Gradient Descent:

↳ Popular & important algorithm, used very widely for its simplicity, computational complexity

But first, what is a descent direction?

Def'n: $\vec{v} \in \mathbb{R}^n$ is a descent direction for $f : \mathbb{R}^n \to \mathbb{R}$ at x if

$$\langle \vec{v}, \nabla f|_x \rangle < 0$$

Why does this def'n make sense?

Recall: Taylor's theorem ⟹

$$f(\vec{x} + \mu\vec{v}) = f(\vec{x}) + \boxed{\mu \vec{v}^T \nabla f(\vec{\xi})}$$

where $\vec{\xi}$ lies between x & $\vec{x} + \mu\vec{v}$

so $\vec{\xi} = \vec{x} + \tilde{\mu}\vec{v}$ where $\tilde{\mu} \in [0, \mu]$

But if ∇f is continous then $\exists \mu$ that is small enough such that

$$\boxed{v^T \nabla f(\vec{x}+\tilde{\mu}\vec{v}) < 0}$$

$$\forall \tilde{\mu} \in [0, \mu]$$

This in turn implies that

$$f(x+\mu v) < f(x)$$

So, moving in the direction of $\vec{v}$ by a small amount decreases the function, hence we are justified in calling v a descent direction.

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

$$\nabla f(x) = (2x_1, 4x_2)$$

$\Rightarrow$ at $\vec{x} = (1,1)$, $\nabla f|_{(1,1)} = (2,4)$

Let $\vec{v} = (0,-1)$, then $\langle \vec{v}, \nabla f|_{(1,1)} \rangle = -4 < 0$

So $\vec{v}$ is a descent direction for f at $(1,1)$.

Let's check $f(x+\mu v)$ compared to $f(x)$

$$f(\underbrace{\vec{x}+\mu\vec{v}}_{=(x_1+\mu v_1,\ x_2+\mu v_2)}) = x_1^2 + 2(x_2-\mu)^2$$

$v_1 \to 0$, $v_2 \to -1$

So $f(\vec{x}+\mu\vec{v})$ at $(1,1)$ is

$$f(\vec{x}+\mu\vec{v}) = 1+2(1-\mu)^2$$
$$= 3+2\mu^2-2\mu$$

while

$$f(\vec{x}) = 3$$

so $f(\vec{x}+\mu\vec{v}) < f(\vec{x})$ whenever $0 \leq \mu < 1$

Remark: If we choose $\vec{v} = -\nabla f(x)$ we always get a descent direction provided $\nabla f(x) \neq 0$.

Idea behind the gradient descent algorithm

Goal: to find a (local) minimizer for the function $f:\mathbb{R}^n \to \mathbb{R}$, i.e. find x^* s.t. $f(x^*) < f(x)$ $\forall x \in \mathcal{N}$

↑ a neighborhood around x^*

Idea: Suppose you make a guess x, and now you want to improve it.
You can pick a descent direction, e.g., $\vec{v} = -\nabla f(x)$ and move ~~by~~ a small amount in that direction

$$\underbrace{f(\vec{x} + \mu \vec{v})}_{\approx\, f(\vec{x}) + \mu \underbrace{v^T \nabla f(x)}_{<0}} < f(\vec{x})$$

Gradient descent (GD)

- Choose $x^{(0)} \in \mathbb{R}^n$
- For $t = 1, 2, \dots,$ (or until a stopping criterion is met)

set
$$x^{(t)} = x^{(t-1)} - \mu^{(t-1)} \nabla f(x^{(t-1)})$$

Example: $f : \mathbb{R}^2 \to \mathbb{R}$

$$f(x) = x_1^2 + 2x_2^2$$

$$\hookrightarrow \nabla f(x) = (2x_1, 4x_2)$$

(here it is clear that $x^* = (0,0)$)

Suppose we start at $x^{(0)} = (2,3)$ and suppose we choose $\mu = 0.1$ then GD gives:

$$x^{(1)} = x^{(0)} - \mu \nabla f(x^{(0)})$$

$$\iff x^{(1)} = (2,3) - 0.1\,(4, 12)$$
$$= (2,3) - (0.4, 1.2)$$
$$= (1.6, 1.8)$$

$$x^{(2)} = (1.6, 1.8) - 0.1\,(3.2, 7.2)$$
$$= (1.28, 1.02)$$

$$\vdots$$

$$x^{(10)} \approx (0.2147, 0.0181)$$

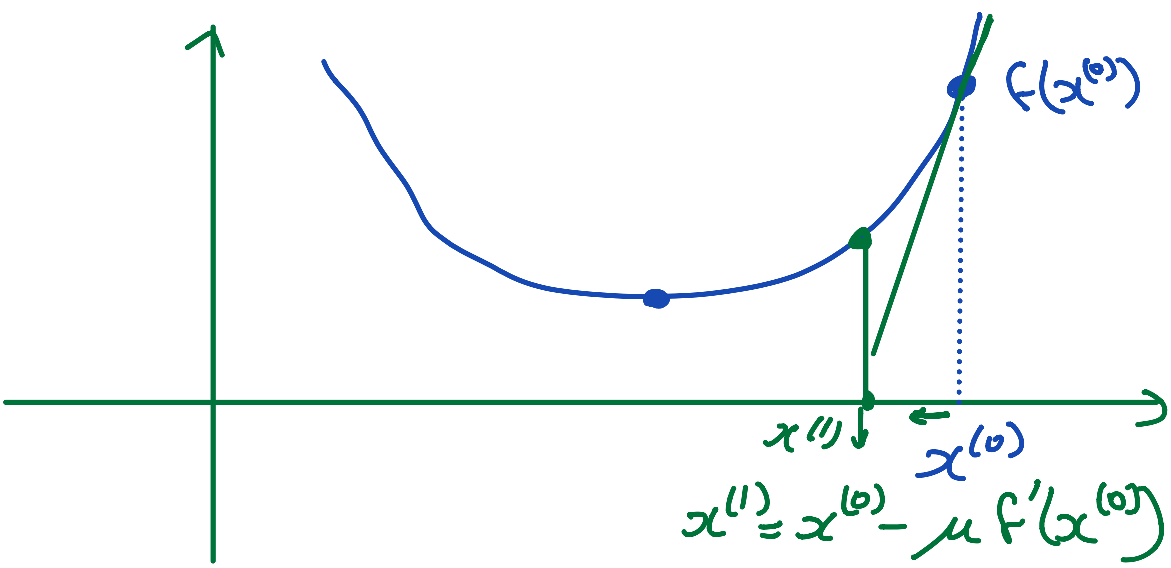
$$\vdots$$

$$x^{(20)} \approx (0.0231, 0.0001)$$

$$\vdots$$

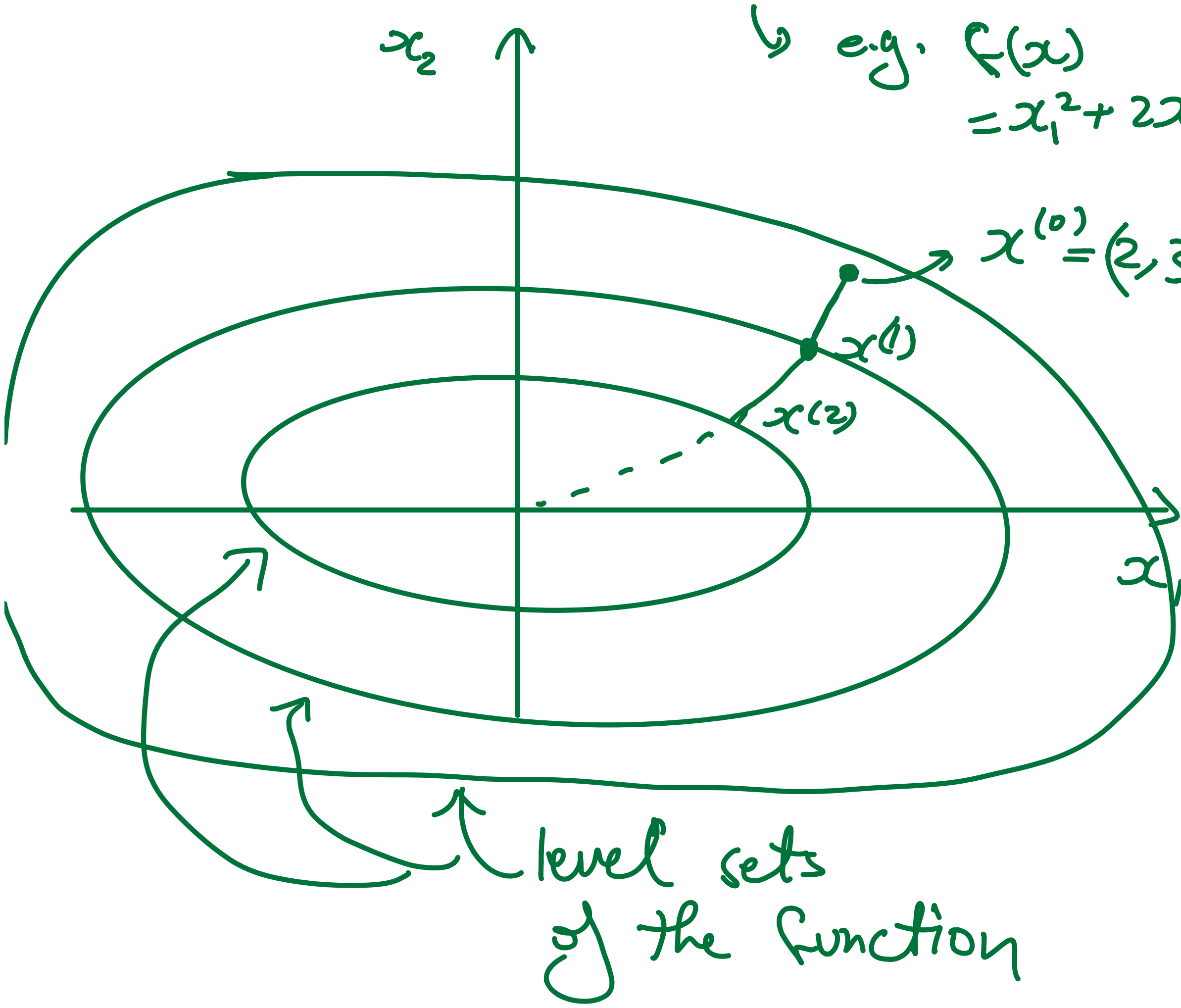
$$x^{(50)} \approx (0.285\times 10^{-4}, 2.4\times 10^{-11})$$

Illustration of what GD is doing (in the single variable case):



What about when $f: \mathbb{R}^2 \to \mathbb{R}$?

↳ e.g. $f(x) = x_1^2 + 2x_2^2$



A couple of issues to consider when applying GD:

1) How do we choose $\mu^{(t)}$?

2) How do we know when to stop?

Let's answer 2) first:

Theorem: Necessary conditions for optimality

(1) If f is continuously differentiable & x^* is a local min then

$$\nabla f(x^*) = 0.$$

(2) If $\nabla^2 f$ is continuous and x^* is a local minimum

$$\nabla^2 f(x^*) \succeq 0$$

notation:
$A \succeq 0$
$\iff$ A is a PSD matrix

This theorem means that if f is cont. differentiable & has a continuous Hessian then we must have

$$\nabla f(x^*) = 0$$
$$\nabla^2 f(x^*) \succeq 0$$