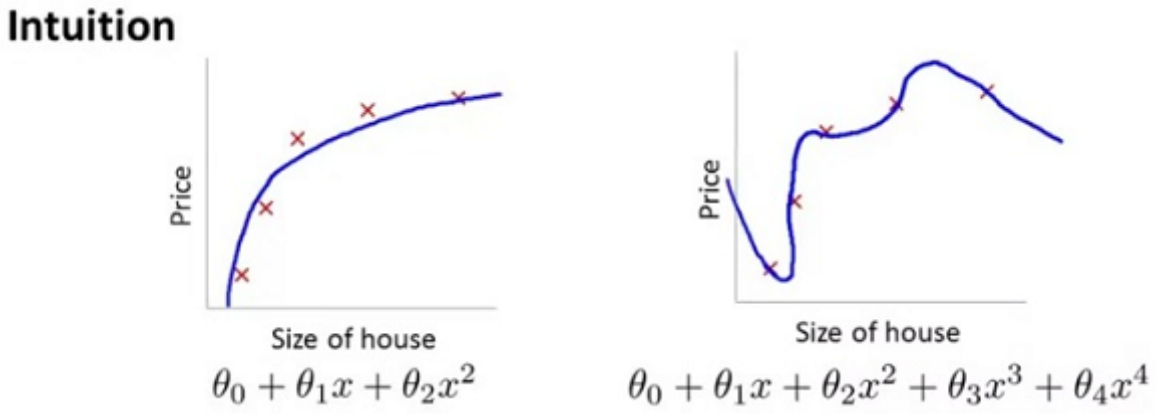
Cost Function (with regularization term)

At the first plot if we were to fit a quadratic function to this data, it gives us a pretty good fit to the data. Whereas, if we were to fit an overly high order degree polynomial, we end up with a curve that may fit the training set very well, but overfit the data poorly, and, not generalize well.



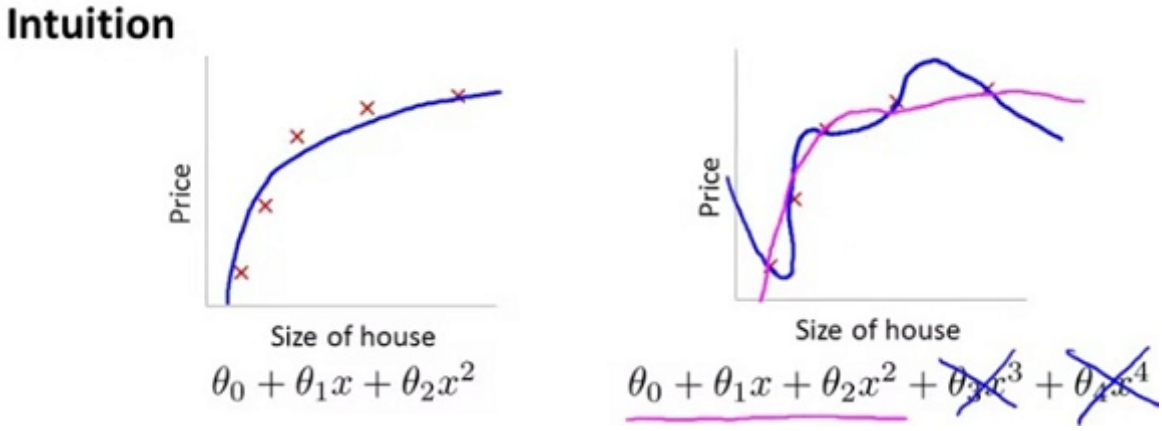
Suppose we were to penalize, and, make the parameters θ_3 and θ_4 really small.

$min_\theta \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Let's say we take this cost function and modify it and add to it: $+1000 \cdot \theta_3^2 + 1000 \cdot \theta_4^2$ (For this example 1000 we're just writing as some huge number).

$min_\theta \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + 1000 \cdot \theta_3^2 + 1000 \cdot \theta_4^2$

If we were to minimize the cost function, the only way to make the new cost function small is if θ_3 and θ_4 are small. So when we minimize this new function we are going to end up with $\theta_3 \approx 0$ and $\theta_4 \approx 0$.



And if we do that, well then, if θ_3 and θ_4 close to 0 then we are being left with a quadratic function, and, so, we end up with a fit to the data, that's, quadratic function plus, tiny contributions from small terms, θ_3 and θ_4, that they may be very close to 0.

Regularization

Small values for parameters $\theta_0, \theta_1, \ldots, \theta_n$

- "Simpler" hypothesis
- Less prone to overfitting

More generally, here is the idea behind regularization. The idea is that, if we have small values for the parameters, then, having small values for the parameters, will somehow, will usually correspond to having a simpler hypothesis.

For example: For housing price prediction we may have our hundred features where may be x_1 is the size, x_2 is the number of bedrooms, x_3 is the number of floors and so on. And we may we may have a hundred features. And unlike the polynomial example, we don't know that θ_3, θ_4, are the high order polynomial terms.

Housing:

- Features: $x_1, x_2, \ldots, x_{100}$
- Parameters: $\theta_0, \theta_1, \theta_2, \ldots, \theta_{100}$

So, if we have just a set of a hundred features, it's hard to pick in advance which are the ones that are less likely to be relevant. So we have a hundred one parameters. And we don't know which parameters to try to pick, to try to shrink.

So, in regularization, what we're going to do, is take our cost function. And what we're going to do is, modify the cost function to shrink all of our parameters, because, we don't know which one or two to try to shrink. So we're going to modify our cost function to add a term at the end.

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

We're going to add an extra regularization term at the end to shrink every single parameter and so this term tend to shrink all of our parameters $\theta_1, \theta_2, \theta_3$ up to θ_{100}. By convention, usually, we regularize only theta through theta 100 (in practice, it makes very little difference, and, whether we include, θ_0 or not, in practice, make very little difference to the results).

Video Question: In regularized linear regression, we choose θ to minimize:

What if λ is set to an extremely large value (perhaps too large for our problem, say $\lambda = 10^{10}$)?

- Algorithm works fine; setting λ to be very large can't hurt it.
- Algorithm fails to eliminate overfitting.

> Algorithm results in underfitting (fails to fit even the training set).

- Gradient descent will fail to converge.

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

Writing our regularized optimization objective, our regularized cost function again. Where, the term on the right is a **regularization term** and lambda is called the **regularization parameter** and what lambda does, is it controls a trade off between two different goals:

- Fit the training set well.
- keeping the parameter small and therefore keeping the hypothesis relatively simple to avoid overfitting.

In regularized linear regression, we choose θ to minimize

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

What if λ is set to an extremely large value (perhaps too large for our problem, say $\lambda = 10^{10}$)?

In regularized linear regression, if the regularization parameter λ is set to be very large, then what will happen is we will end up penalizing the parameters $\theta_1, \theta_2, \theta_3, \theta_4$ very highly (And if we end up penalizing $\theta_1, \theta_2, \theta_3, \theta_4$ very heavily, then we end up with all of these parameters close to zero). And if we do that, it's as if we're getting rid of these terms in the hypothesis so that we're just left with a hypothesis that will say, housing prices are equal to theta zero, and that is akin to fitting a flat horizontal straight line to the data.

And this is an example of **underfitting**, and in particular this hypothesis, this straight line it just fails to fit the training set well. It's just a fat straight line, it doesn't go anywhere near most of the training examples.

Summary

If we have overfitting from our hypothesis function, we can reduce the weight that some of the terms in our function carry by increasing their cost.

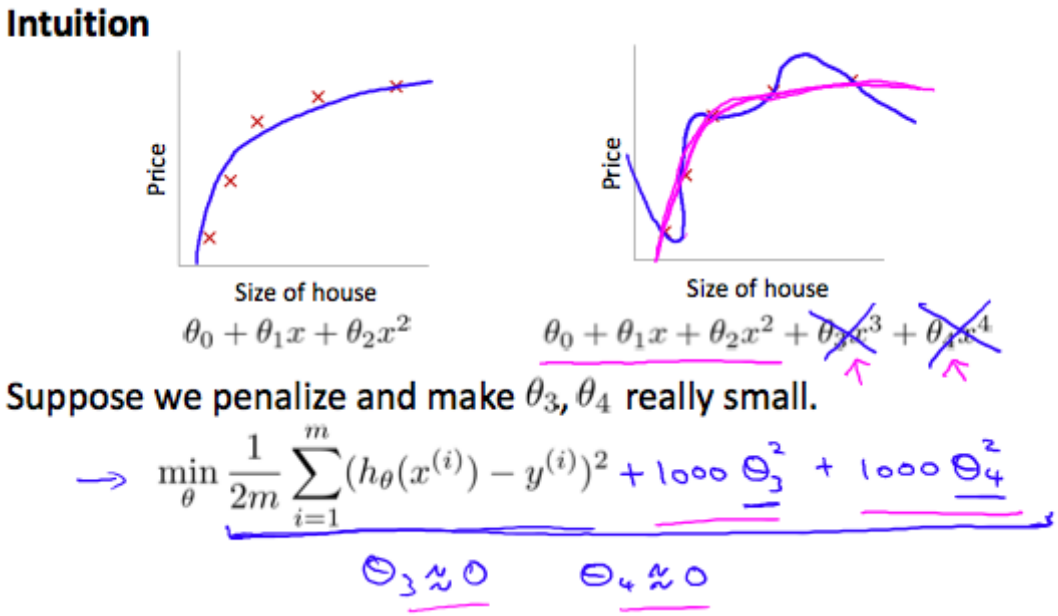
Say we wanted to make the following function more quadratic:

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

We'll want to eliminate the influence of $\theta_3 x^3$ and $\theta_4 x^4$. Without actually getting rid of these features or changing the form of our hypothesis, we can instead modify our **cost function**:

$$min_\theta \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + 1000 \cdot \theta_3^2 + 1000 \cdot \theta_4^2$$

We've added two extra terms at the end to inflate the cost of θ_3 and θ_4. Now, in order for the cost function to get close to zero, we will have to reduce the values of θ_3 and θ_4 to near zero. This will in turn greatly reduce the values of $\theta_3 x^3$ and $\theta_4 x^4$ in our hypothesis function. As a result, we see that the new hypothesis (depicted by the pink curve) looks like a quadratic function but fits the data better due to the extra small terms $\theta_3 x^3$ and $\theta_4 x^4$.



We could also regularize all of our theta parameters in a single summation as:

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

The λ, or lambda, is the **regularization parameter**. It determines how much the costs of our theta parameters are inflated.

Using the above cost function with the extra summation, we can smooth the output of our hypothesis function to reduce overfitting. If lambda is chosen to be too large, it may smooth out the function too much and cause underfitting. Hence, what would happen if $\lambda = 0$ or is too small ?