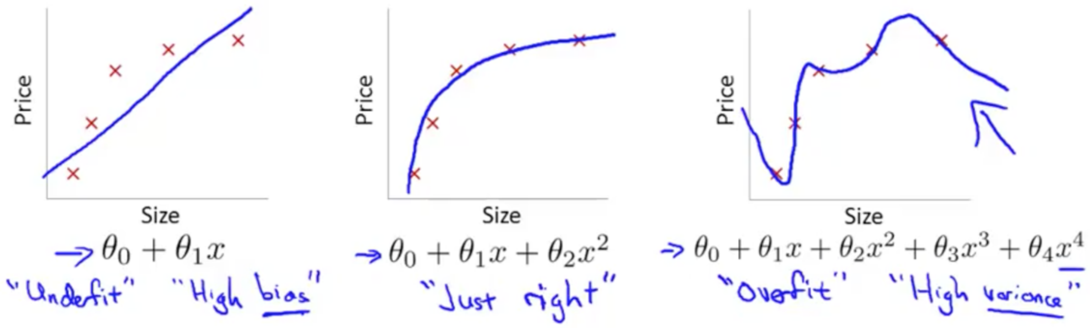
Overfitting Problem

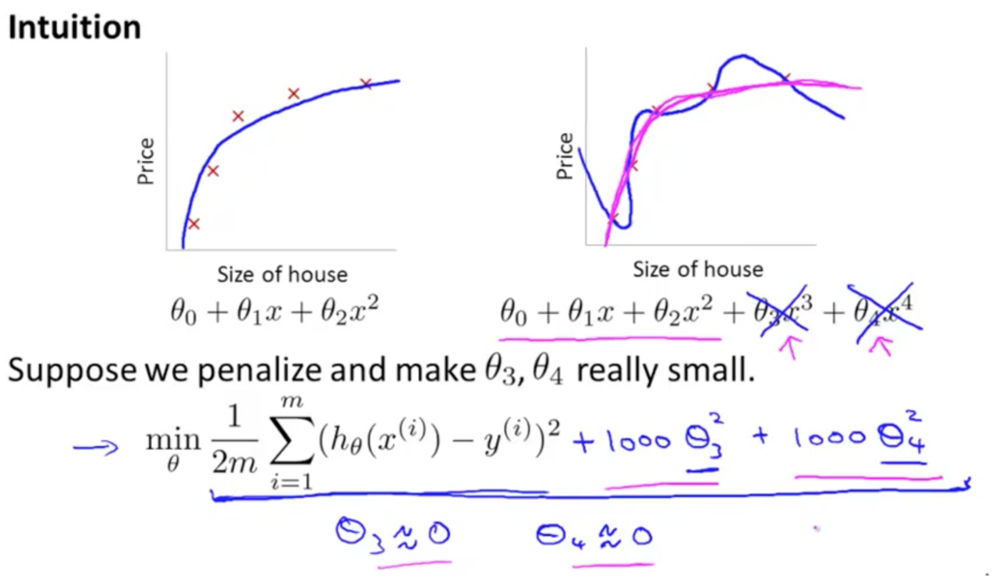
Overfitting: if we have too many features, the learned hypothesis may fit the training set very well, but fail to generalize to new examples.



1. Reduce the number of features:
    - Manually select which features to keep.
    - Use a model selection algorithm (studied later in the course).
2. Regularization
    - Keep all the features, but reduce the magnitude of parameters θj.
    - Regularization works well when we have a lot of slightly useful features.

Cost Function

Essentially, we want to eliminate the weight of extra terms. Look at the following example:



- Using the above cost function with the extra summation, we can smooth the output of our hypothesis function to reduce overfitting.
- The λ, or lambda, is the regularization parameter. It determines how much the costs of our theta parameters are inflated.

- If lambda is chosen to be too large, it may smooth out the function too much and cause under-fitting.

Regularized Linear Regression

Non-Invertible: X is non-invertible if m < n, and may be non-invertible if m = n.

Gradient Descent

$$
\begin{aligned}
&\text{Repeat } \{ \\
&\quad \theta_0 := \theta_0 - \alpha \ \frac{1}{m} \ \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)} \\
&\quad \theta_j := \theta_j - \alpha \ \left[\left(\frac{1}{m} \ \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \right) + \frac{\lambda}{m} \theta_j \right] \qquad\qquad j \in \{1, 2 \ldots n\} \\
&\}
\end{aligned}
$$

- Note: we separate out theta (0) because we don't want to penalize theta (0)

The term $\frac{\lambda}{m}\theta_j$ performs our regularization. With some manipulation our update rule can also be represented as:

$\theta_j := \theta_j(1 - \alpha \frac{\lambda}{m}) - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$

The first term in the above equation, $1 - \alpha \frac{\lambda}{m}$ will always be less than 1. Intuitively you can see it as reducing the value of θ_j by some amount on every update. Notice that the second term is now exactly the same as it was before.

Normal Equation

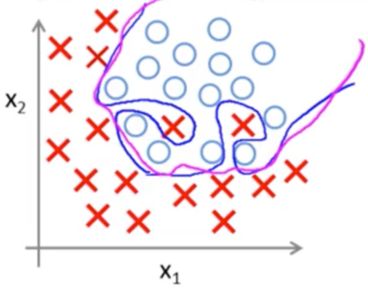
$$
\begin{aligned}
&\theta = \left(X^T X + \lambda \cdot L \right)^{-1} X^T y \\
&\text{where } \ L = \begin{bmatrix} 0 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}
\end{aligned}
$$

L is a matrix with 0 at the top left and 1's down the diagonal, with 0's everywhere else. It should have dimension (n+1)×(n+1). Intuitively, this is the identity matrix (though we are not including x_0), multiplied with a single real number λ.

Recall that if m < n, then $X^T X$ is non-invertible. However, when we add the term λ⋅L, then $X^T X$ + λ⋅L becomes invertible.

Regularized Logistic Regression

Regularized logistic regression.



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \dots)$$

Cost function:

$$J(\theta) = -\left[\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log\left(1-h_\theta(x^{(i)})\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

$\boxed{\theta_1, \theta_2, \dots, \theta_n}$

- Blue line in the chart represent the over-fitting situation
- Purple line is regularized logistic regression and more reasonable

Cost Function

Recall that our cost function for logistic regression was:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\ \log(h_\theta(x^{(i)})) + (1-y^{(i)})\ \log(1-h_\theta(x^{(i)}))]$$

We can regularize this equation by adding a term to the end:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\ \log(h_\theta(x^{(i)})) + (1-y^{(i)})\ \log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

The second sum, means to explicitly exclude the bias term, theta 0. This sum explicitly skips theta 0, by running from 1 to n, skipping 0. Thus, when computing the equation, we should continuously update the two following equations:

Gradient descent

Repeat {

$$\theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_0^{(i)}$$

$$\theta_j := \theta_j - \alpha\underbrace{\left[\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} + \frac{\lambda}{m}\theta_j\right]}_{\frac{\partial}{\partial\theta_j}J(\theta)}$$

$(j = \cancel{0}, 1, 2, 3, \dots, n)$

$\theta_1 \dots \theta_n$

}

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$