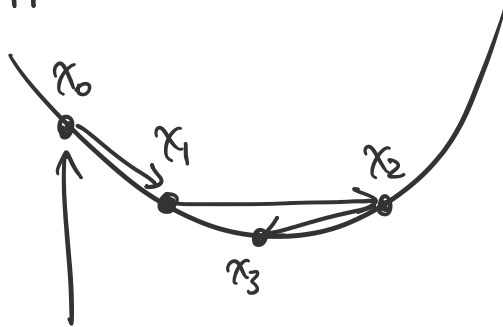


Convex Optimization, Gradient Descent

Motivation: suppose want to minimize a "nice" function (in 1-dim):



Start anywhere.

If slope \searrow (derivative < 0) then go right.

If slope \swarrow (derivative > 0) then go left.

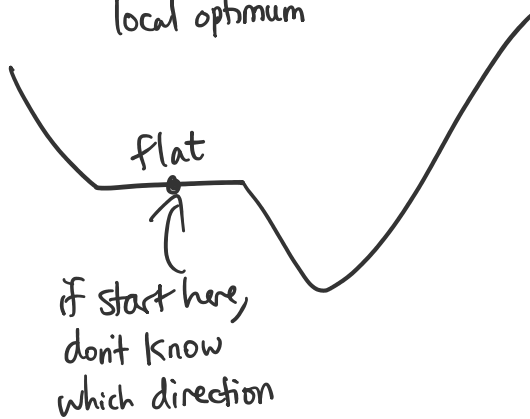
Take progressively smaller steps over time

In fact, scale step size by derivative:

$$x_{t+1} \leftarrow x_t + \eta \cdot f'(x_t)$$

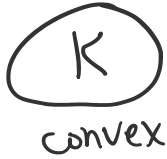
Fact: this "converges" to the minimum when f is convex.

Why convex?

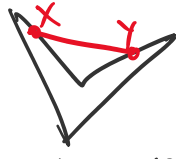


Def: A set $K \subseteq \mathbb{R}^n$ is convex if $\forall x, y \in K, \underbrace{\lambda x + (1-\lambda)y}_{\dots} \in K$

Def: A set $K \subseteq \mathbb{R}^n$ is convex if $\forall x, y \in K$, $\lambda x + (1-\lambda)y \in K$
 any weighted average of x, y



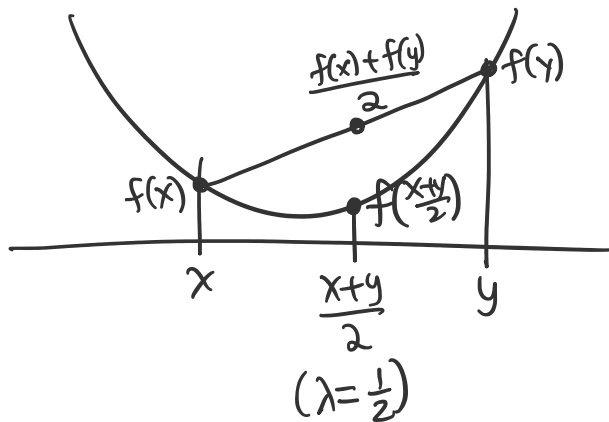
convex



not convex

Def: A convex function $f: K \rightarrow \mathbb{R}$ defined on convex set K is called convex if $\forall x, y \in K$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$



Goal: Solve $\min_{x \in \mathbb{R}} f(x)$ and $\min_{x \in K} f(x)$ for convex f, K .
 "unconstrained" "constrained"

Def: The gradient $\nabla f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the function

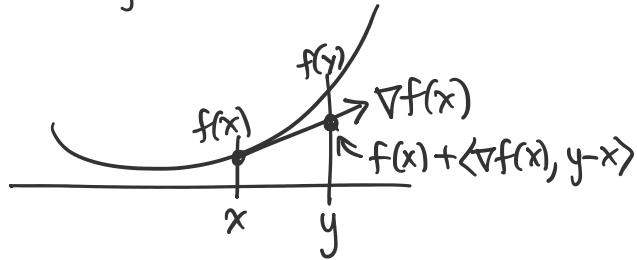
$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right)$$

$$\frac{\partial f}{\partial x_i}(x) = \lim_{\epsilon \rightarrow 0} \frac{f(x + \epsilon e_i) - f(x)}{\epsilon}$$

Fact (First-order condition): A function $f: K \rightarrow \mathbb{R}$ is convex iff

Fact (First-order condition): A function $f: K \rightarrow \mathbb{R}$ is convex iff

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle \quad \forall x, y \in K.$$



Proof: (\Rightarrow) Suppose $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$
 Write as
$$\frac{f(x + \lambda(y-x)) - f(x)}{\lambda} \leq f(y) - f(x)$$

Take limit $\lambda \rightarrow 0^+$: $\langle \nabla f(x), y-x \rangle \leq f(y) - f(x).$

(\Leftarrow for $\lambda = \frac{1}{2}$): $\langle \nabla f(\frac{x+y}{2}), y - \frac{x+y}{2} \rangle \leq f(y) - f(\frac{x+y}{2})$
 $\langle \nabla f(\frac{x+y}{2}), x - \frac{x+y}{2} \rangle \leq f(x) - f(\frac{x+y}{2})$

$$+ \frac{\quad}{0} \leq f(x) + f(y) - 2f(\frac{x+y}{2}).$$

Fact (First-order optimality): For convex $f: K \rightarrow \mathbb{R}$, if $\nabla f(x) = 0$ then $x = \min_{x \in K} f(x)$
 (Converse not true: possible that $\nabla f(x) \neq 0 \quad \forall x \in K$)

Def (Lipschitz): $f: K \rightarrow \mathbb{R}$ is G -Lipschitz w.r.t. norm $\|\cdot\|$ if

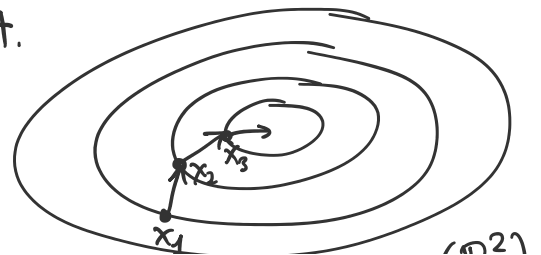
$$|f(x) - f(y)| \leq G \|x - y\| \quad \forall x, y \in K.$$

Fact: $f: K \rightarrow \mathbb{R}^n$ is G -Lipschitz w.r.t. $\|\cdot\|_2$ iff $\|\nabla f(x)\|_2 \leq G \quad \forall x \in K.$

Gradient Descent: unconstrained setting

Idea: move in (opposite) direction of gradient \rightarrow steepest descent,
 and scale step size to gradient.

Algo: Initialize $x_0 \leftarrow$ any point in K
 $t = 0, 1, \dots, T-1$:



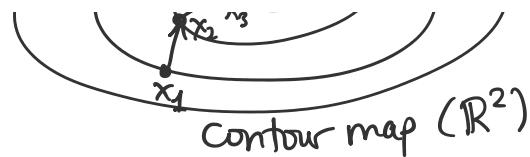
Algo: Initialize $x_0 \leftarrow$ any $p \dots$

For $t=0, 1, \dots, T-1$:

$$x_{t+1} \leftarrow x_t - \eta \cdot \nabla f(x_t)$$

Return $\bar{x} = \frac{1}{T} \sum_{t=1}^T x_t$. \leftarrow Why not just x_T ? Also works, but more complicated analysis.

\uparrow not necessarily in K ! K is "bounding box"



Theorem: Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be convex, differentiable, and G -Lipschitz. Let $L = \max_{x,y \in K} \|x-y\|^2$. Then,

$$\sum_{t=1}^T f(x_t) \leq \underbrace{T f(x^*)}_{\text{optimal play}} + \underbrace{\frac{1}{2} \eta T G^2}_{\text{regret}} + \frac{1}{2\eta} L^2 \quad \forall x^* \in K.$$

Proof: Later.

like εT in MWU like $\frac{\ln N}{\varepsilon}$ in MWU

Corollary: Setting $T = \frac{G^2 L^2}{\varepsilon^2}$ and $\eta = \frac{L}{G\sqrt{T}}$ for any $x^* \in K$,

the solution \bar{x} satisfies $f(\bar{x}) \leq f(x^*) + \varepsilon$

Proof: $f(\bar{x}) = f\left(\frac{1}{T} \sum_{t=1}^T x_t\right) \leq \frac{1}{T} \sum_{t=1}^T f(x_t)$ [by convexity]

$$\leq f(x^*) + \underbrace{\frac{1}{2} \eta G^2 + \frac{1}{2\eta T} L^2}_{\text{balance the errors with } \eta:}$$

balance the errors with η :

$$\frac{1}{2} \eta G^2 = \frac{1}{2} \frac{L G}{\sqrt{T}}$$

$$\frac{1}{2\eta T} L^2 = \text{same}$$

$$= f(x^*) + \frac{L G}{\sqrt{T}}$$

$$= f(x^*) + \varepsilon.$$

Proof of Theorem: Define potential function $\Phi_t = \frac{\|x_t - x^*\|^2}{2\eta}$.

Proof of Theorem: Define potential function $\Phi_t = \frac{\|x_t - x^*\|^2}{2\eta}$.

Lemma: $f(x_t) + (\Phi_{t+1} - \Phi_t) \leq f(x^*) + \frac{1}{2}\eta G^2$.

Proof: Use identity $\|a+b\|^2 = \|a\|^2 + 2\langle a, b \rangle + \|b\|^2$ with

$$a = x_t - x^*, \quad b = x_{t+1} - x_t = -\eta \nabla f(x_t)$$

$$\Phi_{t+1} - \Phi_t = \frac{1}{2\eta} \left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2 \right)$$

$$= \frac{1}{2\eta} \left(\|a+b\|^2 - \|a\|^2 \right)$$

$$= \frac{1}{2\eta} \left(2\langle a, b \rangle + \|b\|^2 \right)$$

$$= \frac{1}{2\eta} \left(2\langle x_t - x^*, -\eta \nabla f(x_t) \rangle + \|\eta \nabla f(x_t)\|^2 \right)$$

$\leq \eta^2 G^2$ since G -Lipschitz

$$\leq \langle x_t - x^*, -\nabla f(x_t) \rangle + \frac{1}{2}\eta G^2$$

By convexity, $f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle \leq f(x^*)$
 $= \langle x_t - x^*, -\nabla f(x_t) \rangle$

$$\leq f(x^*) - f(x_t) + \frac{1}{2}\eta G^2. \quad \square$$

Summing Lemma over all $t=0, 1, \dots, T-1$:

$$\sum_{t=0}^{T-1} \left(f(x_t) + (\Phi_{t+1} - \Phi_t) \right) \leq T f(x^*) + \frac{1}{2}\eta G^2 T$$

telescopes to $\Phi_T - \Phi_0$

$$\Rightarrow \sum_{t=0}^{T-1} f(x_t) \leq T f(x^*) + \frac{1}{2}\eta G^2 T - \underbrace{\Phi_T}_{\geq 0} + \underbrace{\Phi_0}_{\leq \frac{L^2}{2\eta}}$$

$$\leq T f(x^*) + \frac{1}{2}\eta G^2 T + \frac{1}{2\eta} L^2. \quad \square$$

$$\leq T f(x^*) + \frac{1}{2} \eta G^2 T + \frac{1}{2\eta} L^2. \quad \square$$

Online Gradient Descent: Algo plays $x_t \in K$, adversary plays convex f_t .

Same proof gives

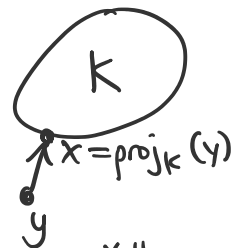
$$\underbrace{\sum_{t=1}^T f_t(x_t)}_{\text{algo score}} \leq \underbrace{\sum_{t=1}^T f_t(x^*)}_{\text{(best) expert score}} + \underbrace{\frac{1}{2} \eta G^2 + \frac{1}{2\eta T} L^2}_{\text{regret}} \quad \forall x^* \in K.$$

Each $x \in K$ is an expert. (Infinite experts!)
But if "loss function" f_t is convex, then still low regret.

Projected Gradient Descent: What if we always $\bar{x} \in K$?

Simple fix: if $x_{t+1} \notin K$, then project it to K .

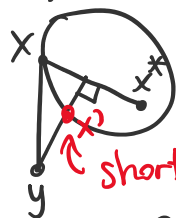
Def (Projection): $\text{proj}_K(y) = \underset{x \in K}{\text{argmin}} \|x - y\|_2.$



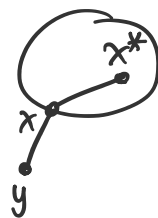
Fact: If $x = \text{proj}_K(y)$ and $x^* \in K$, then $\|x - x^*\| \leq \|y - x^*\|.$

Proof: We claim $y - x - x^*$ forms a non-acute angle:

If not, then



shorter distance $\|x^* - y\|_2$, contradiction



Algo: Initialize $x_0 \leftarrow$ any point in K

For $t = 0, 1, \dots, T-1$:

$$x'_{t+1} \leftarrow x_t - \eta \cdot \nabla f(x_t)$$

$$x_{t+1} \leftarrow \text{proj}_K(x'_{t+1})$$

$$\text{Return } \bar{x} = \frac{1}{T} \sum_{t=1}^T x_t.$$

$$\text{Return } \bar{x} = \frac{1}{T} \sum_{t=1}^T x_t.$$

Same proof works, except at step

$$\bar{\Phi}_{t+1} - \bar{\Phi}_t = \frac{1}{2\eta} \left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2 \right)$$

$$\leq \frac{1}{2\eta} \left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2 \right)$$

⋮
(same)