Calculus Review, Gradient Descent, Chain Rule

- Let's do a refresher on some multivariable calculus. We will need to recall a few things to address things properly.
- Suppose you have a function of several variables $f(x_1, x_2, x_3, x_4) = x_1 + x_2^2 + x_4 x_3^3$. Can differentiate it with respect to each of the variables:

$$\frac{\partial f}{\partial x_1}(x_1, x_2, x_3, x_4) = 1$$

$$\frac{\partial f}{\partial x_2}(x_1, x_2, x_3, x_4) = 2x_2$$

$$\frac{\partial f}{\partial x_3}(x_1, x_2, x_3, x_4) = x_4 3x_3^2$$

$$\frac{\partial f}{\partial x_4}(x_1, x_2, x_3, x_5) = x_3^3$$

- The function f takes four numbers and spits out one number. So do its derivatives.
- Makes sense to talk about things like $\frac{\partial f}{\partial x_3}(1,2,3,4) = 4 \cdot 3 \cdot 3^2 = 108$.

- The gradient of $f(\cdot)$ stacks these up:

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \\ \frac{\partial f}{\partial x_4} \end{pmatrix}$$

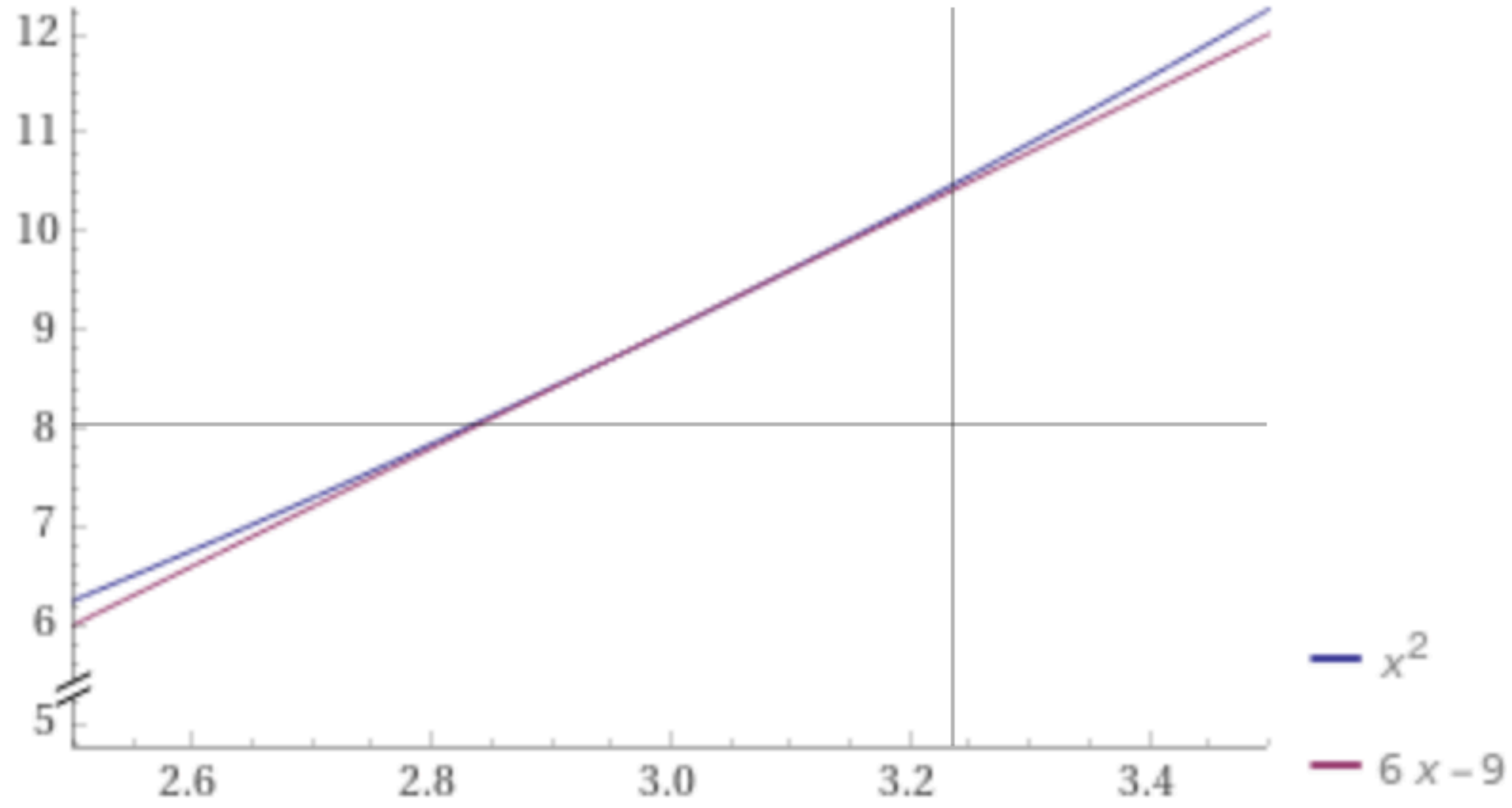
- For $f(x_1, x_2, x_3, x_4) = x_1 + x_2^2 + x_4 x_3^3$, we have

$$\nabla f = \begin{pmatrix} 1 \\ 2x_2 \\ x_4 3x_3^2 \\ x_3^3 \end{pmatrix}$$

- Note: the gradient takes four numbers and spits out a vector.
- In this case $f : \mathbb{R}^4 \to \mathbb{R}, \dfrac{\partial f}{\partial x_i} : \mathbb{R}^4 \to \mathbb{R}, \nabla f : \mathbb{R}^4 \to \mathbb{R}^4$.

- In general, suppose we have $f : \mathbb{R}^n \to \mathbb{R}$. That is, $f(x_1, \ldots, x_n)$ is a scalar.
- We have that $\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$
- We have that $\nabla f : \mathbb{R}^n \to \mathbb{R}$.

- What does the gradient mean?
- Let's go back and ask a question about what the derivative means.
- Suppose $f(x) = x^2$. So $f'(3) = 6$.
- What this means: the nonlinear function $f(x)$ is well-approximated by the line of slope 6 going through the point $(3,3^2)$.
- In other words: when $x \approx 3$, we have that $x^2 \approx 3^2 + 6 \cdot (x - 3) = 6x - 9$



Note how close the two curves are around 3

- Now let's discuss the case of many variables.
- Fix some point $y_1, \ldots, y_n$. Around this point, we have the approximation

$$f(x_1, \ldots, x_n) \approx f(y_1, \ldots, y_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(y_1, \ldots, y_n)(x_i - y_i)$$

- For example, suppose $f(x_1, x_2) = x_1^2 + x_2^2$.
- Since $\frac{\partial f}{\partial x_1}(x_1, x_2) = 2x_1$ and $\frac{\partial f}{\partial x_2} = 2x_2$ we have that around the point $y = (1,2)$, we have

$$\begin{aligned} f(x_1, x_2) &\approx (1^2 + 2^2) + 2 \cdot (x_1 - 1) + 4 \cdot (x_2 - 2) \\ &= 2x_1 + 4x_2 - 5 \end{aligned}$$

- Very easy to get confused here: $\frac{\partial f}{\partial x_1}(y_1, \ldots, y_n)$ means:

— take the function of $x_1, \ldots, x_n$

— differentiate with respect to the first variable to obtain a new function

— then plug in $y_1, \ldots, y_n$

- So we have the approximation

$$f(x_1, \ldots, x_n) \approx f(y_1, \ldots, y_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(y_1, \ldots, y_n)(x_i - y_i)$$

- Standard to write this as

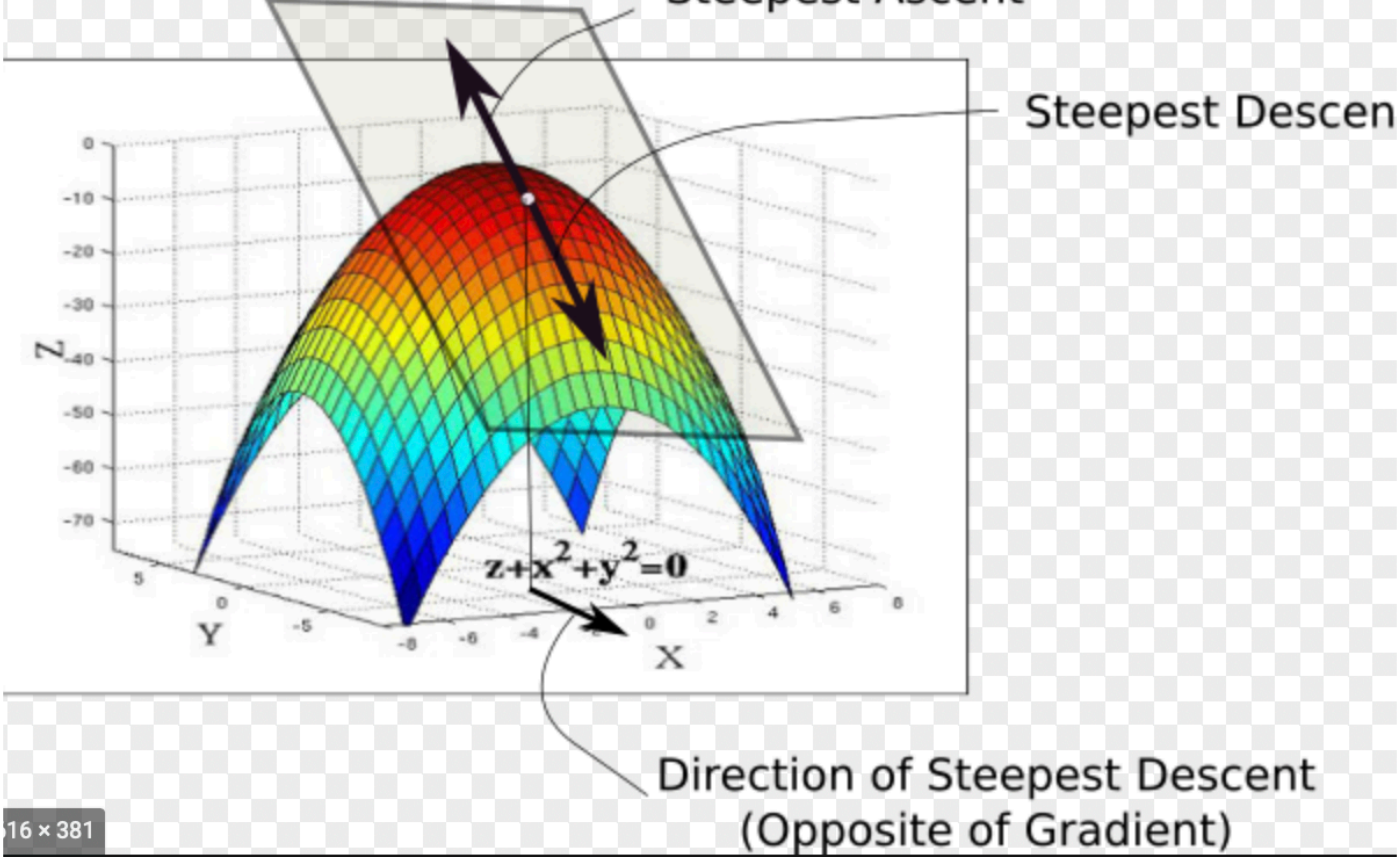
$$f(x) \approx f(y) + \nabla f(y)^T (x - y) \qquad (*)$$

  Here x and y are understood to be vectors by default and the inner product is consistent with our earlier definition of gradient.

- Same story as before: the error is really small when x is close to y.

- Next question: at a point y, what direction of motion offers the quickest increase of a function $f : \mathbb{R}^n \to \mathbb{R}$?
- We use the approximation $f(x) \approx f(y) + \nabla f(y)^T(x - y)$.
- Suppose we want to choose x such that $||x - y||_2 = \Delta$ to maximize $f(x)$. How should we choose x?
- If Δ is small, we are justified in using the approximation above.
- Want: $\max_{||x-y||_2=\Delta} \nabla f(x)^T(x - y)$?
- Or: $\max_{||z||_2=\Delta} \nabla f(x)^T z$
- Solution: choose $z = x - y$ to be proportional to $\nabla f(y)$.
- Conclusion: the direction of steepest increase is proportional to ∇f.

- OK, how about the following: at a point y, what direction of motion offers the quickest **decrease** of a function $f : \mathbb{R}^n \to \mathbb{R}$?
- We use the approximation $f(x) \approx f(y) + \nabla f(y)^T(x - y)$.
- Suppose we want to choose x such that $||x - y||_2 = \Delta$ to **minimize** $f(x)$. How should we choose x?
- Want: $\min\limits_{||x-y||_2=\Delta} \nabla f(x)^T(x - y)$?
- Solution: choose $x - y$ to be proportional to $-\nabla f(y)$.
- Conclusion: $-\nabla f$ is the direction of steepest decrease.



Everything we've been talking about in one picture

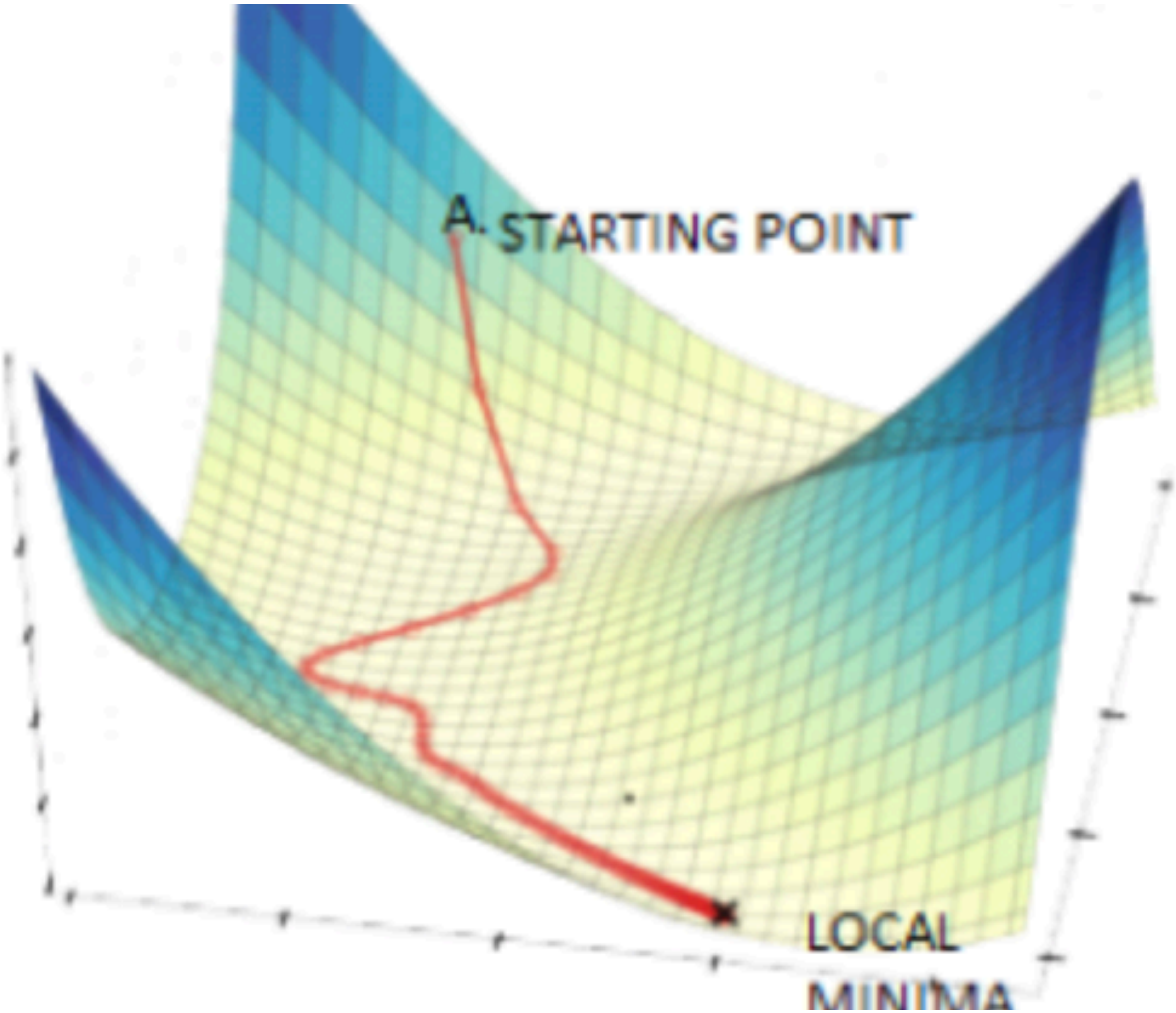
- All this motivates the gradient descent method.
- Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and we want to minimize it, i.e., we want to solve

  $$\min_x f(x)$$

- In general, this is hard. But here is something natural to do:

  Maintain an iterate x_t (initialize x_0 arbitrarily, perhaps $x_0 = 0$)

  Update $x_{t+1} = x_t - \alpha \nabla f(x_t)$
- Intuition: at every point, you are going in the direction of steepest descent.
- The step-size α should be small.



What gradient descent looks like in practice

- All of this is to solve $\min_x f(x)$
- If instead you want to solve $\max_x f(x)$ you instead update as

  $x_{t+1} = x_t + \alpha \nabla f(x_t)$
- Intuition: at every point, you are going in the direction of steepest ascent.
- Can generally go between the two cases by considering the transforming $g = -f$.

  Maximizing g is the same thing as minimizing f.

- One of the main difficulties in updating

  $x_{t+1} = x_t - \alpha \nabla f(x_t)$

  is the choice of step-size α.

- How small should α be?
- There is a tradeoff here. What is it?
- The smaller α is, the less progress this iteration makes towards the optimal solution per step.
- The larger α is, then the worse the approximation $f(x) \approx f(y) + \nabla f(y)^T (x - y)$ is on which the gradient descent method is based.
- In practice: try an α.

  If convergence too slow, but it seems to be converging, increase α.

  If it oscillates wildly, decrease α.

- Another idea: update $x_{t+1} = x_t - \alpha_t \nabla f(x_t)$ and choose α_t to go to zero at a slow rate.
- For example, $\alpha_t = 1/\sqrt{t}$ or $\alpha_t = 1/t$.
- Why? Well, maybe the first iterates will be too big, but the later ones should presumably be OK.
- ...and if it goes to zero slow enough, you'll get many iterations in before your step-size gets unreasonably small.

- This is good time to recall the chain rule.
- For a function of one variable, we have
$$\frac{d}{dx}f(g(x)) = f'(g(x))g'(x). \quad \text{(!)}$$
- You probably remember this as an equation, but it makes sense on a gut level.
- Summary of everything we just said:
$$q(x+\Delta) \approx q(x) + q'(x)\Delta$$
- So we need to approximate $f(g(x+\Delta))$.
- First: $f(g(x+\Delta)) \approx f(g(x) + g'(x)\Delta)$
- Next: $f(g(x+\Delta)) \approx f(g(x)) + f'(g(x))g'(x)\Delta$
- ....this is exactly what you get if I told you that $\dfrac{d}{dx}f(g(x)) = f'(g(x))g'(x)$

- Let's generalize this. Suppose $g : \mathbb{R}^n \to \mathbb{R}^m, f : \mathbb{R}^m \to \mathbb{R}$. Let $h(x) = f(g(x))$.
- Let's introduce the notation $\partial^i q$ to denote the derivative of the function q with respect to its i'th argument.
- Let us also use the notation $g(x) = \begin{pmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{pmatrix}$.
- We have that

  $$(\partial^i h)(x) = \sum_{j=1}^{n} (\partial^j f)(g(x))(\partial^i g_j)(x)$$

- This is the most general form of the chain rule.
- Should make sense:

  — if you perturb $(x_1, \ldots, x_n)$ to $(x_1 + \Delta, \ldots, x_n)$, this affects every g_j

  — to work out the effect on f, you've got to multiply these by the ``sensitivities'' of f with respect to each of these entries and add.