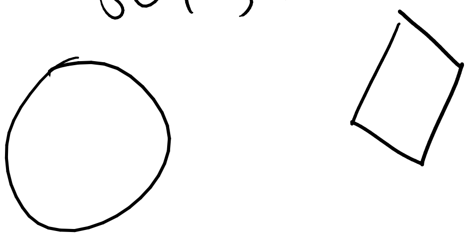
Convexity.

Let V be a vector space over the reals.
A set $S \subset V$ is <u>convex</u> if, for all $x, y \in S$, $\lambda \in [0,1]$,

$$(1-\lambda)x + \lambda y \in S.$$

Convex sets:

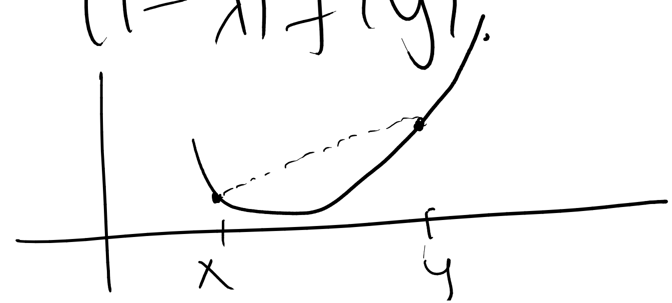


Non-convex:

A function $f: S \to \mathbb{R}$ defined on convex set S is called <u>convex</u> if, for all $x, y \in S$, $\lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda) y) \le \lambda f(x) + (1-\lambda) f(y).$$

"graph of f lies below line segment joining $f(x)$ & $f(y)$"



Exercise: Let $S \subset \mathbb{R}^d$ be convex. Call $x \in S$ a local min if $\exists r > 0$ st, for $B_2(x, r) := \{ y \in \mathbb{R}^d : \|x - y\|_2 \le r \}$,

$f(v) \ge f(x)$ for all $v \in B_2(x, r)$.

a) Show that if $f : S \to \mathbb{R}$ is convex, every local min is a global minimum.

b) Show that this is not true if f is not convex.

Lemma (1). Let $f : \mathbb{R} \to \mathbb{R}$ be twice differentiable. TFAE:

(1) f is convex.

(2) f' is monotone nondecreasing

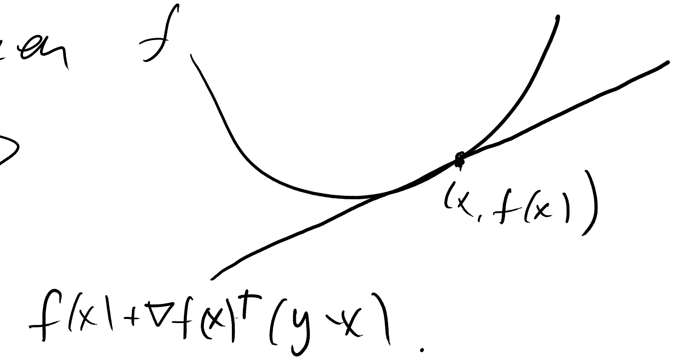
(3) f'' is nonneg.

Pf: exercise.

Lemma (2). If $f:\mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, then $\forall x, y \in \mathbb{R}^d$,

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$$

(Always lies above first-order Taylor appx.)



Remark. This is an "iff".

Pf. Consider $d=1$ first. WTS: $f(y) \geq f(x) + f'(x)(y-x)$.

By convexity, for $0 < t \leq 1$, $f(x + t(y-x)) = f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$.

$$\Rightarrow \quad f(y) \geq \frac{1}{t} f(x + t(y-x)) - \frac{1-t}{t} f(x) = f(x) + \underbrace{\frac{f(x+t(y-x)) - f(x)}{t}}_{\xrightarrow{t \downarrow 0} \langle \nabla f(x), y-x \rangle}.$$

For $d > 1$, consider $g(t) := f(ty + (1-t)x)$,

so $g'(t) = \langle \nabla f(ty + (1-t)x), y-x \rangle$.

By $d=1$ case, $g(1) \geq g(0) + g'(0)$

$$\Leftrightarrow \quad f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle. \quad \boxtimes$$

Lemma (3) Let $f: \mathbb{R}^d \to \mathbb{R}$, $g: \mathbb{R} \to \mathbb{R}$. If $\exists\, x \in \mathbb{R}^d, y \in \mathbb{R}$ s.t.
$f(w) = g(\langle w, x \rangle + y)$, and if g is convex,
then f is convex.

Pf: Exercise.

Example. $f(w) = (\langle w, x \rangle - y)^2$ is convex, since
$g(t) = t^2$ has $g''(t) = 2 > 0$.

Lemma (4). Let $f_i: \mathbb{R}^d \to \mathbb{R}$ be convex, $i = 1, \dots, r$. Then

(1) $x \mapsto \max_{i=1,\dots,r} f_i(x)$ is convex.

(2) For any $w_i \geq 0$, $x \mapsto \sum_{i=1}^{r} w_i f_i(x)$ is convex.

Pf: Exercise.

Def. Let $S \subset \mathbb{R}^d$. A function $f: \mathbb{R}^d \to \mathbb{R}^k$ is <u>L-Lipschitz</u> over S if, $\forall x_1, x_2 \in S$, $\|f(x_1) - f(x_2)\| \le L\|x_1 - x_2\|$.

Def. A differentiable fcn $f: \mathbb{R}^d \to \mathbb{R}$ is <u>β-smooth</u> if, $\forall x_1, x_2 \in \mathbb{R}^d$, $\|\nabla f(x_1) - \nabla f(x_2)\| \le \beta \|x_1 - x_2\|$.

Exercise. (a) If g_i are L_i-Lip., $i = 1, 2$, then $f(x) = g_1(g_2(x))$ is $L_1 L_2$-Lip.

(b) If $f(w) = g(\langle w, x\rangle + b)$ for β-smooth g, $w, x \in \mathbb{R}^d$, $b \in \mathbb{R}$, then f is $\beta\|x\|^2$-smooth.

Lemma 5. If $f: \mathbb{R}^d \to \mathbb{R}$ is β-smooth, $\forall x, y \in \mathbb{R}^d$, $|f(x) - f(y) - \nabla f(y)^T(x-y)| \le \frac{\beta}{2}\|x-y\|^2$.

Pf. By Fundamental Thm of Calculus, $f(x) - f(y) = \int_0^1 \nabla f(y + t(x-y))^T (x-y)\, dt$
($\vec{r}(t) = y + t(x-y)$, $d\vec{r} = (x-y)\,dt$).

$$\to |f(x) - f(y) - \langle \nabla f(y), x-y\rangle| = \left|\int_0^1 \langle \nabla f(y + t(x-y)), x-y\rangle dt - \int_0^1 \langle \nabla f(y), x-y\rangle dt\right|$$

$$\le \int_0^1 \|\nabla f(y + t(x-y)) - \nabla f(y)\| \cdot \|x-y\|\, dt \quad (C\text{-}S)$$

$$\le \int_0^1 \beta t \|x-y\|^2 dt = \frac{\beta}{2}\|x-y\|^2. \quad \square$$

Def. A differentiable function f has gradient descent iterates

$$w_{t+1} = w_t - \alpha_t \nabla f(w_t), \qquad \alpha_t > 0.$$

Lemma 6. If $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth, & $\alpha_t \le \frac{1}{\beta}$ $\forall t$, then gradient descent iterates satisfy

$$\|\nabla f(w_t)\|^2 \le \frac{2}{\alpha_t}\left(f(w_t) - f(w_{t+1})\right).$$

Pf.

$$f(w_{t+1}) \le f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{\beta}{2}\|w_{t+1} - w_t\|^2$$

$$= f(w_t) - \alpha_t \|\nabla f(w_t)\|^2 + \frac{\beta}{2}\alpha_t^2 \|\nabla f(w_t)\|^2$$

$$\Rightarrow \|\nabla f(w_t)\|^2 \alpha_t\left(1 - \frac{\beta \alpha_t}{2}\right) \le f(w_t) - f(w_{t+1}).$$

As $\alpha_t \le \frac{1}{\beta}$, implies $\|\nabla f(w_t)\|^2 \le \dfrac{f(w_t) - f(w_{t+1})}{\alpha_t / 2}$. $\boxed{}$

Convexity not used at all! Also implies $f(w_t)$ is decreasing.

Lemma (7) Let f be β-smooth, not nec. cvx. Assume $\alpha_t = \alpha < \frac{1}{\beta}$.
Then $\forall T \geq 1$,

$$\min_{t < T} \|\nabla f(w_t)\|^2 \leq \frac{2}{\alpha T}\left(f(w_0) - f(w_T)\right).$$

Pf. By previous lemma,

$$\|\nabla f(w_t)\|^2 \leq \frac{2}{\alpha}\left(f(w_t) - f(w_{t+1})\right).$$

$$\Rightarrow \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(w_t)\|^2 \leq \frac{2}{\alpha T}\left(f(w_0) - f(w_T)\right).$$

$$\Rightarrow \min_{t<T} \|\nabla f(w_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(w_t)\|^2$$

$$\leq \frac{2}{\alpha T}\left(f(w_0) - f(w_T)\right). \quad \square$$

→ GD finds ε-stationary points in time $T = \Theta(\alpha^{-1}\varepsilon^{-1})$, if objective is β-smooth and $\alpha < \frac{1}{\beta}$.

Lemma ⑧ Suppose f is β-smooth and convex, and assume $\alpha_t \equiv \alpha < \frac{1}{\beta}$ $\forall t$.
Then for any $z \in \mathbb{R}^d$, and any $t \geq 1$, G.D. iterates satisfy

$$f(w_t) \leq f(z) + \frac{\|w_0 - z\|^2 - \|w_t - z\|^2}{2\alpha t}.$$

Pf.

Let $z \in \mathbb{R}^d$. Let $\Delta_t^2 = \|w_t - z\|^2$.

$$\|w_t - z\|^2 - \|w_{t+1} - z\|^2 = \|w_t - z\|^2 - \|w_t - z - \alpha_t \nabla f(w_t)\|^2$$

$$\Rightarrow \Delta_t^2 - \Delta_{t+1}^2 = \Delta_t^2 - \left(\Delta_t^2 + \alpha_t^2 \|\nabla f(w_t)\|^2 - 2\alpha_t \langle w_t - z, \nabla f(w_t)\rangle\right)$$

$$= 2\alpha_t \langle \nabla f(w_t), w_t - z\rangle - \alpha_t^2 \|\nabla f(w_t)\|^2.$$

convexity

$$\geq 2\alpha_t \left(f(w_t) - f(z)\right) - \alpha_t^2 \|\nabla f(w_t)\|^2$$

Lemma

$$\geq 2\alpha_t \left(f(w_t) - f(z)\right) - \alpha_t^2 \cdot \frac{2}{\alpha_t}\left(f(w_t) - f(w_{t+1})\right)$$

$$= 2\alpha_t \left(f(w_{t+1}) - f(z)\right).$$ For $\alpha_t \equiv \alpha$,

$$\sum_{t=0}^{T-1} \Rightarrow \Delta_0^2 - \Delta_T^2 \geq 2\alpha\left[\sum_{0}^{T-1} f(w_{t+1}) - T f(z)\right].$$

$$\rightarrow \frac{1}{T}\sum_{t=0}^{T-1} f(w_{t+1}) \le f(z) + \frac{\Delta_0^2 - \Delta_T^2}{2\alpha T}.$$

By Lemma, $f(w_{t+1}) - f(w_t) \le 0$, so $f(w_t)$ is decr., & hence

$$f(w_T) = \min_{t<T} f(w_t) \le \frac{1}{T}\sum_{t=0}^{T-1} f(w_t) \le f(z) + \frac{\Delta_0^2 - \Delta_T^2}{2\alpha T}. \quad \boxtimes$$

<u>Corollary</u>. $\left|f(w_t) - \min_w f(w)\right| = O(1/\alpha t)$.

<u>Exercise</u>. In the above results, we assumed $\alpha_t \equiv \alpha < \frac{1}{\beta}$. How large of a fixed step size can we allow for the same analyses to hold?

Discuss reading group logistics.