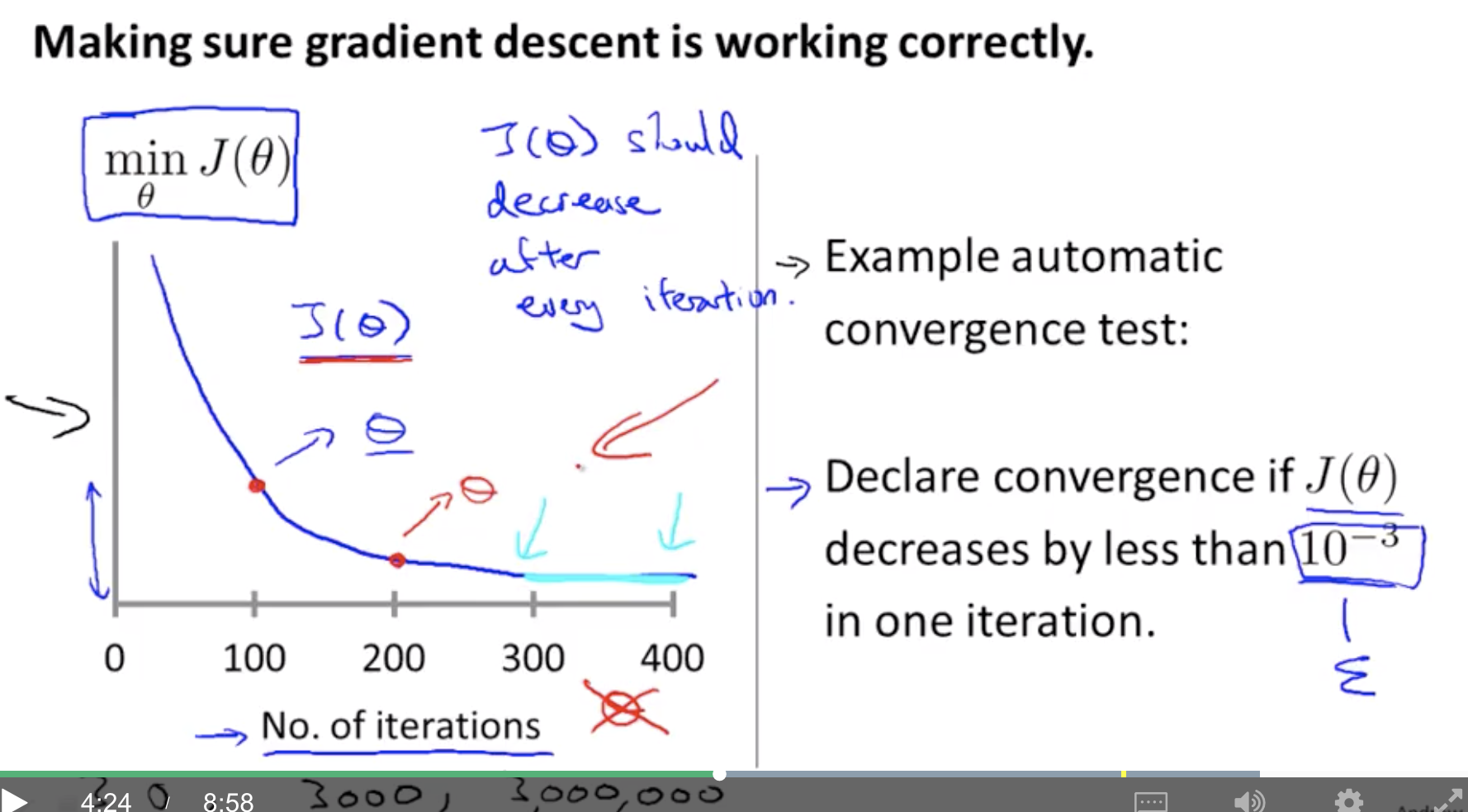
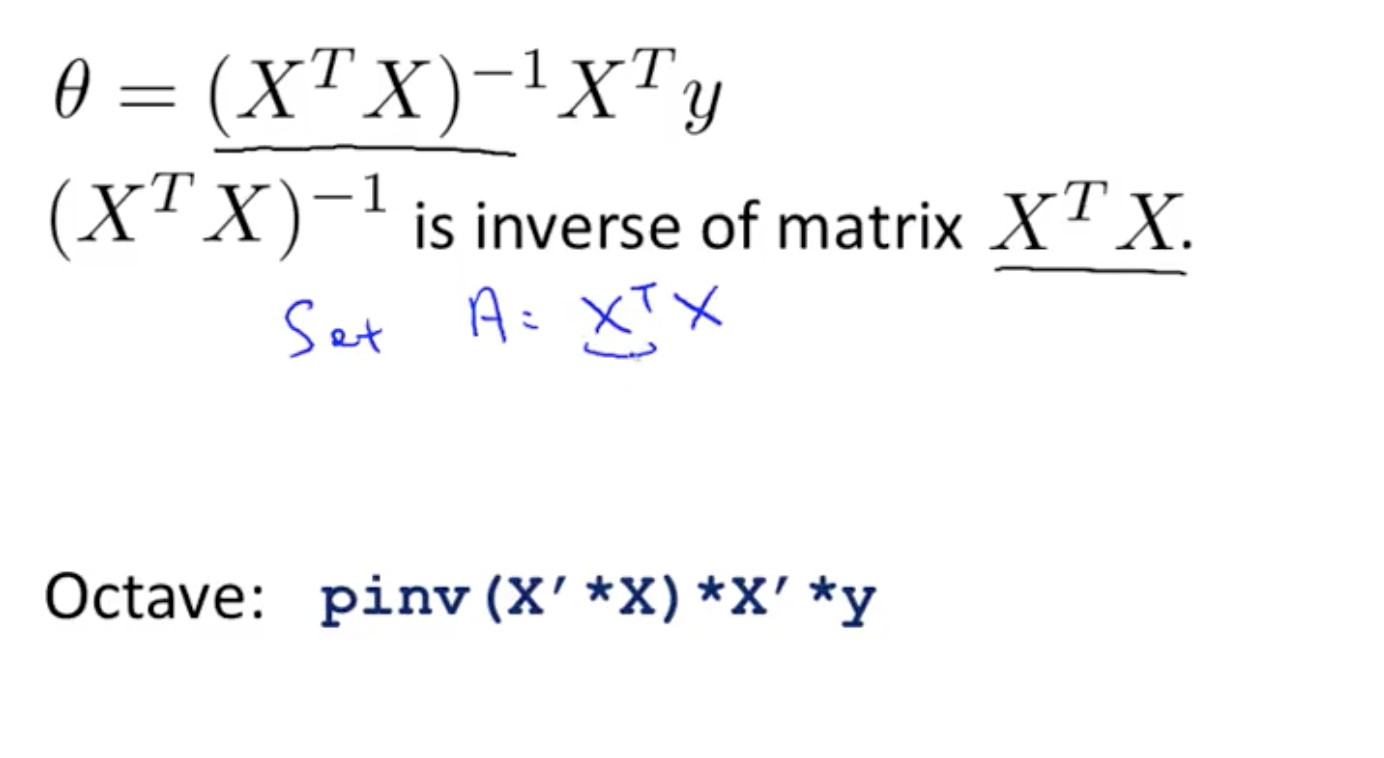
当采用gradient descent进行优化的时候，如果不做scaling，梯度下降的过程将会很慢



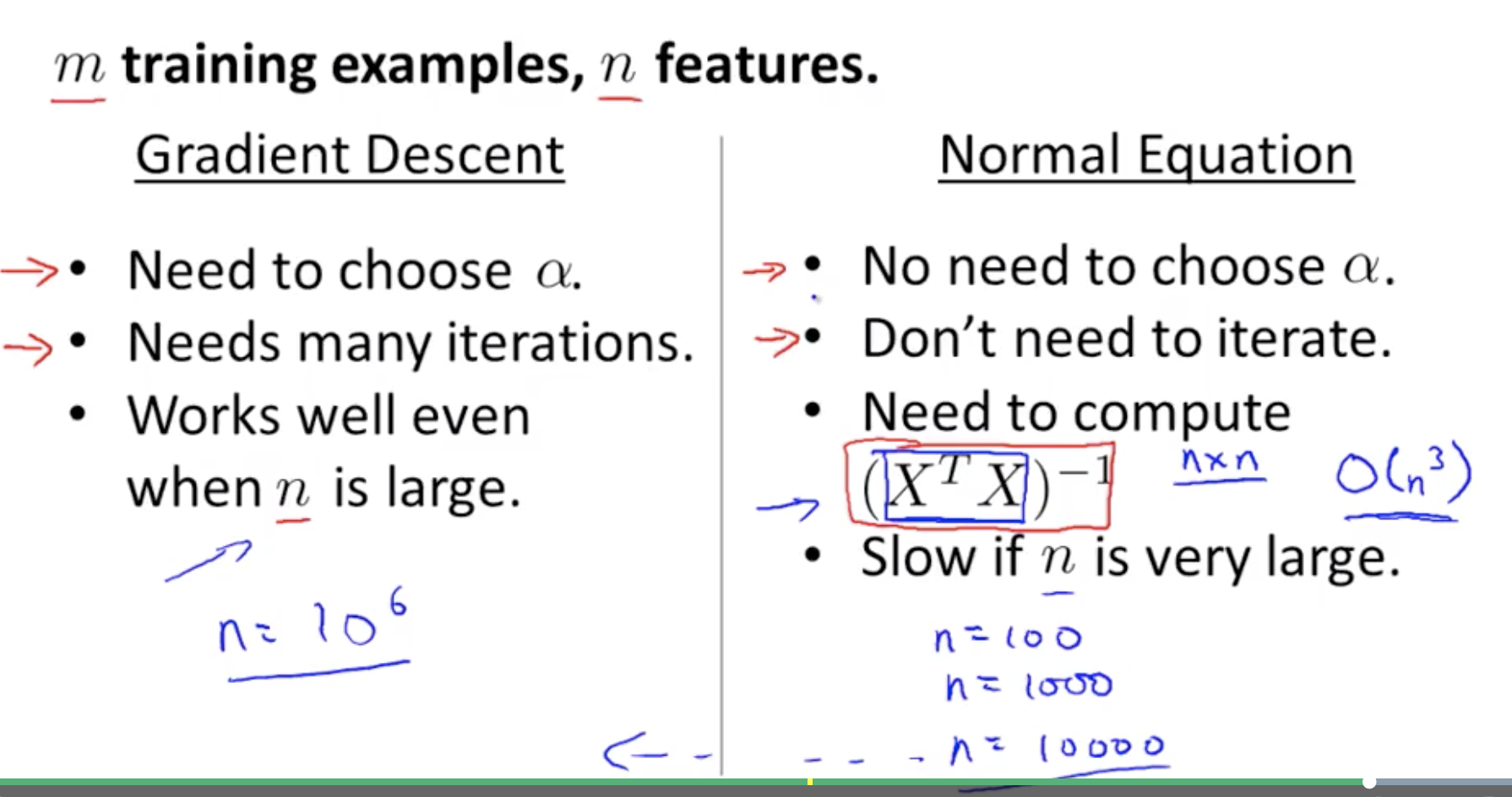
用min Jθ和iter的关系图去判断梯度下降是否正确和iter是否足够

另外，也可以采用automatic convergence。但是很难说有一个标准的要求，最好还是看左边这张图

**Normal Equation & Gradient Descent**



* Normal Equation的公式

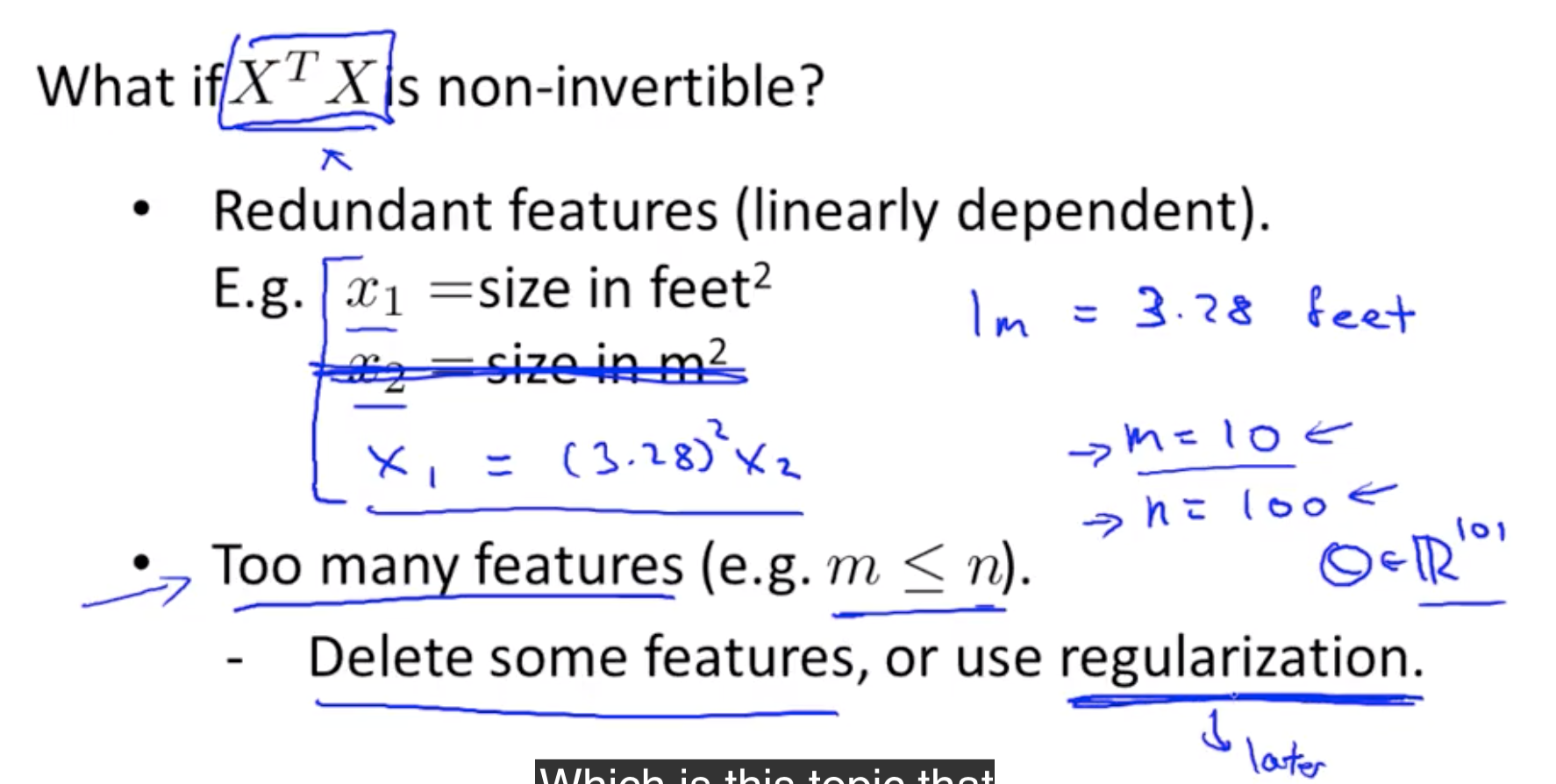


* 当feature特别多的时候，比如达到百万级别，gradient descent应该就会比normal equation更加高效
* 另外normal equation可适用的模型相对少，比如classification模型就不适用了，e.g. LR

<https://blog.csdn.net/nfmsr/article/details/78577121>

关于normal equation的推导可以通过向量投影的方法

不可逆的原因（奇异矩阵or degenerate）



一种情况是因为有自相关的特征，另一个原因可能是特征的数量比样本数量多

