08. Unconstrained Minimization

By Yang Lin[1] (2023 秋季, @NJU)

Contents

[1] Institute: Nanjing University. Email: linyang@nju.edu.cn.

1 Unconstrained minimization

$$\texttt{maximize} \quad f(x)$$

1. f convex, twice continuously differentiable (hence **dom** f open)
2. We assume optimal value $p^* = \inf_x f(x)$ is attained (and finite)

Unconstrained Minimization Methods

1. produce sequence of points $x^{(k)} \in \textbf{dom}\, f,\ \ k = 0, 1, \ldots$ with

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

2. can be interpreted as iterative methods for solving optimality condition

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

1 Unconstrained minimization

Initial point and sublevel set

algorithms in this chapter require a starting point $x^{(0)}$ such that

1. $x^{(0)} \in \mathbf{dom}\, f$
2. Sublevel set $S = \{x | f(x) \leq f(x^{(0)})\}$ is closed

2nd condition is hard to verify, except when all sublevel sets are closed:

1. equivalent to condition that $\mathbf{epi}\, f$ is closed
2. true if $\mathbf{dom}\, f = \mathbf{R}^n$
3. true if $f(x) \to \infty$ as $x \to \mathbf{bd}\ \mathbf{dom}\, f$

examples of differentiable functions with closed sublevel sets:

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

1 Unconstrained minimization

Strong convexity and implications

f is strongly convex on S if there exists an $m > 0$ such that

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

Implications

1. for $x, y \in S$

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

hence, S is bounded

2. for $p^* > -\infty$, and for $x \in S$

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

useful as stopping criterion (if you know m)

2 Descent methods

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

1. other notations: $x^+ = x + t\Delta x$, $x := x + t\Delta x$
2. Δx is the step, or search direction; t is the step size, or step length
3. from convexity, $f(x^+) < f(x)$ implies $\nabla f(x)^T \Delta x < 0$ (i.e., Δx is a descent direction)

General descent method.

given a starting point $x \in \textbf{dom}\, f$.
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 is satisfied.

---

2 Descent methods

Line search types

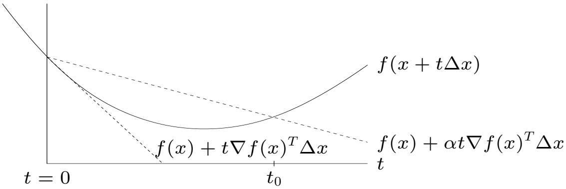
exact line search: $t = \arg\min_{t>0} f(x + t\delta x)$

backtracking line search (with parameters $\alpha \in (0, 1/2), \beta \in (0, 1)$)

1. starting at $t = 1$, repeat $t := \beta t$ until

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

2. graphical interpretation: backtrack until $t \leq t_0$



2 Descent methods

general descent method with $\Delta x = -\nabla f(x)$

---

given a starting point $x \in \mathbf{dom}\, f$.
repeat
1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose step size t via exact or backtracking line search.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

---

1. stopping criterion usually of the form $\|\nabla f(x)\|_2 \leq \epsilon$
2. convergence result: for strongly convex f,

$$f(x^{(k)}) - p^* \leq c^k (f(x^{(0)} - p^*)$$

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

3. very simple, but often very slow; rarely used in practice

2 Descent methods

quadratic problem in $\mathbf{R}^2$

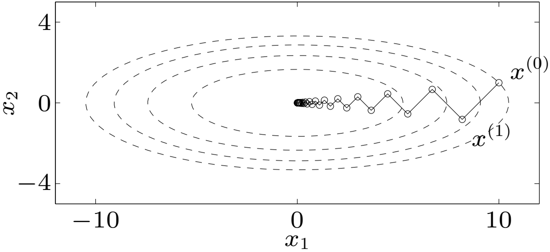
$$f(x) = (1/2)(x_1^2 + \gamma x_2^2) \quad (\gamma > 0)$$

with exact line search, starting at $x^{(0)} = (\gamma, 1)$:

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

very slow if $\gamma \gg 1$ or $\gamma \ll 1$

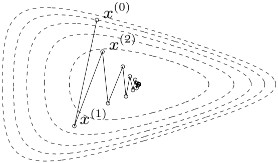
example for $\gamma = 10$:



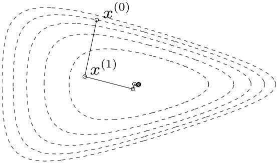
2 Descent methods

nonquadratic example

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



backtracking line search

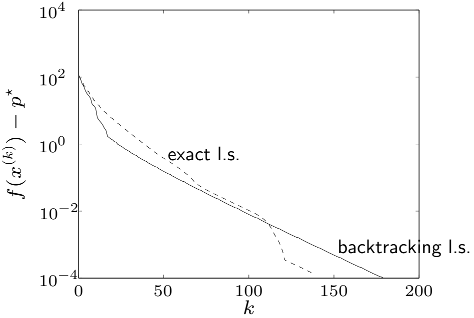


exact line search

2 Descent methods

a problem in $\mathbf{R}^{100}$

$$f(x) = c^T x - \sum_{i=1}^{500} \log(b_i - a_i^T x)$$



'linear' convergence, i.e., a straight line on a semilog plot

3 Steepest descent method

normalized steepest descent direction (at x, for norm $\|\cdot\|$):

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

interpretation: for small v, $f(x+v) \approx f(x) + \nabla f(x)^T v$

direction Δx_{nsd} is unit-norm step with most negative directional derivative

(unnormalized) steepest descent direction

$$\Delta x_{\text{sd}} = \|\nabla f(x)^T\|_* \Delta x_{\text{nsd}}$$

satisfies $\nabla f(x)^T \Delta x_{\text{sd}} = -\|\nabla f(x)^T\|_*^2$

steepest descent method

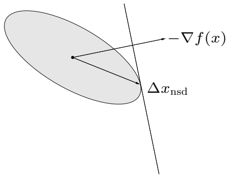
1. general descent method with $\Delta x = \Delta x_{\text{sd}}$
2. convergence properties similar to gradient descent

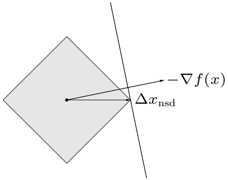
3 Steepest descent method

Examples

1. Euclidean norm: $\Delta x_{\rm sd} = -\nabla f(x)$
2. Quadratic norm $\|x\|_P = (x^T P x)^{1/2} (P \in \mathbf{S}^n_{++})$: $\Delta x_{\rm sd} = -P^{-1} \nabla f(x)$
3. l_1-norm: $\Delta x_{\rm sd} = -(\partial f(x)/\partial x_i) e_i$, where $|\partial f(x)/\partial x_i| = \|\nabla f(x)\|_\infty$

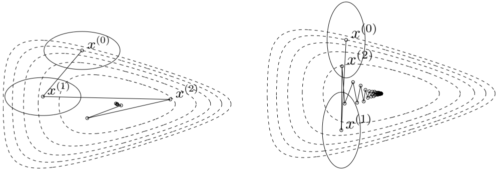
unit balls and normalized steepest descent directions for a quadratic norm and the l_1-norm:





3 Steepest descent method

choice of norm for steepest descent



1. Steepest descent with backtracking line search for two quadratic norms

2. Ellipses show $\{x | \|x - x^{(k)}\|_P = 1\}$

3.Equivalent interpretation of steepest descent with quadratic norm $\| \cdot \|_P$: gradient descent after change of variables $\bar{x} = P^{1/2}x$

Shows choice of P has strong effect on speed of convergence

4 Newton's Method

Newton step

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

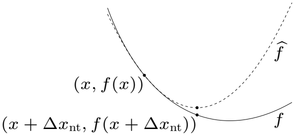
interpretations

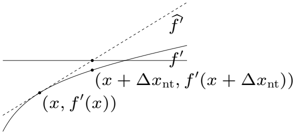
1. $x + \Delta x_{\mathrm{nt}}$ minimizes second order approximation

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

2. $x + \Delta x_{\mathrm{nt}}$ solves linearized optimality condition

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

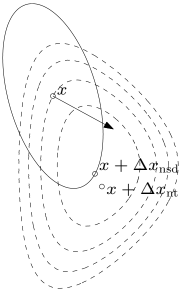




4 Newton's Method

Δx_{nt} is steepest descent direction at x in local Hessian norm

$$\|u\|_{\nabla^2 f(x)} = \left(uT\nabla^2 f(x)u\right)^{1/2}$$



dashed lines are contour lines of f; ellipse is $\{x + v | v^T \nabla^2 f(x) v = 1\}$
arrow shows $-\nabla f(x)$

4 Newton's Method

Newton decrement

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

a measure of the proximity of x to x^*

properties

1. gives an estimate of $f(x) - p^*$, using quadratic approximation $\hat{f}$:

$$f(x) - \inf_y \hat{f}(y) = \frac{1}{2}\lambda(x)^2$$

2. equal to the norm of the Newton step in the quadratic Hessian norm

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

3. directional derivative in the Newton direction: $\nabla f(x)^T \delta x_{\text{nt}} = -\lambda(x)^2$

4. Affine invariant (unlike $\nabla f(x)\|^2$)

4 Newton's Method

---

given a starting point $x \in \mathbf{dom}\, f$, tolerance $\epsilon > 0$.
repeat

1. *Compute the Newton step and decrement.*
   $\Delta x_{\mathrm{nt}} := -\nabla^2 f(x)^{-1} \nabla f(x); \quad \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x).$
2. *Stopping criterion.* **quit** if $\lambda^2/2 \le \epsilon$.
3. *Line search.* Choose step size t by backtracking line search.
4. *Update.* $x := x + t\Delta x_{\mathrm{nt}}$.

---

affine invariant, i.e., independent of linear changes of coordinates: Newton iterates for $\tilde{f}(y) = f(Ty)$ with starting point $y^{(0)} = T^{-1}x^{(0)}$ are

$$y^{(k)} = T^{-1}x^{(k)}$$

4 Newton's Method

Classical convergence analysis

Assumptions

1. f strongly convex on S with constant m
2. $\nabla^2 f$ is Lipschitz continuous on S, with constant $L > 0$:

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

(L measures how well f can be approximated by a quadratic function)

outline: there exist constants $\eta \in (0, m^2/L)$, $\gamma > 0$ such that

1. if $\|\nabla f(x)\|_2 \geq \eta$, then $f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$
2. if $\|\nabla f(x)\nabla_2 < \eta$, then

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

4 Newton's Method

damped Newton phase ($\|\nabla f(x)\nabla_2 \geq \eta$)

1. most iterations require backtracking steps

2. function value decreases by at least γ

3. if $p^* > -\infty$, this phase ends after at most $(f(x^{(0)}) - p^*)/\gamma$ iterations

quadratically convergent phase ($\|\nabla f(x)\nabla_2 < \eta$)

1. all iterations use step size $t = 1$

2. $\|\nabla f(x)\nabla_2$ converges to zero quadratically: if $\|\nabla f(x)\nabla_2 < \eta$, then

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

4 Newton's Method

Conclusion: number of iterations until $f(x) - p^* \leq \epsilon$ is bounded above by

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

1. γ, ϵ_0 are constants that depend on $m, L, x^{(0)}$
2. second term is small (of the order of 6) and almost constant for practical purposes
3. In practice, constants m, L (hence , ϵ_0) are usually unknown
4. provides qualitative insight in convergence properties (i.e., explains two algorithm phases)

5 Self-concordance

shortcomings of classical convergence analysis

1. Depends on unknown constants $(m, L, \ldots)$
2. Bound is not affinely invariant, although Newton' s method is

convergence analysis via self-concordance (Nesterov and Nemirovski)

1. does not depend on any unknown constants
2. gives affine-invariant bound
3. applies tospecial classof convexfunctions('self-concordant' functions)
4. developed to analyze polynomial-time interior-point methods for convex optimization

5 Self-concordance

definition

1. convex $f : \mathbf{R} \to \mathbf{R}$ is self-concordant if $|f'''(x)| \le 2f''(x)^{3/2}$ for all $x \in \mathbf{dom}\ f$
2. $f : \mathbf{R}^n \to \mathbf{R}$ is self-concordant if $g(t) = f(x + tv)$ is self-concordant for all $x \in \mathbf{dom}\ f$, $v \in \mathbf{R}^n$

examples on R

1. linear and quadratic functions
2. negative logarithm $f(x) = -\log x$
3. negative entropy plus negative logarithm: $f(x) = x \log x - \log x$

affine invariance: if $f : \mathbf{R} \to \mathbf{R}$ is s.c., then $\tilde{f}(y) = f(ay + b)$ is s.c.:

$$\tilde{f}'''(y) = a^3 f'''(x)^{3/2}(ay + b), \quad \tilde{f}''(y) = a^2 f''(x)^{3/2}(ay + b)$$

5 Self-concordance

properties

1. preserved under positive scaling $\alpha \geq 1$, and sum
2. preserved under composition with affine function
3. if g is convex with $\mathbf{dom}\ g = \mathbf{R}^{++}$ and $|g'''(x)| \leq 3g''(x)/x$ then

$$f(x) = \log(-g(x)) - \log x$$

is self-concordant

examples: properties can be used to show that the following are s.c.

1. $f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x)$ on $\{x | a_i^T x < b_i, i = 1, \ldots, m\}$
2. $f(X) = -\log\det X$ on $\mathbf{S}_{++}^n$
3. $f(x) = -\log(y^2 - x^T x)$ on $\{(x, y) | \|x\|_2 < y\}$

5 Self-concordance

Convergence analysis for self-concordant functions

summary: there exist constants $\eta \in (0, 1/4]$, $\gamma > 0$ such that

1. if $\lambda(x) > \eta$, then $f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$

1. if $\lambda(x) \leq \eta$, then $2\lambda(x^{(k+1)}) \leq \left(2\lambda(x^{(k)})\right)^2$ (and γ only depend on backtracking parameters α, β)

complexity bound: number of Newton iterations bounded by

$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2 (1/\epsilon)$$

for $\alpha = 0.1$, $\beta = 0.8$, $\epsilon == 10^{-10}$, bound evaluates to $375(f(x^{(0)}) - p^*) + 6$