Lecture 14: Gradient descent – direction and step size

Ciaran Evans

Recap: optimization

- Derivative-free optimization
  - Compass search, Nelder-Mead, etc.
- Derivative-based optimization with closed form solutions
  - Least-squares linear regression, weighted least squares, etc.
- Derivative-based optimization with iterative methods
  - So far: gradient descent

Gradient descent

- Points $\mathbf{x} = (x_1, ..., x_d)^T \in \mathbb{R}^d$
- $f(\mathbf{x}) \in \mathbb{R}$
- Gradient:

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix} \in \mathbb{R}^d$$

- $\alpha > 0$

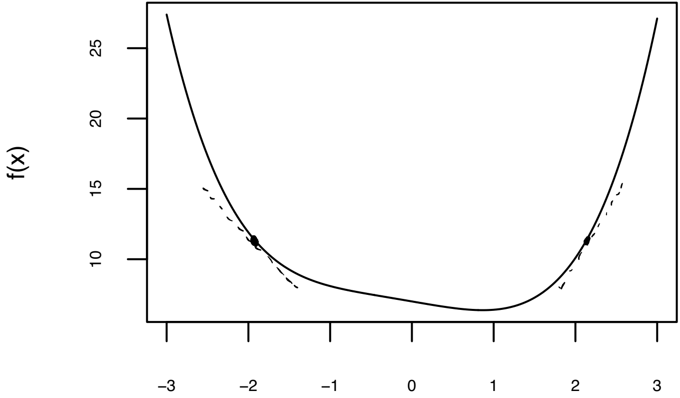
Iterative updates: $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)})$

Questions for today:

1. Why the gradient?
2. How far should we move? (i.e., choosing α)

Question 1: Why the gradient?

In the univariate case: $x^{(k+1)} = x^{(k)} - \alpha f'(x^{(k)})$

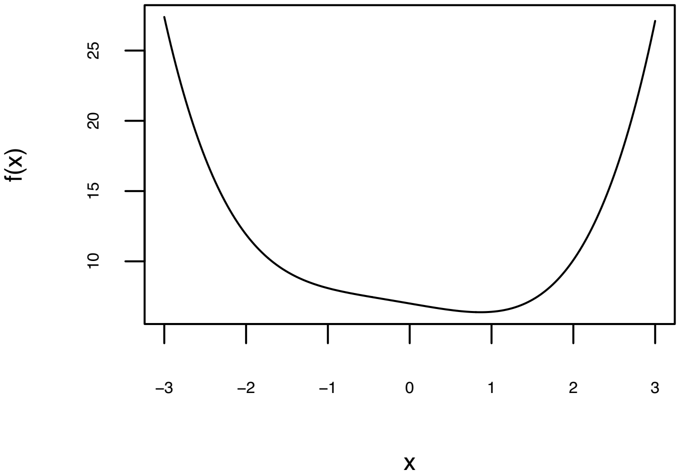


If $f'(x^{(k)}) > 0$, then $x^{(k+1)} < x^{(k)}$ (moved left)

If $f'(x^{(k)}) < 0$, then $x^{(k+1)} > x^{(k)}$ (move right)

Question 1: Why the gradient?

In the univariate case: $x^{(k+1)} = x^{(k)} - \alpha f'(x^{(k)})$

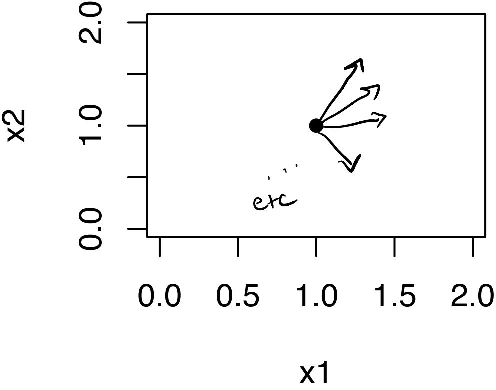


In the univariate case, there are only two possible directions, and the derivative tells us which way to go!

Why the gradient? Multivariate case

Example: $\mathbf{x} = (x_1, x_2)^T$, and $f(\mathbf{x}) = 5x_1^2 + 0.5x_2^2$

Suppose we are at point $\mathbf{x}^{(0)} = (1, 1)$



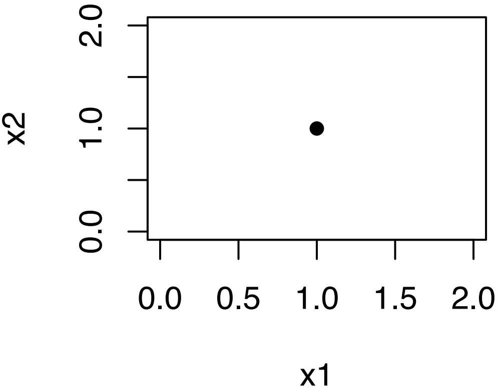
Question: How many directions could we move?

infinitely many!

Why the gradient? Multivariate case

Example: $\mathbf{x} = (x_1, x_2)^T$, and $f(\mathbf{x}) = 5x_1^2 + 0.5x_2^2$

Suppose we are at point $\mathbf{x}^{(0)} = (1, 1)$



Question: What criterion should I use to determine the direction of movement? direction of greatest decrease in f

Recap: what is a derivative?

Suppose we have a differentiable function $f : \mathbb{R} \to \mathbb{R}$. What does the *derivative* f' tell us?

Derivative: instantaneous rate of change at a point

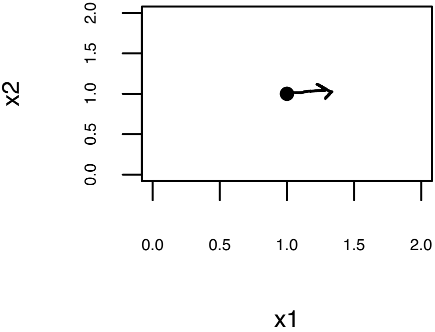
Formally:

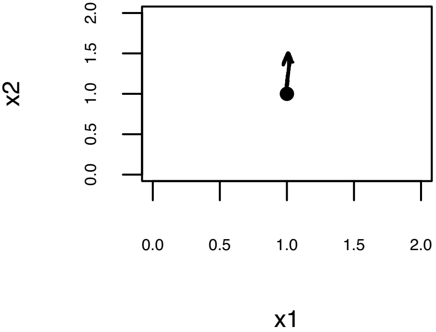
$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Derivatives for functions of multiple variables

Partial derivative: rate of change in f when moving along one of the axes

Example: $f(\mathbf{x}) = 5x_1^2 + 0.5x_2^2$





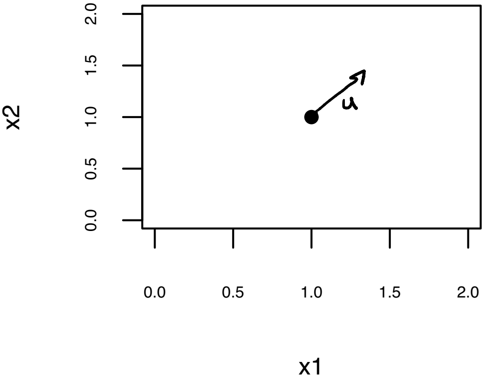
$\frac{\partial f}{\partial x_1} = 10x_1$

$\frac{\partial f}{\partial x_2} = x_2$

Directional derivatives

unit vector: $\|u\| = 1$

At point $\mathbf{x}$, and want to know how fast $f(\mathbf{x})$ changes in direction of unit vector $\mathbf{u}$



Directional derivative: $D_{\mathbf{u}} f(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{u}) - f(\mathbf{x})}{h}$

$D_u f(x) = \nabla f(x)^T u$

Directional derivatives

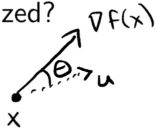
$\nabla f(x)^T u = \nabla f(x) \cdot u$ (dot product)

Directional derivative: $D_{\mathbf{u}} f(\mathbf{x}) = \lim_{h \to 0} \dfrac{f(\mathbf{x} + h\mathbf{u}) - f(\mathbf{x})}{h}$

Turns out:

$$D_u f(\mathbf{x}) = \nabla f(\mathbf{x})^T \mathbf{u}$$

Question: In which direction $\mathbf{u}$ is $D_{\mathbf{u}} f(\mathbf{x})$ maximized?



$\nabla f(x)^T u = \|\nabla f(x)\| \|u\| \cos(\theta)$

$= \|\nabla f(x)\| \cos(\theta)$

$\cos\theta \in [-1, 1]$

$\cos(0) = 1$

$\theta = 0 \Rightarrow \nabla f(x)$ and u are the same direction

$\cos(\pi) = -1 \Rightarrow \nabla f(x)$ & u are opposite directions

$\Rightarrow$ biggest increase : $\theta = 0$ (direction of gradient)
biggest decrease : $\theta = \pi$ (direction opposite gradient)

Directional derivatives

Directional derivative: $D_{\mathbf{u}} f(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{u}) - f(\mathbf{x})}{h}$

Turns out:

$$D_u f(\mathbf{x}) = \nabla f(\mathbf{x})^T \mathbf{u}$$

- Direction of greatest **increase** in f is $\nabla f(\mathbf{x})$
- Direction of greatest **decrease** in f is $-\nabla f(\mathbf{x})$

So: $\mathbf{x} - \alpha \nabla f(\mathbf{x})$ is movement in direction of *greatest decrease* in f

Question 2: How far should we move?

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)})$$

- α too big: sequence diverges
- α too small: takes too many iterations

Question: How would you decide on a “good” value of α to use at each step?

Question 2: How far should we move?

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)})$$

- α too big: sequence diverges
- α too small: takes too many iterations

Idea: maximize benefit:

$$\min_{\alpha > 0} f(\mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)}))$$

Line search

$$\min_{\alpha > 0} f(\mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)}))$$

- Exact minimization is expensive and unnecessary
- Instead: try a limited number of α values until $f(\mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)}))$ is "good enough"

Question: What is "good enough"?

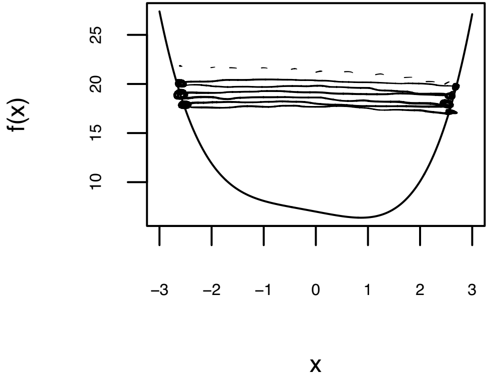
Requirement for α: initial idea

Idea: Choose α such that

requires that $f(x^{(k+1)})$ is smaller than $f(x^{(k)})$

$$f(\mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)})) < f(\mathbf{x}^{(k)})$$

Counterexample: Allows this sort of behavior:



Sufficient decrease condition

Idea: Decrease has to be "big enough"

Step size α must satisfy

$$f(\mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)})) \leq f(\mathbf{x}^{(k)}) - \overbrace{c_1 \alpha ||\nabla f(\mathbf{x}^{(k)})||^2}^{\text{α needs to allow us to decrease f by at least this amount}}$$

for some $c_1 \in (0, 1)$. (In practice, c_1 is pretty small, e.g. 10^{-4})

Backtracking line search

Simple, common way to choose α which often works:

1. Start with initial value of α (often $\alpha = 1$)
2. Check sufficient decrease condition:

$$f(\mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)})) \overset{?}{\leq} f(\mathbf{x}^{(k)}) - c_1 \alpha ||\nabla f(\mathbf{x}^{(k)})||^2$$

3. If sufficient decrease condition satisfied, use current value of α
4. Otherwise, $\alpha = 0.5\alpha$ and go back to step 2

Your turn

Practice questions on the course website:

https://sta379-s25.github.io/practice_questions/pq_14.html

- Try backtracking line search
- Start in class. You are welcome to work with others
- Practice questions are to help you practice. They are not submitted and not graded
- Solutions are posted on the course website