*NOTES:*

**Week1:**

**Definition**

A computer program is said to learn form experience E with respect to some Task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

一个程序被认为能从经验E中学习, 解决任务T, 达到性能度量值P，当且仅当，有了经验E后, 经过P评判, 程序在处理 T 时的性能有所提升.

**Supervised Learning**

In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output.

Supervised learning problems are categorized into "regression" and "classification" problems.

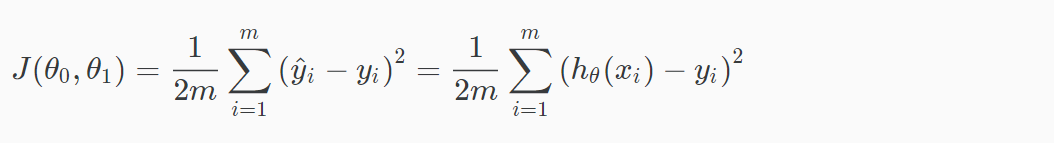
**Unsupervised Learning**

Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive (得到) structure from data where we don't necessarily know the effect of the variables. We can derive this structure by clustering the data based on relationships among the variables in the data.

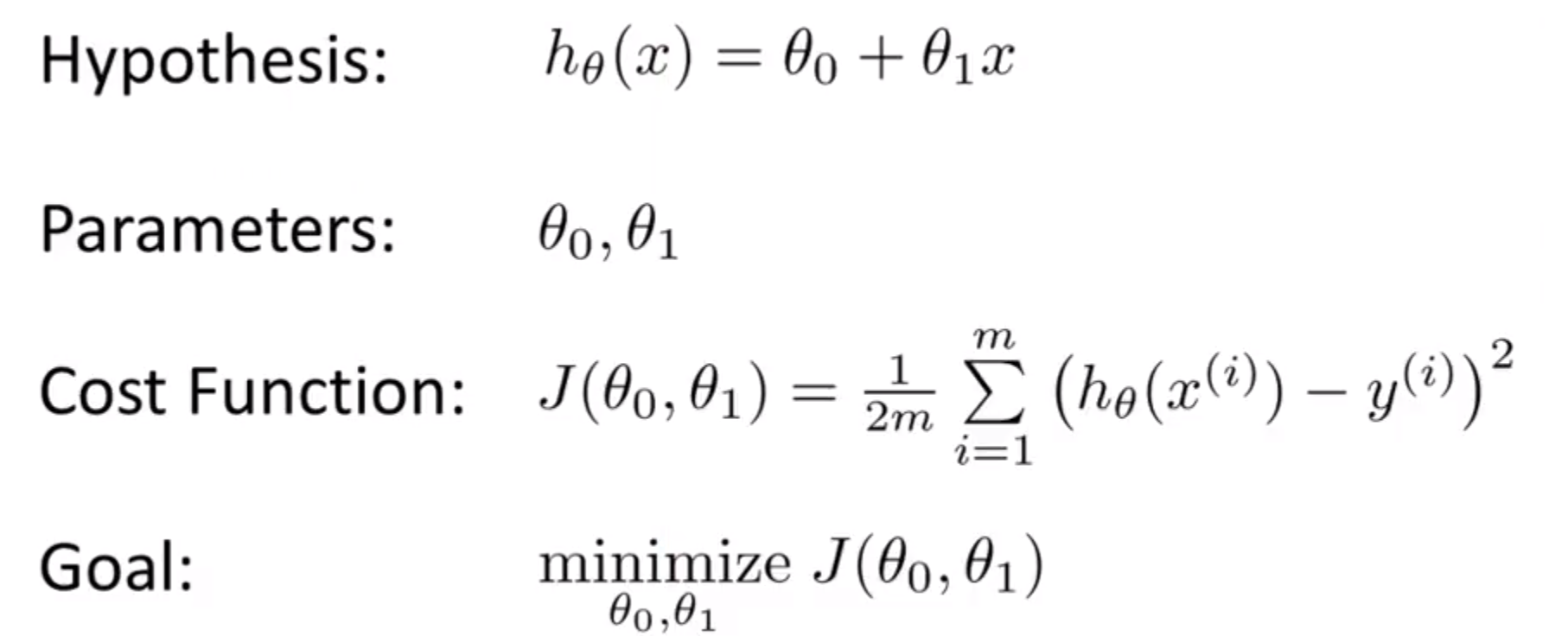
With unsupervised learning there is no feedback based on the prediction results.

**Learning Regression**

Cost function:



The 1/2 is as a convenience for the computation of the gradient descent, as the derivative term of the square function will cancel out the 1/2 term.

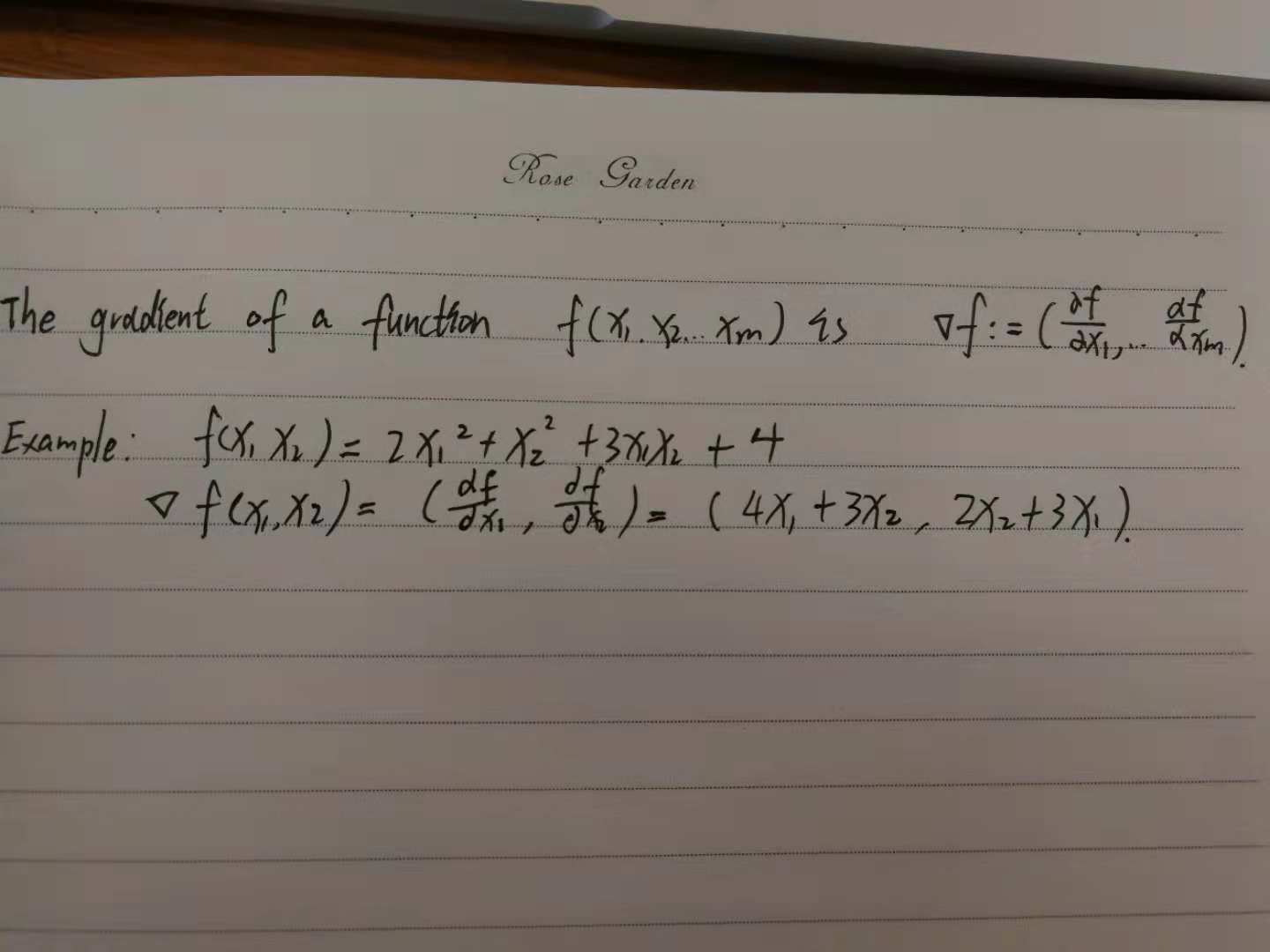


In order to minimize cost function, using gradient descent algorithm.

Hint: Difference between =(判断为真的声明) and :=(赋值)

**Gradient descent algorithm**

Definition of gradient (From neural computation course):

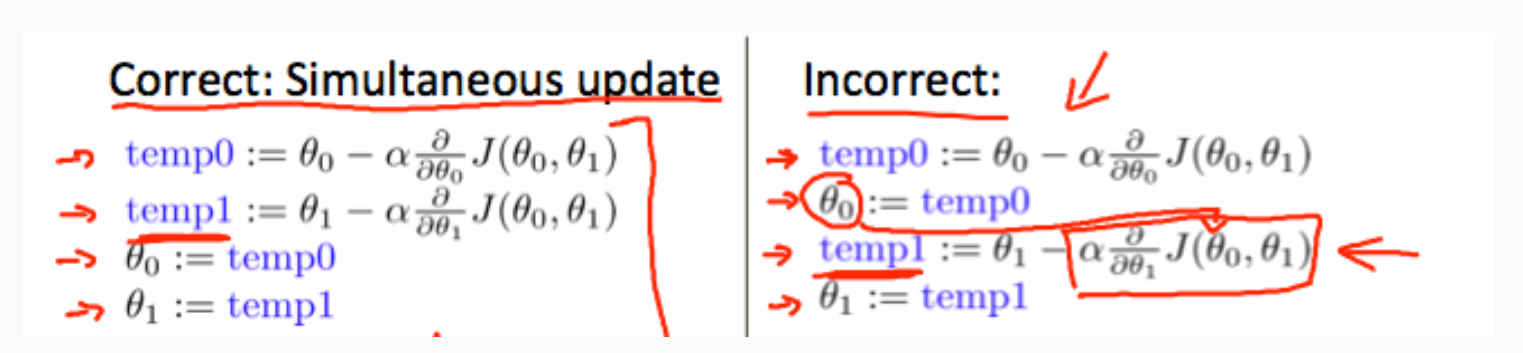


标量场中某一点的梯度指向在这点标量场增长最快的方向

The gradient descent algorithms show that where j = 0,1 and is learning rate. Its needs to repeat until convergence.



The parameters need to update simultaneously because >>



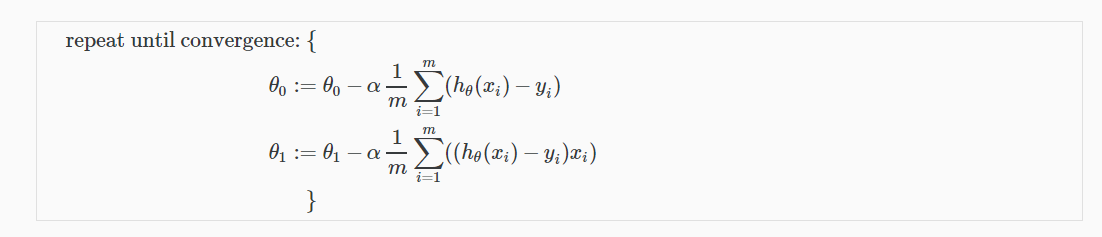
If learning rate is too small, gradient descent can be slow

If leaning rate is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge(偏离，相异).

When point approach a local minimum, gradient descent will automatically take smaller steps because the slope(斜坡，倾斜，斜率) will small.

**Gradient descent for linear regression**

事实证明The cost function for linear regression is always going to be a bow shaped function (凸函数convex function). It does not have any local optimal except for one global optimum.



Batch gradient descent refers to each step of gradient descent uses all the training examples. Using m.

**Linear Algebra Review**

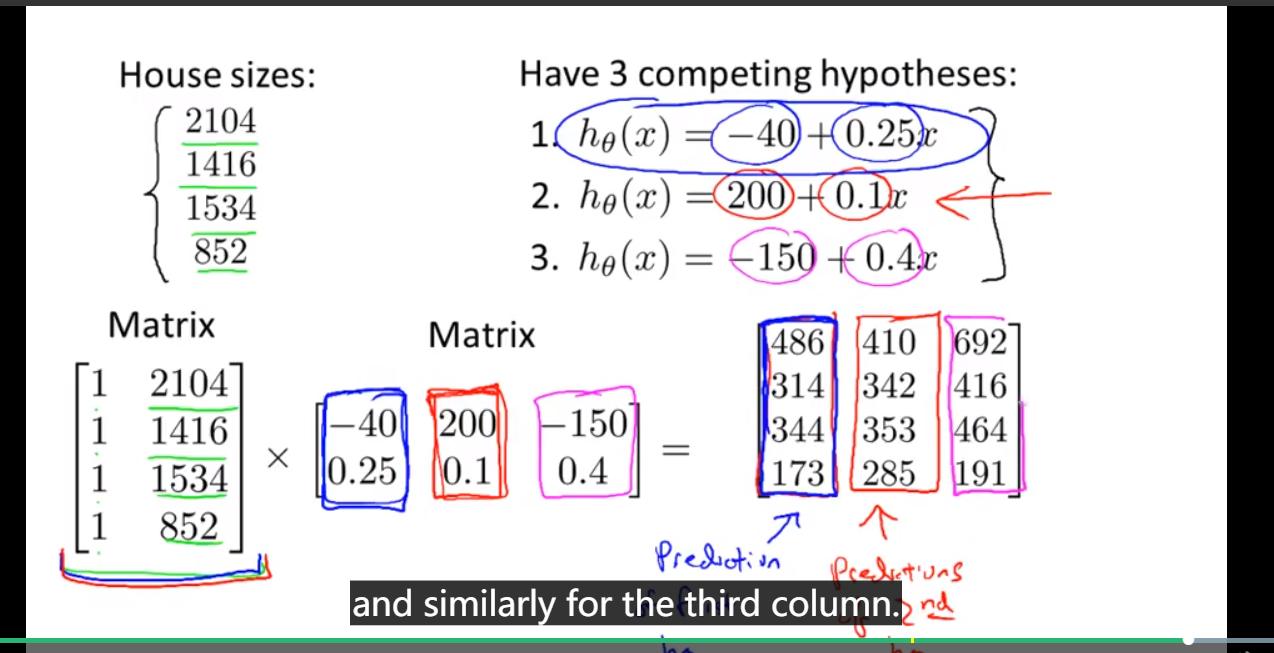
Matrix is rectangular array of numbers。

A vector is turns out to be a special case of matrix. A vector is a matrix that only has one column.

Matrix addition need matrix have same dimension.

Matrix multiplication. 新矩阵的行数是第一个矩阵的行，列数是第二个矩阵的列. 第一个矩阵的列数必须等于第二个矩阵的行数.

矩阵乘法的妙用~~



矩阵乘法不服从交换律, 但是符合结合律

新的加入