MODELLI 6/10/26

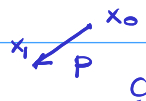
TODAY: Optimization and nonlinear LS

FRI : 2 hrs in the classroom to work (in groups)

TUE : New case study (+ some time for questions on the group work)

FRI : Reports of the results

$\min_{x \in \mathbb{R}^n} f(x)$ ↳ objective function



Goal $f(x_0+p) < f(x_0)$ ↳ descent direction

$$f(x) = f(x_0) + \nabla f(x_0)^T p + \frac{1}{2} p^T \nabla^2 f(x_0) p + \dots$$

$O(\|p\|_2^3)$

$$f(x_0+tp) = f(x_0) + t \underbrace{\nabla f(x_0)^T p}_{<0} + \dots$$

for $f(x_0+tp) < f(x_0)$ for small t

DEF p is a descent direction iff $\nabla f(x)^T p < 0$

GRAD. DESCENT

Obvious choice: $p = -\nabla f(x)^T \Rightarrow \nabla f(x)^T p = -\|\nabla f(x)\|_2^2 < 0$ $B = I$

More general choices: $p = -B \nabla f(x_0)$ with B positive def.

$\nabla f(x_0)^T p = -\nabla f(x_0)^T B \nabla f(x_0) < 0$ NEWTON or NEWTON-like

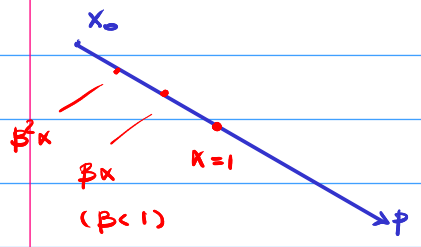
Lemma (Descent) Let $f \in C^1$ and ∇f is Lipschitz with constant L.

Then: $f(x+p) \leq f(x) + \nabla f(x)^T p + \frac{L}{2}\|p\|_2^2$

Proof $\underbrace{f(x+p) - f(x) - \nabla f(x)^T p}_{G(1)} = \int_0^1 [\nabla f(x+tp) - \nabla f(x)]^T p \, dt$

$G(t) = f(x+tp) - f(x) - \nabla f(x)^T tp$

Taking norms $\dots \leq L \int_0^1 t \|p\|_2^2 \, dt = \frac{L}{2}\|p\|_2^2$



($\beta < 1$)

IDEAL: Choose $t>0$: $f(x_0+tp)$ is minimum. EXACT LINE SEARCH (DIFFICULT)

LESSIDEAL: Pick $t = \alpha$ such that

$f(x+\alpha p) < f(x)$

↳ supp. smaller

Strategy: ARMIJO BACKTRACKING ⇝ guarantee that the final α will be large enough $\alpha \geq O(1/L)$ L Lipschitz const. of the gradient.

Thm f is C^1 and with Lipschitz gradient and bounded below; If p are chosen by $-B_k^{-1}\nabla f(x_k)$, with $0 < \mu_B < B_k < M_B < +\infty$, then if the step length are chosen by Armijo:

$$\sum_{k=0}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty \qquad (\text{in particular } \|\nabla f(x_k)\| \to 0)$$

EXAMPLE: $f(x) = e^x$ (likely $x_k \to -\infty$ so not strictly speaking not converging to a min. point)

If $\{f(x) \leq f(x_0)\}$ is cpt $\Rightarrow$ convergence of $x_k \to x_*$ at least with subsequences.

=

Thm Assume f is convex, and ∇f is Lipschitz, with a minimizer x_*; with $\alpha = 1/L$

then $f(x_k) - f(x_*) \leq \dfrac{L\|x_0 - x_*\|^2}{2k}$

Proof In our case $x_{j+1} = x_j + \alpha p = x_j - \alpha \nabla f(x_j)$; the descent lemma gives

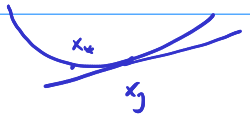
$$f(x_{j+1}) \leq f(x_j) - \alpha \|\nabla f(x_j)\|_2^2 + \frac{L}{2}\alpha^2 \|\nabla f(x_j)\|_2^2$$

$$\overset{!}{\leq} f(x_j) - \frac{\alpha}{2}\|\nabla f(x_j)\|_2^2 \qquad (\text{since } \alpha = 1/L)$$

$$\|x_{j+1} - x_*\|_2^2 = \|x_j - \alpha\nabla f(x_j) - x_*\|_2^2$$

$$\overset{!}{=} \|x_j - x_*\|_2^2 + \alpha^2\|\nabla f(x_j)\|_2^2 - 2\alpha\nabla f(x_j)^T(x_j - x_*)$$

By convexity $f(x_*) \geq f(x_j) + \nabla f(x_j)^T(x_* - x_j)$



$$\nabla f(x_j)^T(x_* - x_j) \leq f(x_*) - f(x_j)$$

$\leq \|x_j + x_*\|_2^2 + \alpha^2 \|\nabla f(x_j)\|_2^2 - 2\alpha (f(x_j) - f(x_*))$

$\leq \|x_j + x_*\|_2^2 - 2\alpha (f(x_{j+1}) - f(x_*))$ (DESCENT LEMMA)

(keep expanding until you get at $\|x_0 - x_*\|_2^2$)

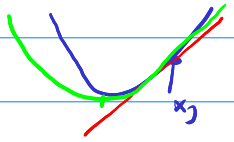
$2\alpha k [f(x_k) - f(x_*)] \leq \|x_0 - x_*\|_2^2$

$\frac{1}{k}$

For convex fcts: grad. descent converges (slowly)

Another viewpoint $f(x_*) = f(x_j) + \nabla f(x_j)^T (x_* - x_j) + \dots$

$+ \frac{1}{2} (x_* - x_j)^T \nabla^2 f(x_j) (x_* - x_j)$



IDEA: Pick the stationary point of the quadratic model

How? Solve $\nabla^2 f(x_j) p = -\nabla f(x_j)$ Newton's method for $\nabla f = 0$

$\rightsquigarrow p = -\underbrace{\nabla^2 f(x_j)^{-1}}_{-B_j^{-1}} \cdot \underbrace{\nabla f(x_j)}_{\nabla f(x_j)} \cong -\frac{g(x_j)}{g'(x_j)}$ (for scalar $g(x) = \nabla f(x)$)

$B_j = \nabla^2 f$ is pos. def if f is cvx.

Otherwise: only local convergence. (quadratic! Easy by thinking of it as rootfinding for $\nabla f = 0$

CHEAP, BUT SLOW CONVERGENCE

More generally

	EXPENSIVE BUT LOCALLY FAST
FIRST-ORDER METHODS	SECOND-ORDER METHODS
gradient-descent	Newton
nonlinear conjugate gradient	quasi-Newton (Approx. the Hessian)
...	BFGS
	L-BFGS
	Trust-region

—

Our case study: two parameters (B, r)

$r: \mathbb{R}^2 \to \mathbb{R}^n$ $\frac{1}{2}\|r(x)\|_2^2$ (up to shift the components)

$\rightsquigarrow f(x)$

$J = \square$

$$f(x) = \frac{1}{2}\sum_{i=1}^{n} r_i(x)^2$$

$J_{ij}(x) := \frac{\partial r_i}{\partial x_j}$

In this case: $\nabla f(x) = \sum_{i=1}^{n} r_i(x)\frac{\partial r_i}{\partial x} = J^T(x)\,r(x)$

$$\nabla^2 f(x) = J(x)^T J(x) + S(x) \qquad S(x) = \sum_{i=1}^{n} r_i(x)\,\nabla^2 r_i(x)$$

GRADIENT DESCENT: Compute $J^T r$ and update

NEWTON'S METHOD: Solve $(J^TJ + S)p = -J^Tr$

↳ not necessarily positive def.

In most cases $\|r(x_*)\|_2$ is small $\Rightarrow$ IDEA: Neglect the $S(x)$ term in $\nabla^2 f$

GAUSS-NEWTON'S METHOD: Solve $J^TJ\,p = -J^Tr$

If $r(x_*) = 0$ → quadratic convergent

If $r(x_*)$ small → linear convergence (but fast)

If J is full-rank → J^TJ is pos. def ⇝ theory ok for convergence.

Equivalent to $\min \|Jp - r\|_2^2$ (linear least squares)

(much better numerically) QR factorization

It could be that J (or J^TJ) is ill-conditioned:

·) Check the scaling!

·) Add regularization (Levenberg-Marquardt)