Convex Optimization

Lecture 11 - Unconstrained Optimization

Instructor: Yuanzhang Xiao

University of Hawaii at Manoa

Fall 2017

Today's Lecture

1. Basic Concepts
2. Descent Methods

Outline

Unconstrained Optimization Problems

unconstrained minimization problem:

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

- $f(x)$ convex, twice continuously differentiable ($\Rightarrow$ $\mathbf{dom} f$ open)
- optimal value $p^\star = f(x^\star) = \inf_x f(x)$ attained and finite

optimality condition:

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

- minimization equivalent to solving n equations

Unconstrained Optimization Algorithms

unconstrained minimization algorithm:

- produce a sequence of points $x^{(k)} \in \mathbf{dom} f, \; k = 0, 1, \dots$

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

starting point $x^{(0)}$:

- $x^{(0)} \in \mathbf{dom} f$
- sublevel set $S = \left\{ x \mid f(x) \le f(x^{(0)}) \right\}$ closed

requirements on starting points will be relaxed later

More on Initial Points

$f(x^{(0)})$-sublevel set closed: hard to check

sufficient conditions for the closedness of $f(x^{(0)})$-sublevel set:

- $\mathbf{dom} f = \mathbb{R}^n$
- $f(x) \to \infty$ as $x \to \mathbf{bd}\ \mathbf{dom} f$

examples:

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

Strong Convexity

we assume that the objective function is strongly convex on S:

$$\nabla^2 f(x) \succeq mI, \ \forall x \in S,$$

for some $m > 0$

for any $x, y \in S$, Taylor expansion:

$$f(y) = f(x) + \nabla f(x)^T (y - x) + (y - x)^T \nabla^2 f(z)(y - x)$$

for some z on the line segment between x and y

therefore, for any $x, y \in S$, we have:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2$$

Implications of Strong Convexity

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2$$

minimize the right-hand side with respect to y:

$$y^\star(x) = x - \frac{1}{m} \nabla f(x)$$

we have

$$f(y) \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2$$

stopping criterion:

$$p^\star = f(x^\star) \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2$$

Implications of Strong Convexity

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|_2^2$$

setting $y = x^\star$, we have:

$$\begin{aligned} p^\star = f(x^\star) &\geq f(x) + \nabla f(x)^T(x^\star - x) + \frac{m}{2}\|x^\star - x\|_2^2 \\ &\geq f(x) - \|\nabla f(x)\|_2 \|x^\star - x\|_2 + \frac{m}{2}\|x^\star - x\|_2^2 \end{aligned}$$

since $p^\star \leq f(x)$, we have

$$-\|\nabla f(x)\|_2 \|x^\star - x\|_2 + \frac{m}{2}\|x^\star - x\|_2^2 \leq 0,$$

distance between x and $x^\star$:

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

A Few Comments

$$p^\star = f(x^\star) \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2$$

stopping criterion:

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

along with $\|x^\star - x\|_2 \leq \frac{2}{m} \|\nabla f(x)\|_2$, we know that

- x close to the optimal solution when $\nabla f(x)$ close to 0

in practice, m is unknown

- conceptually useful
- special functions: convergence analysis independent of m

Outline

1. Basic Concepts

2. Descent Methods

Descent Methods

an algorithm that produces a sequence $x^{(k)}$, $k = 0, 1, \ldots$:

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)} \text{ with } f(x^{(k+1)}) < f(x^{(k)})$$

- $\Delta x^{(k)}$: step, or search direction
- $t^{(k)} > 0$: step size

from convexity of f, we have

$$\begin{aligned} f(x^{(k+1)}) &\geq f(x^{(k)}) + \nabla f(x^{(k)})^T \left(x^{(k+1)} - x^{(k)} \right) \\ &= f(x^{(k)}) + t^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)} \end{aligned}$$

a descent direction at $x^{(k)}$ must satisfy:

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

General Procedure of Descent Methods

a general descent method:

- a starting point $x \in \textbf{dom} f$
- repeat the following steps until stopping criterion is satisfied
  1. determine a descent direction Δx
  2. line search: choose a step size $t > 0$
  3. update $x := x + t\Delta x$

different descent directions $\Rightarrow$ different descent methods

line search crucial to ensure

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

Line Search

exact line search:

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

used when the above minimization can be solved efficiently

backtracking line search:

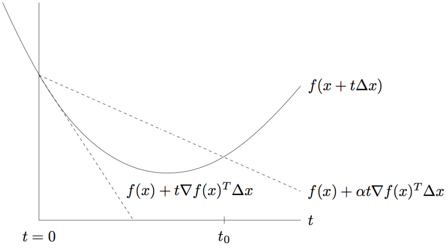
- given Δx, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$
- initial $t = 1$
- repeat $t := \beta t$ until

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

most commonly used

Illustration of Backtracking Line Search

graphical illustration:



impacts of parameters α and β:

- α large: f decreases fast, line search slow
- β large: less crude line search, line search slow

Gradient Descent Method

gradient descent method:

- a starting point $x \in \textbf{dom} f$
- repeat the following steps until $\|\nabla f(x)\|_2 \leq \eta$
  1. $\Delta x = -\nabla f(x)$
  2. exact or backtracking line search
  3. update $x := x + t\Delta x$

convergence result for strongly convex functions:

$$f(x^{(k)}) - p^\star \leq c^k \left(f(x^{(0)}) - p^\star \right)$$

where $c \in (0, 1)$ depends on m, $x^{(0)}$, line search method

linear convergence rate (slow)

Example - A Quadratic Problem in $\mathbb{R}^2$

quadratic objective function:

$$f(x) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)$$

with $\gamma > 0$

starting at $x^{(0)} = (\gamma, 1)$ and using exact line search, we have

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

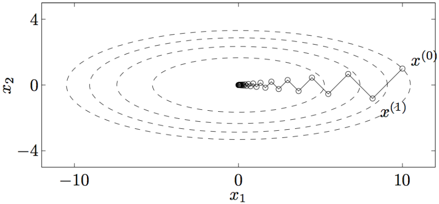
and

$$f(x^{(k)}) = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f(x^{(0)})$$

slow convergence when $\gamma << 1$ or $\gamma >> 1$

Example - A Quadratic Problem in $\mathbb{R}^2$

illustration when $\gamma = 10$:

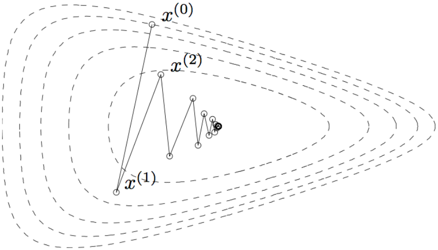


Example - A Nonquadratic Problem in $\mathbb{R}^2$

objective function:

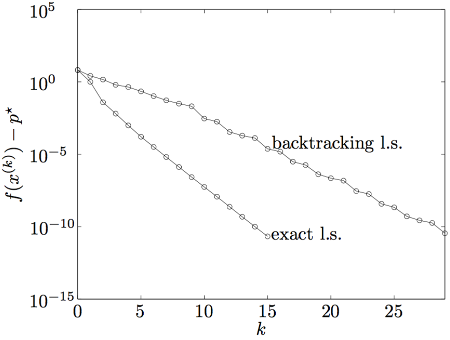
$$f(x) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$

gradient descent method with backtracking line search



Example - A Nonquadratic Problem in $\mathbb{R}^2$

backtracking versus exact line search



Steepest Descent Method

first-order Taylor approximation:

$$f(x+v) \approx f(x) + \nabla f(x)^T v$$

normalized steepest descent direction:

$$\Delta x_{\text{nsd}} = \text{argmin} \left\{ \nabla f(x)^T v \mid \|v\| = 1 \right\}$$

(unnormalized) steepest descent direction: Δx_{sd}

linear convergence rate (slow)

Steepest Descent Method

steepest descent for Euclidean norm:

$$\Delta x_{\text{sd}} = -\nabla f(x)$$

(gradient descent)

steepest descent for quadratic norm $\|z\|_P = \sqrt{z^T P z}$:

$$\Delta x_{\text{sd}} = -P^{-1}\nabla f(x)$$

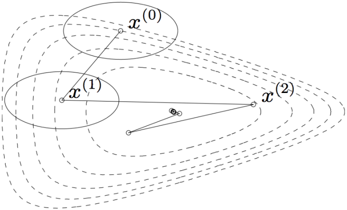
steepest descent for ℓ_1-norm:

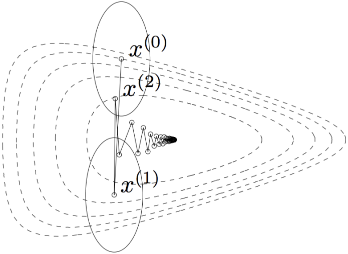
$$\Delta x_{\text{sd}} = -\frac{\partial f(x)}{\partial x_i} e_i, \text{ where } \left|\frac{\partial f(x)}{\partial x_i}\right| = \|\nabla f(x)\|_\infty$$

(may simplify line search)

Choice of Norm

Nonquadratic example using steepest descent with quadratic norm:

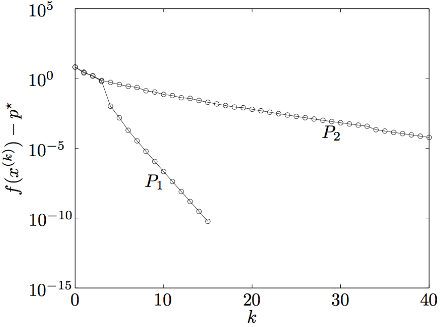




left: $P_1 = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}$; right: $P_2 = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}$

Choice of Norm

Nonquadratic example using steepest descent with quadratic norm:



choice of norm has large impact on steepest descent methods

Newton's Method

second-order Taylor approximation:

$$f(x+v) \approx \hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

Newton step:

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

steepest descent direction in Hessian norm $\|\cdot\|_{\nabla^2 f(x)}$

Newton Decrement

Newton decrement at x:

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

interpretation:

$$\frac{1}{2}\lambda(x)^2 = f(x) - \inf_v \hat{f}(x+v) = f(x) - \hat{f}(x + \Delta x_{\mathrm{nt}})$$

$\frac{1}{2}\lambda(x)^2$ is an estimate of $f(x) - p^\star$

Newton's Method

Newton's method:

- a starting point $x \in \textbf{dom} f$
- repeat the following steps
  1. Newton step and decrement
  $$\Delta x_{\text{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x), \quad \lambda(x)^2 = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$
  2. quit if $\frac{\lambda^2}{2} \leq \epsilon$
  3. exact or backtracking line search
  4. update $x := x + t\Delta x$

minor difference:
check stopping criterion after computing the search direction

Convergence Results of Newton's Method

assume that

- f is strongly convex with $\nabla^2 f(x) \succeq mI$
- $\nabla^2 f(x)$ is Lipschitz continuous with constant L

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \leq L \left\|x - y\right\|_2$$

convergence result: there exists $\eta \in (0, m^2/L)$ and $\gamma > 0$ such that

- when $\left\|\nabla f(x^{(k)})\right\|_2 \geq \eta$, we have

$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$$

- when $\left\|\nabla f(x^{(k)})\right\|_2 < \eta$, we have $t^{(k)} = 1$ and

$$\frac{L}{2m^2} \left\|\nabla f(x^{(k+1)})\right\|_2 \leq \left(\frac{L}{2m^2} \left\|\nabla f(x^{(k)})\right\|_2\right)^2$$

Convergence Results of Newton's Method

damped Newton phase ($\left\|\nabla f(x^{(k)})\right\|_2 \geq \eta$)

- most iterations require backtracking line search
- function value decreases by at least γ
- this phase ends after at most $\frac{f(x^{(0)})-p^\star}{\gamma}$ iterations

quadratically convergent phase ($\left\|\nabla f(x^{(k)})\right\|_2 < \eta$)

- no backtracking line search $t^{(k)} = 1$
- norm of gradient $\|\nabla f(x)\|_2$ converges to zero quadratically:

$$\frac{L}{2m^2}\left\|\nabla f(x^{(\ell)})\right\|_2 \leq \left(\frac{L}{2m^2}\left\|\nabla f(x^{(k)})\right\|_2\right)^{2^{\ell-k}}, \ \forall \ell \geq k$$

Convergence Results of Newton's Method

total number of iterations bounded by

$$\frac{f(x^{(0)}) - p^{\star}}{\gamma} + \log_2 \log_2(\epsilon_0/\epsilon)$$

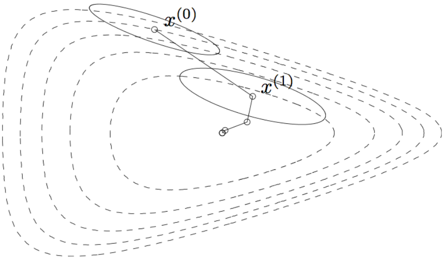
- γ and ϵ_0 are constants that depend on m, L, $x^{(0)}$
- the second term is almost constant (≈ 6)

Revisit The Nonquadratic Example in $\mathbb{R}^2$

objective function:

$$f(x) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$

Newton's method with backtracking line search



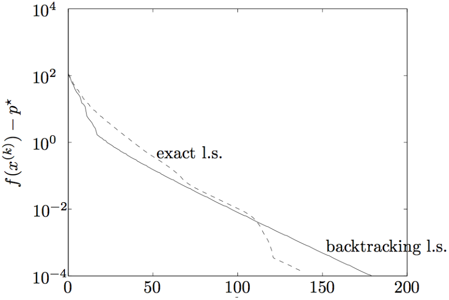
Scalability of Newton's Method

objective function in $\mathbb{R}^{100}$:

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

with $m = 500$ and $n = 100$

gradient descent:



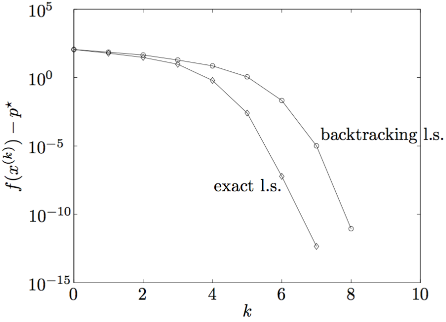
Scalability of Newton's Method

objective function in $\mathbb{R}^{100}$:

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

with $m = 500$ and $n = 100$

Newton's method:



Scalability of Newton's Method

objective function in $\mathbb{R}^{10000}$:

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

with $m = 500$ and $n = 10000$

Newton's method: