XXX Congreso Nacional – Sociedad Matemática Mexicana
September 29 - October 3, 1997

Optimization Algorithms

From Line Searches to Trust Regions

Jorge Moré

Mathematics and Computer Science Division
Argonne National Laboratory
http://www.mcs.anl.gov/home/more/

Background

The fundamental problem is to find a local minimizer of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, that is, solve

$$\min \{ f(x) : x \in \Omega \},$$

where Ω is an open set in $\mathbb{R}^n$.

Facts

- If x^* is a local minimizer of f then $\nabla f(x^*) = 0$.
- If x^* is a local minimizer of f and f is twice differentiable at x^* then the Hessian matrix $\nabla^2 f(x^*)$ is positive semidefinite.
- If $\nabla f(x^*) =$ and $\nabla^2 f(x^*)$ is positive definite then x^* is an *isolated* local minimizer of f.

Theoretical Goals

- $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$
- Every limit point x^* of $\{x_k\}$ satisfies $\nabla f(x^*) = 0$
- $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$
- If $\{x_k\}$ has a limit point x^* and $\nabla^2 f(x^*)$ is positive definite, then $\{x_k\}$ converges to x^*
- If $\{x_k\}$ converges to x^* and $\nabla^2 f(x^*)$ is positive definite, then $\{x_k\}$ converges Q-linearly (Q-superlinearly, Q-quadratically).
- If $\{x_k\}$ converges to x^*, then $\nabla^2 f(x^*)$ is positive semidefinite

Optimal Design with Composites Problem

Determine a minimizer $v : \mathbb{R}^2 \mapsto \mathbb{R}$ of the problem

$$\min\{f_\lambda(v) : v \in H_0^1(\mathcal{D})\},$$

where

$$f_\lambda(v) = \int_{\mathcal{D}} \Big\{ \psi_\lambda\left(\|\nabla v(x)\|\right) + v(x) \Big\}\, dx$$

$$\psi_\lambda(t) = \begin{cases} \frac{1}{2}\mu_2 t^2, & 0 \le t \le t_1, \\ \mu_2 t_1 (t - \frac{1}{2} t_1), & t_1 \le t \le t_2, \\ \frac{1}{2}\mu_1 (t^2 - t_2^2) + \mu_2 t_1 (t_2 - \frac{1}{2} t_1), & t_2 \le t, \end{cases}$$

$$t_1 = \left(2\lambda \frac{\mu_1}{\mu_2}\right)^{\frac{1}{2}}, \qquad t_2 = \left(2\lambda \frac{\mu_2}{\mu_1}\right)^{\frac{1}{2}}.$$

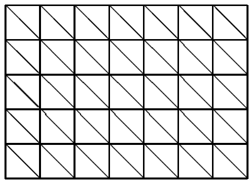
Discretization of the ODC Problem

A finite element formulation leads to

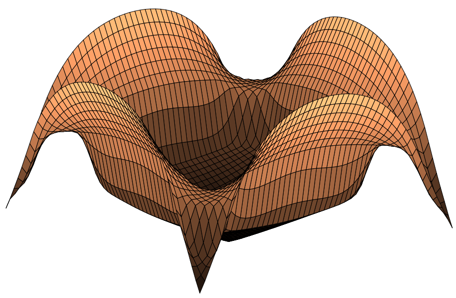
$$\min \{ f(x) : x \in \mathbb{R}^n \}$$

with `nx*ny` variables. The mapping $f : \mathbb{R}^n \to \mathbb{R}$ and the gradient ∇f can be computed in order n operations.

The Hessian matrix $\nabla^2 f$ is sparse but piecewise continuous.



Optimal Design Problem



$\|\nabla v\|$ for $\lambda = 0.008$

The Ginzburg-Landau Model for Superconductivity

$$\min\{f(v,A) : v \in K_1,\ A \in K_2\},$$

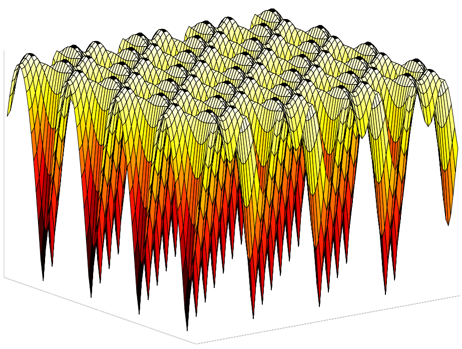
$$f(v,A) = \int_{\mathcal{D}} \left\{ -|v(x)|^2 + \tfrac{1}{2}|v(x)|^4 + \phi(v,A) \right\}\ dx$$

$$\phi(v,A) = \left\| \left[\nabla - iA(x)\right] v(x) \right\|^2 + \kappa^2 \left\| (\nabla \times A)(x) \right\|^2$$

$v : \mathbb{R}^2 \to \mathbb{C}$ is the order parameter

$A : \mathbb{R}^2 \to \mathbb{R}^2$ is the vector potential

The Ginzburg-Landau Model for Superconductivity



Surface plot of v for $\kappa = 5$

Line Searches: Motivation

Given $x \in \mathbb{R}^n$, determine a descent direction p in the sense that

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

for all $\alpha > 0$ sufficiently small. Now replace x by $x + \alpha p$ and repeat the process.

Facts

- The direction p is a descent direction if $\nabla f(x)^T p < 0$.
- $\nabla f(x)^T p < 0$ if and only if $p = -B^{-1} \nabla f(x)$ for some positive definite B.

Line search algorithms

- Compute $\nabla f(x_k)$ and determine a positive definite B_k.
- Solve for $p_k = -B_k^{-1} \nabla f(x_k)$
- Determine an α_k that satisfies some decrease condition. For example, the strong Wolfe conditions require the *sufficient decrease* condition ($\mu = 0.001$)

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \mu \alpha_k \nabla f(x_k)^T p_k$$

  and the *curvature* condition ($\eta = 0.9$)

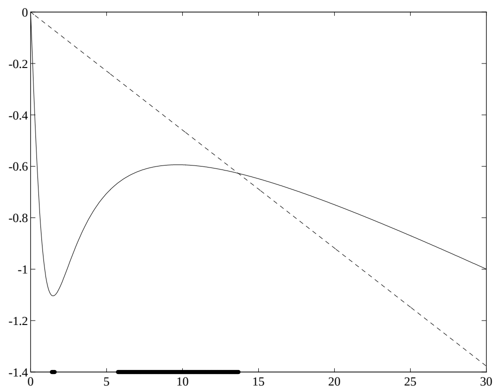
$$\left| \nabla f(x_k + \alpha_k p_k)^T p_k \right| \leq \eta \left| \nabla f(x_k)^T p_k \right|$$

- Set $x_{k+1} = x_k + \alpha_k p_k$
- Update B_k

The Strong Wolfe Line Search Conditions

If $\phi(\alpha) = f(x + \alpha p)$ then the Strong Wolfe conditions are

$$\phi(\alpha) \le \phi(0) + \mu\phi'(0)\alpha, \qquad |\phi'(\alpha)| \le \eta|\phi'(0)|$$



Solid plot is ϕ; dashed plot is $l(\alpha) = \phi(0) + \mu\phi'(0)\alpha$.

Convergence Results

Assume that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable on

$$\mathcal{L}(x_0) = \{f(x) \le f(x_0)\},$$

and that $\mathcal{L}(x_0)$ is compact. If $p_k = -B_k^{-1}\nabla f(x_k)$ and B_k is positive definite with bounded condition numbers, that is,

$$\|B_k\| \, \|B_k^{-1}\| \le \kappa, \qquad k \ge 0$$

then

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$$

Note The algorithm terminates if $\nabla f(x_k) = 0$ at some iterate.

FAQ for Line Search Algorithms

- How do we determine α_k?
  - Line search strategies.
  - Non-monotone line searches.
- How do we determine a positive definite B_k?
  - Variable Metric methods.
  - Modified Cholesky factorizations.
  - Limited Memory - Variable Metric methods.
- Can we guarantee convergence with $\nabla^2 f(x^*)$ positive semidefinite?
  - Directions of negative curvature.
- How do we determine p_k in large-scale problems?
  - Truncated Newton methods.

Modified Cholesky Factorizations - Newton's Method

$$f(x+p) \approx f(x) + \nabla f(x)^T p + \tfrac{1}{2} p^T \nabla^2 f(x) p$$

If $\nabla^2 f(x)$ is positive definite then the quadratic

$$q(p) = \nabla f(x)^T p + \tfrac{1}{2} p^T \nabla^2 f(x) p$$

has a global minimizer given by the Newton direction

$$p^N = -\nabla^2 f(x)^{-1} \nabla f(x)$$

If $\nabla^2 f(x)$ is not positive definite then we define

$$p = -B^{-1} \nabla f(x)$$

where B is a positive definite approximation to $\nabla^2 f(x)$.

Modified Cholesky Factorizations

Given a symmetric matrix A, a modified Cholesky factorization produces a symmetric diagonal perturbation E such that

- If A is *sufficiently positive definite* then $E = 0$.
- If A is indefinite then

  $$\|E\| \le \beta \min\{\|S\| : A + S \text{ is positive definite}\}$$

  for some $\beta > 1$.
- The matrix $A + E$ is well conditioned.
- The cost of the algorithm should be comparable to cost of a standard (unmodified) Cholesky factorization.

Trust Region Methods: Motivation

$$f(x + w) \approx f(x) + \nabla f(x)^T w + \tfrac{1}{2} w^T \nabla^2 f(x) w$$

Determine the step w as an approximate solution of the subproblem

$$\min\{q(w) : \|w\| \le \Delta\}, \qquad (\mathcal{P})$$

where

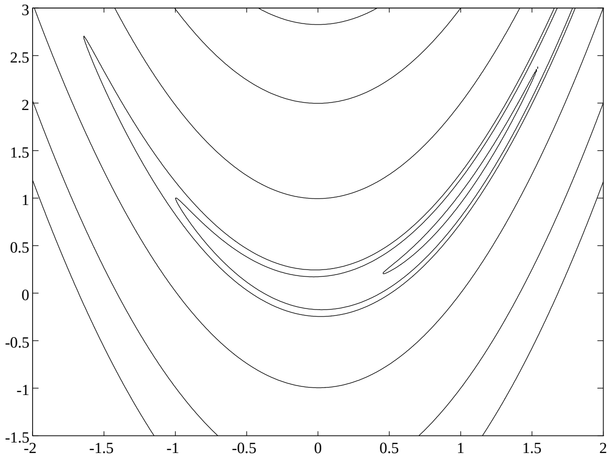
$$q(w) = \nabla f(x)^T w + \tfrac{1}{2} w^T \nabla^2 f(x) w$$

Update the trust region radius Δ depending on progress.

Fact

If $q : \mathbb{R}^n \mapsto \mathbb{R}$ is a general quadratic and $\|\cdot\|$ is the l_2 norm, then we can determine the global minimum of $(\mathcal{P})$ in order n^3 operations. Problem $(\mathcal{P})$ is NP-hard if $\|\cdot\|$ is the l_∞ norm.

Trust Region Methods – A Picture



Contours of $f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1^2)^2$

Trust Region Methods

- Compute a scaling matrix D_k and the *quadratic* model

$$q_k(w) = \nabla f(x_k)^T w + \tfrac{1}{2} w^T B_k w,$$

- Compute an approximate solution s_k to the *subproblem*

$$\min \{ q_k(w) : \| D_k w \| \le \Delta_k \}$$

- Compute the ratio of actual to predicted reduction

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{q_k(0) - q_k(s_k)}$$

- Set $x_{k+1} = x_k + s_k$ if $\rho_k \ge \sigma_0$ and $x_{k+1} = x_k$ if $\rho_k < \sigma_0$.
- Update B_k, D_k and Δ_k.

Convergence Results

Assume that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable on

$$\mathcal{L}(x_0) = \{ f(x) \le f(x_0) \} ,$$

and that $\mathcal{L}(x_0)$ is compact. If $\{ \|B_k\| \}$ is bounded, then

$$\lim_{k \to \infty} \|\nabla f(x_k)\| = 0$$

If s_k is determined by the optimal step and $\{x_k\}$ converges to x^*, then $\nabla^2 f(x^*)$ is positive semidefinite

FAQ for Trust Region Algorithms

- How do we determine s_k?
  - Dogleg algorithm
  - Optimal algorithm
- How do we update Δ_k
- How do we compute the scaling matrix D_k?
  - A diagonal scaling matrix.
  - An incomplete Cholesky factorization.
- How do we determine the step s_k in large-scale problems?
  - Truncated Newton methods.

Algorithms for the Subproblem

$$\min\{q(w) : \|w\| \le \Delta\}, \qquad q(w) = g^T w + \tfrac{1}{2} w^T B w$$

The optimal algorithm

$$q(s) \le \beta \min\{q(w) : \|w\| \le \Delta\}$$

The dogleg algorithm

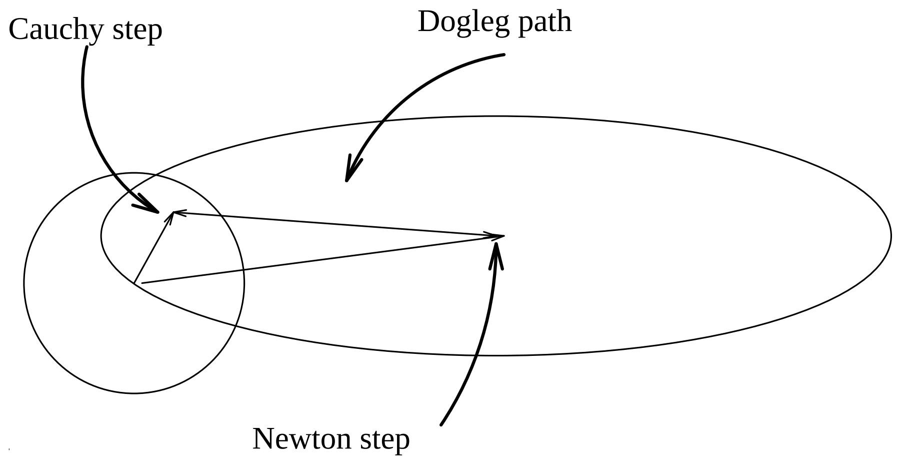
$$q(s) = \min\{q(s(\alpha)) : \|s(\alpha)\| \le \Delta\}, \qquad s(\alpha) = \alpha s_C + (1-\alpha) s_N$$

Newton step $s^N = -B^{-1} g$

Cauchy step s^C

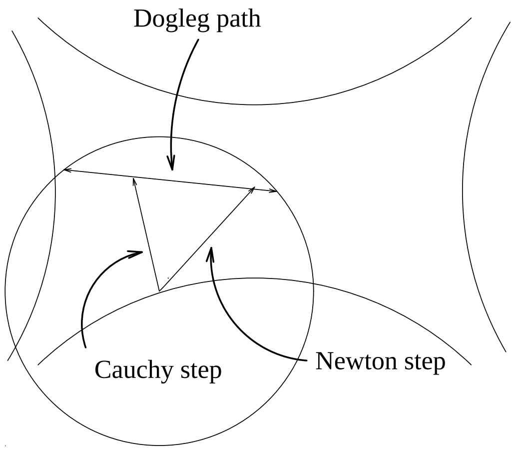
$$q(s^C) = \min\{q(w) : w = \nu g,\ \|w\| \le \Delta\}$$

The Dogleg Algorithm: The Convex Case



The ellipse is the level set of the quadratic $q(w) = g^T w + \frac{1}{2} w^T B w$

The Dogleg Algorithm: The Non-Convex Case



The hyperbola is the level set of $q(w) = g^T w + \frac{1}{2} w^T B w$

The Optimal Algorithm

A vector $w^* \in \mathbb{R}^n$ is a **global** minimizer of

$$\min \left\{ g^T w + \tfrac{1}{2} w^T B w : \|w\|_2 \le \Delta \right\}$$

if and only if there is a λ^* such that

$$(B + \lambda^* I)\, w^* = -g, \quad \lambda^* \ge 0$$
$$B + \lambda^* I \text{ positive semidefinite}$$
$$\|w^*\|_2 \le \Delta, \text{ and } \|w^*\|_2 = \Delta \text{ if } \lambda^* > 0$$

If $B + \lambda^* I$ is positive definite then either $\lambda^* = 0$ or $\lambda^* \ge 0$ is the largest root of the equation

$$\left\| (B + \lambda^* I)^{-1} g \right\|_2 = \Delta$$

Truncated Newton Methods

Steihaug [1983] suggested using the PCG method on the model

$$q(w) = g^T w + \tfrac{1}{2} w^T B w$$

to generate a step s for the subproblem.

The CG algorithm on q generates a sequence $\{w_l\}$ such that

$$w_{l+1} = w_l + \alpha_l p_l$$

The iterate w_{l+1} is well defined if $p_l^T B p_l > 0$. Moreover,

$$\|w_l\| \le \|w_{l+1}\|$$

Truncated Newton Methods

Negative curvature: If $p_l^T B p_l \le 0$ then choose β such that

$$s = w_l + \beta p_l, \qquad \|s\| = \Delta.$$

The CG iterates exit the trust region: If

$$\|w_l\| \le \Delta, \qquad \|w_{l+1}\| > \Delta$$

then choose $\beta \in (0,1)$ such that

$$s = w_l + \beta(w_{l+1} - w_l), \qquad \|s\| = \Delta.$$

The CG iterates converge: Set $s = w_l$ if

$$\|B w_l + g\| \le \eta \|g\|, \qquad \|w_l\| \le \Delta, \qquad \eta \in (0,1).$$