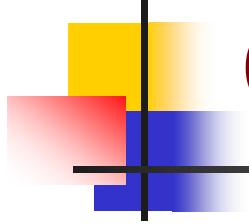




## MORE ON PERCEPTRON LEARNING

INT301 Bio-computation, Week 3, 2025





# Outline

- **Gradient Descent Rule**

# The Gradient Descent Rule

- Perceptron rule fails **if data is not linearly separable** ↗ GD 不管数据是否线性可分都在空间内寻找“误差最小”
- **Idea:** uses **gradient descent** to search the **hypothesis space** ↗ 阶跃函数  $H(\cdot)$  不可微 (要么0要么不存在)
  - perceptron rule cannot be used (not differentiable)
  - hence, an unthresholded linear unit is an appropriate error measure: ↗ 无阈值单元 却保留  $w_i \cdot x_i$

$$E(w) = \frac{1}{2} \sum_e (y_e - o_e)^2$$

- To understand gradient descent, it is helpful to visualize the entire hypothesis space with
  - all possible weight vectors
  - associated E values

# The Gradient Descent Rule

- The objective is to minimize the following error:

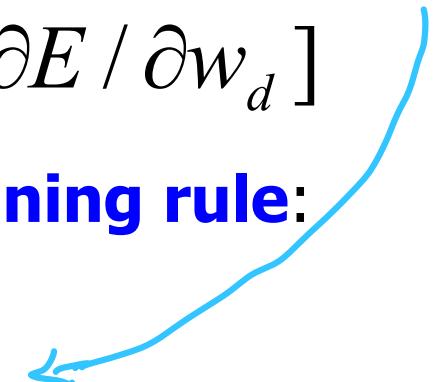
$$E(w) = \frac{1}{2} \sum_e (y_e - o_e)^2$$

- The training is a process of minimizing the error  $E(w)$  in the steepest direction (most rapid decrease), **that is *in direction opposite to the gradient:*** 由于求导代表增长率，所以用减去代表反方向减少。

$$\nabla E(w) = [\underbrace{\partial E / \partial w_0, \partial E / \partial w_1, \dots, \partial E / \partial w_d}]$$

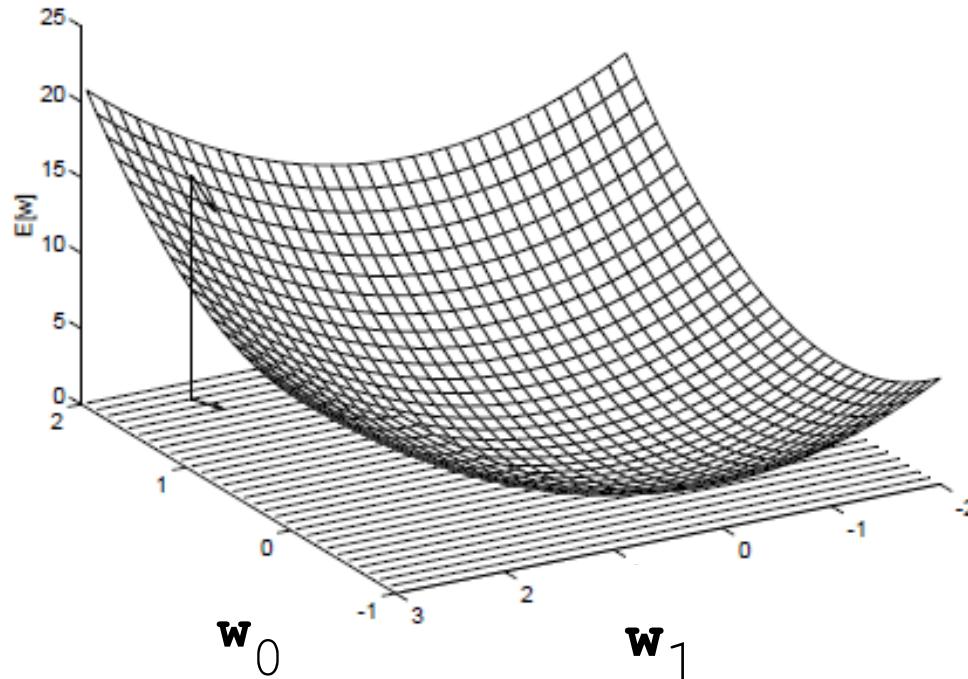
which leads to the **gradient descent training rule:**

$$w_i = w_i - \eta \partial E / \partial w_i$$



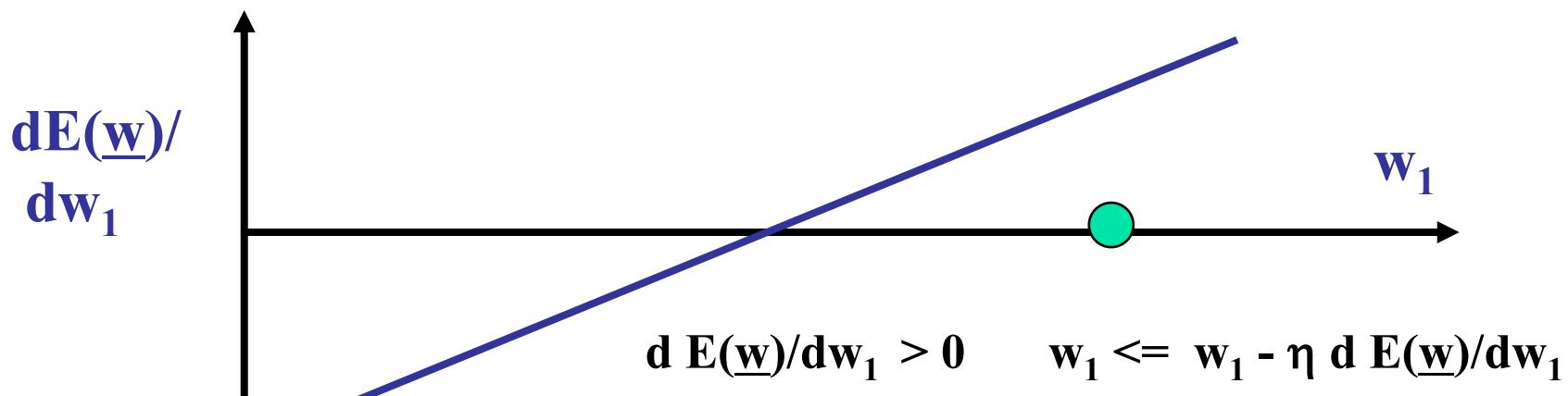
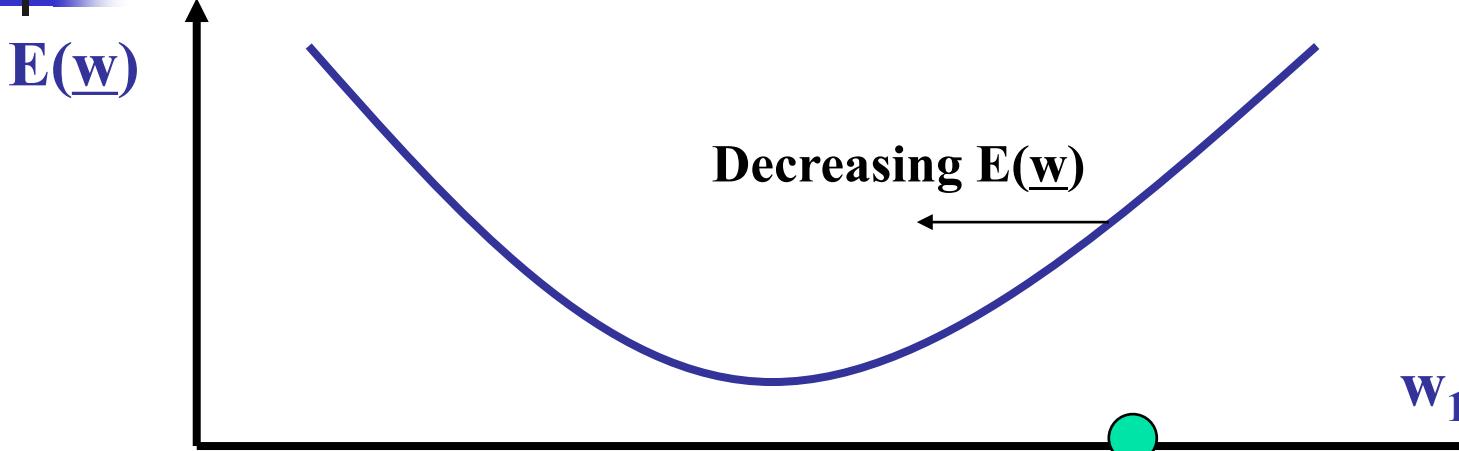
# Error Surface

- the axes  $w_0, w_1$  represent possible values for the two weights of a simple linear unit

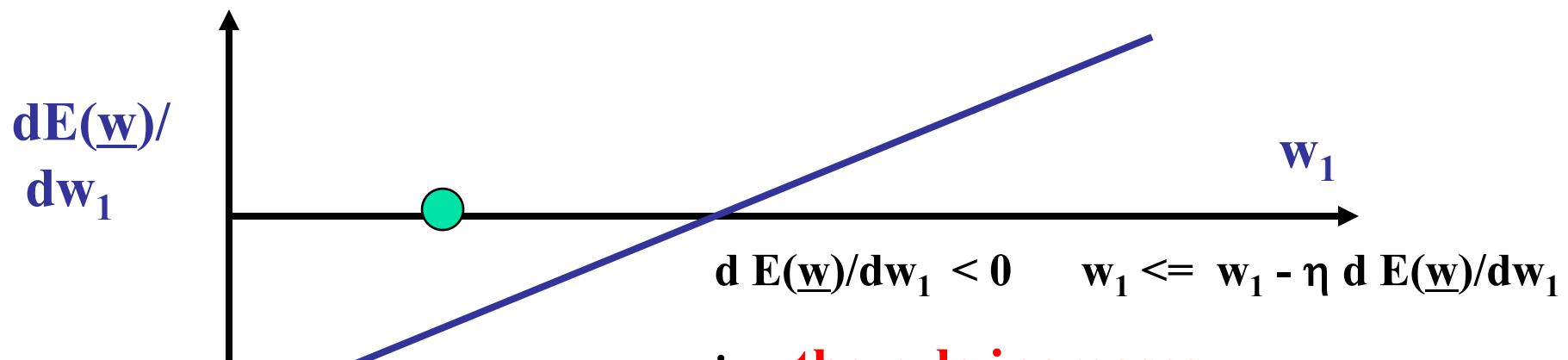
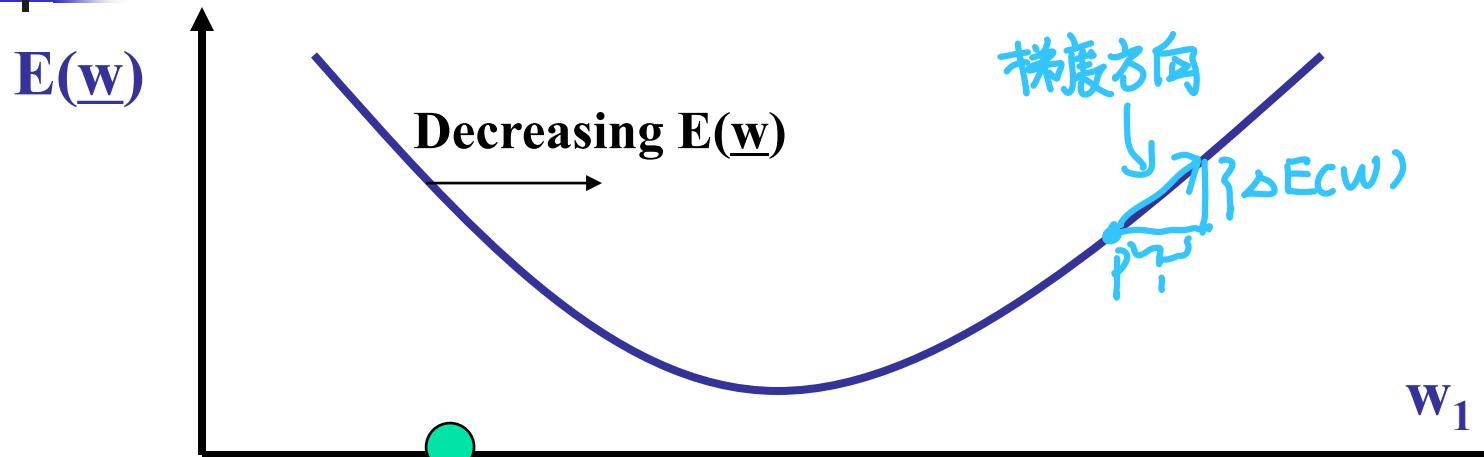


error surface must be **parabolic** with a **single** global minimum

# Moving Downhill: Move in direction of negative derivative

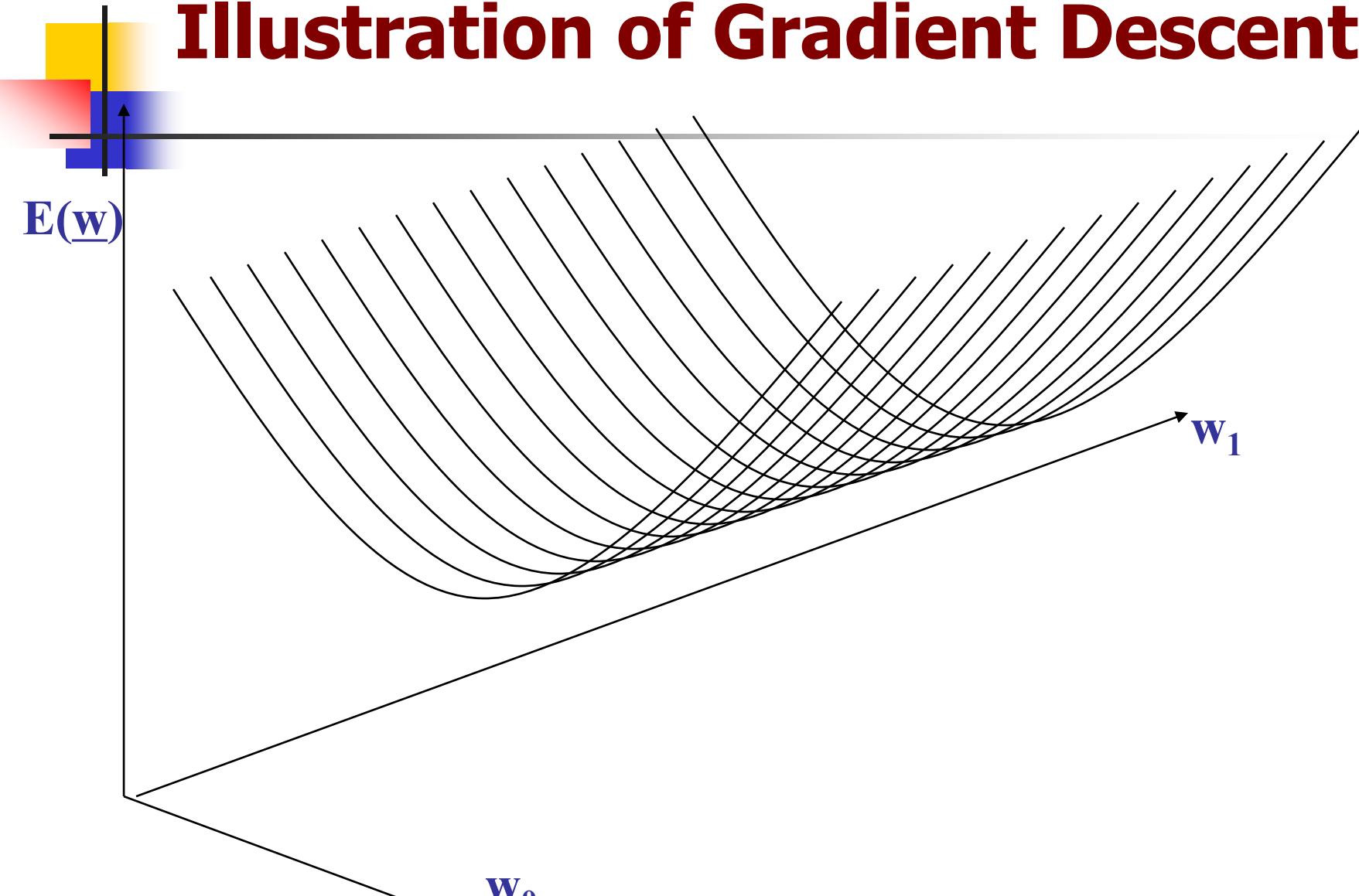


# Moving Downhill: Move in direction of negative derivative

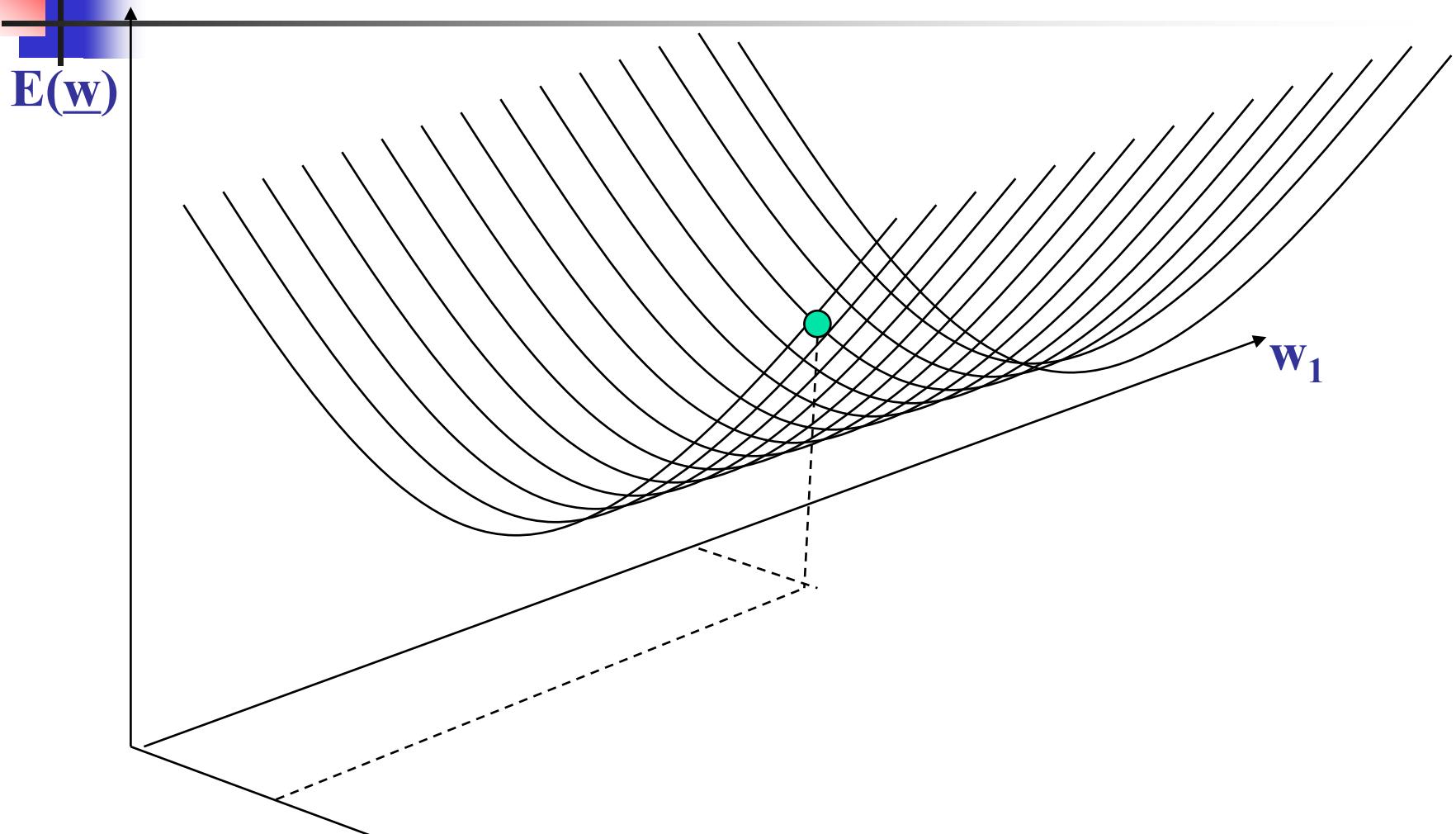




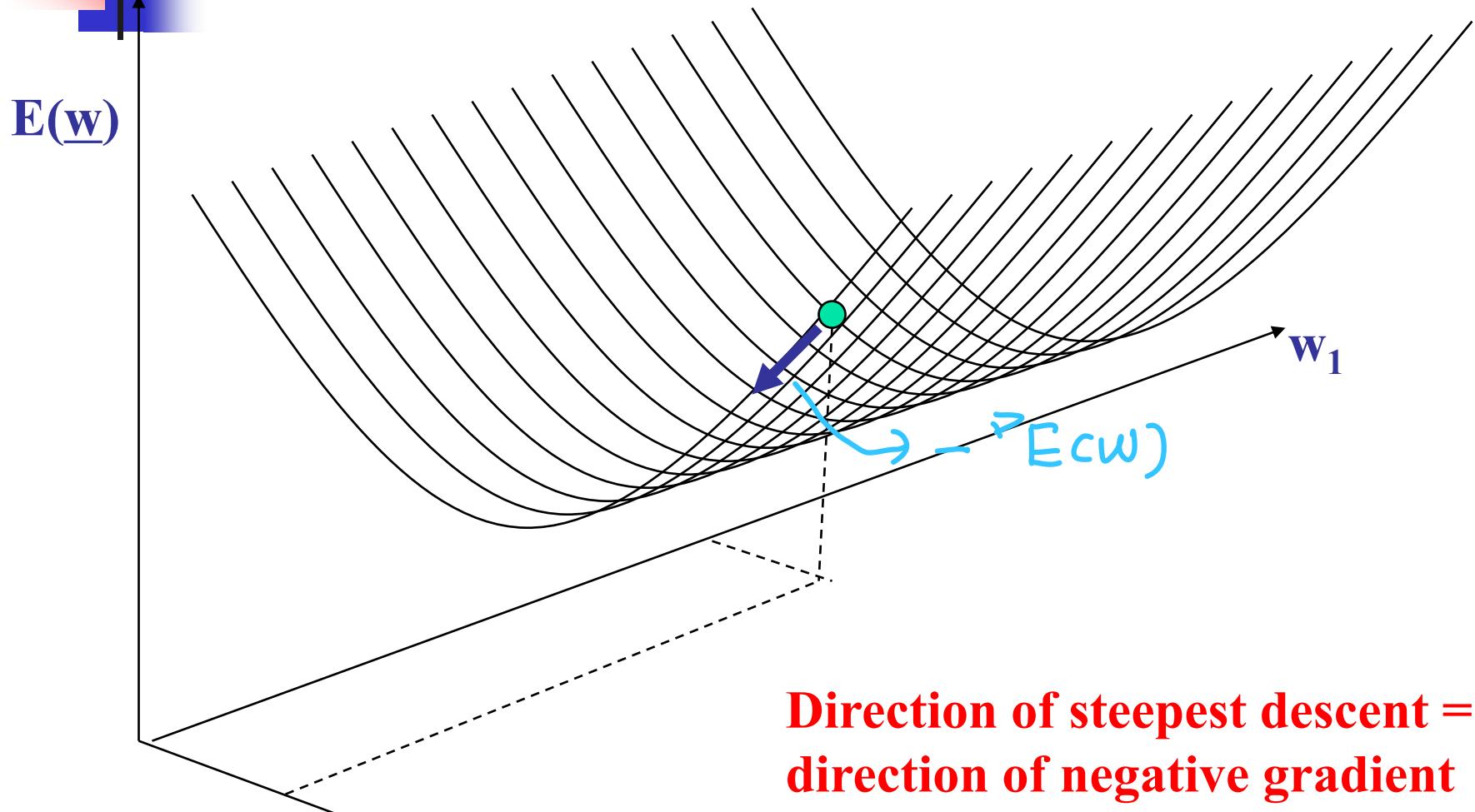
# Illustration of Gradient Descent



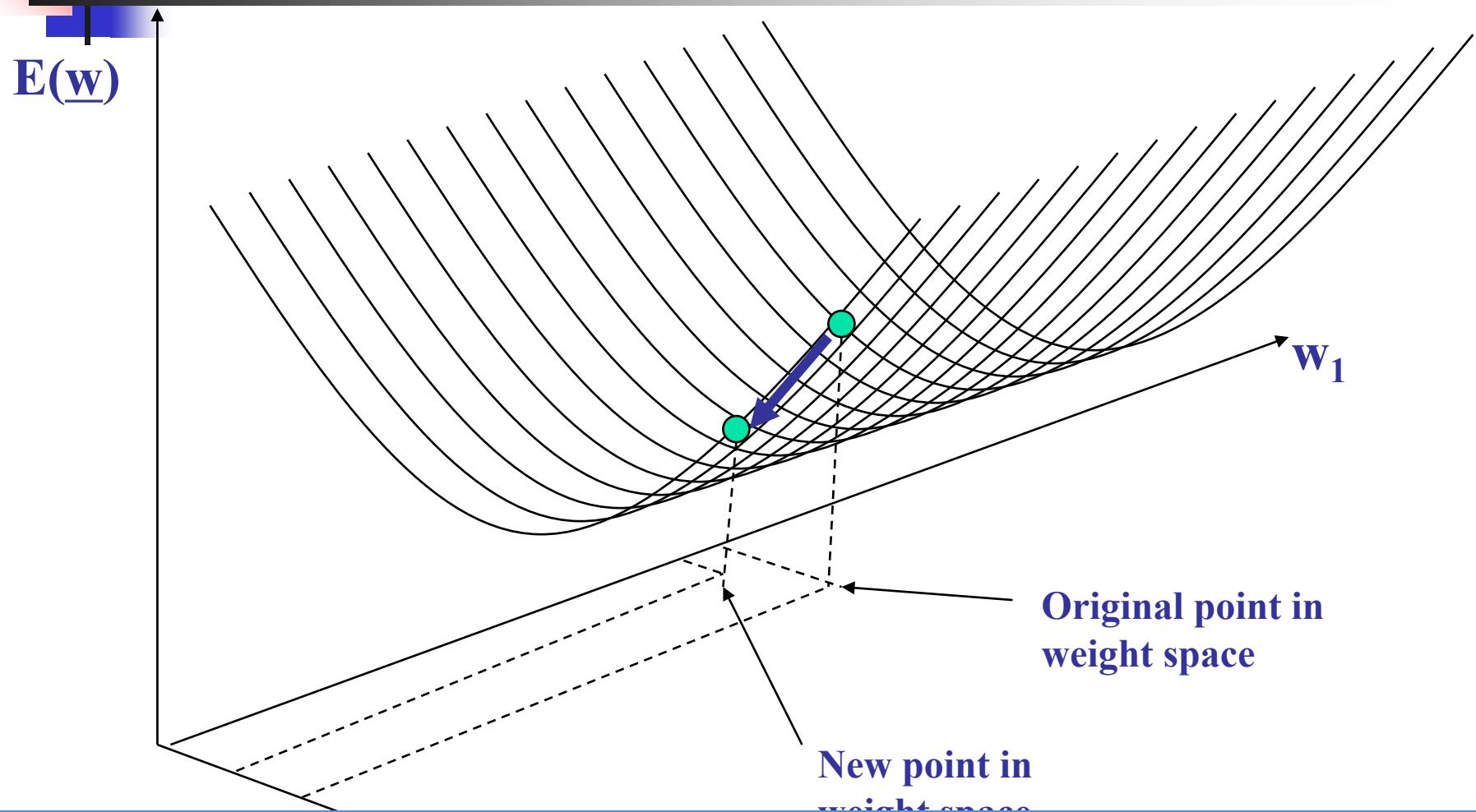
# Illustration of Gradient Descent



# Illustration of Gradient Descent



# Illustration of Gradient Descent



# The Gradient Descent Rule

- The weight update can be derived as follows:

$$\partial E / \partial w_i = \partial \left( \frac{1}{2} \sum_e (y_e - o_e)^2 \right) / \partial w_i$$

$$= \frac{1}{2} \sum_e \partial (y_e - o_e)^2 / \partial w_i$$

$$= \frac{1}{2} \sum_e 2(y_e - o_e) \partial (y_e - o_e) / \partial w_i$$

$$= \sum_e (y_e - o_e) \partial (y_e - w_i x_{ie}) / \partial w_i$$

$$= \sum_e (y_e - o_e) (-x_{ie})$$

$$\Rightarrow w_i = w_i - \eta \frac{\partial E}{\partial w_i}$$

where  $x_{ie}$  denotes the  $i$ -th component of the example  $e$ .

The gradient descent training rule becomes:

$$w_i = w_i + \eta \sum_e (y_e - o_e) x_{ie}$$

$$= w_i + \eta \sum_e (y_e - o_e) x_e$$

# Gradient Descent Learning Algorithm

- *Initialization: Examples  $\{(x_e, y_e)\}_{e=1}^N$ , initial weights  $w_i$  set to small random values, learning rate parameter  $\eta$*
- *Repeat*

*for each training example ( $x_e, y_e$ )*

*- calculate the network output:* 
$$o_e = \sum_{i=0}^d w_i x_{ie}$$

*- if the Perceptron does not respond correctly,  
compute weight corrections:*

$$\Delta w_i = \Delta w_i + \eta(y_e - o_e)x_{ie}$$

注意  $\Delta w$  要全部累加完  
最后再用  $w = w + \Delta w$

*update the weights with the accumulated error from all examples*

$$w_i = w_i + \Delta w_i$$

Gradient  
Descent Rule

*until termination condition is satisfied.*

# Example

- Suppose an example of Perceptron which accepts two inputs  $x_1$  and  $x_2$ , with weights  $w_1 = 0.5$  and  $w_2 = 0.3$  and  $w_0 = -1$ , learning rate = 1.
- Let the example is given:  $x_1 = 2, x_2 = 1, y = 0$   
The network output of the Perceptron is :
- The weight updates according to the **gradient descent algorithm** will be:

$$\begin{aligned}
 & \text{Diagram of Perceptron: } \\
 & \quad \text{Inputs } x_1 \text{ and } x_2 \text{ with weights } w_1 = 0.5 \text{ and } w_2 = 0.3 \text{ connect to a bias node } w_0 = -1. \\
 & \quad \text{The output is calculated as: } 0 = 0.5x_1 + 0.3x_2 - 1. \\
 & \quad \text{Error: } 0 - 0 = 0. \\
 & \quad \text{Learning Rate: } \eta = 1. \\
 & \quad \text{Weight Updates:} \\
 & \quad \Delta w_1 = \eta(0 - 0) * x_1 = 1 * 2 - 0.5 = 0.5 \\
 & \quad \Delta w_2 = \eta(0 - 0) * x_2 = 1 * 1 - 0.3 = 0.3 \\
 & \quad \Delta w_0 = \eta(0 - 0) * 1 = 1 * 1 - 1 = 0
 \end{aligned}$$

$\eta = \gamma$

$$\begin{aligned}
 \Delta w_1 &= \eta(0 - 0.3) * 2 = -0.6 \\
 \Delta w_2 &= \eta(0 - 0.3) * 1 = -0.3 \\
 \Delta w_0 &= \eta(0 - 0.3) * 1 = -0.3
 \end{aligned}$$

# Example

- Let another example is given:  $x_1 = 1, x_2 = 2, y = 1$
- The network output of the Perceptron is :

$$o = 1 * 0.5 + 2 * 0.3 - 1 = 0.1$$

The weight updates according to the gradient descent algorithm will be:

$$\Delta w_1 = -0.6 + (1 - 0.1) * 1 = 0.3$$

$$\Delta w_2 = -0.3 + (1 - 0.1) * 2 = 1.5$$

$$\Delta w_0 = -0.3 + (1 - 0.1) * 1 = 0.6$$

If there are no more examples, the weights will be modified as follows:

$$w_1 = 0.5 + 0.3 = 0.8$$

$$w_2 = 0.3 + 1.5 = 1.8$$

$$w_0 = 1 + 0.6 = 0.4$$

# Incremental gradient descent

(增量GD): 每算完一个样本更新一次

传统GD：算完全部梯度更新一次

The gradient descent rule faces two difficulties in practice:

- it converges very slowly
- if there are multiple local minima in the error surface, then there is no guarantee that it will find the global minimum

- That is why, a **stochastic version** called *incremental gradient descent* rule is developed to overcome these difficulties.  
*Whereas the gradient descent rule updates the weights after calculating the whole error accumulated from all examples, the incremental version approximates the gradient descent error decrease by updating the weights after each training example.*

# Incremental gradient descent

- Incremental gradient descent is implemented

$$w_i = w_i + \eta(y_e - o_e)x_{ie} \quad \text{where} \quad o_e = \sum_{i=0}^d w_i x_{ie}$$

少了求和符号

## Incremental Gradient Descent Learning Algorithm

**Initialization:** Examples  $\{(x_e, y_e)\}_{e=1}^N$ , initial weights  $w_i$  set to small random values, learning rate parameter  $\eta$

## *Repeat*

 *for each training example ( $x_e, y_e$ )*

- *calculate the network output:*

$$o_e = \sum_{i=0}^d w_i x_{ie}$$

- *if the Perceptron does not respond correctly update the weights:*

$$w_i = w_i + \eta (y_e - o_e) x_{ie}$$

*until termination condition is satisfied.*

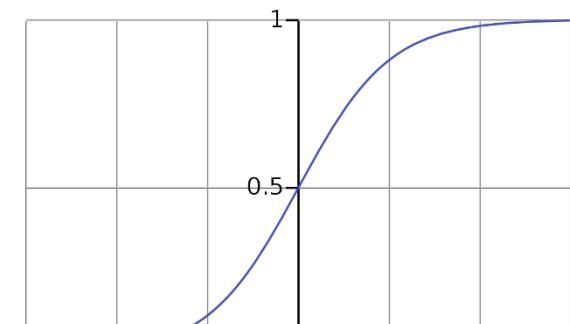
# Sigmoidal Perceptrons

- The simple single-layer Perceptrons with threshold or linear activation functions are not generalizable to more powerful learning mechanisms like multilayer neural networks.
  - That is why, single-layer Perceptrons with sigmoidal activation functions are developed.
- The sigmoidal Perceptron produces output:

$$o = \sigma(S) = \frac{1}{1 + e^{-S}},$$

where:

$$S = \sum_{i=1}^d w_i x_i$$



# Training Sigmoidal Perceptrons

- The gradient descent rule for training sigmoidal Perceptrons is again:

$$w_i = w_i - \eta \partial E / \partial w_i$$

- The difference is in the error derivative  $\partial E / \partial w_i$  which due to the use of the sigmoidal function  $\sigma(s)$  becomes:

$$\partial E / \partial w_i = \partial ((\frac{1}{2}) \sum_e (y_e - o_e)^2) / \partial w_i$$

$$= (\frac{1}{2}) \sum_e \partial (y_e - o_e)^2 / \partial w_i$$

$$= (\frac{1}{2}) \sum_e 2(y_e - o_e) \partial (y_e - o_e) / \partial w_i$$

$$= \sum_e (y_e - o_e) \partial (y_e - \sigma(s)) / \partial w_i$$

$$= \sum_e (y_e - o_e) \sigma'(s) (-x_{ie})$$

# Training Sigmoidal Perceptrons

- The *Gradient descent training rule* for training

sigmoidal Perceptrons is:

$$\frac{d \text{GCS})}{ds} = \frac{d}{ds} \frac{1}{1+e^{-s}}$$
$$= -C(1+e^{-s})^{-2} \cdot$$

$$w_i = w_i + \eta \sum_e (y_e - o_e) \sigma'(S) x_{ie}$$

$$\frac{d}{ds} C(1+e^{-s})$$
$$= +e^{-s} \cdot +\frac{1}{(1+e^{-s})^2}$$
$$= \frac{1}{1+e^{-s}} \cdot \frac{e^{-s}}{1+e^{-s}}$$
$$= GCS(1 - GCS)$$

where:

$$\sigma'(S) = \sigma(S)(1 - \sigma(S))$$



# Gradient Descent Learning Algorithm for Sigmoidal Perceptrons

- *Initialization: Examples  $\{(x_e, y_e)\}_{e=1}^N$ , initial weights  $w_i$  set to small random values, learning rate parameter  $\eta$*
- *Repeat*
  - for each training example  $(x_e, y_e)$* 
    - calculate the network output:  $o = \sigma(s)$  where  $s = \sum_{i=0}^d w_i x_{ie}$*
    - if the Perceptron does not respond correctly compute weight corrections:*

$$\Delta w_i = \Delta w_i + \eta(y_e - o_e)\sigma(s)(1 - \sigma(s))x_{ie}$$

*update the weights with the accumulated error from all examples*  $w_i = w_i + \Delta w_i$

# Example

$$\begin{array}{r} \text{1 } x_1 \text{ } \cancel{0.3} \text{ } 0 \\ \text{1 } -1 \\ \hline y = 1 \end{array}$$

- Suppose an example of Perceptron which accepts two inputs  $x_1$  and  $x_2$ , with weights  $w_1 = 0.5$  and  $w_2 = 0.3$  and  $w_0 = -1$ , learning rate = 1.
- Let the following example is given:  $x_1 = 2, x_2 = 1, y = 0$

The output of the Perceptron is :

$$O = \sigma(-1 + 2 * 0.5 + 1 * 0.3) = \sigma(0.3) = 0.5744$$

- The weight updates according to the gradient descent algorithm will be:  $(y - 0e) * 6cs * (1 - 6cs) * x_i$
- $$\Delta w_0 = (0 - 0.5744) * 0.5744 * (1 - 0.5744) * 1 = -0.1404$$
- $$\Delta w_1 = (0 - 0.5744) * 0.5744 * (1 - 0.5744) * 2 = -0.2808$$
- $$\Delta w_2 = (0 - 0.5744) * 0.5744 * (1 - 0.5744) * 1 = -0.1404$$

# Example

Let another example is given:  $x_1 = 1, x_2 = 2, y = 1$

The output of the Perceptron is :

$$O = \sigma(-1 + 1 * 0.5 + 2 * 0.3) = \sigma(0.1) = 0.525$$

The weight updates according to the gradient descent algorithm will be:

$$\Delta w_0 = -0.1404 + (1 - 0.525) * 0.525 * (1 - 0.525) * 1 = -0.0219$$

$$\Delta w_1 = -0.2808 + (1 - 0.525) * 0.525 * (1 - 0.525) * 1 = -0.1623$$

$$\Delta w_2 = -0.1404 + (1 - 0.525) * 0.525 * (1 - 0.525) * 2 = 0.0966$$

If there are no more examples in the batch, the weights will be modified as follows:

$$w_0 = -1 + (-0.0219) = -1.0219$$

$$w_1 = 0.5 + (-0.1623) = 0.3966$$

$$w_2 = 0.3 + 0.0966 = 0.3966$$

# Incremental Gradient Descent Learning

## Algorithm for Sigmoidal Perceptrons

### ***Initialization:***

Examples  $\{(x_e, y_e)\}$ , initial weights  $w$ , set to small random values, learning rate parameter  $\eta$

### ***Repeat***

*for each training example ( $x_e, y_e$ )*

- calculate the network output:  $o = \sigma(s)$*   
*where  $s = \sum_{i=0}^d w_i X_{ie}$*
- If the Perceptron does not respond correctly update the weights:  $w_i = w_i + \eta(y_e - o_e)\sigma(s)(1 - \sigma(s))x_{ie}$*

*until termination condition is satisfied.*

# Perceptron vs. Gradient Descent

- Gradient descent finds the decision boundary which minimizes the *sum squared error* of the (target - net) value rather than the (target - output) value
  - Perceptron rule will find the decision boundary which minimizes the classification error – *if the problem is linearly separable*
  - Gradient descent decision boundary may leave more instances misclassified as compared to the perceptron rule: could have a higher misclassification rate than with the perceptron rule
- Perceptron rule (target - thresholded output) guaranteed to converge to a separating hyperplane if the problem is linearly separable.

# The error surface

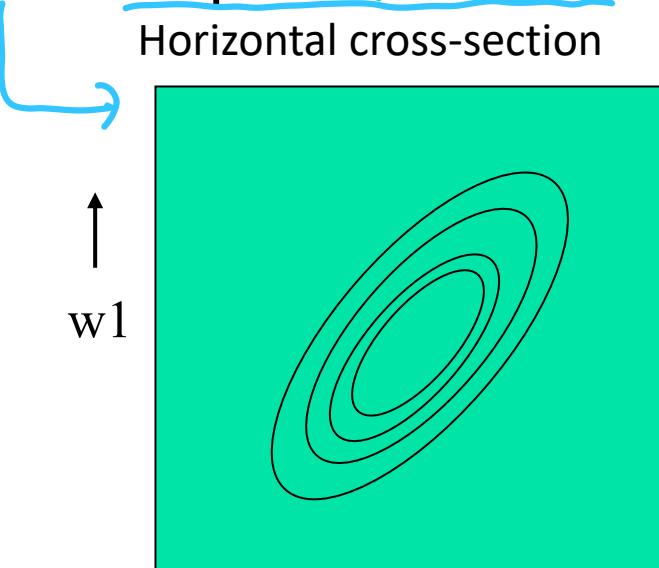
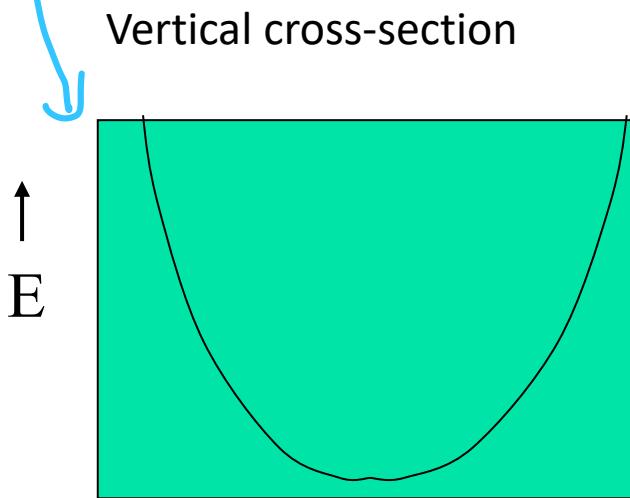
误差平面

- The error surface lies in a space with a horizontal axis for each weight and one vertical axis for the error.

$E = \frac{1}{2} (y - \hat{y})^2$  展开

$$w_1^2 x_1^2 + w_2^2 x_2^2 + \dots$$

- For a linear neuron, it is a quadratic bowl.
- Vertical cross-sections are parabolas. (抛物线)
- Horizontal cross-sections are ellipses. (椭圆形)

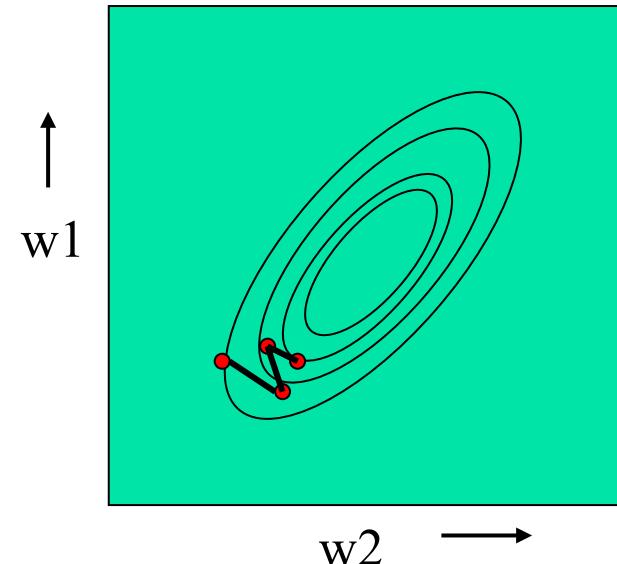
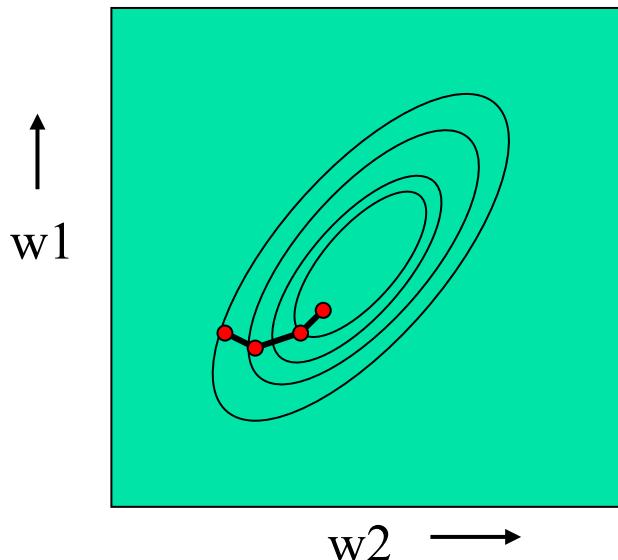


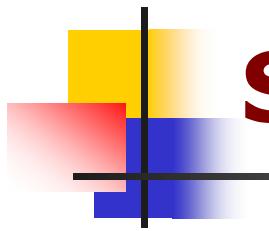
## All Examples

11

# Batch vs incremental learning

- Batch learning does steepest descent on the error surface
- Incremental learning zig-zags around the direction of steepest descent





# Summary

- ◆ **Perceptron training:**

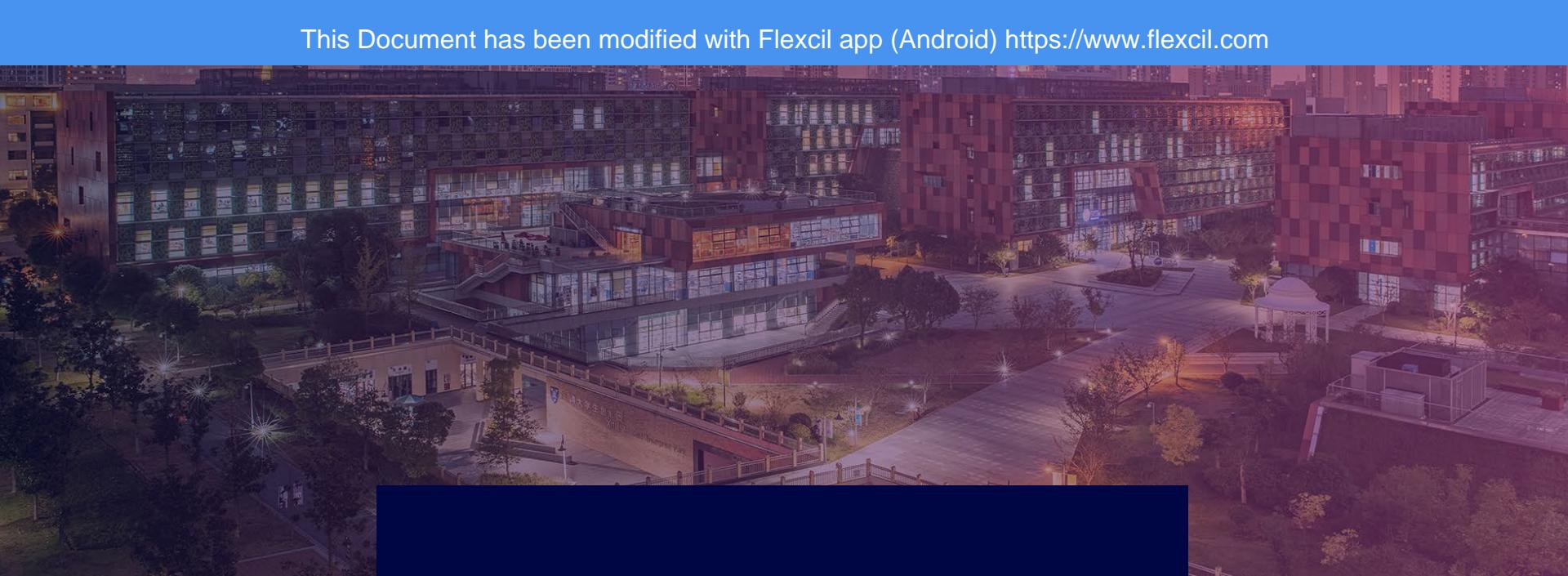
- uses thresholded unit
- converges after a finite number of iterations
- output hypothesis classifies training data perfectly
- linearly separability necessary

- ◆ **Gradient descent:**

- uses unthresholded linear unit
- converges asymptotically toward a minimum error hypothesis
- termination is not guaranteed 训练结束无法得到保证
- linear separability not necessary

# The fall of the Perceptron

- Researchers begun to discover the Perceptron's limitations.
- Unless input categories were "linearly separable", a perceptron could not learn to discriminate between them.
- Unfortunately, it appeared that many important categories were not linearly separable.
- E.g., those inputs to an XOR gate that give an output of 1 (namely 10 & 01) are not linearly separable from those that do not (00 & 11).



# THANK YOU



**VISIT US**

[WWW.XJTLU.EDU.CN](http://WWW.XJTLU.EDU.CN)



**FOLLOW US**

@XJTLU



Xi'an Jiaotong-Liverpool University  
西交利物浦大学