DISCOVERING ACCELERATION

- MOTIVATION
- PROPERTIES of f
- ALGORITHM & CONVERGENCE of SD
- BACKTRACKING (ARMIJO)
- DERIVING ACCELERATION
- ACCELERATED GRADIENT

① MOTIVATION

"THE SIMPLEST OPTIMIZATION METHOD", one of those topic of which you might say "OK, I know this", BUT actually there is MUCH MORE WEALTH BENEATH THE SURFACE!

- I ALREADY MENTIONED THE GRADIENT DESCENT METHOD LAST YEAR IN THE CONTEXT of THE MATRIX COMPLETION PROBLEM.
- THERE ARE MANY VARIANTS of GRADIENT DESCENT! e.g. PROJECTED STEEPEST DESC., ACCELERATED, CONJUGATE, COORDINATE WISE, STOCHASTIC...
  BUT SIMPLE UNDERLYING COMMON PATTERNS!
- MUCH RESEARCH IN OPTIMIZATION FOCUSES ON CONVERGENCE RATES, BUT OTHER PROPERTIES ARE ALSO IMPORTANT, e.g. ROBUSTNESS.
- BASIC GRADIENT DESCENT IS ROBUST TO NOISE IN SEVERAL IMPORTANT WAYS, WHILE ACCELERATED GRADIENT DESCENT IS MUCH MORE BRITTLE. TRADE-OFF!
- We will fix our ideas on the specific case of: UNCONSTRAINED CONVEX OPTIMIZATION, i.e.,

$$\boxed{\min_{x \in \mathbb{R}^n} f(x) \quad \text{, called CONVEX PROGRAM.}}$$

② SMOOTH AND STRONGLY CONVEX f

$f : \mathbb{R}^n \to \mathbb{R}$, TWICE DIFFERENTIABLE ($C^2$) AND ($\alpha$-)STRONGLY CONVEX,

i.e., $\exists\, \alpha > 0,\ \forall x \quad \nabla^2 f(x) \succeq \alpha I$

$\iff \exists\, \alpha$ s.t. $\forall x, y \quad f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2} \|y - x\|^2$

Roughly speaking, we want the function f to be "convex enough"

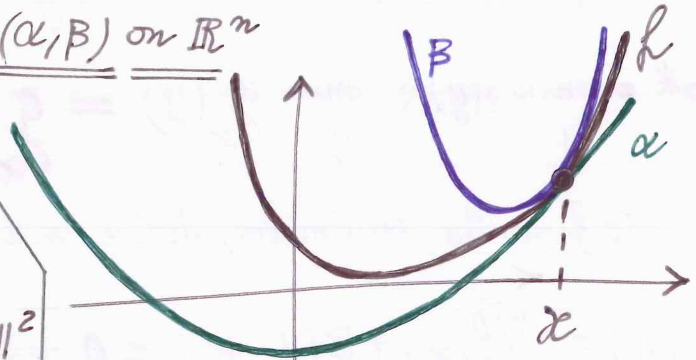
At the same time, we do not want f to be "too convex":

<u>(β-)smoothness</u>: $\nabla^2 f(x) \preceq \beta I, \quad \beta < \infty$

$$\iff f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{\beta}{2}\|y-x\|^2$$

In short, <u>f IS OF TYPE (α,β) on $\mathbb{R}^n$</u>

> **<u>Thm 1</u> (<u>estimate on p^*</u>)**
> Let f be of type (α,β).
> Then:
> $$\frac{1}{2\beta}\|\nabla f(x)\|^2 \leq f(x) - p^* \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2$$



<u>ROUGHLY SPEAKING</u>: f CAN BE SQUEEZED BETWEEN TWO PARABOLAS !!

(THIS GIVES A WAY TO STOP THE ITERATIONS) STOPPING CONDITION of ITERATIVE ALGORITHM: $\|\nabla f(x_k)\| \leq \sqrt{2\alpha\varepsilon} \Rightarrow f(x_k) - p^* \leq \varepsilon$

"SMALL GRADIENT $\approx$ SOLVED PROBLEM"

③ <u>STEEPEST DESCENT (SD)</u> [CAUCHY, 1847]

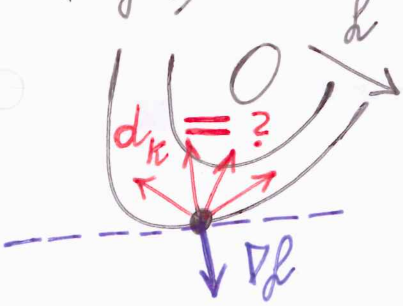
- Many methods (like SD) are of the form $x_{k+1} = x_k + t_k d_k$.
  (SEARCH DIR. d_k; $t_k > 0$, STEPSIZE)
- DESCENT TYPE: $f(x_{k+1}) < f(x_k)$

A) <u>How to choose</u> d_k?

For differentiable f: (Taylor)

$$f(x_{k+1}) - f(x_k) = \underbrace{\nabla f^T_{(x_k)} (t_k d_k)}_{\text{WE WANT DESCENT! } \leq 0} + o(\|t_k d_k\|)$$

SO: $\nabla f(x_k)^T d_k \leq 0$



<u>Greedy choice</u>: most decrease of f at x

$$\begin{cases} \max\limits_{d_k} -\nabla f(x)^T d_k \\ \text{s.t. } \|d_k\| \leq 1 \end{cases}$$

solution is $d_k = -\nabla f(x)$

<u>DIRECTION of SD</u>

B) <u>How to calculate</u> t_k? <u>line-search</u> (<u>LS</u>)

- <u>Exact LS</u>: $\min\limits_{t \geq 0} f(x_k + t d_k) \rightsquigarrow t_k^{EX}$ is the unique minimizer if f is stricly convex.

Can sometimes be computed. Good for theory.

Exact LS is important for theoretical analysis:

Thm2 Let f be of type (α, β). Then SD with exact LS satisfies:

the optimal value

$$(*) \quad f(x_k) - p^* \leq \gamma^k \left(f(x_0) - p^* \right)$$

CONVG. FACTOR $\gamma := 1 - \frac{\alpha}{\beta}$

So SD converges to the exact solution x_* for any x_0.

Proof. $x_+ = x - t \nabla f(x)$

Def. β-smooth: $f(x_+) \leq f(x) + \underbrace{\nabla f(x)^T(-t\nabla f(x))}_{-t\|\nabla f(x)\|^2} + \frac{\beta}{2} t^2 \|\nabla f(x)\|^2$

Rewrite:

$$f(x - t\nabla f(x)) \leq f(x) + \left(\frac{\beta}{2}t^2 - t\right)\|\nabla f(x)\|^2, \quad \text{valid } \forall t.$$

Minimizing both sides over $t \geq 0$

$\partial/\partial t(\dots) = 0$

$\Rightarrow (\beta t - 1)\|\nabla f(x)\|^2 = 0$

$\Rightarrow t^* = \frac{1}{\beta}$

min. over t gives $t = t^*$, $t = t^*$

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x)\|^2$$

$$f(x_{k+1}) - p^* \leq f(x_k) - p^* - \frac{1}{2\beta}\|\nabla f(x_k)\|^2$$

FROM THM.1 WE HAVE $-\|\nabla f(x_k)\|^2 \leq -2\alpha(f(x_k) - p^*)$

$$f(x_{k+1}) - p^* \leq \underbrace{\left(1 - \frac{\alpha}{\beta}\right)}_{\gamma!!}(f(x_k) - p^*), \text{ use it recursively and get } (*)$$

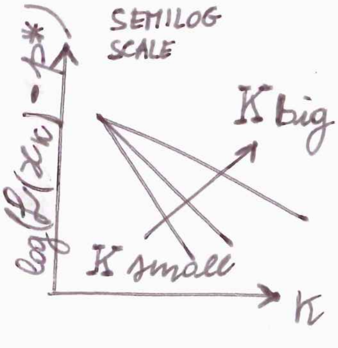
- Let's bound the CONVERGENCE FACTOR. Call $K := \beta/\alpha$ the "condition number", $1 \leq K < \infty$

$$\gamma^k = \left(1 - \frac{1}{K}\right)^k = (e^a)^k, \quad a = \log\left(1 - \frac{1}{K}\right) \leq -\frac{1}{K}$$

$$\leq \exp\left(-\frac{k}{K}\right)$$

$K \to 1$, very good

$K \to \infty$, very bad



- So we have **exponential convergence** (for some reason called **linear convergence** in optimization)

- If we use $t_k = \frac{1}{\beta}, \forall k$ (SD with CONSTANT STEPSIZE), we would get the same convergence bound.

IN GENERAL, FOR GENERIC f (NOT of TYPE (α, β)), WE WILL NOT USE EXACT LS!!!

④ ARMIJO BACKTRACKING. (what we do IN PRACTICE) [LARRY ARMIJO, 1966]

Goal: replace exact LS with something computationally cheaper, but still effective.

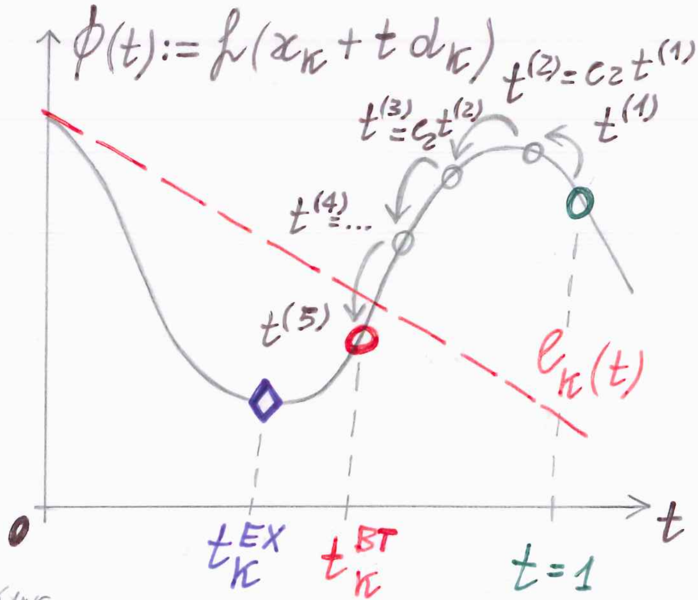
$-\nabla f(x_k)$ DIRECTION of SD

$$\ell_k(t) = f(x_k) + c_1 t \nabla f(x_k)^T d_k$$

$$0 < c_1 < 1/2$$

$$0 < c_2 < 1$$

NB: if $c_1 = 1$, we get the tangent line to $f(x_k + t d_k)$ at $f(x_k)$



Sufficient decrease condition:

$$\boxed{f(x_k + t_k d_k) < \ell_k(t_k)}$$

⇒ SHOW SLIDES ON ARMIJO BACKTRACKING.

Thm 3 Let f be of type (α, β). Then SD with Armijo LS (c_1, c_2) satisfies

$$f(x_k) - p_* \leq \gamma^k (f(x_0) - p_*),$$

with $\gamma := 1 - 2 c_1 \min(\alpha, c_2/K)$, $K := \beta/\alpha$.

⟹ SHOW MOVIE SD ON QUADRATIC WITH ARMIJO.

5 SD ON A QUADRATIC f

- Let $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$, and consider the quadratic objective function $f(x) = \frac{1}{2} x^T A x - b^T x$

[WHEN $A \succ 0$, THIS QUADRATIC IS STRICTLY CONVEX AND HAS A UNIQUE GLOBAL MINIMUM x^*]

- IN PARTICULAR, Let f be of type (α, β) $\Longrightarrow$ SPECTRAL CONDITION $\boxed{\alpha I \preceq A \preceq \beta I}$

$\nabla^2 f(x) = A$

Since $\nabla f(x) = Ax - b$, we get $\min f(x) \iff Ax = b$; sol.: $x^* = A^{-1} b$.

- I.e., SD computes the sol. of the linear system $Ax = b$.
- One (not me in this 45 min. talk) can show that SD on this f with constant "OPTIMAL" stepsize $t = \frac{2}{\alpha + \beta}$ satisfies:

$$f(x_k) - p^* \leq \underbrace{\left(\frac{K-1}{K+1}\right)^{2k}}\left(f(x_0) - p^*\right)$$

- Doing the same exercise (BOUND THE CONVERGENCE FACTOR) we get

$$\gamma^k = \left(1 - \frac{2}{K+1}\right)^{2k} \leq \exp\left(-\frac{4k}{K+1}\right)$$

$\Longrightarrow$ SHOW SD ON QUADRATIC WITH $K=1$. CONVERGES IN 1 ITERATION!

- BEST RATE FOR SD? NUMERICAL EXPERIMENTS SAY YES!
- IS THIS THE BEST METHOD? $\rightsquigarrow$ DEFINE...

$\Longrightarrow$ SHOW ZIGZAG BEHAVIOR of SD. THIS MOTIVATES US TO LOOK FOR A BETTER METHOD!!

6 DERIVING ACCELERATION

- Let's start SD with constant stepsize $t = 1/\beta$ at $x_0 = \frac{1}{\beta} b$.

One can check by induction that

$$x_k = \left(\sum_{j=0}^{k} (I - A')^j\right) b',$$

where $A' = \frac{1}{\beta} A$ and $b' = \frac{1}{\beta} b$.

WHY DOES THIS CONVERGE TO $A^{-1}b$?

Recall that for all scalars $|x|<1$,

$$\sum_{j=0}^{\infty}(1-x)^j = \frac{1}{x} \qquad ⊛$$

$\alpha I \preceq A \preceq \beta I$, $A' = tA$ and $t=\frac{1}{\beta}$; so $\frac{\alpha}{\beta} I \preceq A' \preceq I$, i.e., the eigenvalues of A' lie within $(0,1]$. Hence ⊛ extends to the matrix case. I.E., GRADIENT DESCENT IS COMPUTING A DEGREE K (MATRIX POLYNOMIAL) APPROXIMATION of THE INVERSE FUNCTION of A !!!

- APPROXIMATION ERROR when truncating ⊛ to K is $O((1-x)^K)$
- IN THE MATRIX CASE, this translates to $O(\|(I-A')^K\|) =$

$$= O(\|I-A'\|_2^K) = O\left(\left(1-\frac{1}{K}\right)^K\right) \Leftarrow$$

This is exactly the convergence rate of SD that we determined earlier !!!

$$\|I-tA\|_2 \underset{\uparrow}{=} \lambda_{max}\left(I-\frac{1}{\beta}A\right) = 1-\frac{1}{\beta}\lambda_{min}(A) = 1-\frac{\alpha}{\beta} = 1-\frac{1}{K}$$

$I-A'$ is symm. and $t = 1/\beta$

- Why we went through this exercise? Because now you see that TO IMPROVE THE CONVERGENCE RATE of GRADIENT DESCENT is equivalent to FIND A BETTER LOW-DEGREE POLYNOMIAL APPROXIMATION TO THE SCALAR FUNCTION $1/x$!!! we'll be able to save a square root in the degree while achieving the same error!

- WE WANT TO MINIMIZE THE RESIDUAL:

$$r_K := \|(I - A\underline{q_K(A)})b\|$$

where $q_K(A)$ is a matrix polynomial approximation to A^{-1}: $q_K(A) \approx A^{-1}$

$$\leq \underbrace{\|I - Aq_K(A)\|}_{=: p_K(A)}\cdot\|b\| = \max_{\mu\in\lambda(A)} |p_K(\mu)|\cdot\|b\| \qquad \boxed{f(A) = Q f(\Lambda) Q^T}$$

↑ corresponding scalar polynomial

Relaxing this condition:

$$\leq \max_{\mu\in[\alpha,\beta]} |p_K(\mu)|\cdot\|b\|$$

NB: IN GENERAL WE MAY NOT KNOW THE SPECTRUM $\lambda(A)$, BUT WE DO KNOW THAT ALL EIGENVALUES $\mu \in [\alpha,\beta]$.

- MINIMIZE THE RESIDUAL :

$$\boxed{\min_{\substack{p_k \in \mathbb{P}_k \\ p_k(0)=1}} \max_{\mu \in [\alpha,\beta]} |p_k(\mu)|}$$

VERY HARD OPTIMIZATION PROBLEM!
We are looking for a polynomial of degree k that is as small as possible on the location of the eigenvalues of A, namely on the interval $[\alpha, \beta]$.
At the same time, we have the normalization constraint $p_k(0)=1$.

"THERE IS ONLY ONE BULLET IN THE GUN: IT'S CALLED THE CHEBYSHEV POLYNOMIAL."

"Sur les questions de minima qui se rattachent à la représentation approximative des fonctions"

(Č.P.'s) **CHEBYSHEV POLYNOMIALS** (of 1st kind) Чебышёв [1859]

Def. $T_0(x)=1$, $T_1(x)=x$, $T_{n+1}(x) = 2x\,T_n(x) - T_{n-1}(x)$, $n \geq 1$.

LEMMA Let $n \geq 1$ and $q(x) = 2^{n-1}x^n + b_{n-1}x^{n-1} + \dots + b_0 \not\equiv T_n(x)$,

"OPTIMALITY PROPERTY of CHEBYSHEV POLYNOMIALS"

THEN $\max_{x\in[-1,1]} |q(x)| > \max_{x\in[-1,1]} |T_n(x)| = 1$.

PUT DIFFERENTLY, THE POLYNOMIAL HAVING THE FORM of $q(x)$ THAT DEVIATES THE LEAST POSSIBLE FROM ZERO ON $[-1;1]$ IS THE CHEBYSHEV POLYNOMIAL !!!

$\Rightarrow$ Show slide on Č.P.'s.

- Suitably rescaled, Č.P.'s MINIMIZE THE ABSOLUTE VALUE of p_k IN A DESIRED INTERVAL $[\alpha, \beta]$ WHILE SATISFYING $p_k(0)=1$:

$$P_n(x) := T_n\left(\frac{\alpha+\beta-2x}{\beta-\alpha}\right) \Big/ T_n\left(\frac{\alpha+\beta}{\beta-\alpha}\right)$$

WHICH IS EXACTLY WHAT WE WANTED !!!

$\Rightarrow$ Show plot of this polynomial.

(7) ACCELERATED GRADIENT METHOD

ERROR BOUND THAT COMES OUT of THE CHEBYSHEV POL.'s:
(RATE OF NESTEROV'S FGM or AGM)

$$\| x_{k+1} - x^* \| \leq 2\exp\left(-k\sqrt{\frac{2}{K}}\right) \| x_0 - x^* \|$$

(quite technical derivation)

THIS MEANS THAT, FOR LARGE K, WE GET QUADRATIC SAVINGS IN THE DEGREE WHILE ACHIEVING THE SAME ERROR !

CHEBYSHEV RECURRENCE RELATION

Due to the recursive def. of C. Pol.'s, we get an iterative algorithm out of it. Transferring the recursive def. to our rescaled C. Pol's, we have:

$$P_{k+1}(\mu) = (t_k \mu + \gamma_k) P_k(\mu) + \delta_k P_{k-1}(\mu)$$

(the coeff.'s t_k, γ_k, δ_k can be worked out from the recurrence def.). Moreover, since $P_k(0) = 1$, we must have $\gamma_k + \delta_k = 1$. $\forall k$

$\Longrightarrow$ UPDATE RULE:

$$x_{k+1} = x_k - t_k \underbrace{(Ax_k - b)}_{\nabla f(x_k)} + \underbrace{\delta_k (x_k - x_{k-1})}_{\text{ONLY THIS ADDITIONAL TERM!}}$$

$\Longrightarrow$ SHOW VIDEO of ACCELERATED GRADIENT!

RKs

- FINDING THE BEST POSSIBLE coeff.'s LEADS TO THE ABOVE CONVERGENCE RATE. WE USED ONLY 1st ORDER INFORMATION!
- WORKS FOR ANY FUNCTION of TYPE (α, β), AND NOT JUST THE SPECIAL of A QUADRATIC OBJECTIVE f THAT WE SHOW HERE!
- THIS IS WHAT NESTEROV SHOWED IN 1983!
- THE POLYNOMIAL APPROX. METHOD WAS KNOWN MUCH EARLIER IN THE CONTEXT of EIGENVALUE METHODS!

REFERENCES