A Brief Survey of Approaches for Unconstrained Optimization Problems

Xin Liu

State Key Laboratory of Scientific and Engineering Computing
Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences, China

Courant Institute of Mathematical Sciences
New York University

Deep Learning Seminar
March 24, 2017

Outline

1. Basic Conceptions
   - Problem Description
   - Optimality Conditions
   - Iterative Methods
2. Classical Optimization Methods
   - Gradient Methods
   - Newton Methods
   - Trust Region Methods
   - Methods for Nonlinear Least Squares
   - Block Coordinate Descent
3. Global Optimization Strategies
   - Overview
   - Undeterministic Methods

Section 1. Basic Conceptions

Problem Description

Unconstrained optimization models

$$\min_{x \in \mathbb{R}^n} \quad f(x).$$

- $f : \mathbb{R}^n \longmapsto \mathbb{R}$.
- convex or nonconvex
- differentiable or nondifferentiable
- acquirable information: function value, derivative[1]
- constrained optimization

$$\min_{x \in \mathbb{R}^n} \quad f(x), \quad \text{s. t.} \quad x \in C.$$

  - equivalent: $\min_{x \in \mathbb{R}^n} f(x) + \delta_C(x)$, where $\delta_C(x) := \begin{cases} 0, & \text{if } x \in C; \\ 1, & \text{otherwise.} \end{cases}$
  - exact penalty functions: ℓ_1 penalty, augmented Lagrangian, ...

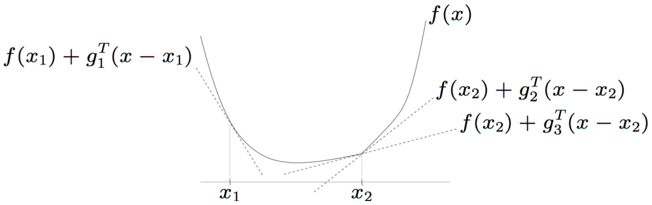
[1] Derivative Free Optimization (DFO) is out of the scope of this presentation.

Optimality Conditions

First-order optimality conditions

- f is differentiable: $\nabla f(x) = 0$.
- f is nondifferentiable but convex:

$$0 \in \partial f(x) := \{g \mid f(y) \geq f(x) + g^\top (y - x), \ \forall\, y\}.$$



Second-order necessary (sufficient) optimality conditions

- f is second-order differentiable: $\nabla^2 f(x) \succeq (\succ) 0$.

Optimality Conditions (Cont'd)

Optimization condition (differentiability is assumed)

- f is convex:
  - x^* is a global minimizer $\Leftrightarrow$ $\nabla f(x^*) = 0$
- f is nonconvex:
  - x^* is a first-order stationary point $\Leftrightarrow$ $\nabla f(x^*) = 0$ ★
  - x^* is a local minimizer $\Rightarrow$ $\nabla f(x^*) = 0$
  - x^* is a second-order stationary point $\Leftrightarrow$ ★ and $\nabla^2 f(x^*) \succeq 0$
  - x^* is a local minimizer $\Rightarrow$ $\nabla^2 f(x^*) \succeq 0$
  - x^* is a local minimizer $\Leftarrow$ ★ and $\nabla^2 f(x^*) \succ 0$

Optimality Conditions (Cont'd)

Finding a minimizer (nonconvexness is assumed)

- finding global minimizer is numerically impossible
- finding global minimizer for quartic polynomial is already NP-hard
- finding local minimizer is not easier

The task of numerical optimization methods

- first-order methods: finding first-order stationary point
- second-order methods: finding second-order stationary point
- only when f is structured, finding global minimizer or local minimizer becomes possible

Iterative Methods (Cont'd)

Stopping criterions

- first-order criterion: $\|\nabla f(x)\| < \epsilon$
- second-order criterion: $\lambda_{\min}(\nabla^2 f(x)) > -\epsilon$

Iterative methods – framework

(1) Input: initial guess $x^{(0)}$, tolerance $\epsilon > 0$, $k := 0$;

(2) Main iteration: $x^{(k+1)} = h(x^{(k)})$;

(3) Check stopping criterion, if satisfied, then terminate and return $x^{(k+1)}$; otherwise, set $k := k + 1$ and goto step (2).

Iterative Methods (Cont'd)

Iterative methods – choosing h

- line search: $x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$.
  - gradient methods;
  - Newton methods;
  - ... ...
- trust region methods
- block coordinate descent methods
- ... ...

Fixed-point convergence – contraction

- $\|\mathcal{J}_h(x)\| < 1$ holds for a given norm $\|\cdot\|$ and any $x \in \mathbb{R}^n$, where $\mathcal{J}_h$ stands for the Jacobian of h.
- $\rho\left(\mathcal{J}_h(x)\right) < 1$ is not sufficient for nonstationary iteration, e.g. $\mathcal{J}_h(x^{(2k-1)}) = \begin{bmatrix} 0.5 & 10 \\ 0 & 0.5 \end{bmatrix}$, $\mathcal{J}_h(x^{(2k)}) = \begin{bmatrix} 0.5 & 0 \\ 10 & 0.5 \end{bmatrix}$, $\forall\, k = 1, \ldots$

Iterative Methods (Cont'd)

Global convergence – to stationarity

- objective is bounded below: $f(x) > -\infty$.
- sufficient function value reduction:

$$f(x^{(k)}) - f(x^{(k+1)}) \geq c\|\nabla f(x^{(k)})\|_2^2.$$

- convergence to first-order stationarity: $\lim_{k \to +\infty} \nabla f(x^{(k)}) = 0$
- if iterate sequence is bounded, subsequence convergence to a stationary point

Iterative Methods (Cont'd)

Local convergence

$$\lim_{k\to+\infty} x^{(k)} = x^*, \qquad q^{(k)} = \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|^p}.$$

- $p = 1$, $\lim_{k\to+\infty} q^{(k)} = q = 1$: local Q-sublinear convergence
- $p = 1$, $\lim_{k\to+\infty} q^{(k)} = q \in (0, 1)$: local Q-linear convergence
- $p = 1$, $\lim_{k\to+\infty} q^{(k)} = q = 0$: local Q-superlinear convergence
- $p > 1$, $\lim_{k\to+\infty} q^{(k)} = q$: local convergence with order p
  - $p = 2$, quadratic
  - $p = 3$, cubic

$$\lim_{k\to+\infty} x^{(k)} = x^*, \qquad \|x^{(k)} - x^*\| \le cr^k.$$

- $r \in (0, 1)$, local R-linear convergence rate

Iterative Methods (Cont'd)

Wost case complexity/Global convergence rate

- global linear convergence:
  get ϵ–solution after $O\left(\log \frac{1}{\epsilon}\right)$ iterations
- global sublinear convergence:

$$\lim_{k \to +\infty} f(x^{(k)}) = f^*, \qquad f(x^{(k)}) - f^* < \frac{c}{k^q}, \quad q > 0.$$

get ϵ–solution after $O\left(\frac{1}{\epsilon^{1/q}}\right)$ iterations

Iterative Methods (Cont'd)

Global convergence – iterate convergence

- Sufficient reduction:

$$f(x^{(k)}) - f(x^{(k+1)}) \geq c_1 \|x^{(k)} - x^{(k+1)}\|_2^2.$$

- Asmptotic small stepsize safe-guard:

$$\|x^{(k)} - x^{(k+1)}\|_2 \geq c_2 \|g^{(k)}\|_2, \qquad g^{(k)} \in \partial f(x^{(k)}).$$

- Łojasiewicz property: $\exists\, \theta \in [0, 1)$ such that

$$|f(x) - f(x^*)|^{\theta} \leq c_3 \|g\|_2, \qquad \forall x \in \mathcal{B}(x^*, \epsilon), \quad \forall g \in \partial f(x).$$

- iterate convergence: $\sum_{k=1}^{\infty} \|x^{(k)} - x^{(k+1)}\|_2 < +\infty.$
- local convergence rate
  - if $\theta = 0$, the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ finite termination;
  - if $\theta \in \left(0, \frac{1}{2}\right]$, there exist $c > 0$ and $Q \in [0, 1)$ such that $\|x^{(k)} - x^*\|_2 \leq c \cdot q^k$;
  - if $\theta \in \left(\frac{1}{2}, 1\right)$, there exist $c > 0$ such that $\|x^{(k)} - x^*\|_2 \leq c \cdot k^{-\frac{1-\theta}{2\theta-1}}$.

Section 2. Classical Optimization Methods

Gradient Methods

Line search

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}.$$

- exact line search: $\alpha^{(k)} = \arg\min_{\alpha \in \mathbb{R}} f(x^{(k)} + \alpha d^{(k)})$
- Armijo line search (back tracking):
  - set $c_1 \in (0,1)$, $\tau \in (0,1)$, $\alpha_0 > 0$, and $j := 0$;
  - if $f(x^{(k)}) - f(x^{(k)} + \alpha_j d^{(k)}) \geq -\alpha_j c_1 \nabla f(x^{(k)})^\top d^{(k)}$, return $\alpha^{(k)} := \alpha_j$;
  - otherwise, set $j := j + 1$ and $\alpha_j = \tau \alpha_{j-1}$.
- Wolfe condition: additional curvature condition with $c_2 \in (c_1, 1)$,

$$-\nabla f(x^{(k)} + \alpha_j d^{(k)})^\top d^{(k)} \leq -c_2 \nabla f(x^{(k)})^\top d^{(k)}.$$

Gradient Methods (Cont'd)

Gradient methods

$$d^{(k)} = -\nabla f(x^{(k)}).$$

- steepest descent: exact line search
- gradient descent with inexact line search
  global convergence and local linear rate related to $\kappa(\nabla^2 f(x^*))$.
- Barzilai-Borwein (BB) stepsize:

$$\alpha^{(k)} = \frac{{s^{(k)}}^\top y^{(k)}}{{y^{(k)}}^\top y^{(k)}}, \quad \text{or} \quad \alpha^{(k)} = \frac{{s^{(k)}}^\top s^{(k)}}{{s^{(k)}}^\top y^{(k)}}.$$

where $s^{(k)} = x^{(k)} - x^{(k-1)} = \alpha^{(k-1)} d^{(k-1)}$, $y^{(k)} = \nabla f(x^{(k)}) - \nabla f(x^{(k-1)})$,

global convergence and local linear convergence only for $f(x) = \frac{1}{2} x^\top A x + b^\top x$ with $A \succ 0$; local superlinear convergence in the case $n = 2$; global convergence if combined with nonmonotone line search.

Gradient Methods (Cont'd)

Conjugate gradient methods

$$d^{(k)} = -\nabla f(x^{(k)}) + \beta^{(k)} d^{(k-1)}.$$

- originally proposed for solving linear system
- $\alpha^{(k)}$: exact line search
- updating rules for $\beta^{(k)}$
  - Fletcher-Reeves: $\beta^{(k)} = \nabla f(x^{(k)})^\top \nabla f(x^{(k)}) \Big/ \nabla f(x^{(k-1)})^\top \nabla f(x^{(k-1)})$;
  - Polak-Ribière: $\beta^{(k)} = \nabla f(x^{(k)})^\top y^{(k)} \Big/ \nabla f(x^{(k-1)})^\top \nabla f(x^{(k-1)})$;
  - Hestenes-Stiefel: $\beta^{(k)} = \nabla f(x^{(k)})^\top y^{(k)} \Big/ {d^{(k-1)}}^\top y^{(k)}$;
  - Dai-Yuan: $\beta^{(k)} = \nabla f(x^{(k)})^\top \nabla f(x^{(k)}) \Big/ {d^{(k-1)}}^\top y^{(k)}$.
- subspace strategy:

$$x^{(k+1)} := \mathop{\arg\min}_{x - x^{(k)} \in \operatorname{span}\{\nabla f(x^{(k)}), d^{(k-1)}\}} f(x).$$

global convergence if combined with line search, local linear convergence rate not related to $\kappa(\nabla^2 f(x^*))$.

Newton Methods

Newton methods

$$d^{(k)} = -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$$

$$= \arg\min_{d \in \mathbb{R}^n} f(x^{(k)}) + \nabla f(x^{(k)})^\top (x^{(k)} + d) + \frac{1}{2}(x^{(k)} + d)\nabla^2 f(x^{(k)})(x^{(k)} + d).$$

- original ones: $\alpha^{(k)} = 1$ or exact line search

  local quadratic convergence.
- hybrid Newton method: $d^{(k)} = -\beta \nabla f(x^{(k)}) - \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$
- negative curvature descent: set $d^{(k)} = d$ if $d^\top \nabla^2 f(x^{(k)}) d < 0$.
- damped Newton method:

  $$\alpha^{(k)} = 1 \Bigg/ \left(1 + \sqrt{\nabla f(x^{(k)})^\top \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})}\right)$$

  global convergence.

Newton Methods (Cont'd)

Motivation of quasi-Newton methods

$$d^{(k)} = -{B^{(k)}}^{-1} \nabla f(x^{(k)}).$$

- $B^{(k)}$ is an approximation of $\nabla^2 f(x^{(k)})$
- easy to calculate, possess the essential characteristics of Hessian, descent direction (positive definiteness of $B^{(k)}$)
- solution: the secant equation

$$B^{(k)} s^{(k)} = y^{(k)}.$$

- SR-1 (symmetric rank-1 update) can not guarantee the positive definiteness
- rank-2 update is more favorable
  - start from $B^{(0)}$, (e.g. αI.)
  - in each iteration, add rank-2 update $B^{(k+1)} = B^{(k)} + \alpha u u^\top + v v^\top$;
  - choose $u = y^{(k)}$, $v = B^{(k)} s^{(k)}$, we arrive at – BFGS.

Newton Methods (Cont'd)

BFGS (Broyden-Fletcher-Goldfarb-Shanno)

$$B^{(k+1)}_{\text{BFGS}} = B^{(k)} + \frac{y^{(k)}{}^\top y^{(k)}}{y^{(k)}{}^\top s^{(k)}} - \frac{B^{(k)} s^{(k)} s^{(k)}{}^\top B^{(k)}}{s^{(k)}{}^\top B^{(k)} s^{(k)}}.$$

- consider the update for inverse $H^{(k)} = B^{(k)}{}^{-1}$

$$H^{(k+1)}_{\text{BFGS}} = \left(I - \frac{s^{(k)} y^{(k)}{}^\top}{s^{(k)}{}^\top y^{(k)}}\right) H^{(k)} \left(I - \frac{s^{(k)} y^{(k)}{}^\top}{s^{(k)}{}^\top y^{(k)}}\right) + \frac{s^{(k)} s^{(k)}{}^\top}{s^{(k)}{}^\top y^{(k)}}.$$

- minimum change property:

$$H^{(k+1)} = \min_{H \in \mathbb{SR}^{n\times n}} \|H - H^{(k)}\|_G, \quad \text{s. t.} \quad H y^{(k)} = s^{(k)}$$

where $\|A\|_G = \|G^{\frac{1}{2}} A G^{\frac{1}{2}}\|_{\mathrm{F}}$, $G \in \{G \mid G s^{(k)} = y^{(k)}\}$,

e.g. $G = \int_0^1 \nabla^2 f(x^{(k)} + \tau \alpha^{(k)} d^{(k)}) d\tau$

global convergence if combined with line search; local linear convergence if f is strict convex; local superlinear convergence if f is strongly convex.

Newton Methods (Cont'd)

DFT (Davidon-Fletcher-Powell)

$$B_{\mathsf{DFP}}^{(k+1)} = \left(I - \frac{s^{(k)} y^{(k)\top}}{s^{(k)\top} y^{(k)}}\right) B^{(k)} \left(I - \frac{s^{(k)} y^{(k)\top}}{s^{(k)\top} y^{(k)}}\right) + \frac{y^{(k)} y^{(k)\top}}{s^{(k)\top} y^{(k)}}.$$

- consider the update for inverse $H^{(k)} = B^{(k)^{-1}}$

$$H_{\mathsf{DFP}}^{(k+1)} = H^{(k)} + \frac{s^{(k)\top} s^{(k)}}{y^{(k)\top} s^{(k)}} - \frac{H^{(k)} y^{(k)} y^{(k)\top} H^{(k)}}{y^{(k)\top} H^{(k)} y^{(k)}}.$$

global convergence if combined with line search and local linear convergence if f is strict convex; local superlinear convergence if f is strongly convex.

The Broyden family

$$B^{(k+1)} = (1 - \phi^{(k)}) B_{\mathsf{BFGS}}^{(k+1)} + \phi^{(k)} B_{\mathsf{DFP}}^{(k+1)}, \qquad \phi^{(k)} \in [0, 1].$$

$\phi^{(k)} \in [0, 1)$ same convergence property with BFGS.

Newton Methods (Cont'd)

Limited memory quasi-Newton method

- if the storage of $B^{(k)}$ ($H^{(k)}$) is not affordable[2]
- rank-2 update provides a limited memory strategy
  - store $\mathcal{L} := \{s^{(k)}, s^{(k-1)}, \dots, s^{\max\{k-m+1,0\}}, y^{(k)}, y^{(k-1)}, \dots, y^{\max\{k-m+1,0\}}\}$;
  - $H^{(k)}$ is built up from $H^{(0)}$ by a rank$-2 \max\{m, k\}$ update
  - reduce the storage from $O(n^2)$ to $O(mn)$ at a cost of $O(mn)$ arithmetic operation
  - reduce the computational cost from $O(n^2)$ to $O(mn)$, if there is no structure
- numerically successful
  - BFGS update
  - $m = 10$

global convergence if combined with line search and local linear convergence.

[2] The difference between using $B^{(k)}$ or $H^{(k)}$ appears at the computational cost, and the storage is a whole other story.

Newton Methods (Cont'd)

The explanation of BB stepsize

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)}), \quad \text{with } \alpha^{(k)} = \frac{{s^{(k)}}^\top y^{(k)}}{{y^{(k)}}^\top y^{(k)}}, \text{ or } \alpha^{(k)} = \frac{{s^{(k)}}^\top s^{(k)}}{{s^{(k)}}^\top y^{(k)}}.$$

- Using $\frac{1}{\alpha} \cdot I$ to approximate $\nabla^2 f(x^{(k)})$

$$\alpha^{(k)} = 1 \bigg/ \arg\min_{\beta \in \mathbb{R}} \left\| \beta s^{(k)} - y^{(k)} \right\|_2^2 .$$

- Using $\alpha \cdot I$ to approximate $\nabla^2 f(x^{(k)})^{-1}$

$$\alpha^{(k)} = \arg\min_{\alpha \in \mathbb{R}} \| \alpha y^{(k)} - s^{(k)} \|_2^2.$$

Trust Region Methods

$$
\begin{aligned}
x^{(k+1)} &= x^{(k)} + s^{(k)}, \\
s^{(k)} &= \arg\min_{s \in \mathbb{R}} \quad m^{(k)}(s), \quad \text{s. t.} \quad \|s\|_2 \le \Delta^{(k)}.
\end{aligned}
$$

- $m^{(k)}(s)$ quadratic approximation of $f(x^{(k)} + s)$ at $x^{(k)}$

$$
m^{(k)}(s) := \nabla f(x^{(k)})^\top s + \frac{1}{2} s^\top B^{(k)} s.
$$

- solving subproblem
  - exactly solver: Moré-Sorensen
  - approximate: Chauchy point, dog-leg
  - inexact solver: truncated CG, 2–D subspace minimization
- the choice of $B^{(k)}$
  - $\nabla^2 f(x^{(k)})$
  - quasi-Newton update
  - other approximation of $\nabla^2 f(x^{(k)})$

Trust Region Methods (Cont'd)

- approximation ratio

$$\eta^{(k)} = \frac{\text{red}_{\text{real}}}{\text{red}_{\text{pred}}} = \frac{f(x^{(k)}) - f(x^{(k)} + s^{(k)})}{m(0) - m(s^{(k)})}.$$

- accept trial step of not:

$$x^{(k+1)} = \begin{cases} x^{(k)} + s^{(k)}, & \text{if } \eta^{(k)} > 0; \\ x^{(k)}, & \text{otherwise.} \end{cases}$$

- updating trust region radius $\Delta^{(k)}$

$$\Delta^{(k+1)} = \begin{cases} b_2 \Delta^{(k)}, & \text{if } \eta^{(k)} > c_2; \\ \Delta^{(k)}, & \text{if } c_2 \geq \eta^{(k)} > c_1; \\ b_1 \Delta^{(k)}, & \text{otherwise.} \end{cases}$$

where $0 < c_1 < c_2 < 1$, $0 < b_1 < 1 < b_2$.

global convergence only requires subproblem inexactly solved; convergence to second-order stationary point if $B^{(k)} = \nabla^2 f(x^{(k)})$ and subproblem exactly solved.

Methods for Nonlinear Least Squares

Nonlinear least squares

$$f(x) = \|F(x)\|_2^2 = \sum_{i=1}^{m} f_i^2(x)$$

- $F(x) := (f_1(x), ..., f_m(x))^\top$, each $f_i(x) : \mathbb{R}^n \mapsto \mathbb{R}$ $(i = 1, ..., m)$
- Jacobian matrix: $\mathcal{J}_F(x) = (\nabla f_1(x), ..., \nabla f_m(x))^\top$
- gradient: $\nabla f(x) = \mathcal{J}_F(x)^\top F(x)$
- Hessian: $\nabla^2 f(x) = \mathcal{J}_F(x)^\top \mathcal{J}_F(x) + \sum_{i=1}^{m} f_i(x) \nabla^2 f_i(x)$
- linear approximation: $F(x) \approx F(x^{(k)}) + \mathcal{J}_F(x^{(k)})(x - x^{(k)})$
- new approximation of Hessian: $\mathcal{J}_F(x)^\top \mathcal{J}_F(x)$
  - approximation quality depends on residuals $f_i(x)$ $(i = 1, ..., m)$
  - obtain partial Hessian information by collecting derivatives
  - positive definiteness

Methods for Nonlinear Least Squares (Cont'd)

Gauss Newton method

$$d^{(k)} = -\left(\mathcal{J}_F(x^{(k)})^\top \mathcal{J}_F(x^{(k)})\right)^{-1} \nabla f(x^{(k)})$$

- similar performance as Newton method if small residual
- similar performance as gradient method if large residual
- numerically unstable if $\mathcal{J}_F(x^{(k)})$ is singular or close to singular

Levenberg-Marquardt method

$$s^{(k)} = -\left(\mathcal{J}_F(x^{(k)})^\top \mathcal{J}_F(x^{(k)}) + \mu^{(k)} \cdot I\right)^{-1} \nabla f(x^{(k)})$$

- regularization parameter $\mu^{(k)}$ can be tuned
  - in the same manner as trust region radius
  - $\|F(x^{(k)})\|_2^t$ $(t = [1, 2])$

global convergence; quadratic local convergence rate if $\mu^{(k)} \to 0$ and zero residual at solution

Block Coordinate Descent

$$
\begin{cases}
x_1^{(k+1)} = \arg\min_{x_1 \in \mathbb{R}^{n_1}} f(x_1, x_2^{(k)}, ..., x_p^{(k)}); \\
x_2^{(k+1)} = \arg\min_{x_2 \in \mathbb{R}^{n_2}} f(x_1^{(k+1)}, x_2, x_3^{(k)} ..., x_p^{(k)}); \\
\dots\dots \\
x_p^{(k+1)} = \arg\min_{x_p \in \mathbb{R}^{n_p}} f(x_1^{(k+1)}, ..., x_{p-1}^{(k+1)}, x_p).
\end{cases}
$$

- $x = (x_1^\top, x_2^\top, ..., x_p^\top)^\top$, $x_i \in \mathbb{R}^{n_i}$ $(i = 1, ..., p)$, $n_1 + \cdots + n_p = n$
- convergence under strongly convex
- essentially Gauss-Seidel iteration: $f = \frac{1}{2}x^\top Ax - b^\top x$ with $A \succ 0$ [3]
- question: does Jacobi iteration work? linear proximal variant:

$$
x_i^{(k+1)} = \arg\min_{x_i \in \mathbb{R}^{n_i}} \nabla_{x_i} f(x^{(k)})^\top x_i + \frac{\beta^{(k)}}{2}\|x_i - x_i^{(k)}\|_2^2, \quad i = 1, ..., p.
$$

[3] This condition can be relaxed to $A \succeq 0$, $A_{ii} \succeq 0$ $(i = 1, ..., p)$.

Section 3. Global Optimization Strategies

Overview

A few strategies

- deterministic methods[4]
  - branch and bound
  - cutting plane
- undeterministic methods
  - homotopy
  - randomly multi-start
  - simulated annealing
  - genetic algorithm
  - ant colony algorithm
- approximation methods
  - SDP relaxation: $x^\top Ax = \langle A, xx^\top \rangle, \quad xx^\top \Rightarrow X \succeq 0$
- problems have nice properties
  - special quartic objective: phase retrieval, matrix completion, ...
  - problem input obeys a certain distribution
  - no nonglobal local minimizer: stationary $\Leftrightarrow$ global or saddle

[4] Combinatorial optimization can be modeled as binary variable programming. Since $x \in \{0,1\} \Leftrightarrow x^2 = x$, it can be viewed as a special nonlinear programming.

Undeterministic Methods

Homotopy (Global continuation)

- let $g(x)$ be a convex relaxation[5] of $f(x)$
- define the homotopy function: $F(x,t) : \mathbb{R}^n \times [0,1] \mapsto \mathbb{R}$
  - $F(x,0) = f(x)$;
  - $F(x,1) = g(x)$;
  - e.g. $F(x,t) = (1-t) \cdot f(x) + t \cdot g(x)$.
- main idea – solving

$$\min_{x \in \mathbb{R}^n} \quad F(x,t),$$

  with t varying from 1 to 0.
- particularly useful for problems
  - one main valley
  - surrounded by side valleys
  - side valleys occur by oscillation

[5] Usually, it means that the epigraph of $g(x)$, $\{(x,v) \mid v \geq f(x)\}$, completely contains the epigraph of $f(x)$.

Undeterministic Methods (Cont'd)

Randomly multi-start

- different with multi-start from grids or other patterns
- main procedure
  1. input: MaxL $\in \mathbb{N}$, MaxW $\in \mathbb{N}$.
  2. set CL $:= 0$, CW $:= 0$, $x^{\text{rec}} := 0$, $f^{\text{rec}} = +\infty$.
  3. certain random sampling procedure: obtain x^{sp}.
  4. certain local search procedure: obtain x^{loc}, CL := CL $+ 1$.
  5. if $f(x^{\text{loc}}) < f^{\text{rec}}$, set $x^{\text{rec}} := x^{\text{loc}}$, $f^{\text{rec}} = f(x^{\text{loc}})$, CW $:= 0$, goto 3.
  6. otherwise, CW := CW $+ 1$.
  7. if CL = MaxL or CW = MaxW, terminate and return x^{rec}.
  8. otherwise, goto 3.
- trade off between sampling phase and local search phase
- convergence
  - finding global minimizer in a compact domain
  - locally Lipschitz
  - when MaxL $\to +\infty$, probability approaches 1

Undeterministic Methods (Cont'd)

Simulated annealing

- inspiration comes from annealing in metallurgy
- main framework
  1. input: initial temperature $T \gg 1$, initial point x, $L \in \mathbb{N}$, MaxW $\in \mathbb{N}$; set CW $:= 0$, $i := 0$.
  2. if $i = L$, goto Step 7; otherwise, goto Step 3.
  3. find a new point x' by certain simple procedure.
  4. evaluate the incremental $\Delta' := f(x') - f(x)$.
  5. if $\Delta' \leq 0$, $x := x'$, CW $= 0$; else if, set $x := x'$, CW $= 0$ in probability $\exp(-\Delta'/(kT))$[6]; otherwise, CW $:=$ CW $+ 1$.
  6. if CW $\geq$ MaxW and $T = 0$, terminate; otherwise, set $i := i + 1$ and goto Step 2.
  7. decrease temperature T slowly, set $i := 0$ and goto Step 2.

[6] k takes Boltzmann constant.

References

1. Jorge Nocedal and Stephen J. Wright, *Numerical Optimization*, Springer, 2006.
2. Ya-xiang Yuan, *Computational Methods for Nonlinear Optimization (in Chinese)*, Science China Press, 2008.
3. Stephen Boyd and Lieven Vandenberghe, *Convex Optimization* Cambridge University Press, 2004.
4. R. Horst, Panos M. Pardalos and Nguyen Van Thoai, *Introduction to Global Optimization*, Kluwer Academy Publishers, 2008.
5. Hédy Attouch and Jérôme Bolte, *On the Convergence of the Proximal Algorithm for Nonsmooth Functions Involving Analytic Features*, Mathematical Programming, 116(2009), pp. 5–16.
6. Amir Beck and Luba Tetruashvili, *On the Convergence of Block Coordinate Descent Type Methods*, SIAM Journal on Optimization 23(2013), pp. 2037–2060.
7. Emmanuel J. Candes, Xiaodong Li, and Mahdi Soltanolkotabi, *Phase retrieval via wirtinger flow: Theory and algorithms*, IEEE Transactions on Information Theory, 61(2014), pp. 1985–2007.
8. Rong Ge, Jason D. Lee and Tengyu Ma, *Matrix Completion has No Spurious Local Minimum*, NIPS, 2016.

Thanks for your attention!

Email: liuxin@lsec.cc.ac.cn