Optimization Theory and Algorithms

Instructor: Prof. LIAO, Guocheng （廖国成）

Email: liaogch6@mail.sysu.edu.cn

School of Software Engineering
Sun Yat-sen University

Outline

- General descent method
- Line search method
- Gradient descent method
- Newton’s method

Unconstrained minimization problem

$$\begin{aligned} &\text{min} \;\; f(x) \\ &\text{s.t.} \;\;\; x \in \mathbb{R}^n \end{aligned}$$

- f is convex and twice continuously differentiable
- Assume optimal point $x^\star$ exists. Let $p^\star = f(x^\star)$ be the optimal value.

Necessary and sufficient condition of optimality:

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

- Special case: directly solve $\nabla f(x^\star) = 0$ and obtain a closed-form solution
- General case: an iterative algorithm

A sequence of points $x^{(0)}, x^{(1)}, \ldots \in \mathbf{dom}\, f$ with $f\left(x^{(k)}\right) \to p^\star$ as $k \to \infty$

Examples

Convex quadratic minimization

$$\min \ f(x) = \frac{1}{2}x^T Q x + b^T x$$

Optimality condition: $\nabla f(x^\star) = Qx^\star + b = 0$

- Case 1: unique solution. $x^\star = -Q^{-1}b$ (closed-form solution)
- Case 2: infinitely many solutions.
- Case 3: no solution, i.e., $\min \ f_0(x) = -\infty$.

Convex geometric programming

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

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

No closed-form solution. Rely on an iterative algorithm to find the solution.

Descent methods

Minimizing sequence

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}, \ k = 1,2,\ldots$$

- k is the iterative number
- $\Delta x^{(k)}$ is the step, or search direction at iteration k
- $t^{(k)}$ is the step size at iteration k
- $x^{(k+1)}$ is the output of iterative method at iteration k

Descent method:

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

- For convex f, $f\left(x^{(k+1)}\right) < f(x^{(k)})$ implies $\nabla f(x^{(k)})^T \underbrace{\boxed{\Delta x^{(k)}}}_{\text{descent direction}} < 0$

Algorithm

- Given a starting point $x \in \mathbf{dom}\ f$
- **Repeat**

1. Determine a descent direction Δx
2. *Line search*. Choose a step size t
3. Update $x \leftarrow x + t\Delta x$

- Until stopping criterion is satisfied (convergence)

Gradient descent method

Set descent direction as $\Delta x = -\nabla f(x)$ $\qquad$ $\nabla f(x)^T \Delta x = -\nabla f(x)^T \nabla f(x) < 0$

- Given a starting point $x \in \mathbf{dom}\ f$
- **Repeat**

1. Determine a descent direction: $\Delta x = -\nabla f(x)$
2. Line search. Choose a step size t.
3. Update $x \leftarrow x - t\nabla f(x)$

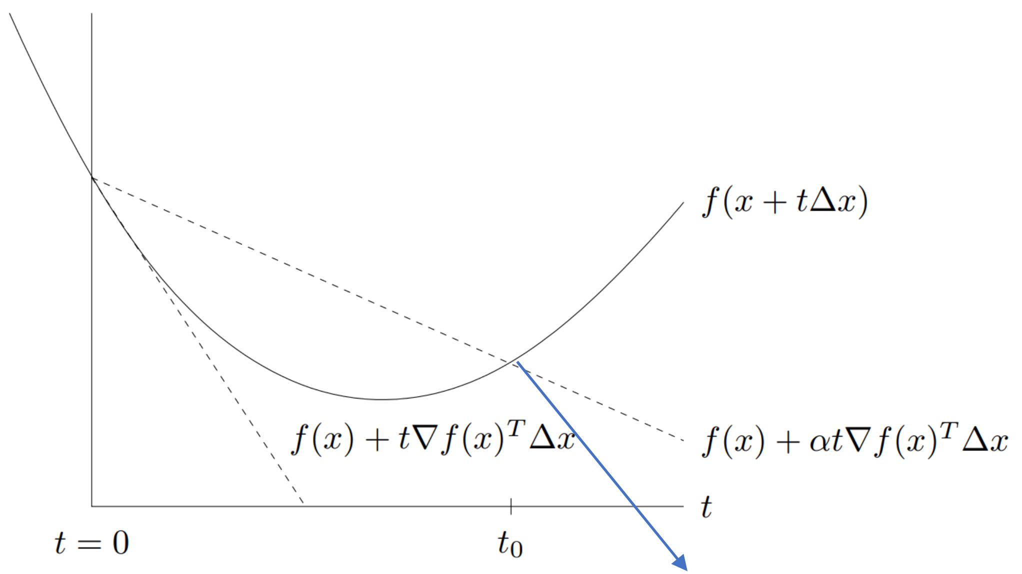
- Until stopping criterion is satisfied (convergence)

Line search

Exact line search $$t = \text{argmin}_{s \geq 0} f(x + s\Delta x)$$

Backtracking line search (inexact method)

- Given a descent direction Δx for f at $x \in \mathbf{dom}\, f$, $\alpha \in (0{,}0.5)$, $\beta \in (0{,}1)$
- $t = 1$
- While $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$:
  $t = \beta t$



Terminates when $f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$

Gradient descent method

- Given a starting point $x \in \mathbf{dom}\, f$
- **Repeat**

1. Determine a descent direction: $\Delta x = -\nabla f(x)$
2. *Line search*. Choose a step size t via *exact line search* or *backtracking line search*.
3. Update $x \leftarrow x - t\nabla f(x)$

- Until stopping criterion is satisfied (convergence)

Stopping criterion: $\|\nabla f(x)\|_2 \leq \varepsilon$ for small ε.

Interpretation of gradient descent

Quadratic approximation:

$$f(x+\Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2}\Delta x^T \boxed{\nabla^2 f(x)} \Delta x$$

Replaced by $\frac{1}{t} I$

$$\min_{\Delta x} f(x+\Delta x) \approx \boxed{f(x) + \nabla f(x)^T \Delta x + \frac{1}{2t}\Delta x^T \Delta x}$$

Set the gradient with respect to Δx as zero

$$\nabla f(x) + \frac{1}{t}\Delta x = 0$$

$$\Delta x = -t\nabla f(x)$$

The next point that minimizes quadratic approximation is $x + \Delta x = x - t\nabla f(x)$

Newton's method

Minimize quadratic approximation:

$$\min_{\Delta x} f(x+\Delta x) \approx \boxed{f(x) + \nabla f(x)^T \Delta x + \frac{1}{2}\Delta x^T \nabla^2 f(x) \Delta x}$$

Set the gradient with respect to Δx as zero

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

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

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

Newton's method

Interpretation: solution of linearized optimality condition

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \mathbb{R}^n \end{array}$$

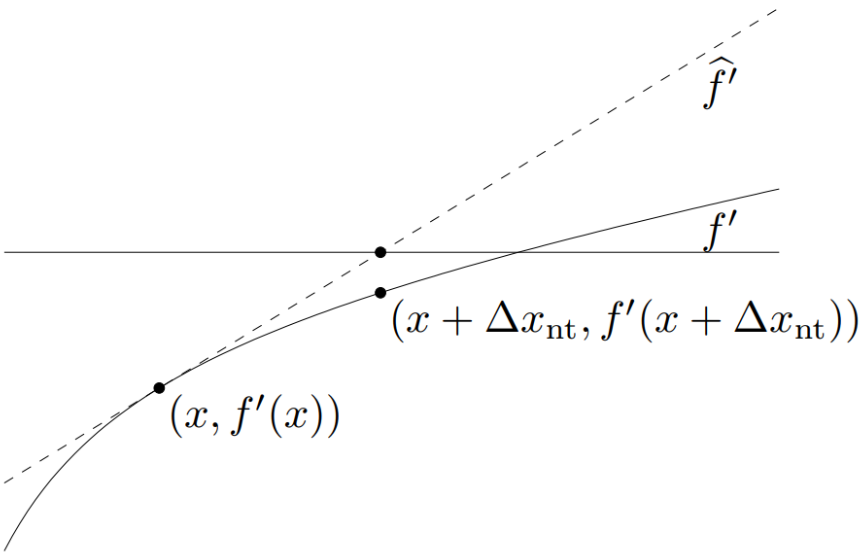
Optimality condition:

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

Taylor approximation

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

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



Newton's method

- Given a starting point $x \in \mathbf{dom}\, f$
- **Repeat**

1. Compute the *Newton step* $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$
2. *Line search*. Choose a step size t via *backtracking line search*.
3. Update $x \leftarrow x + t\Delta x_{nt}$

Until stopping criterion is satisfied (convergence)

Rely on Newton decrement

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

- Gives an estimate of $f(x) - p^\star$, using quadratic approximation $\hat{f}$

$$f(x) - \left(f(x) + \nabla f(x)^T \Delta x_{nt} + \frac{1}{2} \Delta x_{nt}{}^T \nabla^2 f(x) \Delta x_{nt}\right) = \frac{1}{2} \lambda(x)^2$$

$$f(x) - p^\star \approx \frac{1}{2} \lambda(x)^2$$

Newton's method

- Given a starting point $x \in \mathbf{dom}\ f$
- **Repeat**

1. Compute the *Newton step* and decrement

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

2. Stopping criterion: if $\lambda^2/2 \le \varepsilon$, break
3. *Line search*: choose a step size t via *backtracking line search*.
4. Update: $x \leftarrow x + t\Delta x_{nt}$

Drawback: high complexity of computing Hessian matrix
Solution: Quasi-Newton method