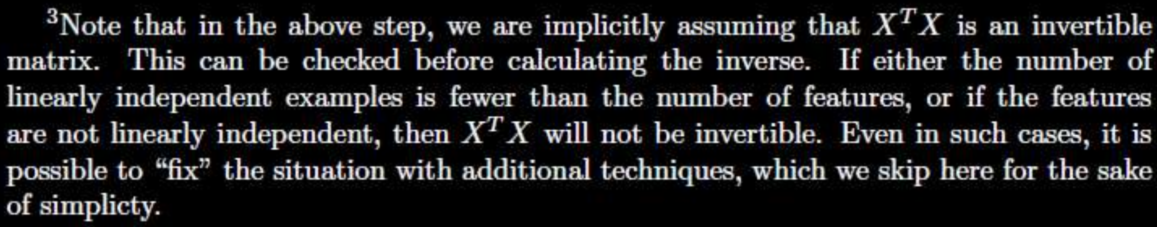
- Gradient descent (batch, stochastic, and mini-batch)
    - CS229 notes (9-13)

$$\theta := \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x^{(i)}$$

    - Often, stochastic gradient descent gets theta close to the minimum much faster than batch gradient descent (Note: however, it may never converge to the minimum, and the parameters theta will keep oscillating around the minimum of J(theta); but in practice most of the values near the minimum will be reasonably good approximations to the true minimum, therefore, when training set is large, stochastic is often preferred over batch)
- Normal equation
    - CS229 notes (13-15)

$\theta = (X^TX)^{-1}X^T\vec{y}.$[3]



[3] Note that in the above step, we are implicitly assuming that X^TX is an invertible matrix. This can be checked before calculating the inverse. If either the number of linearly independent examples is fewer than the number of features, or if the features are not linearly independent, then X^TX will not be invertible. Even in such cases, it is possible to "fix" the situation with additional techniques, which we skip here for the sake of simplicty.

    -