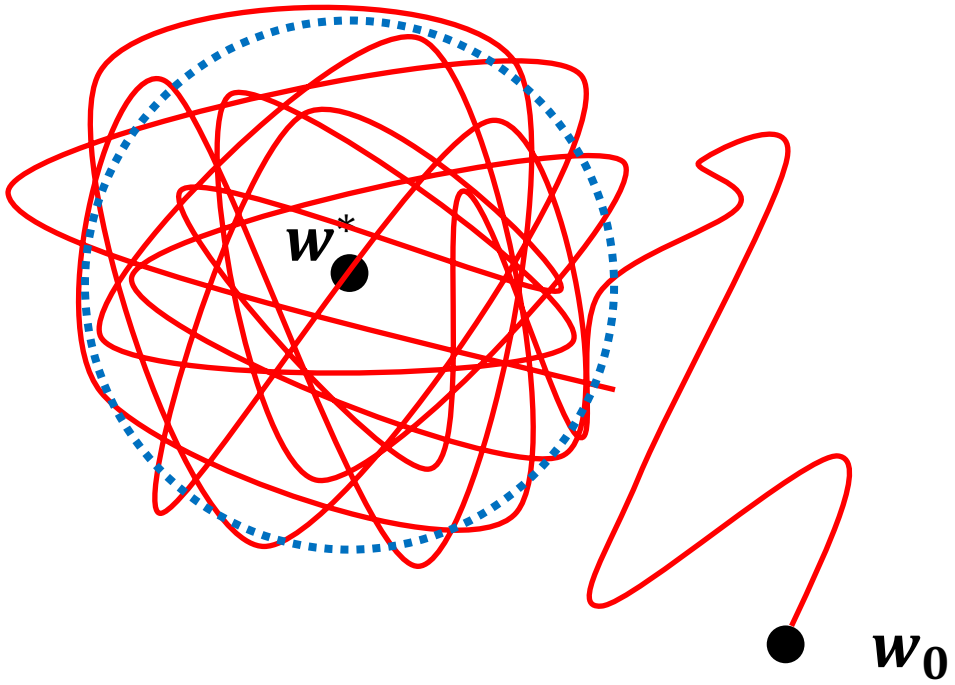
3. Theory and Limitation

Limitation of Stochastic Gradient descent

Gradient descent



Stochastic gradient descent



Erratic behavior confined to a “circle” around solution. The radius of the circle is proportional to the step-size.

Gradient Descent

Hao, Qi
School of Astronomy and Space Science

THANKS