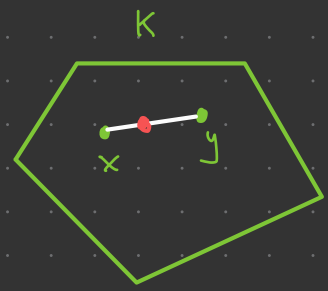
ADFOCS 2021

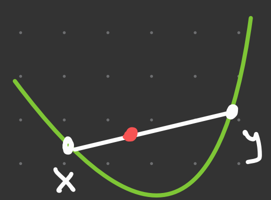
Lecture 3: Gradient Descent

Alejandro Cassis

Convexity: $K \subseteq \mathbb{R}^n$ is convex if $\forall x,y \in K \ \forall \alpha \in [0,1]$
$(1-\alpha)x + \alpha y \in K$

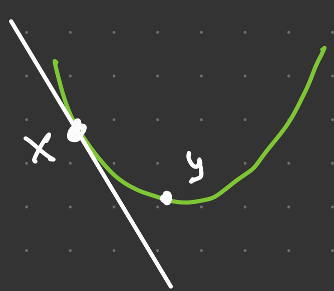


$f: K \to \mathbb{R}$ is convex if $\forall x,y \in K \ \forall \alpha \in [0,1]$
$f((1-\alpha)x + \alpha y) \leq (1-\alpha) f(x) + \alpha f(y)$



if f is differentiable: $f(x + \alpha(y-x)) \leq (1-\alpha) f(x) + \alpha f(y)$

$$\Leftrightarrow \quad \frac{f(x+\alpha(y-x)) - f(x)}{\alpha} \leq f(y) - f(x)$$

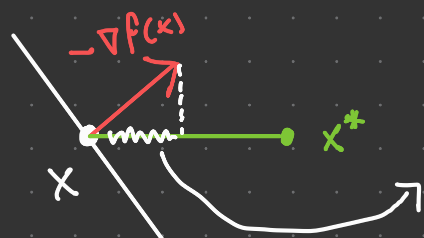


as $\alpha \to 0$: $\nabla f(x)^T (y-x) \leq f(y) - f(x)$

Usefulness: let $x^* = \underset{x}{\operatorname{argmin}} f(x)$. Plug $y := x^*$:

$$0 \leq \underbrace{f(x) - f(x^*)}_{\text{optimality gap}} \leq \nabla f(x)^T (x - x^*) = -\nabla f(x)^T (x^* - x)$$

→ neg. gradient is positively correlated with direction to x^*



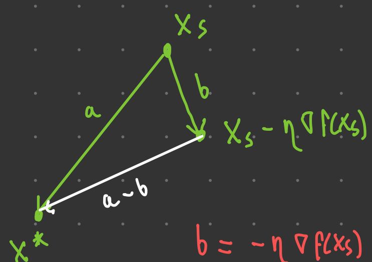
progress is lower bounded by $f(x) - f(x^*)$

GD Take 1: since moving in the negative gradient direction brings us closer to x^*, this suggests the following algorithm: (Assume $K = \mathbb{R}^n$)

let $x_1 \in K$ be an initial point
For $s = 1$ to $T-1$:
$\quad | \; x_{s+1} := x_s - \eta \nabla f(x_s)$

$\eta \in \mathbb{R}$ learning rate or step size.

To analyze: use $\|x^* - x_s\|^2$ as potential



change of potential:

$$\|a\|^2 - \|a-b\|^2 = 2 \underbrace{a^T b}_{\text{pos: corr progress}} - \underbrace{\|b\|^2}_{\text{error term}}$$

$$\therefore \quad \|x^* - x_s\|^2 - \|x^* - x_{s+1}\|^2 = 2(x^* - x_s)^T(-\eta \nabla f(x_s)) - \eta^2 \|\nabla f(x_s)\|^2$$

$$\Longleftrightarrow \quad \nabla f(x_s)^T (x_s - x^*) = \frac{\|x^* - x_s\|^2 - \|x^* - x_{s+1}\|^2}{2\eta} + \frac{\eta}{2} \|\nabla f(x_s)\|^2$$

Theorem: if $\|\nabla f(x)\| \le L \;\; \forall x \in K$, then

$$f\left(\frac{1}{T} \sum_{s=1}^{T} x_s\right) - f(x^*) \le \frac{\|x_1 - x^*\|^2}{2\eta T} + \frac{\eta L^2}{2}$$

$$f\left(\frac{1}{T} \sum_{s=1}^{T} x_s\right) - f(x^*) \le \frac{1}{T} \sum_{s=1}^{T} f(x_s) - f(x^*) \le \frac{1}{T} \sum_{s=1}^{T} \nabla f(x_s)^T (x_s - x^*)$$

$$\le \frac{\|x_1 - x^*\|^2}{2\eta T} + \frac{\eta L^2}{2} \qquad \square$$

Theorem: if $\|\nabla f(x)\| \le L \quad \forall x \in K$, then

$$f\left(\frac{1}{T}\sum_{s=1}^{T} x_s\right) - f(x^*) \le \frac{\|x_1 - x^*\|^2}{2\eta T} + \frac{\eta L^2}{2}$$

$R := \|x_1 - x^*\|^2$ "radius"

optimize $\eta = \frac{R}{L\sqrt{T}} \Rightarrow f\left(\frac{1}{T}\sum_{s=1}^{T} x_s\right) - f(x^*) \le \frac{RL}{\sqrt{T}}$

Corollary: Can find x st $f(x) - f(x^*) \le \varepsilon$ in $T = O\left(\frac{R^2 L^2}{\varepsilon^2}\right)$ iterations

Remarks:

- $\|\nabla f(x)\| \le L$ is equivalent to f being L-Lipschitz i.e. $\forall x, y \in K$

  e.g: $\|\cdot\|_2$ is 1-Lipschitz

  $$|f(x) - f(y)| \le L\|x - y\|$$

- For some applications R, L are known. Read convergence rate as $1/\sqrt{T}$

- This rate is optimal in black box model

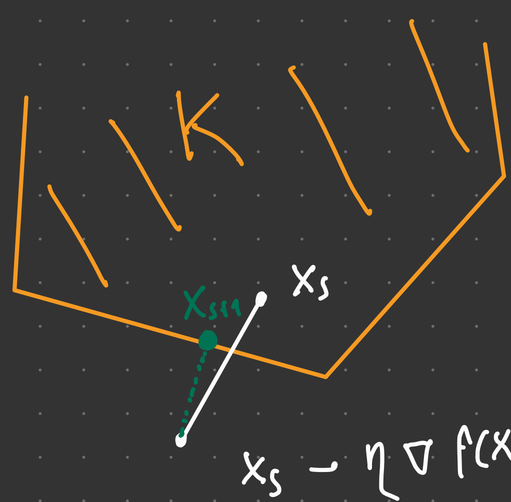
- Works for non-diff. fns. e.g. $\|x\|_1$ or $\|x\|_\infty$

What if K is not $\mathbb{R}^n$?

Projected Gradient Descent:

$$x_{s+1} := \text{Proj}_K\left(x_s - \eta \nabla f(x_s)\right)$$

where $\text{Proj}_K(y) = \underset{x \in K}{\text{argmin}} \|x - y\|$



Same analysis goes through! Recall that potential was $\|x^* - x_s\|^2$.

$y_{s+1} := x_s - \eta \nabla f(x_s)$
$x_{s+1} := \text{Proj}_K(y_{s+1})$

check: $\|x^* - x_{s+1}\|^2 \leq \|x^* - y_{s+1}\|^2$
i.e: projection only brings us closer to x^*.

(only works for convex K!!!)

Smoothness:

f is β-smooth if its gradient is β-Lipschitz.

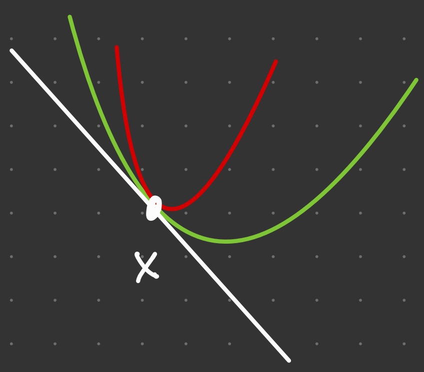
i.e. $\forall x, y \in K \quad \|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$

if f is twice differentiable: β-smooth is equivalent to $\nabla^2 f(x) \preceq \beta \cdot I \quad \forall x \in K$

i.e. $\lambda_{max}(\nabla^2 f(x)) \leq \beta$.

smoothness is useful to obtain upperbounds:

lemma: $\forall x, y: \quad f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|x - y\|^2$



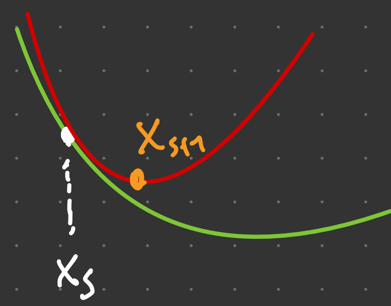
proof of lem: By Taylor's thm $\exists\ z = (1-\alpha)x + \alpha y$ st

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \underbrace{\frac{1}{2}(y - x)\nabla^2 f(z)(y - x)}_{\leq \frac{\beta}{2}\|y - x\|^2}$$

$\square$

How to exploit smoothness?

Minimize the upper bound!



$$x_{s+1} := \underset{z}{\operatorname{argmin}} \Big\{ f(x_s) + \underbrace{\nabla f(x_s)^T (z - x_s) + \frac{\beta}{2} \|z - x_s\|^2}_{g(z):=} \Big\}$$

$$\nabla g(z) = \nabla f(x_s) + \beta (z - x_s) = 0$$

$$z = x_s - \frac{1}{\beta} \nabla f(x_s)$$

$\Rightarrow$ update rule: $x_{s+1} = x_s - \frac{1}{\beta} \nabla f(x_s)$

What progress do we make?

[smoothness lemma: $f(x_{s+1}) - f(x_s) \leq - \frac{1}{2\beta} \|\nabla f(x_s)\|^2$]

proof: by smoothness:

$$f(z) \leq f(x_s) + \nabla f(x_s)^T (z - x_s) + \frac{\beta}{2} \|z - x_s\|^2$$

plugging $z := x_{s+1} = x_s - \frac{1}{\beta} \nabla f(x_s)$

$$\Rightarrow \quad f(x_{s+1}) - f(x_s) \leq - \frac{1}{\beta} \|\nabla f(x_s)\|^2 + \frac{1}{2\beta} \|\nabla f(x_s)\|^2$$

$\square$

(Strong) Convexity:

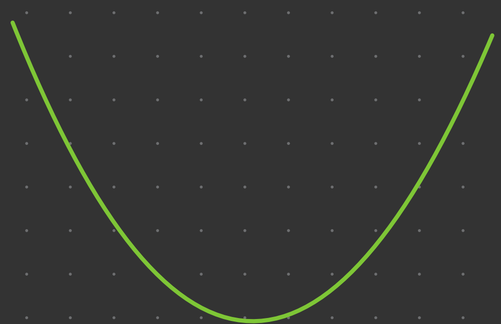
Recall that f convex gives us lower bounds:

$$\forall x, y \quad f(y) \geq f(x) + \nabla f(x)^T (y-x)$$

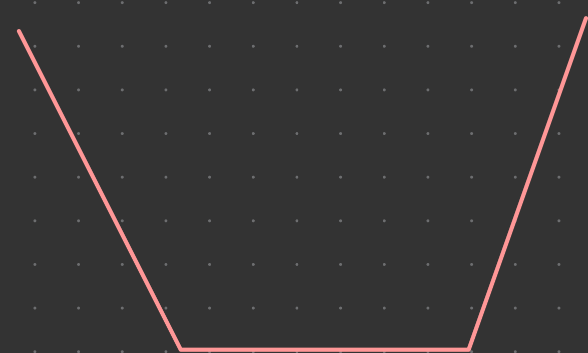
f is α-convex if $f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\alpha}{2} \|x-y\|^2$

[Similar as smoothness, for twice diff f's we have that this is equivalent to $\nabla^2 f(x) \succeq \alpha \cdot I \;\; \forall x$]

- strongly convex functions have unique minimizers



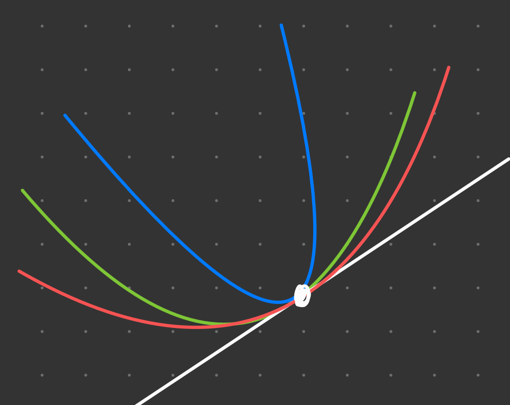
strongly convex



convex

β-Smoothness + α-convex $\equiv$ quadratic upper and lower bounds

$$\forall x, y: f(x) + \nabla f(x)^T (y-x) + \frac{\alpha}{2} \|x-y\|^2 \leq f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{\beta}{2} \|x-y\|^2$$



f is α-convex if $f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\alpha}{2}\|x-y\|^2$

$\Big[$ α-convex lemma: $\forall x \in \mathbb{R}^n$

$$f(x) - f(x^*) \leq \frac{1}{2\alpha} \|\nabla f(x)\|^2 \Big]$$

proof. by α-convex: $f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\alpha}{2}\|x^* - x\|^2$

$$\geq f(x) + \min_u \left\{ \nabla f(x)^T u + \frac{\alpha}{2}\|u\|^2 \right\}$$

minimizer is $u = -\frac{1}{\alpha}\nabla f(x)$

$$= f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|^2$$

$\square$

[Smoothness lemma: $f(X_{s+1}) - f(X_s) \le -\frac{1}{2\beta}\|\nabla f(X_s)\|^2$]

[α-convex lemma: $f(x) - f(x^*) \le \frac{1}{2\alpha}\|\nabla f(x)\|^2$]

Analyzing gradient descent: $X_{s+1} := X_s - \frac{1}{\beta}\nabla f(X_s)$

$$
\begin{aligned}
f(X_{s+1}) - f(x^*) &\le f(X_s) - f(x^*) - \frac{1}{2\beta}\|\nabla f(X_s)\|^2 \\
&\le f(X_s) - f(x^*) - \frac{\alpha}{\beta}\left[f(X_s) - f(x^*)\right] \\
&= \left(1 - \frac{\alpha}{\beta}\right)\left[f(X_s) - f(x^*)\right] \\
&\le \left(1 - \frac{\alpha}{\beta}\right)^s\left[f(X_1) - f(x^*)\right] \\
&\le \exp\left(-\frac{\alpha}{\beta}\cdot s\right)\left[f(X_1) - f(x^*)\right]
\end{aligned}
$$

[Theorem: Let f be α-convex and β-smooth.
Then, T steps of G.D. w/ step size $\frac{1}{\beta}$ satisfy

$$f(X_{T+1}) - f(x^*) \le \exp\left(-\frac{\alpha}{\beta}\cdot T\right)\left[f(X_1) - f(x^*)\right]$$
]

$$\Rightarrow \underbrace{\frac{\beta}{\alpha}}_{K}\log\left(\frac{f(X_1) - f^*}{\varepsilon}\right) \text{ to get } \varepsilon\text{-close}$$

K: condition number

Theorem: Let f be α-convex and β-smooth. Then, T steps of G.D. w/ step size $\frac{1}{\beta}$ satisfy

$$f(x_{T+1}) - f(x^*) \leq \exp\left(-\frac{\alpha}{\beta} \cdot T\right) \left[f(x_1) - f(x^*)\right]$$

Remarks:

- Same result holds for projected G.D. analysis is a bit more involved see [Bubeck 3.2]

- Progress is measured wrt $f(x_1) - f(x^*)$.

  Using β-smoothness: $f(x_{T+1}) - f(x^*) \leq \beta \exp\left(-\frac{\alpha}{\beta} \cdot T\right) \|x_1 - x^*\|^2$

- Important application: solving $Ax = b$ when $\alpha I \preceq A \preceq \beta I$ e.g. Laplacian L

  $f(x) = \frac{1}{2} x^T A x - x^T b$

  $\nabla f(x) = Ax - b$

  $\therefore$ unique minimizer satisfies $\nabla f(x^*) = Ax^* - b = 0$

More in Rasmus' lectures!

Excursion: Preconditioning

We saw that if $\alpha I \preceq \nabla^2 f(x) \preceq \beta I$ and $\frac{\beta}{\alpha}$ is "small" then G.D. converges exponentially.

$\rightarrow$ Suppose that instead $\alpha \cdot H \preceq \nabla^2 f(x) \preceq \beta \cdot H$ for some p.d. H.

intuition: f is well condition in a different basis

Consider $g(x) := f(Mx)$ where M is p.d.

$$\nabla g(x) = M^T \nabla f(Mx)$$

$$\nabla^2 g(x) = M^T \nabla^2 f(Mx) M$$

$\rightarrow$ setting $M := H^{-1/2}$: $\nabla^2 g(x) = H^{-1/2} \nabla^2 f(H^{-1/2} x) H^{-1/2}$

$$\therefore \quad \alpha H \preceq \nabla^2 f(H^{-1/2} x) \preceq \beta \cdot H$$

recall $A \succeq 0 \Leftrightarrow B^T A B \succeq 0 \ \forall B \in \mathbb{R}^{n \times n}$

$$\Longleftrightarrow \quad \alpha \cdot I \preceq \underbrace{H^{-1/2} \nabla^2 f(H^{-1/2} x) H^{-1/2}}_{= \nabla^2 g(x)} \preceq \beta \cdot I$$

i.e. g is α-convex and β-smooth!

Gradient Descent on g: $$x_{s+1} = x_s - \frac{1}{\beta} H^{-\frac{1}{2}} \nabla f(H^{-1/2} x_s)$$

$$\Longleftrightarrow \quad H^{-1/2} x_{s+1} = \underbrace{H^{-1/2} x_s}_{y_s} - \frac{1}{\beta} H^{-1} \nabla f(\underbrace{H^{-1/2} x_s}_{y_s})$$

$$\therefore \quad y_{s+1} = y_s - \frac{1}{\beta} H^{-1} \nabla f(y_s)$$

Newton's Method: $$x_{s+1} = x_s - \left(\nabla^2 f(x_s)\right)^{-1} \nabla f(x_s)$$

i.e: preconditioning using the Hessian at x_s!

[See Vishnoi's Chapter 9 for (much) more details on Newton's Method]

G.D. for smooth functions.

What if f is 0-convex? i.e. only β-smooth.

Reduction to strongly convex case:

Fix $x_1 \in \mathbb{R}^n$ and define

$$g(x) := f(x) + \underbrace{\frac{\alpha}{2}\|x_1 - x\|^2}_{\text{regularization}}$$

check: g is α-convex and $(\alpha+\beta)$-smooth

$\therefore$ can find x s.t. $g(x) - g(x^*) \leq \varepsilon$

in $O\left(\frac{\alpha+\beta}{\alpha} \log\left(\frac{1}{\varepsilon}\right)\right)$ iterations

let $x^* = \underset{x}{\operatorname{argmin}}\, f(x)$, $\tilde{x} = \underset{x}{\operatorname{argmin}}\, g(x)$

$$f(x) - f(x^*) = g(x) - g(x^*) + \frac{\alpha}{2}\left(\|x - x_1\|^2 - \|x^* - x_1\|^2\right)$$

$$\leq \varepsilon + \frac{\alpha}{2} R^2 \leq 2\varepsilon$$

$\alpha := \varepsilon / R^2$

$\therefore$ can find ε-optimal pt in $O\left(\frac{\beta \cdot R^2}{\varepsilon} \cdot \log\left(\frac{f(x_1) - f(x^*)}{\varepsilon}\right)\right)$

$[R = \|x^* - x_1\|]$

Summary:

	Rate	# iters
L-Lipschitz	$LR/\sqrt{T}$	L^2R^2/ε^2
β-smooth	$\beta R^2/T$	$\beta R^2/\varepsilon$
α-convex + β-smooth	$R^2 \exp\left(-\frac{T}{K}\right)$ $\left(K = \frac{\beta}{\alpha}\right)$	$K \log\left(R/\varepsilon\right)$

- Various parameters: β, L, R. Adaptive G.D: obtain optimal rates w/o knowing these. See Alina's Lectures!

- Rate for L-Lipschitz is optimal. See [Bubeck 3.5] for lower bound.

β-smooth can be improved to $\frac{\beta R^2}{T^2}$

α-cvx + β-smooth to $R^2 \exp\left(-T/\sqrt{K}\right)$

via "acceleration". See Alina's L.2.
also Vishnoi's Ch 8