Unconstrained Optimization Algorithms

Chee Wei Tan

Convex Optimization and its Applications to Computer Science

Outline

- Unconstrained minimization problems
- Gradient method
- Newton method
- Equality constrained minimization problems

Unconstrained Minimization Problems

Given $f : \mathbf{R}^n \to \mathbf{R}$ convex and twice differentiable:

$$\text{minimize} \quad f(x)$$

Optimizer x^*. Optimized value $p^* = f(x^*)$

Necessary and sufficient condition of optimality:

$$\nabla f(x^*) = 0$$

Solve a system of nonlinear equations: n equations in n variables

Iterative algorithm: computes a sequence of points $\{x^{(0)}, x^{(1)}, \ldots\}$ such that

$$\lim_{k \to \infty} f(x^{(k)}) = p^*$$

Terminate algorithm when $f(x^{(k)}) - p^* \leq \epsilon$ for a specified $\epsilon > 0$

Examples

- Least-squares: minimize

$$\|Ax - b\|_2^2 = x^T(A^T A)x - 2(A^T b)^T x + b^T b$$

Optimality condition (called normal equations for least-squares):

$$A^T A x^* = A^T b$$

- Unconstrained geometric programming: minimize

$$f(x) = \log\left(\sum_{i=1}^{m} \exp(a_i^T x + b_i)\right)$$

Optimality condition has no analytic solution:

$$\nabla f(x^*) = \frac{1}{\sum_{j=1}^m \exp(a_j^T x^* + b_j)} \sum_{i=1}^m \exp(a_i^T x^* + b_i) a_i = 0$$

- Unconstrained quadratic programming:

  Suppose C is positive definite and $A \in \mathbf{R}^{m \times n}$ with rank n: minimize

  $$\frac{1}{2}(Ax - b)^T C (Ax - b) + x^T d$$

  Optimality condition is related to equilibrium of potential energy and may not have analytic solution

Strong Convexity

f assumed to be strongly convex: there exists $m > 0$ such that

$$\nabla^2 f(x) \succeq mI$$

which also implies that there exists $M \geq m$ such that

$$\nabla^2 f(x) \preceq MI$$

Bound optimal value:

$$f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2 \leq p^* \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$$

Suboptimality condition:

$$\|\nabla f(x)\|_2 \leq (2m\epsilon)^{1/2} \Rightarrow f(x) - p^* \leq \epsilon$$

Distance between x and optimal x^*:

$$\|x - x^*\|_2 \leq \frac{2}{m}\|\nabla f(x)\|_2$$

Descent Methods

Minimizing sequence $x^{(k)}, k = 1, \ldots,$ (where $t^{(k)} > 0$)

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$$

$\Delta x^{(k)}$: search direction

$t^{(k)}$: step size

Descent methods:

$$f(x^{(k+1)}) < f(x^{(k)})$$

By convexity of f, search direction must make an acute angle with negative gradient:

$$\nabla f(x^{(k)})^T \Delta x^{(k)} < 0$$

Because otherwise, $f(x^{(k+1)}) \geq f(x^{(k)})$ since
$$f(x^{(k+1)}) \geq f(x^{(k)}) + \nabla f(x^{(k)})^T (x^{k+1}) - x^{(k)})$$

General Descent Method

GIVEN a starting point $x \in \mathbf{dom}\, f$

REPEAT

1. Determine a descent direction Δx
2. Line search: choose a step size $t > 0$
3. Update: $x := x + t\Delta x$

UNTIL stopping criterion satisfied

Line Search

- Exact line search:

$$t = \operatorname*{argmin}_{s \geq 0} f(x + s\Delta x)$$

- Backtracking line search:

  GIVEN a descent direction Δx for f at x, $\alpha \in (0, 0.5), \beta \in (0, 1)$

  $t := 1$

  WHILE $f(x) - f(x + t\Delta x) < \alpha |\nabla f(x)^T (t\Delta x)|$, $t := \beta t$

Caution: t such that $x + t\Delta x \in \mathbf{dom}\, f$

Gradient Descent Method

GIVEN a starting point $x \in \mathbf{dom}\, f$

REPEAT

1. $\Delta x := -\nabla f(x)$
2. Line search: choose a step size $t > 0$
3. Update: $x := x + t\Delta x$

UNTIL stopping criterion satisfied

Theorem: we have $f(x^{(k)}) - p^* \leq \epsilon$ after at most

$$\frac{\log((f(x^{(0)}) - p^*)/\epsilon)}{\log\left(\frac{1}{1-m/M}\right)}$$

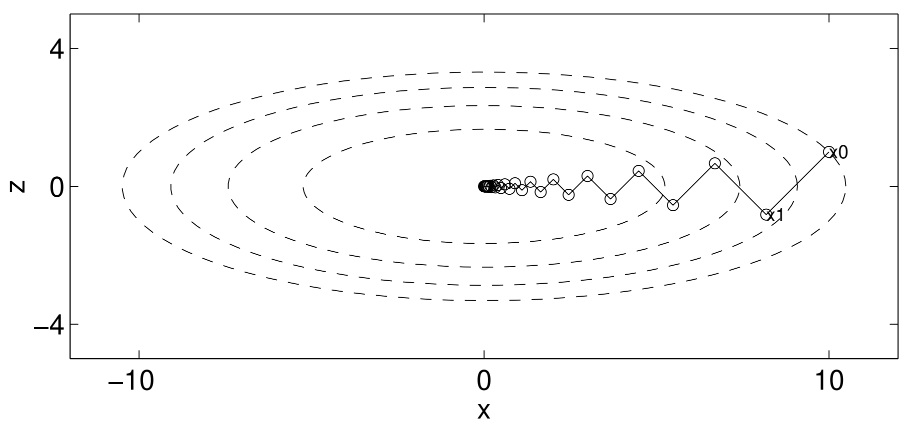
iterations of gradient method with exact line search

Example in $\mathbf{R}^2$

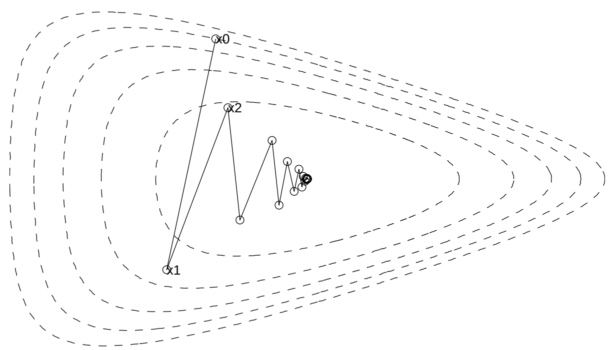
$$\text{minimize} \quad f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2), \quad x^\star = (0, 0)$$

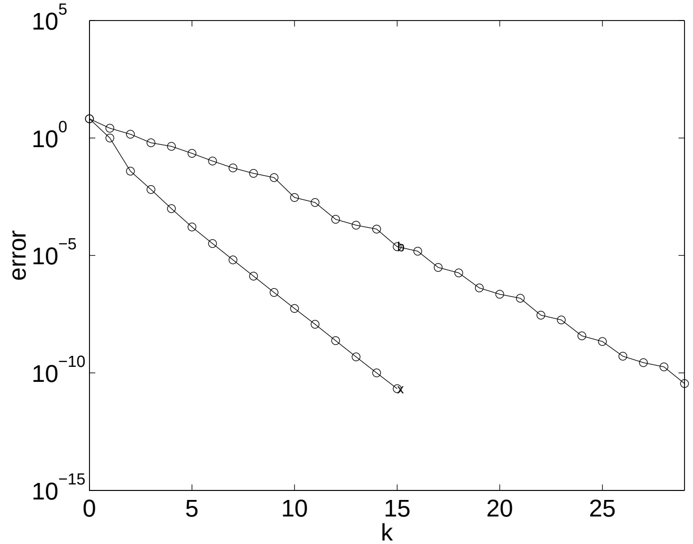
Gradient descent with exact line search:

$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1} \right)^k, \quad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1} \right)^k$$



Example in $\mathbf{R}^2$





Which error decay curve is by backtracking and which is by exact line search?

Observations

- Exhibits approximately linear convergence (error $f(x^{(k)}) - p^*$ converges to zero as a geometric series)

- Choice of α, β in backtracking line search has a noticeable but not dramatic effect on convergence speed

- Exact line search improves convergence, but not always with significant effect

- Convergence speed depends heavily on condition number of Hessian

Newton Method

Newton step:

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

Positive definiteness of $\nabla^2 f(x)$ implies that Δx_{nt} is a descent direction

Interpretation: linearize optimality condition $\nabla f(x^*) = 0$ near x,

$$\nabla f(x + v) \approx \nabla f(x) + \nabla^2 f(x) v = 0$$

Solving this linear equation in v, obtain $v = \Delta x_{nt}$. Newton step is the addition needed to x to satisfy linearized optimality condition

Main Properties

- Affine invariance: given nonsingular $T \in \mathbf{R}^{n\times n}$ and let $\bar{f}(y) = f(Tx)$. Then Newton step for $\bar{f}$ at y:
$$\Delta y_{nt} = T^{-1}\Delta x_{nt}$$
and
$$x + \Delta x_{nt} = T(y + \Delta y_{nt})$$

- Newton decrement:
$$\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2} = \left(\Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt}\right)^{1/2}$$

Let $\hat{f}$ be second order approximation of f at x. Then
$$f(x) - p^* \approx f(x) - \inf_y \hat{f}(y) = f(x) - \hat{f}(x + \Delta x_{nt}) = \frac{1}{2}\lambda(x)^2$$

Newton Method

GIVEN a starting point $x \in \mathbf{dom}\, f$ and tolerance $\epsilon > 0$

REPEAT

1. Compute Newton step and decrement: $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$ and $\lambda = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2}$

2. Stopping criterion: QUIT if $\frac{\lambda^2}{2} \leq \epsilon$

3. Line search: choose a step size $t > 0$

4. Update: $x := x + t\Delta x$

Advantages of Newton method: Fast, Robust, Scalable

Equality Constrained Problems

Solve a convex optimization with equality constraints:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

$f : \mathbf{R}^n \to \mathbf{R}$ is twice differentiable

$A \in \mathbf{R}^{p \times n}$ with rank $p < n$

Optimality condition: KKT equations with $n + p$ equations in $n + p$ variables x^*, ν^*:

$$Ax^* = b, \quad \nabla f(x^*) + A^T \nu^* = 0$$

Approach 1: Can be turned into an unconstrained optimization, after eliminating the equality constraints

Example With Analytic Solution

Convex quadratic minimization over equality constraints:

$$\begin{array}{ll} \text{minimize} & (1/2)x^T P x + q^T x + r \\ \text{subject to} & Ax = b \end{array}$$

Optimality condition:

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \nu^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

If KKT matrix is nonsingular, there is a unique optimal primal-dual pair x^*, ν^*

If KKT matrix is singular but solvable, any solution gives optimal x^*, ν^*

If KKT matrix has no solution, primal problem is unbounded below

Approach 2: Dual Solution

Dual function: $g(\nu) = -b^T\nu - f^*(-A^T\nu)$

Dual problem: maximize $\;-b^T\nu - f^*(-A^T\nu)$

Example: Let us solve the following primal problem using dual

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^n \log x_i \\ \text{subject to} & Ax = b \end{array}$$

Dual problem:

$$\text{maximize} \quad -b^T\nu + \sum_{i=1}^n \log(A^T\nu)_i$$

Recover primal variable from dual variable: $x_i(\nu) = 1/(A^T\nu)_i$

Approach 3: Direct Derivation of Newton Method

Make sure initial point is feasible and $A\Delta x_{nt} = 0$

Replace objective with second order Tayler approximation near x:

$$\begin{array}{ll} \text{minimize} & \hat{f}(x+v) = f(x) + \nabla f(x)^T v + (1/2) v^T \nabla^2 f(x) v \\ \text{subject to} & A(x+v) = b \end{array}$$

Find Newton step Δx_{nt} by solving:

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{nt} \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$

where w is associated optimal dual variable of $Ax = b$

Newton's method (Newton decrement, affine invariance, and stopping criterion) stay the same

Summary

- Iterative algorithm with descent steps for unconstrained minimization problems

- Gradient method and Newton method

- Convert equality constrained optimization into unconstrained optimization

Reading assignment: Sections 9.1-9.3, 9.5 and 10.1-10.2 of textbook.