Lecture 2: Methods for local unconstrained optimization. Linesearch methods

Coralia Cartis, Mathematical Institute, University of Oxford

C6.2/B2: Continuous Optimization

Methods for local unconstrained optimization

minimize $f(x)$ subject to $x \in \mathbb{R}^n$ (UP) $[f \in \mathcal{C}^1(\mathbb{R}^n)$ or $f \in \mathcal{C}^2(\mathbb{R}^n)]$

A Generic Method (GM)

Choose $\epsilon > 0$ and $x^0 \in \mathbb{R}^n$.
While (TERMINATION CRITERIA not achieved), REPEAT:

- compute the change

  $$x^{k+1} - x^k = F(x^k, \text{problem data}),$$ [linesearch, trust-region]

  to ensure $f(x^{k+1}) \le f(x^k)$.
- set $x^{k+1} := x^k + F(x^k, \text{prob. data})$, $k := k + 1$. □

- TC: $\|\nabla f(x^k)\| \le \epsilon$; maybe also, $\lambda_{\min}(\nabla^2 f(x^k)) \ge -\epsilon$.
- e.g., $x^{k+1} \equiv$ minimizer of some (simple) model of f around x^k $\longrightarrow$ linesearch, trust-region methods.
- if $F = F(x_k, x_{k-1}, \text{problem data}) \longrightarrow$ conjugate gradients mthd.

Issues to consider about GM

Finite termination of GM: for any $\epsilon > 0$, there exists k such that $\|\nabla f(x^k)\| \le \epsilon$? $\Longleftrightarrow \liminf_{k\to\infty} \|\nabla f(x^k)\| = 0$

Global convergence of GM: if $\epsilon := 0$ and any $x^0 \in \mathbb{R}^n$: $\nabla f(x^k) \to 0$, as $k \to \infty$? all limit points of $\{x^k\}$ are then stationary. [maybe also, $\liminf_{k\to\infty} \lambda_{\min}(\nabla^2 f(x^k)) \ge 0$?]

Local convergence of GM:

if $\epsilon := 0$ and x^0 sufficiently close to $x^* \equiv$ stationary/local minimizer of f: $x^k \to x^*$, $k \to \infty$?

Global/local complexity of GM: count number of iterations and their cost required by GM to generate x^k within desired accuracy $\epsilon > 0$, e.g., such that $\|\nabla f(x^k)\| \le \epsilon$.

[connection to convergence and its rate]

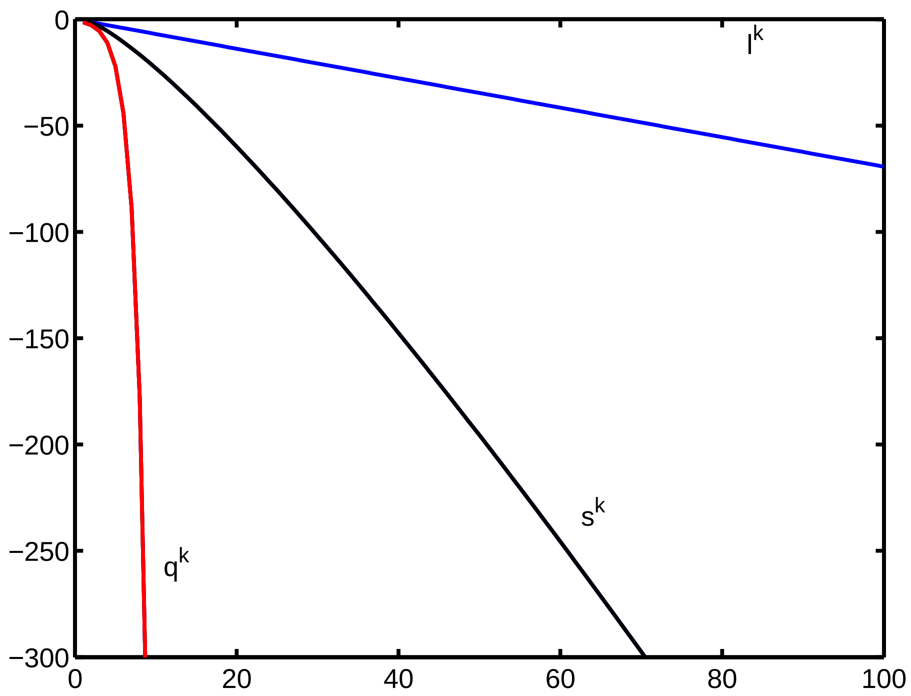
Rate of global/local convergence of GM.

Rates of convergence of sequences: an example

$l^k := (1/2)^k \longrightarrow 0$ linearly,

$q^k := (1/2)^{2^k} \longrightarrow 0$ quadratically,

$s^k := k^{-k} \longrightarrow 0$ superlinearly as $k \longrightarrow \infty$.



Rates of convergence on a log scale.

k	l^k	q^k
0	1	0.5
1	0.5	0.25
2	0.25	$0.6 \cdot (-1)$
3	0.12	$0.4 \cdot (-2)$
4	$0.6 \cdot (-2)$	$0.1 \cdot (-4)$
5	$0.3 \cdot (-2)$	$0.2 \cdot (-9)$
6	$0.2 \cdot (-2)$	$0.5 \cdot (-19)$

Notation: $(-i) := 10^{-i}$.

Rates of convergence of sequences

$\{x^k\} \subset \mathbb{R}^n, x^* \in \mathbb{R}^n; \quad x^k \to x^*$ as $k \to \infty$.

p-Rate of convergence: $x^k \to x^*$ with rate $p \geq 1$ if $\exists \rho > 0$ and $k_0 \geq 0$ such that

$$\|x^{k+1} - x^*\| \leq \rho \|x^k - x^*\|^p, \quad \forall k \geq k_0.$$

- ρ convergence factor; $e^k := x^k - x^*$ error in $x^k \approx x^*$.

Linear convergence: $p = 1 \quad \Rightarrow \rho < 1$; (asymptotically,) no of correct digits grows linearly in the number of iterations.

Quadratic convergence: $p = 2$; (asymptotically,) no of correct digits grows exponentially in the number of iterations.

Superlinear convergence: $\|x^{k+1} - x^*\| / \|x^k - x^*\| \to 0$ as $k \to \infty$. [faster than linear, slower than quadratic; practically very acceptable]

Summary: methods for local unconstrained probs.

Consider (UP), with $f \in \mathcal{C}^1(\mathbb{R}^n)$ or $\mathcal{C}^2(\mathbb{R}^n)$.

Methods:

- iterative: start from any initial 'guess' x^0, generate x^k, $k \geq 0$.
- find (approximate) local solutions, unless special structure (convexity, etc.)
- terminate when iterate within ϵ of local optimality.

Issues: global convergence, local convergence, rate of convergence, complexity.

Information employed on each iteration:

current x^k: linesearch and trust-region methods

current+previous: conjugate-gradients method etc

A generic linesearch method

(UP): minimize $f(x)$ subject to $x \in \mathbb{R}^n$, where $f \in \mathcal{C}^1$ or $\mathcal{C}^2(\mathbb{R}^n)$.

A Generic Linesearch Method (GLM)

Choose $\epsilon > 0$ and $x^0 \in \mathbb{R}^n$.
While $\|\nabla f(x^k)\| > \epsilon$, REPEAT:

- compute a descent search direction $s^k \in \mathbb{R}^n$,

$$\nabla f(x^k)^T s^k < 0;$$

- compute a stepsize $\alpha^k > 0$ along s^k such that

$$f(x^k + \alpha^k s^k) < f(x^k);$$

- set $x^{k+1} := x^k + \alpha^k s^k$ and $k := k + 1$. □

Recall property of descent directions (Lemma 1, Lecture 1).

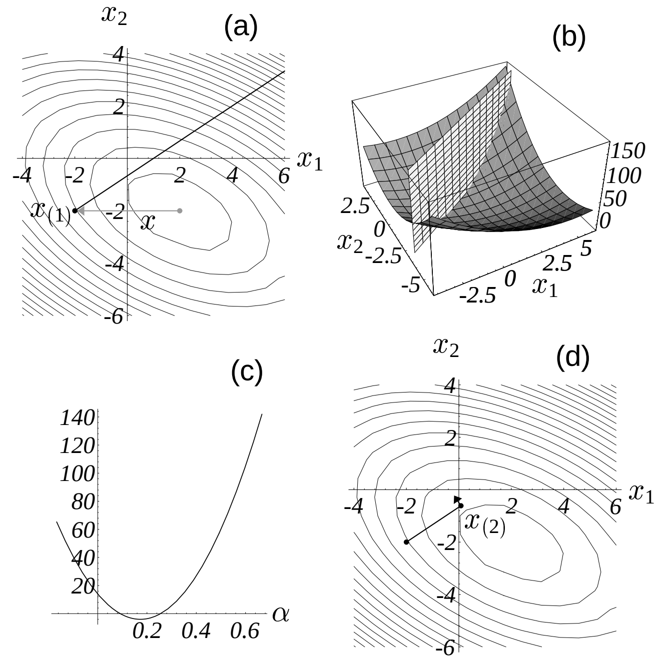
Performing a linesearch

How to compute α^k?

Exact linesearch:

$\alpha^k := \arg\min_{\alpha>0} f(x^k + \alpha s^k)$.

- computationally expensive for nonlinear objectives.



Exact linesearch for quadratic functions

Example: $q(x) = \frac{1}{2} x^T \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} x + (-2 \quad 8)^T x$, where $x \in \mathbb{R}^2$.

Let $x^1 := (-2 \; -2)^T$ and $s^1 := -\nabla q(x^1) = (12 \; 8)^T$.

Figure (a): contours of q and the line $x^1 + \alpha s^1$; (b): the plane $z(\alpha) = x^1 + \alpha s^1$ is shown cutting the q-surface; (c): plot of $\phi(\alpha)$; (d): x^2 is shown and $\phi'(\alpha^*) = 0$.(see next slide) □

Exact linesearches for quadratic objectives

$$q(x) = g^T x + \frac{1}{2} x^T H x, \quad x \in \mathbb{R}^n,$$

and let $\phi_k(\alpha) := q(x^k + \alpha s^k)$. Then

$$\phi'(\alpha) = \frac{d}{d\alpha}\phi(\alpha) = \sum_{i=1}^n \frac{dx_i}{d\alpha} \cdot \frac{\partial}{\partial x_i}\phi(\alpha)$$

$$= \sum_{i=1}^n s_i^k \frac{\partial}{\partial x_i} q(x^k + \alpha s^k) = (s^k)^T \nabla q(x^k + \alpha s^k).$$

- $\nabla q(x) = g + Hx$ and $\nabla q(x^k + \alpha s^k) = g + H(x^k + \alpha s^k)$.

$$\Longrightarrow \quad \phi'(\alpha) = (s^k)^T \nabla q(x^k) + \alpha (s^k)^T H s^k.$$

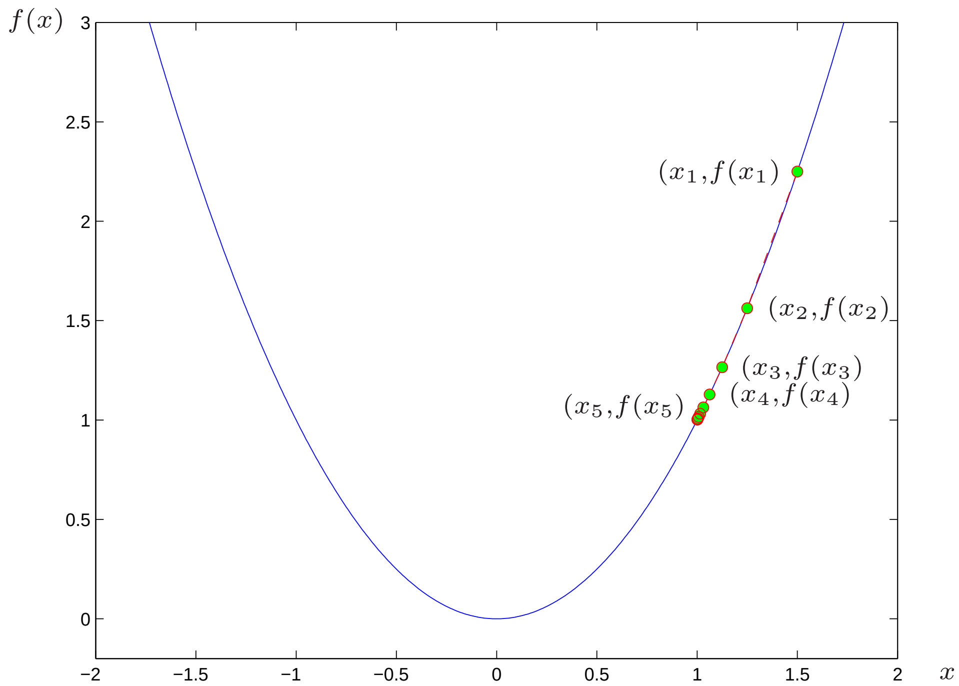
Thus α^* stationary point of $\phi(\alpha)$ iff $(s^k)^T H s^k \neq 0$ and $\phi'(\alpha^*) = 0 \Longrightarrow \quad \alpha^* = -(s^k)^T \nabla q(x^k) / (s^k)^T H s^k.$

- α^* global minimizer of $\phi(\alpha)$ if $(s^k)^T H s^k > 0$.
- for general f, no explicit expression of α^k; approximate minimizers of $f(x^k + \alpha s^k)$ may be used instead. [see Pb Sheet 1]

Inexact linesearch

- want stepsize α^k not “too short”.

Example: $f(x) = x^2$; $x^0 = 2$; $s^k = -1$ and $\alpha^k = 1/(2^{k+1})$ for all k. Then GLM gives $x^k \longrightarrow 1$ as $k \longrightarrow \infty$. [see Pb Sheet 1]



Inexact linesearch ...

- want stepsize α^k not “too short”.

A backtracking linesearch algorithm

Choose $\alpha_{(0)} > 0$ and $\tau \in (0, 1)$.

While $f(x^k + \alpha_{(i)} s^k)'' \geq'' f(x^k)$, REPEAT:

- set $\alpha_{(i+1)} := \tau\alpha_{(i)}$ and $i := i + 1$.

END.

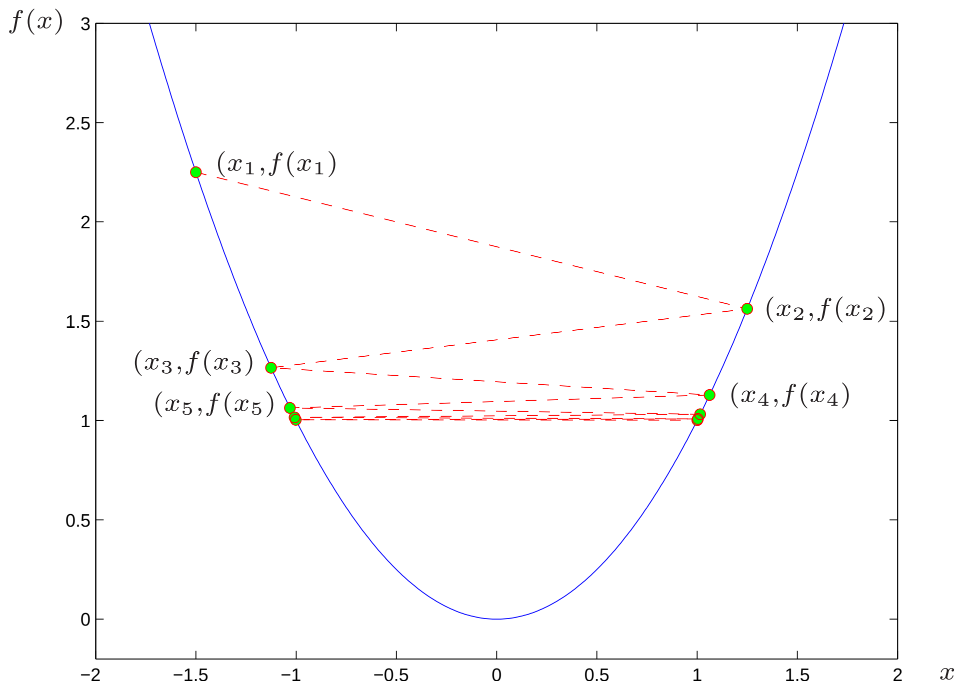
Set $\alpha^k := \alpha_{(i)}$. $\square$

- $\alpha_{(0)} := 1; \tau := 0.5 \implies \alpha_{(0)} := 1, \alpha_{(1)} := 0.5, \alpha_{(2)} := 0.25, \ldots$
- “<”: simple or more sophisticated decrease in f at x^k.

Inexact linesearch ...

- want stepsize α^k not "too long" compared to the decrease in f.

Example: $f(x) = x^2$; $x^0 = 2$; $s^k = (-1)^{k+1}$ and $\alpha^k = 2 + 3/2^{k+1}$ for all k. Then GLM gives $x^k \longrightarrow \pm 1$ as $k \longrightarrow \infty$. [see Pb Sheet 1]



Inexact linesearch ...

- want stepsize α^k not "too long" compared to the decrease in f.

The Armijo condition

`Choose` $\beta \in (0,1)$.
`Compute` $\alpha^k > 0$ `such that`

$$f(x^k + \alpha^k s^k) \leq f(x^k) + \beta \alpha^k \nabla f(x^k)^T s^k \qquad (*)$$

`is satisfied.` $\square$

- in practice, $\beta := 0.1$ or even $\beta := 0.001$.

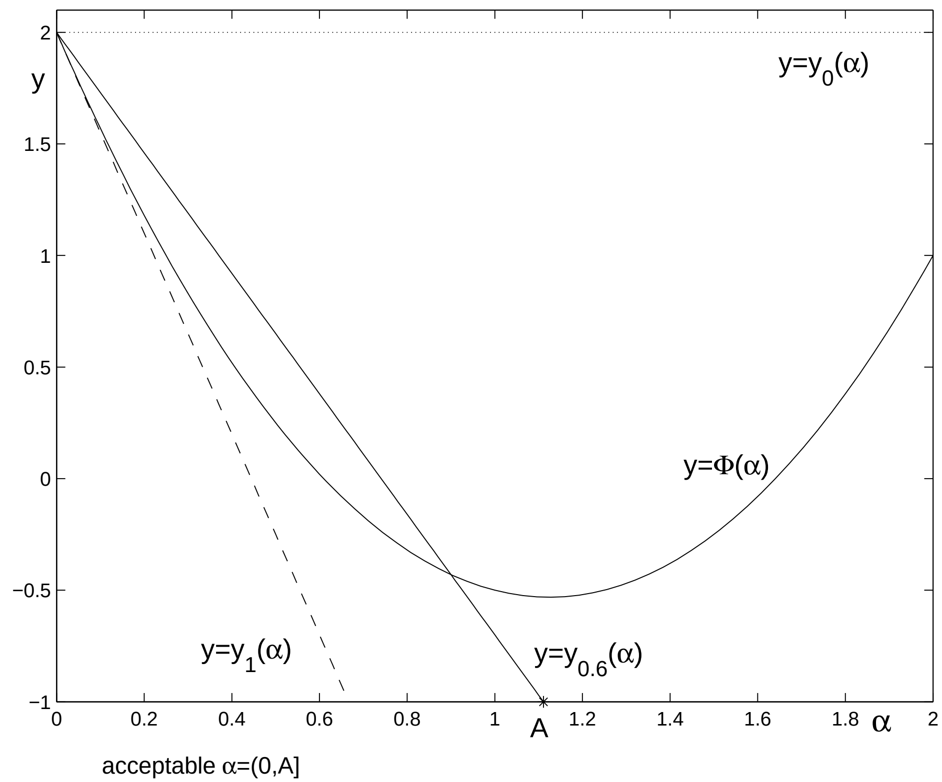
- due to the descent condition, $\exists\, \overline{\alpha}^k > 0$ (unknown explicitly in general) such that $(*)$ holds for all $\alpha \in [0, \overline{\alpha}^k]$. [see Pb Sheet 2]
Choose α^k as large as possible in $(0, \overline{\alpha}^k]$ or in other (greater) intervals of positive α-values that may satisfy $(*)$.

Inexact linesearch ...

- $\Phi_k : \mathbb{R} \to \mathbb{R}, \quad \Phi_k(\alpha) := f(x^k + \alpha s^k), \quad \alpha \geq 0$. Then

  Armijo $\iff \Phi_k(\alpha^k) \leq \Phi_k(0) + \beta\alpha^k\Phi'(0)$.

  Let $y_\beta(\alpha) := \Phi_k(0) + \beta\alpha\Phi'(0), \quad \alpha \geq 0$.



Inexact linesearch ...

The backtracking-Armijo (bArmijo) linesearch algorithm

Choose $\alpha_{(0)} > 0$, $\tau \in (0,1)$ and $\beta \in (0,1)$ [at the start of GLM].

While $f(x^k + \alpha_{(i)} s^k) > f(x^k) + \beta \alpha_{(i)} \nabla f(x^k)^T s^k$, REPEAT:

- set $\alpha_{(i+1)} := \tau \alpha_{(i)}$ and $i := i + 1$.

END.

Set $\alpha^k := \alpha_{(i)}$. $\square$

- $\alpha_{(0)}$, β and τ chosen as before.

- on each GLM iteration k, the bArmijo linesearch algorithm terminates in a finite number of steps with $\alpha^k > 0$, due to the descent condition. [see Pb Sheet 2]

  [without any additional assumptions on $f \in \mathcal{C}^1$]

- other popular/useful inexact linesearch techniques: Wolfe, Goldstein-Armijo, etc.

Global convergence of GLM

- $f \in \mathcal{C}^1(\mathbb{R}^n)$; ∇f is Lipschitz continuous (on $\mathbb{R}^n$) iff $\exists L > 0$,
$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|, \quad \forall x, y \in \mathbb{R}^n.$$

Lemma 2. Let $f \in \mathcal{C}^1(\mathbb{R}^n)$ with ∇f Lipschitz continuous with Lipschitz constant L. Then the Armijo condition is satisfied:
$$f(x^k + \alpha s^k) \leq f(x^k) + \beta\alpha\nabla f(x^k)^T s^k \text{ for all } \alpha \in [0, \alpha^k_{\max}],$$
where $\alpha^k_{\max} = (\beta - 1)\nabla f(x^k)^T s^k / [L\|s^k\|^2]$.

Proof. First-order Taylor: for any $\alpha > 0$ and some $\tilde{\alpha} \in (0, \alpha)$,
$$f(x^k+\alpha s^k) = f(x^k)+\alpha\nabla f(x^k)^T s^k+\alpha[\nabla f(x^k+\tilde{\alpha}s^k)-\nabla f(x^k)]^T s^k$$
$$\leq f(x^k) + \alpha\nabla f(x^k)^T s^k + \alpha\|\nabla f(x^k + \tilde{\alpha}s^k) - \nabla f(x^k)\| \cdot \|s^k\|$$
$$\leq f(x^k) + \alpha\nabla f(x^k)^T s^k + \alpha L\tilde{\alpha}\|s^k\|^2 \leq f(x^k) + \alpha\nabla f(x^k)^T s^k + \alpha^2 L\|s^k\|^2.$$
Thus Armijo condition (*) satisfied for all $\alpha \geq 0$ such that
$$f(x^k) + \alpha\nabla f(x^k)^T s^k + \alpha^2 L\|s^k\|^2 \leq f(x^k) + \beta\alpha\nabla f(x^k)^T s^k,$$
which is equivalent to $\alpha \in [0, \alpha^k_{\max}]$. $\Box$

Global convergence of GLM ...

Lemma 3. Let $f \in \mathcal{C}^1(\mathbb{R}^n)$ with ∇f Lipschitz continuous with Lipschitz constant L. Then the bArmijo stepsize α^k satisfies

$$\alpha^k \geq \min\{\alpha_{(0)}, \tau\alpha^k_{\max}\} \text{ for all } k \geq 0.$$

Proof of Lemma 3. If $\alpha_{(0)}$ satisfies the Armijo condition (*), bArmijo terminates with $i = 0$ and $\alpha^k = \alpha_{(0)}$. Else, it will terminate as soon as $\alpha^k \leq \alpha^k_{\max}$. Then, let $(i-1)$ be the last iteration such that $\alpha_{(i-1)} > \alpha^k_{\max}$ and $\alpha_{(i)} \leq \alpha^k_{\max}$. It follows that $\alpha^k = \alpha_{(i)} = \tau\alpha_{(i-1)} > \tau\alpha^k_{\max}$. Note that if $\alpha_{(0)} > \alpha^k_{\max}$, then $\alpha_{(i)} = \tau^i\alpha_{(0)} \leq \alpha^k_{\max}$ for any $i \geq \log(\alpha_{(0)}/\alpha^k_{\max})/|\log\tau|$. □

- for the global convergence of GLM, we need a lower bound on the bArmijo stepsize (as in Lemma 3), not just to know that such a stepsize exists; hence we required stronger assumptions on f.

(global convergence of GLM to be continued ...)