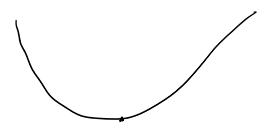
* Minima of convex functions:

(P) $\min_{x \in \mathcal{X}} f(x)$: f is convex, $\mathcal{X}$ is convex, closed, and nonempty

* <u>Thm 2.6</u> Consider (P). We have the following:

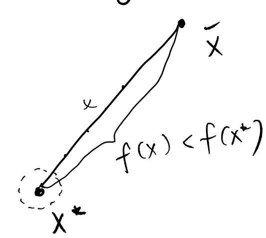
(a) Any local solution to (P) is also a global solution.



(b) The set of global solutions to (P) is convex.

<u>Proof:</u>

(a) Suppose f.p.o.c. that x^* is a local but not a global solution. Then $\exists\, \bar{x} \in \mathcal{X}$ s.t. $f(\bar{x}) < f(x^*)$.



As $\mathcal{X}$ is convex, $\forall\, \alpha \in (0,1)$:

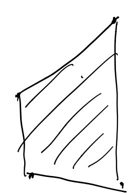
$(1-\alpha)\, x^* + \alpha \bar{x} \in \mathcal{X}$.

As f is convex, $\forall\, \alpha \in (0,1)$:

$$f((1-\alpha)x^* + \alpha \bar{x}) \leq (1-\alpha) f(x^*) + \alpha f(\bar{x}) < f(x^*).$$

$\Rightarrow$ Every neighborhood of x^* must include a point $(1-\alpha)x^* + \alpha \bar{x}$ for some $\alpha > 0$ that will have a <u>strictly</u> lower function value. $\Rightarrow$ x^* cannot be a local solution.

(b) Let $x^*, \bar{x} \in \mathcal{X}$ be any two global solutions.



$\mathcal{X}$ is convex $\Rightarrow \forall\, \alpha \in (0,1)$: $(1-\alpha)\, x^* + \alpha \bar{x} \in \mathcal{X}$.

f is convex $\Rightarrow \forall\, \alpha \in (0,1)$:

$$f((1-\alpha)x^* + \alpha \bar{x}) \leq (1-\alpha) f(x^*) + \alpha f(\bar{x})$$
$$= f(x^*) = f(\bar{x})$$

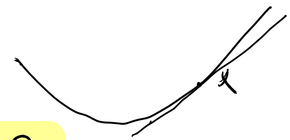
$\Rightarrow f((1-\alpha)\, x^* + \alpha \bar{x}) = f(x^*)$

$\Rightarrow (1-\alpha)\, x^* + \alpha \bar{x}$ is a global solution.

$\Rightarrow$ the set of global solutions must be convex. $\blacksquare$

* Thm.

(a) Let f be cont.'ly diff.'able. f is convex if and only if $\forall x, y: \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

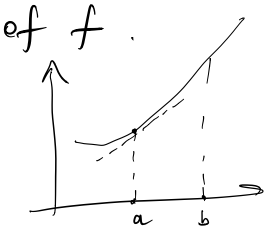


(b) Let f be twice cont.'ly diff.'able. f is convex if and only if $\forall x: \nabla^2 f(x) \succeq 0$.

* Thm 2.7 Let f be cont.'ly diff.'able and convex. If $\nabla f(x^*) = 0$, then x^* is a global min of f.

Pf: Use Part (a) of the Thm above:

$$\forall x: \quad f(x) \geq f(x^*) + \langle \underbrace{\cancel{\nabla f(x^*)}}_{0}, x - x^* \rangle.$$



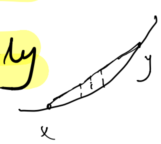
(for constrained setups, we would use $\langle \nabla f(x^*), x - x^* \rangle \geq 0, \ \forall x$)

* Strongly convex functions:

* Def. Given $m > 0$, we say that $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ is m-strongly convex (or strongly convex w/ modulus m), if $\forall x, y \in \mathbb{R}^d$:

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha) f(x) + \alpha f(y) - \frac{m}{2}(1-\alpha)\alpha \|y - x\|^2.$$

* Ex: 1) When f is cont.'ly diff.'able, equivalently



$$\forall x, y: \ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|^2.$$

2) When f is twice cont.'ly diff.'able, equivalently:

$$\forall x: \quad \nabla^2 f(x) \succeq m I.$$

* Thm 2.8. Suppose that $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ is cont.'ly diff.'able and m-strongly convex for some $m > 0$. If $\nabla f(x^*) = 0$, then x^* is the unique global min of f.

Proof: From Ex 1):

$$\forall x: \ f(x) \geq f(x^*) + \langle \overset{0}{\cancel{\nabla f(x^*)}}, x - x^* \rangle + \underbrace{\frac{m}{2}\|x - x^*\|^2}_{>0 \text{ unless } x = x^*}$$

∎

* **Growth of sequences:**

$\{a_k\}_{k\geq 1}$, $\{b_k\}_{k\geq 1}$, $\forall k: a_k, b_k \geq 0$.

$a_k \leq 10 b_k$
$a_k = \mathcal{O}(b_k)$

* "Big-Oh" notation:

$a_k = \mathcal{O}(b_k) \iff (\exists M > 0)(\exists K < \infty)(\forall k \geq K): a_k \leq M b_k$.

$\left(\text{E.g., } k = \mathcal{O}\left(\frac{1}{10} k^2\right), \; k = \mathcal{O}\left(\frac{1}{10!} k\right) \right)$

* If $a_k = \mathcal{O}(b_k)$ and $b_k = \mathcal{O}(a_k)$, we write $a_k = \Theta(b_k)$.

* "Little-Oh" notation:

$a_k = o(b_k) \iff \lim_{k \to \infty} \frac{a_k}{b_k} = 0$.

---

* **Algorithmic setup:**

1) first-order oracle model:

$$x \longrightarrow \boxed{\text{oracle}} \longrightarrow \underline{f(x)}, \underline{\nabla f(x)}$$

2) second-order oracle model:

$$x \longrightarrow \boxed{\text{oracle}} \longrightarrow f(x), \nabla f(x), \nabla^2 f(x)$$

* All algorithms we consider in this class are iterative:
  → start w/ some x_0, get oracle answers for x_0, choose x_1
  → at iteration k, get oracle answers for x_k, choose x_{k+1}

* **Basic Descent Methods:**

* Assumptions for this part:

(A1) f is L-smooth for some $L < \infty$ (thus also cont.'ly diff.'able)

(A2) $\mathcal{X} = \mathbb{R}^d$, i.e., the problem is unconstrained

<u>Note</u>: for now, and until explicitly stated otherwise, we are <u>not</u> assuming that f is convex.

* **Def.** $p \in \mathbb{R}^d$ is a descent direction for f at x if $f(x+tp) < f(x)$ for all suff. small $t > 0$.

* **Prop 3.2.** If f is cont.'ly diff.'able (in a neighborhood of x), then any p s.t. $\langle \nabla f(x), p \rangle < 0$ is a descent direction.

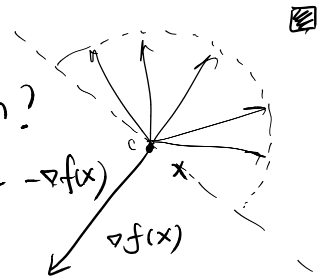
Proof: TT + continuity of ∇f: $(y = x + tp)$

$$f(x+tp) = f(x) + t \langle \nabla f(x + \mu tp), p \rangle \text{ for some } \mu \in (0,1).$$

We know that $\langle \nabla f(x), p \rangle < 0$. As ∇f is continuous, for all suff. small $t > 0$:

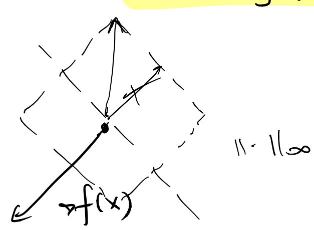
$$t \langle \nabla f(x + \mu tp), p \rangle < 0.$$

$$\Rightarrow f(x+tp) < f(x).$$



* What would be a good descent direction?
  - could try to move in the direction of $-\nabla f(x)$
  - justification:
    Look at all p w/ $\|p\|_2 = 1$. Then:

$$\inf_{\|p\|_2 = 1} \langle \nabla f(x), p \rangle = -\|\nabla f(x)\|_2 \text{ attained for } p = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$



* "Simplest" descent algorithm:

$$x_{k+1} = x_k - \underbrace{\alpha_k}_{\text{step size}} \nabla f(x_k)$$

α_k is chosen small enough so that $f(x_{k+1}) < f(x_k)$, assuming $\nabla f(x_k) \neq 0$.

"gradient method", "gradient descent", "steepest descent"

* From Lemma 2.2:

$$\boxed{\forall x,y: \quad f(y) \le f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|_2^2 .}$$

$x = x_k$

$$\forall y: f(y) \le \underbrace{f(x_k) + \langle \nabla f(x_k), \overset{=0}{y-x_k}\rangle + \frac{L}{2}\|\overset{=0}{y-x_k}\|_2^2}$$ for $y = x_k$

$$x_{k+1} = \operatorname*{argmin}_{y\in\mathbb{R}^d} \{ \qquad \} .$$

$$\Rightarrow \quad f(x_{k+1}) \le f(x_k) .$$

$$\nabla f(x_k) + L(x_{k+1} - x_k) = 0$$

$$\Leftrightarrow \quad x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) .$$

$$\Rightarrow \quad f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2 .$$

* Ex. If $x_{k+1} = x_k - \alpha \nabla f(x_k)$, $\alpha \in (0, \frac{1}{L}]$, then:

$$\boxed{f(x_{k+1}) \le f(x_k) - \frac{\alpha}{2}\|\nabla f(x_k)\|_2^2} \quad \text{"Descent Lemma"}$$

Using Descent Lemma, if $\boxed{x_{k+1} = x_k - \alpha \nabla f(x_k), \forall k}$, (GD)

$$\begin{aligned} f(x_{k+1}) &\le f(x_k) - \frac{\alpha}{2}\|\nabla f(x_k)\|_2^2 \\ &\le \underbrace{f(x_{k-1}) - \frac{\alpha}{2}\|\nabla f(x_{k-1})\|_2^2} - \frac{\alpha}{2}\|\nabla f(x_k)\|_2^2 \\ &\ \vdots \\ &\le f(x_0) - \frac{\alpha}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2 . \end{aligned}$$

$$\frac{\alpha}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2 \le f(x_0) - f(x_{k+1})$$

Let's assume $f(x) \ge f_* > -\infty$, $\forall x$. Then:

$$(*) \qquad \underline{\frac{\alpha}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2} \le \underline{f(x_0) - f_*}$$

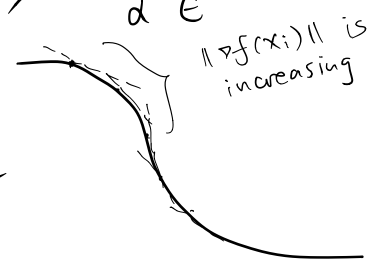
(**) $\frac{\alpha}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2 \geq \frac{\alpha}{2}(k+1)\cdot\min_{0\le i\le k}\|\nabla f(x_i)\|_2^2$

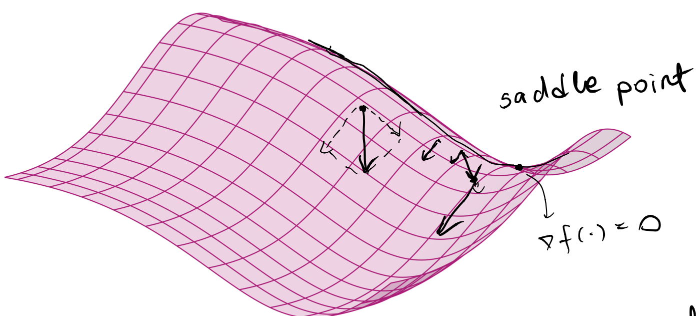
(*) + (**) $\Rightarrow$ $\min_{0\le i\le k}\|\nabla f(x_i)\|_2^2 \le \frac{2(f(x_0)-f_*)}{\alpha(k+1)}$.

$$\min_{0\le i\le k}\|\nabla f(x_i)\|_2 \le \sqrt{\frac{2(f(x_0)-f_*)}{\alpha(k+1)}} \le \epsilon$$

For any target error $\epsilon > 0$, GD satisfies

$\min_{0\le i\le k}\|\nabla f(x_i)\|_2 \le \epsilon$ for $k+1 \geq \frac{2(f(x_0)-f_*)}{\alpha\,\epsilon^2}$.





"randomly initialized GD is unlikely to converge to a saddle point"

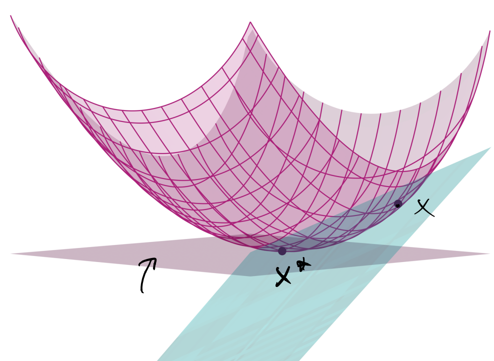
* The convex case:

How does the convexity help?

Let $x^* \in \arg\min_{x\in\mathbb{R}^d} f(x)$

might not be unique; assume all minimizers are from $\mathbb{R}^d$.

$\forall x$: $f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x\rangle$



Want: bound $\underbrace{f(x)-f(x^*)}_{\text{optimality gap}}$.

GD: $x_{k+1} = x_k - \alpha \nabla f(x_k)$, $\alpha \in (0, \frac{1}{L}]$

Know from before: $f(x_{k+1}) \le f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|_2^2$.

$$f(x^*) \geq f(x_k) + \langle \underbrace{\nabla f(x_k)}_{\frac{1}{\alpha}(x_k - x_{k+1})}, x^* - x_k \rangle$$

$$= f(x_k) + \frac{1}{\alpha} \langle x_k - x_{k+1}, x^* - x_k \rangle$$

$$\left[(a-b)\cdot(c-a) = \frac{1}{2}(c-b)^2 - \frac{1}{2}(a-b)^2 - \frac{1}{2}(c-a)^2 \right]$$

$$= f(x_k) + \frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2\alpha} \|x_k - x^*\|_2^2 - \frac{1}{2\alpha} \|\underbrace{x_k - x_{k+1}}_{-\alpha \nabla f(x_k)}\|_2^2$$

$$= f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|_2^2 + \frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2\alpha} \|x_k - x^*\|_2^2.$$

$$\geq f(x_{k+1}) + \frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2\alpha} \|x_k - x^*\|_2^2.$$

1) $\|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|_2^2 \le 2\alpha (f(x^*) - f(x_{k+1}))$.

≤ 0, w/ strict ineq whenever $f(x_{k+1}) \ne f(x^*)$.

$$\inf_{x^* \in \arg\min_x f(x)} \|x_k - x^*\|_2 \xrightarrow{k \to \infty} 0.$$

2) $f(x_{k+1}) - f(x^*) \le \frac{1}{2\alpha} \left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 \right)$;

$$\sum_{k=0}^{K} (f(x_{k+1}) - f(x^*)) \le \frac{1}{2\alpha} (\|x_0 - x^*\|_2^2 - \|x_{K+1} - x^*\|_2^2)$$

$$\le \frac{1}{2\alpha} \|x_0 - x^*\|_2^2.$$

$$\sum_{k=0}^{K} (f(x_{k+1}) - f(x^*)) \geq (K+1) (f(x_{K+1}) - f(x^*))$$

$$f(x_{K+1}) - f(x^*) \le \frac{\|x_0 - x^*\|_2^2}{2\alpha(K+1)}.$$

$\forall \epsilon > 0$: $f(x_K) - f(x^*) \leq \epsilon$ after at most

$$K = \left\lceil \frac{\|x_0 - x^*\|_2^2}{2\alpha\epsilon} \right\rceil \quad \text{iterations.}$$

* The strongly convex case:

$\forall k$:

$$f(x^*) \geq f(x_k) + \langle \underbrace{\nabla f(x_k)}_{\frac{1}{\alpha}(x_k - x_{k+1})}, x^* - x_k \rangle + \boxed{\frac{m}{2} \|x^* - x_k\|_2^2}$$

$$\geq f(x_{k+1}) + \frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2\alpha} \|x_k - x^*\|_2^2 + \frac{m}{2} \|x^* - x_k\|_2^2$$

$$= f(x_{k+1}) + \frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 \boxed{- \left(\frac{1}{2\alpha} - \frac{m}{2}\right) \|x_k - x^*\|_2^2}$$

1) $$\frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 \leq \boxed{\left(\frac{1}{2\alpha} - \frac{m}{2}\right) \|x_k - x^*\|_2^2} + \underbrace{f(x^*) - f(x_{k+1})}_{\leq 0}$$

$$\boxed{\|x_{k+1} - x^*\|_2^2 \leq (1 - m\alpha) \|x_k - x^*\|_2^2}.$$

Ex. $\boxed{m\alpha \in (0, 1]}$. $\left(\alpha \in (0, \tfrac{1}{L}]\right)$

$$\|x_{k+1} - x^*\|_2^2 \leq (1 - m\alpha)^{k+1} \|x_0 - x^*\|_2^2.$$

$$\|x_{k+1} - x^*\| \leq \epsilon \quad \text{for} \quad k = O\left(\frac{1}{m\alpha} \log\left(\frac{\|x_0 - x^*\|}{\epsilon}\right)\right)$$