Lecture 3: Linesearch methods (continued). Steepest descent methods

Coralia Cartis, Mathematical Institute, University of Oxford

C6.2/B2: Continuous Optimization

Global convergence of GLM (continued)

Theorem 4. Let $f \in \mathcal{C}^1(\mathbb{R}^n)$ be bounded below on $\mathbb{R}^n$ by f_{low}. Let ∇f Lipschitz continuous. Apply GLM with bArmijo linesearch to minimizing f with $\epsilon := 0$. Then

either

there exists $l \geq 0$ such that $\nabla f(x^l) = 0$

or

$$\lim_{k\to\infty} \min\left\{ \frac{|\nabla f(x^k)^T s^k|}{\|s^k\|}, |\nabla f(x^k)^T s^k| \right\} = 0.$$

Proof of Theorem 4. Assume $\nabla f(x^k) \neq 0$ for all k so GLM does not terminate finitely. Then Armijo condition (*) gives

$$f(x^k) - f(x^{k+1}) \geq \beta \alpha^k (-\nabla f(x^k))^T s^k \text{ for all } k \geq 0.$$

Summing this up from $k = 0$ to $k = i$, consecutive terms on the left-hand side cancel to give

$$f(x^0) - f(x^{i+1}) \geq \beta \sum_{k=0}^{i} \alpha^k (-\nabla f(x^k))^T s^k \text{ for all } i \geq 0.$$

As f is bounded below by f_{low}, $f(x^{i+1}) \geq f_{\text{low}}$ for all $i \geq 0$.

Global convergence of GLM ...

Proof of Theorem 4. Thus we deduce from the above that

$$\infty > f(x^0) - f_{\rm low} \geq \beta \sum_{k=0}^{\infty} \alpha^k |\nabla f(x^k))^T s^k|, \quad (1)$$

where we also used that $\nabla f(x^k)^T s^k < 0$ so that $(-\nabla f(x^k))^T s^k = |\nabla f(x^k))^T s^k|$. We deduce from the convergence of the series in (1) that

$$\lim_{k \longrightarrow \infty} \alpha^k |\nabla f(x^k))^T s^k| = 0. \quad (2)$$

Let $\mathcal{K}_1 = \{k : \alpha_{(0)} \geq \tau \alpha^k_{\max}\}$ and $\mathcal{K}_2 = \{k : \alpha_{(0)} < \tau \alpha^k_{\max}\}$. For all $k \in \mathcal{K}_1$, we have from Lemmas 2 & 3 that

$$\alpha^k |\nabla f(x^k))^T s^k| \geq \frac{(1-\beta)\tau}{L} \cdot \left(\frac{|\nabla f(x^k)^T s^k|}{\|s^k\|} \right)^2 \geq 0$$

and so (2) implies $\lim_{k \to \infty, k \in \mathcal{K}_1} |\nabla f(x^k)^T s^k| / \|s^k\| = 0$. Lemma 3 gives that $\alpha^k \geq \alpha_{(0)}$ for all $k \in \mathcal{K}_2$ and so (2) provides $\lim_{k \to \infty, k \in \mathcal{K}_2} |\nabla f(x^k)^T s^k| = 0$. These two limits and the property $\min\{a_k, b_k\} \leq a_k,\ b_k,\ \forall k$, give the required limit.$\square$

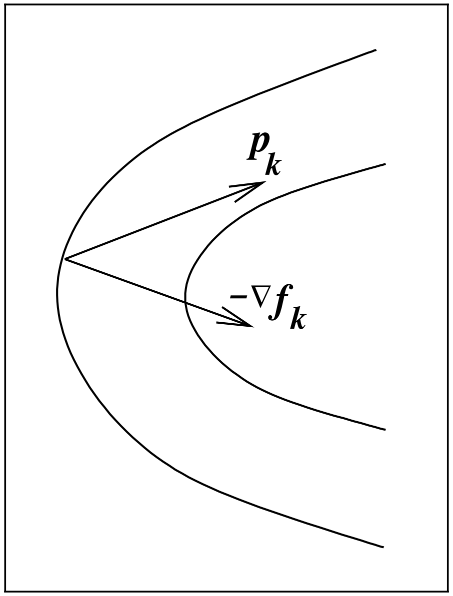
Global convergence of GLM ...

Interpretation of Theorem 4: Recall

$$\cos\theta_k = \frac{(-\nabla f(x^k))^T s^k}{\|\nabla f(x^k)\| \cdot \|s^k\|} = \frac{|\nabla f(x^k)^T s^k|}{\|\nabla f(x^k)\| \cdot \|s^k\|}.$$

Then Th 4 gives, if $\nabla f(x^k) \neq 0$ for all k,

$$\lim_{k\to\infty} \|\nabla f(x^k)\| \cdot \cos\theta_k \cdot \min\{1, \|s^k\|\} = 0.$$



A descent direction p_k.

Thus to ensure global convergence of GLM, namely, $\|\nabla f(x^k)\| \longrightarrow 0$ as $k \to \infty$, it is not sufficient to have s^k be descent for each k; we need $\cos\theta_k \geq \delta > 0$ for all k, so that s^k is prevented from becoming orthogonal to the gradient as k increases.

Summary and a look ahead

Linesearch methods:

- Linesearch: how to choose the stepsize α^k, from any x^k and along any descent direction s^k.
- How to choose a descent direction s^k? What are the important such choices of s^k?
  - Steepest descent direction (next).
  - Newton direction.

Steepest descent method

Steepest descent (SD) direction: set $s^k := -\nabla f(x^k)$, $k \geq 0$, in Generic Linesearch Method (GLM).

- s^k descent direction whenever $\nabla f(x^k) \neq 0$:

$$\nabla f(x^k)^T s^k < 0 \iff \nabla f(x^k)^T(-\nabla f(x^k)) < 0 \iff -\|\nabla f(x^k)\|^2 < 0.$$

- s^k steepest descent: unique global solution of

$$\text{minimize}_{s \in \mathbb{R}^n} f(x^k) + s^T \nabla f(x^k) \quad \text{subject to} \quad \|s\| = \|\nabla f(x^k)\|.$$

Cauchy-Schwarz: $|s^T \nabla f(x^k)| \leq \|s\| \cdot \|\nabla f(x^k)\|$, $\forall s$, with equality iff s is proportional to $\|\nabla f(x^k)\|$.

Steepest descent methods

Method of steepest descent (SD): GLM with $s^k ==$ SD direction; any linesearch.

Steepest Descent (SD) Method

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

- compute $s^k = -\nabla f(x^k)$.
- 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 $k := k + 1$. $\square$

- SD-e :== SD method with exact linesearches;
- SD-bA :== SD method with bArmijo linesearches.

Global convergence of steepest descent methods

- $f \in \mathcal{C}^1(\mathbb{R}^n)$; ∇f is Lipschitz continuous (on $\mathbb{R}^n$) iff $\exists L > 0$,

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|, \quad \forall x, \, y \in \mathbb{R}^n.$$

Theorem 5   Let $f \in \mathcal{C}^1(\mathbb{R}^n)$ be bounded below on $\mathbb{R}^n$.
Let ∇f be Lipschitz continuous. Apply the SD-e or the SD-bA method to minimizing f with $\epsilon := 0$.
Then both variants of the SD method have the property:

either

there exists $l \geq 0$ such that $\nabla f(x^l) = 0$

or

$\|\nabla f(x^k)\| \rightarrow 0$ as $k \rightarrow \infty$.

Proof for SD-bA.   Let $s^k = -\nabla f(x^k)$ for all k in Th 4. $\square$

SD methods have excellent global convergence properties (under weak assumptions).

Some disadvatanges of steepest descent methods

- SD methods are scale-dependent.

  poorly scaled problem/variables $\Longrightarrow$ SD direction gives little progress.

- Usually, SD methods converge very slowly to solution, asymptotically.

The scale-dependence of steepest descent

Example of a poorly scaled quadratic.

$$f(x) = \frac{1}{2}(ax_1^2 + x_2^2) = \frac{1}{2}x^T \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} x, \quad x = (x_1 \;\; x_2)^T, \quad (\Diamond)$$

where $a > 0$. Note $x^* = (0 \;\; 0)^T$ unique global minimizer.

- $a \gg 1 \quad \longrightarrow \quad f$ poorly scaled (or poorly conditioned).
- apply SD-e to ($\Diamond$) starting at $x^0 := (1 \;\; a)^T$. Then[see Pb Sheet 2]

$$x^k = \left(\frac{a-1}{a+1}\right)^k \begin{pmatrix} (-1)^k \\ a \end{pmatrix}, \quad k \geq 0.$$

$\Longrightarrow \quad x^k \to 0$ as $k \to \infty$, linearly with $\rho := |(a-1)/(a+1)|$ convergence factor.

- $a \gg 1 \implies \rho$ closer to $1 \implies$ SD-e converges very slowly.

The scale-dependence of steepest descent

Example of a well-scaled quadratic.

Linear transformation of variables:

$$y = \begin{pmatrix} a^{1/2} & 0 \\ 0 & 1 \end{pmatrix} x.$$

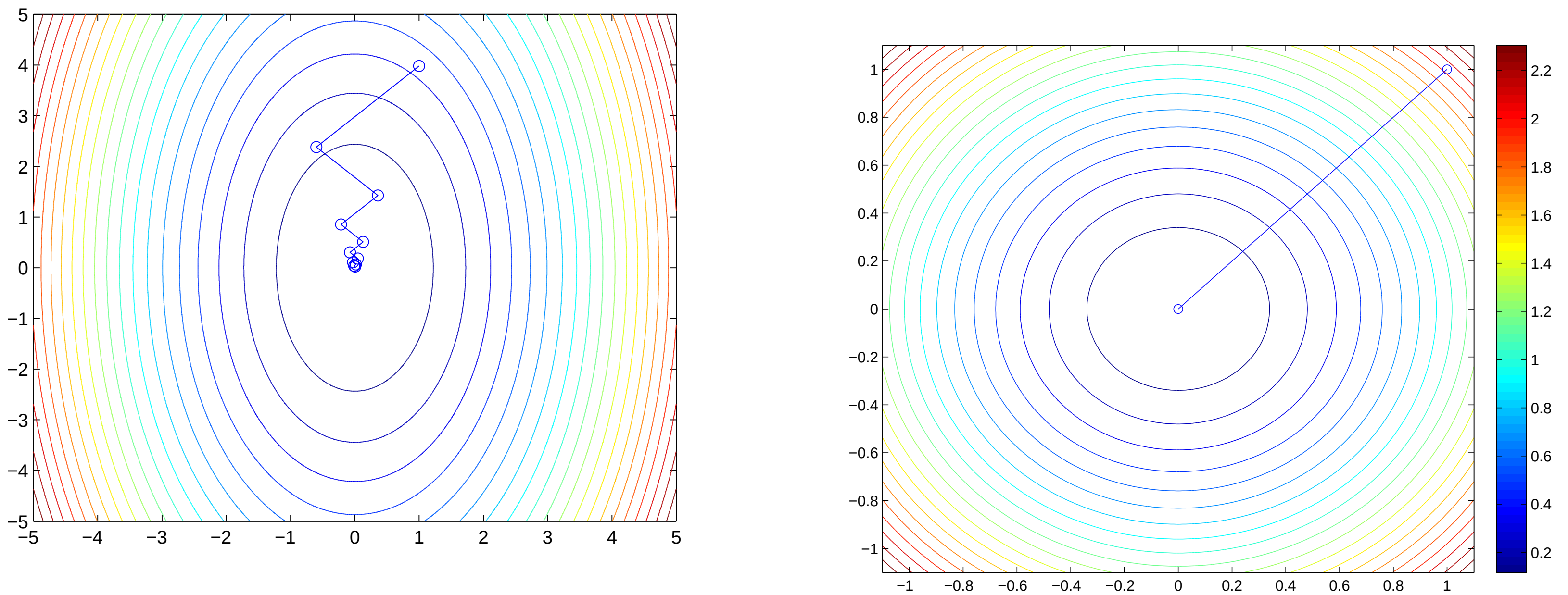
- let $\overline{f}(y) := f(x(y))$, namely f in the new coordinates y.

$\Longrightarrow \overline{f}(y) = \frac{1}{2} y^T y = \frac{1}{2}(y_1^2 + y_2^2)$.

$\longrightarrow \quad \overline{f}$ well-scaled.

- $y^* = (0\ 0)^T$ unique global minimizer.
- apply SD-e to $\overline{f}$ from any $y^0 \in \mathbb{R}^2$: $y^1 = (0\ 0)^T = y^*$.

The scale-dependence of steepest descent



The effect of problem scaling on SD-e performance.
Left figure: $a = 10^{0.6}$ (mildly poor scaling).
Right figure: $a = 1$ ("perfect" scaling).

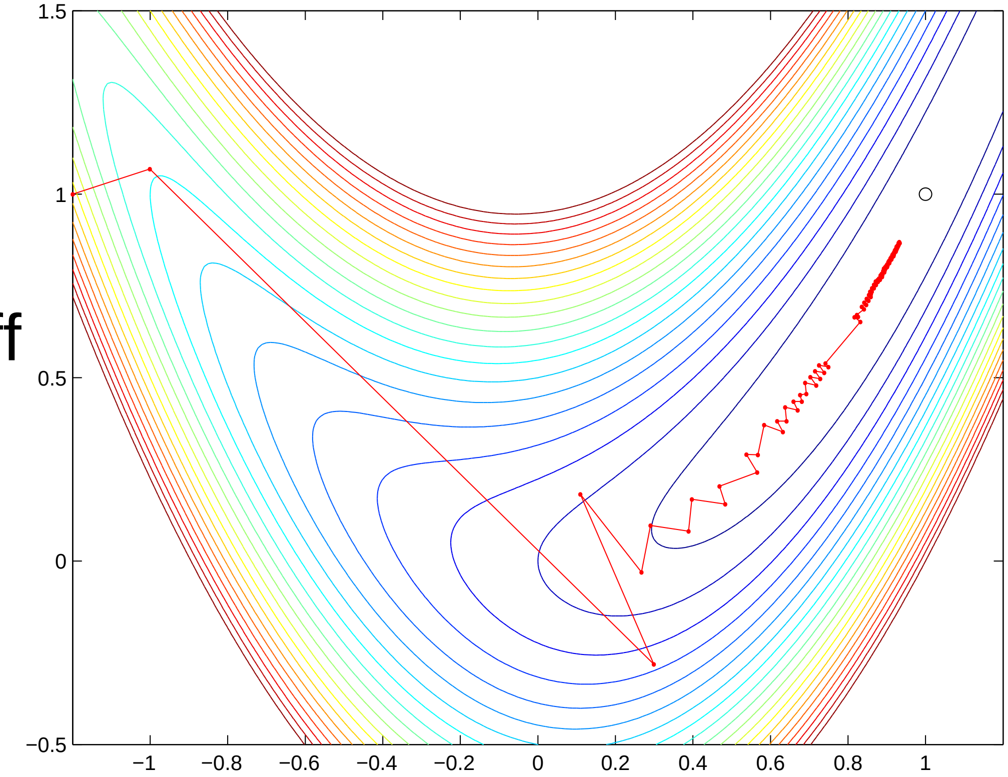
Local rate of convergence for steepest descent

- Usually, SD methods converge very slowly to solution, asymptotically.

theory: very slow conv.

numerics: break-down (cumulation of round-off and ill-conditioning).

$$f(x_1, x_2) = 10(x_2 - x_1^2)^2 + (x_1 - 1)^2.$$



SD-bA applied to the Rosenbrock function f.

Local rate of convergence for steepest descent

Asymptotically, SD converges linearly to a solution,

$$|f(x^{k+1}) - f(x^*)| \leq \rho|f(x^k) - f(x^*)|, \forall k \text{ suff. large}$$

BUT convergence factor ρ v. close to 1 usually!

Theorem 6 $f \in \mathcal{C}^2$; x^* local minimizer of f with $\nabla^2 f(x^*)$ positive definite $\longrightarrow \lambda^*_{\max}$, $\lambda^*_{\min}$ eigenvalues.
Apply SD-e to $\min f$. If $x^k \to x^*$ as $k \to \infty$, then $f(x^k)$ converges linearly to $f(x^*)$,

$$\rho \leq \left(\frac{\kappa(x^*)-1}{\kappa(x^*)+1}\right)^2 := \rho_{SD},$$

where $\kappa(x^*) = \lambda^*_{\max}/\lambda^*_{\min}$ condition number of $\nabla^2 f(x^*)$.

- practice: $\rho = \rho_{SD}$;

for Rosenbrock f: $\kappa(x^*) = 258.10$, $\rho_{SD} \approx 0.984$.

- $\kappa(x^*) = 800$, $f(x^0) = 1$, $f(x^*) = 0$. SD-e gives

$f(x^k) \approx 0.007$ after 1000 iterations!

Summary: steepest descent methods

- first-order method $\longrightarrow$ inexpensive.
- global convergence under weak assumptions, but no second-order optimality guarantees for the generated solution.
- scale-dependent; too expensive, or impossible, to make a function well-scaled.
- when the objective is poorly scaled, very very slow convergence to a solution; hence, not used in general.
- useful sometimes: for example, for some convex problems with special structure that are very well conditioned (compressed sensing, etc).