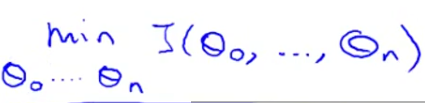
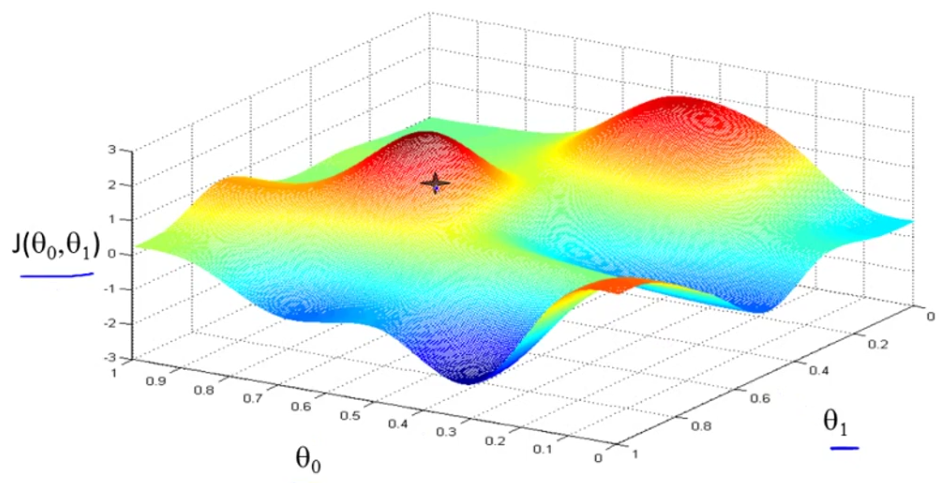
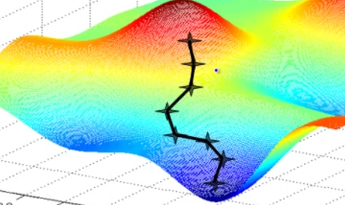
***PARAMETER LEARNING***

**I. GRADIENT DESCENT**

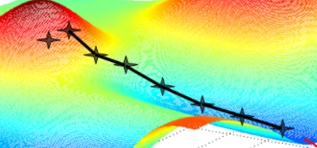
* **Gradient Descent** = algorithm for minimizing the cost function J(θ0,θ1)
* It’s a more general algorithm used not only in linear regression, but all over in ML to minimize other functions besides J(θ0,θ1)
* Assume we have some function J(θ0,θ1) + we want an algorithm for minimizing it
* **NOTE:** It turns out that gradient descent actually applies to more general functions.
* i.e. minimize the value of J(θ0…..θ(n)) over θ0…..θ(n) 
* Idea for gradient descent = to start w/ some initial guesses for θ0 + θ1 (common choice 🡪θ0 = 0, θ1 = 0)
* Then we keep changing these values little by little bit to try to reduce J(θ0,θ1), until hopefully, we wind up at a minimum (or a local minimum)
* θ0 + θ1 = horizontal axes, J(θ0,θ1) = height of the surface shows J(θ0,θ1) + we want to minimize it.



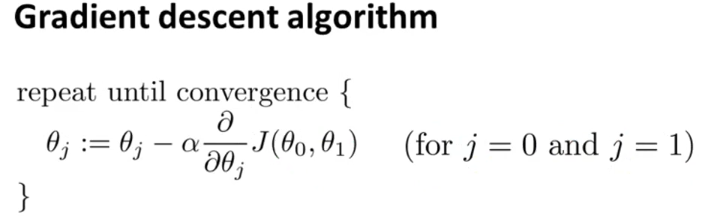
* Imagine that this figure shows 2 hills + you ‘re physically standing at the point on the hill
* Start off w/ {θ0, θ} at some point on the surface (black cross above)
* In gradient descent, we spin 360 degrees, look around, + ask “if I were to take a little baby step in *some* direction, + I want to go downhill as quickly as possible, in what direction do go?”
* To go down, want to physically walk down this hill as rapidly as possible.
* Then you take that “best direction” + end up new point on the hill
* Then, again, look all around + see what direction to step in order to take a little baby step downhill + take another step in that direction + then keep going until you converge to a **local minimum**.



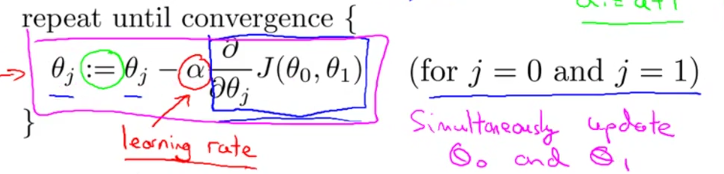
* Now imagine we initialized gradient descent just a couple steps to the right instead, + then repeated the process from that point
* Now gradient descent would've taken you to a 2nd local minimum, more to the right.



* This is an interesting property of gradient descent
* This is the definition of the gradient descent algorithm.



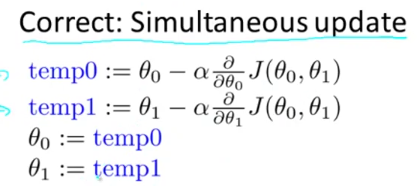
* Repeatedly do the formula above until **convergence**
* update parameter θj by taking θj + subtracting from it **α**\*a derivative term
* **:=** 🡪 denotes assignment
* a **:=** b 🡪 in a CPU, this means take the value in b + use it to overwrite whatever value is a.
* “set a to be equal to the value of b”
* Whereas in contrast, w/ an equal 🡪 a = b, this is a **truth assertion.**
* **Alpha (α)** = the **learning rate**.
* This controls *how big a step we take* downhill w/ gradient descent.
* Very large α = very aggressive gradient descent procedure (huge steps downhill)
* Very small α = taking baby steps downhill.
* Then we have a **derivative term** in blue

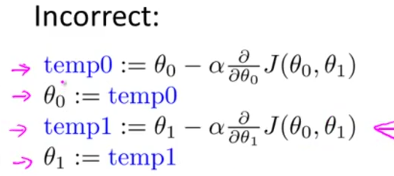


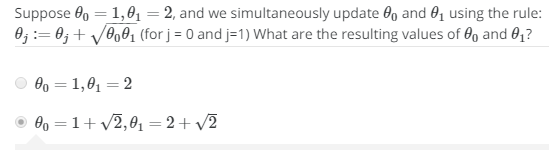
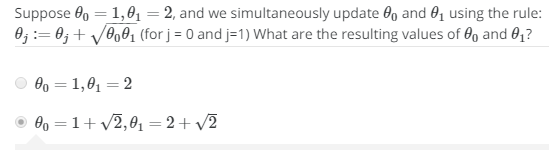
* There's 1 more subtlety about gradient descent 🡪 in gradient descent we're going to update θ0 + θ1,
* “the update takes place for j = 0 and j = 1”
* So when updating θ0 + θ1, we ***simultaneously update*** θ0 + θ1.
* Meaning in this update equation, we're going update θ0 := …. + update θ1 := …..
* The way to implement = compute the right-hand for θ0 + θ1 + then simultaneously update θ0 + θ1

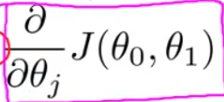


* Ex: Set *temp0* + *temp1* to compute the right-hand sides + store them into variables temp0 + temp1 + then update θ0 + θ1 simultaneously = CORRECT

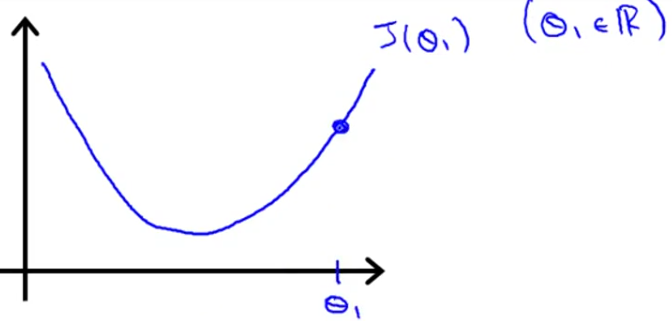


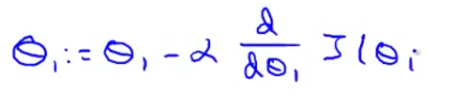


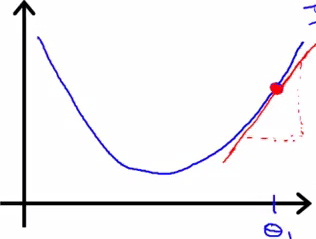
* In this 2nd attempt, by the time we tried to update θ1 (3rd step), we’d be using a new value of θ0 (from 2nd step), which gives a different value of temp1 than what we *should* have
* The way gradient descent is usually naturally implemented is w/ simultaneous updates.
* When people talk about gradient descent, they always mean simultaneous update
* If you implement the NON-simultaneous update, it will probably work anyway, but the algorithm won’t be right + is actually some other algorithm w/ different properties that can, for various reasons, behave in slightly stranger ways
* 
* The **derivative term:**



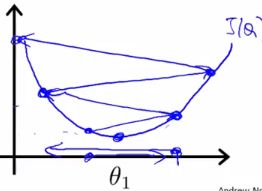
* **Derivative =** measures the sensitivity to change of a function (output) value w/ respect to a change in its argument (input value)
* **Partial derivative** = derivative of a function w/ respect to ONE of its variables, with the others held constant (as opposed to the **total derivative**, in which all variables are allowed to vary)
* Think of it as how much we expect the value of the function to change is we make a small change to ONE of its arguments, leaving the others unchanged
* Say we have a cost function of just 1 parameter, j(θ1), where θ1 = a real number
* Let's try to understand what gradient decent would do on this function.



* So we initialize gradient descent w/ θ1 to the right of the x-axis
* Gradient descent will update θ1 as 
* We compute the derivative of (take the tangent to) the J(θ1) + look at the slope of this line



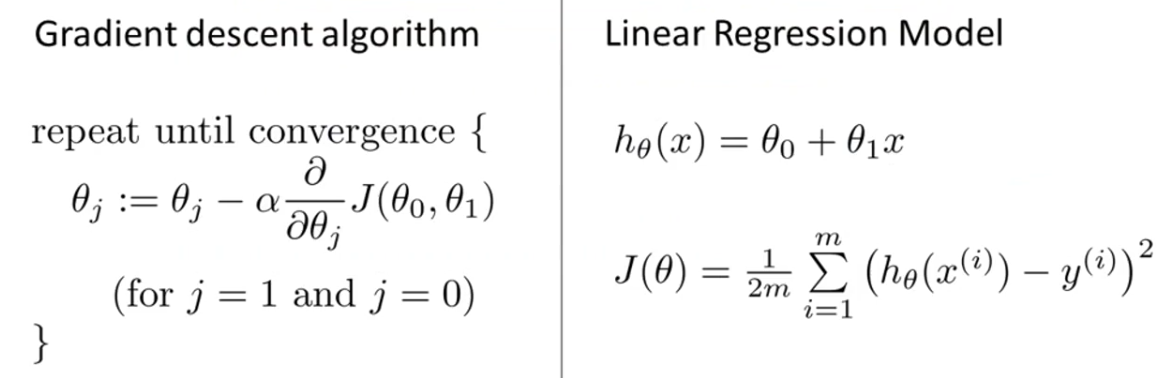
* If slope = positive = positive derivative 🡺 so my update ( **:=** ) to θ1 is going to be θ1 – **α**\*some positive number (**α** *is always a positive number*)
* So we end up w/ θ1 minus a positive 🡪 we move to a new θ value to the left (decrease θ1)
* We *want* to head in this direction 🡪 towards minimum J(θ1) 🡪 dip of the curve
* If the initial θ1 was to the left of the minimum, we’d have a negative tangent slope = negative derivative = θ1 minus a negative # 🡺 increase θ1 🡺 update θ1 to the right towards the minimum J(θ1)
* If **α** is too small, we multiply updated θj by some small number = taking a small step = we’ll need a lot of steps to get to the minimum = gradient descent is *very slow*
* If **α** is too large= gradient descent can *overshoot* the minimum + may even *fail* to converge or even actually *diverge*



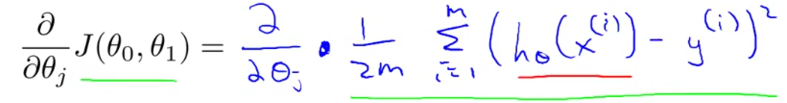
* See we end up w/ a worse value of θ1
* Can continue to overshoot actually get further + further away from the minimum.
* What if parameter θ1 is *already at a local minimum*?
* Initialize θ1 at a local minimum/optimum 🡪 the derivative = 0 so the tangent slope = 0 🡪 derivative term = 0 🡪 gradient descent update = θ1 – 0
* *If you're already at the local optimum, gradient descent does absolutely nothing*
* This also explains why gradient descent can converge to the local minimum even w/ a fixed learning rate/**α**.
* Imagine cost function J(θ) + we start gradient descent w/ a steep derivative
* Next step = derivative = less steep b/c as we approach the minimum, the derivative gets closer + closer to 0
* Now we take that smaller step + are now even closer to global minimum (derivative is even smaller)
* As gradient descent runs, you will automatically take smaller + smaller steps until eventually you're taking very small steps + finally converge to the local minimum.
* So as we approach local minimum, the *derivative term* will automatically *get smaller*, so gradient descent will automatically take smaller steps = NO *need to decrease* **α**.
* This gradient descent algorithm can be used to try to minimize any cost function J

**I. GRADIENT DESCENT FOR LINEAR REGRESSION**

* Put together, gradient descent w/ a cost function gives an algorithm for linear regression



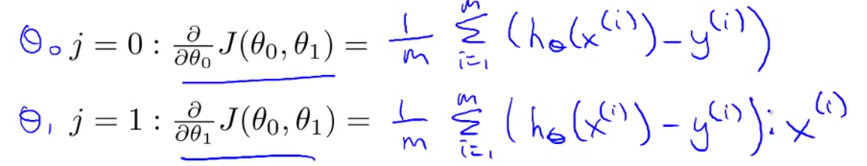
* We want to *apply gradient descent to minimize squared error/cost function*.
* In order to apply gradient descent, the key term needed is the *derivative term,* into which we plug in the definition of J(θ0,θ1)

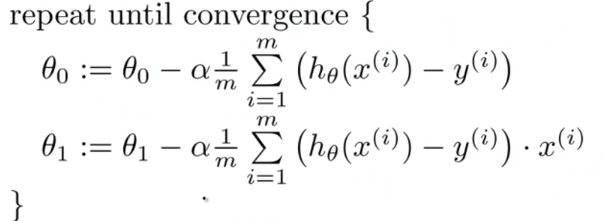
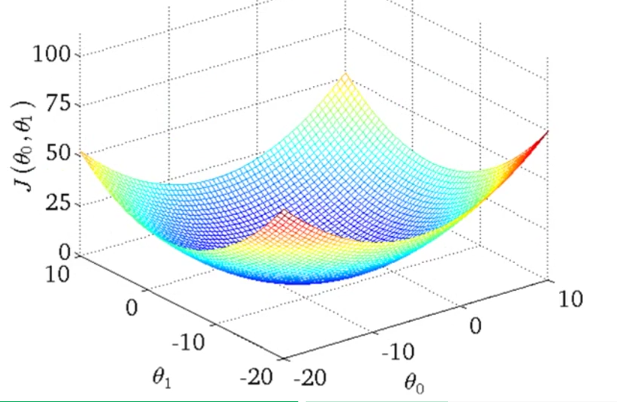


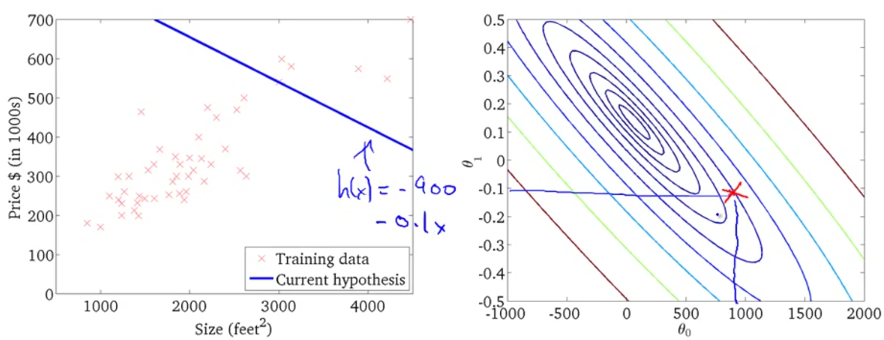
* Then we plug in the definition for hypothesis(x(i))

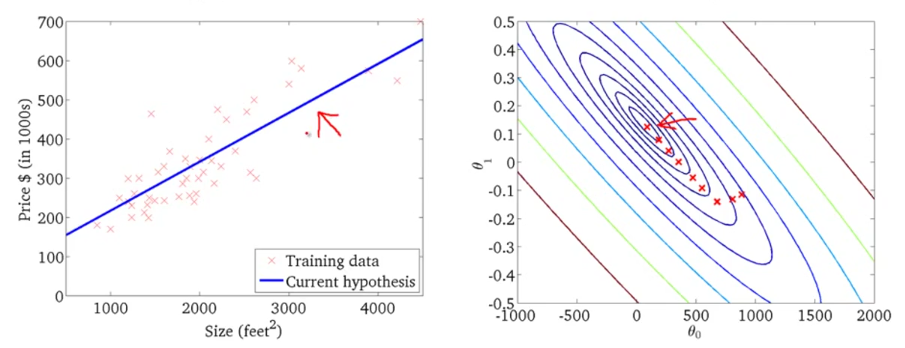
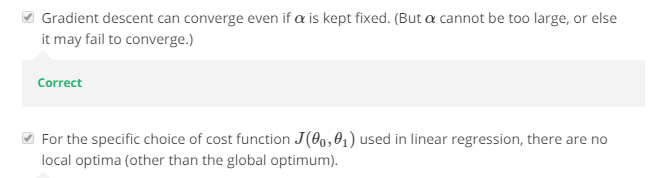


* Turns out we need to figure out the **partial derivative** for TWO cases 🡪 for J = θ0 and J = θ1.



* \*\*\* Computing these partial derivative terms requires some multivariate calculus
* So armed w/ these definitions, we worked the derivatives = the slope of the cost function J + can now plug them back in to our gradient descent algorithm.
* **Gradient Descent For Linear Regression**
* 
* W/ alpha times the partial derivatives w/ respect to θ0 and respect to θ1 worked out before
* REMINDER: you should be implementing this so the update θ 0 + θ 1 are simultaneous
* 1 of the issues we saw with gradient descent is it can be susceptible to local optima = depending on where you initialize it, you can end up at different local optima.
* It turns out the cost function for linear regression is *always going to be a bow shaped function =* a **convex function**.
* 
* This function only has 1 global optimum and no local optima
* Gradient descent on this type of cost function (which you get whenever using linear regression) will always converge to the global optimum (b/c there are no other local optimum)
* Now, we usually initialize parameters at (0,0), but for the demonstration we’ll do θ0 = 900 + θ1 = ~-0.1 🡪 corresponds to h(x)= 900-0.1x



* Now, if we take 1 step in gradient descent, we move down and left + the h(x) lines changes a bit
* As I take further steps of gradient descent, we go down in cost (move parameter data point close and closer to middle/minimum) and h(x) line better and better fits to the data
* Eventually we wind up at the global minimum = gets a good fit to the data set of housing prices
* 
* Can now use it to predict the price to sell a 1250 square ft. house = ~$250k
* The algorithm just used is sometimes called **batch gradient descent** (ML people were not always great at giving names to algorithms)
* *Batch Gradient Descent* refers to the fact that in every step of gradient descent, we're looking at ALL training examples 🡪 when computing the derivatives, we're computing the sums over our “m” training examples 🡪 refers to the fact that we're looking at the *entire batch* of training examples
* There are sometimes other versions of gradient descent that are not batch versions + do not look at the entire training set, but instead look at small subsets of the training set at a time.
* If you've seen advanced linear algebra before, you might know that there exists a solution for numerically solving for the minimum of the cost function J w/out needing to use an iterative algorithm like gradient descent = the **normal equations method**.
* But gradient descent will scale better to larger data sets than that normal equation method, + we can use it in lots of different contexts + in lots of different ML problems.
* 

1. A computer program is said to learn from 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**. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. What would be a reasonable choice for P?

* **The probability of it correctly predicting a future date's weather.**

1. The amount of rain that falls in a day is usually measured in either millimeters (mm) or inches. Suppose you use a learning algorithm to predict how much rain will fall tomorrow. Would you treat this as a classification or a regression problem?

* **Regression**

1. Suppose you are working on stock market prediction, and you would like to predict the price of a particular stock tomorrow (measured in dollars). You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?

* **Regression**

1. Some of the problems below are best addressed using a supervised learning algorithm (*we teach CPU*), and the others with an unsupervised learning algorithm (*CPU teaches itself)*. In each case, assume some appropriate dataset is available for your algorithm to learn from.

* Given data on crop yields over the last 50 years, learn to predict next year's yields 🡪 **Supervised**
* Given data on how 1000 patients respond to an experimental drug (effectiveness of treatment, side effects, etc.), discover whether there are different categories or "types" of patients in terms of how they respond to the drug, and if so what these categories are. 🡪 **Unsupervised**
* Examine a web page, and classify whether the content on the page should be considered "child friendly" (e.g., non-pornographic, etc.) or "adult." 🡪 **Supervised**
* Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments. 🡪 **Unsupervised**

1. Which of these is a reasonable definition of machine learning?

* **Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.**

1. An email program watches which emails we mark as spam or not spam, and based on this activity learns how to better filter spam. What is the task, T, here?

* **Classify emails as spam or not spam**
* **E = watching you classify spam and not spam**
* **P = probability the program will correctly ID /not spam, OR the # or fraction/% of emails correctly classified as spam/not spam**

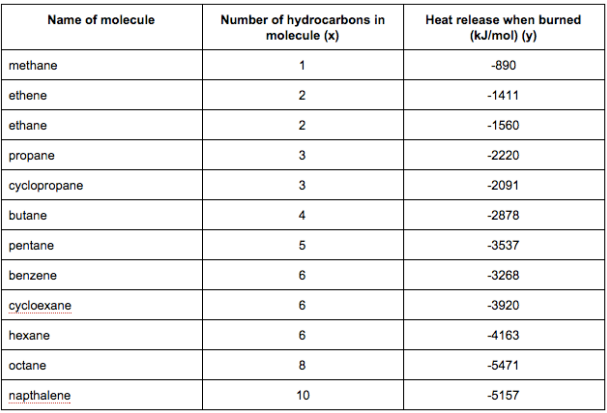
1. Consider the problem of predicting how well a student does in her 2nd year of college, given how well she did in her 1st year. Specifically, let x = # of "A" grades (including A-. A, A+) a student receives in their 1st year of college. We would like to predict the value of y = the # of "A" grades they get in their 2nd year. Here each row is 1 training example. Recall that in linear regression, our hypothesis is hθ(x)=θ0+θ1x + we use m to denote the # of training examples.



For the training set given above, what is the value of m?

* **4**

1. Many substances that can burn (such as gasoline + alcohol) have a chemical structure based on carbon atoms + for this reason they’re called hydrocarbons. A chemist wants to understand how the # of carbon atoms in a molecule affects how much energy is released when that molecule combusts (is burned). The chemist obtains the dataset below. In the column on the right, “kJ/mol” is the unit measuring the amount of energy released.



You would like to use linear regression (hθ(x)=θ0+θ1x) to estimate the amount of energy released (y) as a function of the # of carbon atoms (x). Which of the following do you think will be the values you obtain for θ0 and θ1? You should be able to select the right answer without actually implementing linear regression.

* **θ0=−569.6,θ1=−530.9**

1. Suppose we set θ0=−1,θ1=0.5. What is hθ(4)?

* **-1 + (0.5\*5) = 1**

1. Let f be some function so that f(θ0,θ1) outputs a #. For this problem, f is some arbitrary/unknown smooth function (not necessarily the cost function of linear regression, so f may have local optima). Suppose we use gradient descent to try to minimize f(θ0,θ1) as a function of θ0 and θ1. Which of the following statements are true?

* **If θ0 and θ1 are initialized at a local minimum, then one iteration will not change their values.**
* **If the learning rate is too small, then gradient descent may take a very long time to converge.**
* **If θ0 and θ1 are initialized at the global minimum, then one iteration will not change their values.**
* **If the first few iterations of gradient descent cause f(θ0,θ1) to increase rather than decrease, then the most likely cause is that we have set the learning rate α to too large a value.**

1. Suppose that for some linear regression problem (say, predicting housing prices as in the lecture), we have some training set, and for our training set we managed to find some θ0, θ1 such that J(θ0,θ1)=0. Which of the statements below must then be true?

* **Our training set can be fit perfectly by a straight line**
* **We can perfectly predict the value of y even for new examples that we have not yet seen. (e.g., we can perfectly predict prices of even new houses that we have not yet seen.)**

1. Consider the following training set of m=4 training examples:

x y

1 0.5

1. 1

4 2

0 0

Consider the linear regression model hθ(x)=θ0+θ1x. What are the values of θ0 and θ1 that you would expect to obtain upon running gradient descent on this model? (Linear regression will be able to fit this data perfectly.)

* **θ0=0,θ1=0.5**

1. Suppose we set θ0=0,θ1=1.5 in the linear regression hypothesis from Q1. What is hθ(2)?

* **3**