Gradient method

Acknowledgement: this slides is based on Prof. Lieven Vandenberghes lecture notes

- gradient method, first-order methods
- quadratic bounds on convex functions
- analysis of gradient method

Algorithms will be covered in this course

first-order methods

- gradient method, line search
- subgradient, proximal gradient methods
- accelerated (proximal) gradient methods

decomposition and splitting

- first-order methods and dual reformulations
- alternating minimization methods

interior-point methods

- conic optimization
- primal-dual methods for symmetric cones

semi-smooth Newton methods

给定x^k，先通过某种算法选取d^k，之后确定正数α_k，则下一步迭代点

$$x^{k+1} = x^k + \alpha_k d^k, \tag{1}$$

其中d^k是一个**下降方向**，即$(d^k)^{\mathrm{T}}\nabla f(x^k) < 0$。$\alpha_k$为相应的**步长**.

- 精确线搜索:
$$\alpha_k = \operatorname*{argmin}_{\alpha>0} f(x^k + \alpha d^k)$$
- 固定步长
- 回退法：找到第一个满足Armijo 准则的点．选取
$$\alpha_k = \gamma^{j_0}\hat{\alpha},$$
其中
$$j_0 = \min\{j = 0, 1, \cdots \mid f(x^k + \gamma^j \hat{\alpha} d^k) \le f(x^k) + c_1\gamma^j\hat{\alpha}\nabla f(x^k)^{\mathrm{T}} d^k\},$$
参数$\gamma \in (0, 1)$为一个给定的实数．

Gradient method

To minimize a convex function differentiable function f: choose $x^{(0)}$ and repeat

$$x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k-1)}), \quad k = 1, 2, \ldots$$

Step size rules

- Fixed: t_k constant
- Backtracking line search
- Exact line search: minimize $f(x - t\nabla f(x))$ over t

Advantages of gradient method

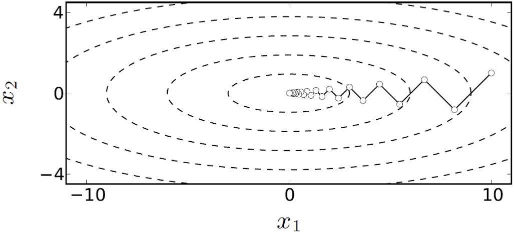
- Every iteration is inexpensive
- Does not require second derivatives

Quadratic example

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

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

$$\frac{||x^{(k)} - x^*||_2}{||x^{(0)} - x^*||_2} = (\frac{\gamma - 1}{\gamma + 1})^k$$



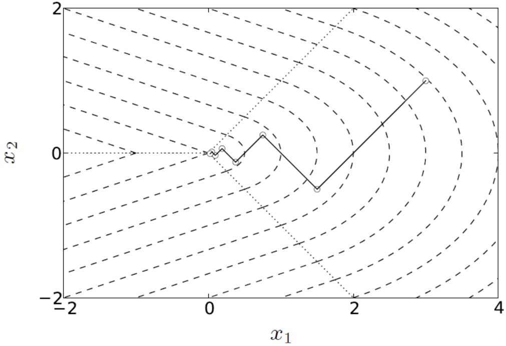
Disadvantages of gradient method

- Gradient method is often slow
- Very dependent on scaling

Nondifferentiable example

$$f(x) = \sqrt{x_1^2 + \gamma x_2^2} \quad (|x_2| \le x_1), \qquad f(x) = \frac{x_1 + \gamma |x_2|}{\sqrt{1+\gamma}} \quad (|x_2| > x_1)$$

with exact line search, $x^{(0)} = (\gamma, 1)$, converges to non-optimal point



gradient method does not handle nondifferential problems

First-order methods

address one or both disadvantages of the gradient method

methods with improved convergence

- quasi-Newton methods
- conjugate gradient method
- accelerated gradient method

methods for nondifferentiable or constrained problems

- subgradient methods
- proximal gradient method
- smoothing methods
- cutting-plane methods

Outline

- gradient method, first-order methods
- **quadratic bounds on convex functions**
- analysis of gradient method

Convex function

f is convex if **dom** f is a convex set and Jensen's inequality holds:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) \quad \forall x, y \in \mathbf{dom}\, f$$

First-order condition

for (continuously) differentiable f, Jensen's inequality can be replaced with

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) \quad \forall x, y \in \mathbf{dom}\, f$$

Second-order condition

for twice differentiable f, Jensen's inequality can be replaced with

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathbf{dom}\, f$$

Strictly convex function

f is strictly convex if **dom** f is convex set and

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta)f(y) \quad \forall x, y \in \mathbf{dom}\, f, x \neq y, \theta \in (0,1)$$

hence, if a minimizer of f exists, it is unique

First-order condition
for differentiable f, Jensen's inequality can be replaced with

$$f(y) > f(x) + \nabla f(x)^\top (y-x) \quad \forall x, y \in \mathbf{dom}\, f, x \neq y$$

Second-order condition
note that $\nabla^2 f(x) \succ 0$ is not necessary for strict convexity($cf.,f(x) = x^4$)

Monotonicity of gradient

differentiable f is convex if and only if **dom** f is convex and

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq 0 \quad \forall x, y \in \mathbf{dom}\, f$$

i.e., $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$ is a *monotone* mapping

differentiable f is strictly convex if and only if **dom** f is convex and

$$(\nabla f(x) - \nabla f(y))^\top (x - y) > 0 \quad \forall x, y \in \mathbf{dom}\, f, x \neq y$$

i.e., $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$ is a *strictly monotone* mapping

Proof.

- if f is differentiable and convex, then

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x), \quad f(x) \geq f(y) + \nabla f(y)^\top (x - y)$$

combining the inequalities gives $(\nabla f(x) - \nabla f(y))^\top (x - y) \geq 0$

- if ∇f is monotone, then $g'(t) \geq g'(0)$ for $t \geq 0$ and $t \in \textbf{dom}\, g$, where

$$g(t) = f(x + t(y - x)), \quad g'(t) = \nabla f(x + t(y - x))^\top (y - x)$$

hence,

$$\begin{aligned} f(y) = g(1) = g(0) + \int_0^1 g'(t)dt &\geq g(0) + g'(0) \\ &= f(x) + \nabla f(x)^\top (y - x) \end{aligned}$$

□

Lipschitz continuous gradient

gradient of f is Lipschitz continuous with parameter $L > 0$ if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y \in \mathbf{dom}\, f$$

- Note that the definition does not assume convexity of f
- We will see that for convex f with $\mathbf{dom}\, f = \mathbf{R}^n$, this is equivalent to

$$\frac{L}{2}x^\top x - f(x) \quad is \quad convex$$

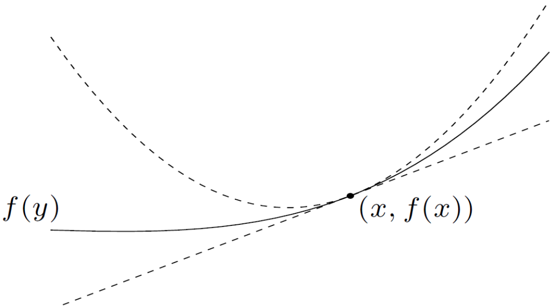
(i.e., if f is twice differentiable, $\nabla^2 f(x) \preceq LI$ for all x)

Quadratic upper bound

suppose ∇f is Lipschitz continuous with parameter L and $\mathbf{dom}\, f$ is convex

- Then $g(x) = (L/2)x^\top x - f(x)$, with $\mathbf{dom}\, g$, is convex
- convexity of g is equivalent to a quadratic upper bound on f:

$$f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2 \quad \forall x, y \in \mathbf{dom}\, f$$



Proof.

- Lipschitz continuity of ∇f and Cauchy-Schwarz inequality imply

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \leq L\|x - y\|_2^2 \; \forall x, y \in \mathbf{dom}\, f$$

  this is monotonicity of the gradient $\nabla g(x) = Lx - \nabla f(x)$

- hence, g is a convex function if its domain $\mathbf{dom}\, g = \mathbf{dom}\, f$

- the quadratic upper bound is the first-order condition for the convexity of g

$$g(y) \geq g(x) + \nabla g(x)^\top (y - x) \quad \forall x, y \in \mathbf{dom}\, g$$

□

Consequence of quadratic upper bound

if $\mathbf{dom}\, f = \mathbf{R}^n$ and f has a minimizer x^*, then

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{L}{2}\|x - x^*\|_2^2 \quad \forall x$$

- Right-hand inequality follows from quadratic upper bound at $x = x^*$
- Left-hand inequality follows by minimizing quadratic upper bound

$$\begin{aligned} f(x^*) &\leq \inf_{y \in \mathbf{dom}\, f} \left(f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2 \right) \\ &= f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2 \end{aligned}$$

minimizer of upper bound is $y = x - (1/L)\nabla f(x)$ because $\mathbf{dom}\, f = \mathbf{R}^n$

Co-coercivity of gradient

if f is convex with $\mathbf{dom}\, f = \mathbf{R}^n$ and $(L/2)x^\top x - f(x)$ is convex then

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2 \quad \forall x, y$$

this property is known as *co-coercivity* of ∇f(with parameter $1/L$)

- Co-coercivity implies Lipschitz continuity of ∇f(by Cauchy-Schwarz)
- Hence, for differentiable convex f with $\mathbf{dom} f = \mathbf{R}^n$

$$\begin{aligned} \text{Lipschitz continuity of } \nabla f &\Rightarrow \text{convexity of } (L/2)x^\top x - f(x) \\ &\Rightarrow \text{co-coervivity of } \nabla f \\ &\Rightarrow \text{Lipschitz continuity of } \nabla f \end{aligned}$$

therefore the three properties are equivalent.

proof of co-coercivity: define convex functions f_x, f_y with domain $\mathbf{R}^n$:

$$f_x(z) = f(z) - \nabla f(x)^\top z, \quad f_y(z) = f(z) - \nabla f(y)^\top z$$

the functions $(L/2)z^\top z - f_x(z)$ and $(L/2)z^\top z - f_y(z)$ are convex

- $z = x$ minimizes $f_x(z)$; from the left-hand inequality on page 16,

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^\top (y - x) &= f_x(y) - f_x(x) \\ &\geq \frac{1}{2L} \|\nabla f_x(y)\|_2^2 \\ &= \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2 \end{aligned}$$

- similarly, $z = y$ minimizes $f_y(z)$; therefore

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

combining the two inequalities shows co-coercivity

Strongly convex function

f is strongly convex with parameter $m > 0$ if

$$g(x) = f(x) - \frac{m}{2}x^\top x \quad \text{is convex}$$

Jensen's inequality: Jensen's inequality for g is

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{m}{2}\theta(1-\theta)\|x-y\|_2^2$$

monotonicity: monotonicity of ∇g gives

$$(\nabla f(x) - \nabla f(y))^\top (x-y) \geq m\|x-y\|_2^2 \quad \forall x, y \in \mathbf{dom} f$$

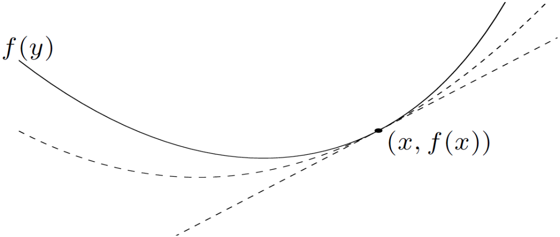
this is called $strong\ monotonicity (covercivity)$ of ∇f

second-order condition: $\nabla^2 f(x) \succeq mI$ for all $x \in \mathbf{dom}\, f$

Quadratic lower bound

form 1st order condition of convexity of g:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2}\|y - x\|_2^2 \quad \forall x, y \in \mathbf{dom}\, f$$



- Implies sublevel sets of f are bounded
- If f is closed(has closed sublevel sets), it has a unique minimizer x^* and

$$\frac{m}{2}\|x - x^*\|_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2m}\|\nabla f(x)\|_2^2 \quad x \in \mathbf{dom}\, f$$

Extension of co-coercivity

if f is strongly convex and ∇f is Lipschitz continuous, then

$$g(x) = f(x) - \frac{m}{2}\|x\|_2^2$$

is convex and ∇g is Lipschitz continuous with parameter $L - m$.

co-coercivity of g gives

$$\begin{aligned}&(\nabla f(x) - \nabla f(y))^\top (x - y)\\ &\geq \frac{mL}{m+L}\|x-y\|_2^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|_2^2\end{aligned}$$

for all $x, y \in \mathbf{dom}\, f$

Outline

- gradient method, first-order methods
- quadratic bounds on convex functions
- **analysis of gradient method**

Analysis of gradient method

$$x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k-1)}), \quad k = 1, 2, \ldots$$

with fixed step size or backtracking line search

assumptions

1. f is convex and differentiable with $\mathbf{dom}\, f = \mathbf{R}^n$

2. $\nabla f(x)$ is Lipschitz continuous with parameter $L > 0$

3. Optimal value $f^* = \inf_x f(x)$ is finite and attained at x^*

Analysis for constant step size

from quadratic upper bound with $y = x - t\nabla f(x)$:

$$f(x - t\nabla f(x)) \leq f(x) - t(1 - \frac{Lt}{2})\|\nabla f(x)\|_2^2$$

therefore, if $x^+ = x - t\nabla f(x)$ and $0 < t \leq 1/L$,

$$\begin{aligned} f(x^+) &\leq f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ &\leq f^* + \nabla f(x)^\top (x - x^*) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ &= f^* + \frac{1}{2t}\left(\|x - x^*\|_2^2 - \|x - x^* - t\nabla f(x)\|_2^2\right) \\ &= f^* + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2) \end{aligned}$$

take $x = x^{(i-1)}, x^+ = x^{(i)}, t_i = t$, and add the bounds for $i = 1, \cdots, k$:

$$\begin{aligned}\sum_{i=1}^{k}(f(x^{(i)}) - f^*) &\leq \frac{1}{2t}\sum_{i=1}^{k}\left(\|x^{(i-1)} - x^*\|_2^2 - \|x^i - x^*\|_2^2\right) \\ &= \frac{1}{2t}\left(\|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2\right) \\ &\leq \frac{1}{2t}\|x^{(0)} - x^*\|_2^2\end{aligned}$$

since $f(x^{(i)})$ is non-increasing,

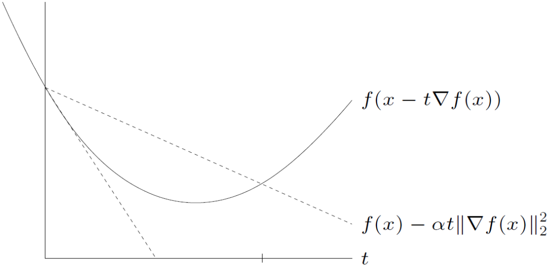
$$f(x^{(k)}) - f^* \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)}) - f^*) \leq \frac{1}{2kt}\|x^{(0)} - x^*\|_2^2$$

conclusions: iterations to reach $f(x^{(k)}) - f^* \leq \epsilon$ is $O(1/\epsilon)$

Backtracking line search

initialize t_k at $\hat{t} > 0$(for example, $\hat{t} = 1$); take $t_k := \beta t_k$ until

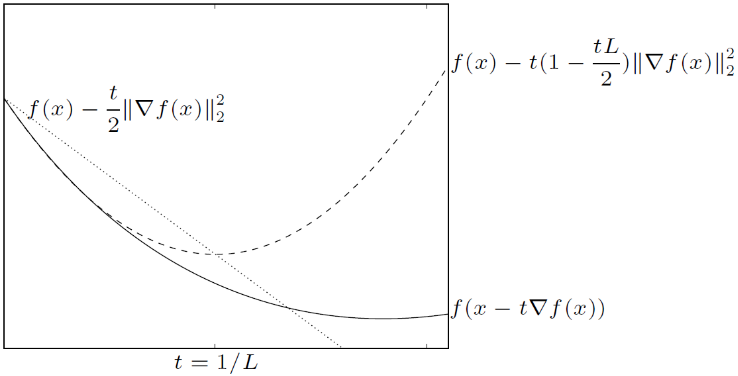
$$f(x - t_k \nabla f(x)) < f(x) - \alpha t_k \|\nabla f(x)\|_2^2$$



$0 < \beta < 1$; we will take $\alpha = 1/2$(mostly to simplify proofs)

Analysis for backtracking line search

line search with $\alpha = 1/2$ if f has a Lipschitz continuous gradient



selected step size satisfies $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$

Convergence analysis

- from page 24:

$$
\begin{aligned}
f(x^{(i)}) &\leq f^* + \frac{1}{2t_i}\left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2\right) \\
&\leq f^* + \frac{1}{2t_{\min}}\left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2\right)
\end{aligned}
$$

- add the upper bounds to get

$$
f(x^{(k)}) - f^* \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)}) - f^*) \leq \frac{1}{2kt_{\min}}\|x^{(0)} - x^*\|_2^2
$$

conclusion: same $1/k$ bound as with constant step size

Gradient method for strongly convex function

better results exist if we add strong convexity to the assumptions

analysis for constant step size

if $x^+ = x - t\nabla f(x)$ and $0 < t \leq 2/(m+L)$:

$$\begin{aligned}
\|x^+ - x^*\|_2^2 &= \|x - t\nabla f(x) - x^*\|_2^2 \\
&= \|x - x^*\|_2^2 - 2t\nabla f(x)^\top (x - x^*) + t^2\|\nabla f(x)\|_2^2 \\
&\leq (1 - t\frac{2mL}{m+L})\|x - x^*\|_2^2 + t(t - \frac{2}{m+L})\|\nabla f(x)\|_2^2 \\
&\leq (1 - t\frac{2mL}{m+L})\|x - x^*\|_2^2
\end{aligned}$$

(step 3 follows from result on page 21)

distance to optimum

$$\|x^{(k)} - x^*\|_2^2 \leq c^k \|x^{(0)} - x^*\|_2^2, \quad c = 1 - t\frac{2mL}{m+L}$$

- implies (linear) convergence
- for $t = \frac{2}{m+L}$, get $c = \frac{(\gamma-1)}{(\gamma+1)}^2$ with $\gamma = L/m$

bound on function value(from page 16),

$$f(x^{(k)}) - f^* \leq \frac{L}{2}\|x^{(k)} - x^*\|_2^2 \leq \frac{c^k L}{2}\|x^{(0)} - x^*\|_2^2$$

conclusion: iterations to reach $f(x^{(k)}) - f^* \leq \epsilon$ is $O(log(1/\epsilon))$

Limits on convergence rate of first-order methods

first-order method: any iterative algorithm that selects $x^{(k)}$ in

$$x^{(0)} + span\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \cdots, \nabla f(x^{(k-1)})\}$$

problem class: any function that satisfies the assumptions on p. 23

theorem(Nesterov): for every integer $k \leq (n-1)/2$ and every $x^{(0)}$, there exist functions in the problem class such that for any first-order method

$$f(x^{(k)}) - f^* \geq \frac{3}{32} \frac{L\|x^{(0)} - x^*\|_2^2}{(k+1)^2}$$

- suggests $1/k$ rate for gradient method is not optimal
- recent fast gradient methods have $1/k^2$ convergence(see later)

Barzilar-Borwein (BB) gradient method

Consider the problem

$$\min\ f(x)$$

- Steepest gradient descent method: $x^{k+1} := x^k - \alpha^k g^k$:

$$a^k := \arg\min_{\alpha} f(x^k - \alpha g^k)$$

- Let $s^{k-1} := x^k - x^{k-1}$ and $y^{k-1} := g^k - g^{k-1}$.
- BB: choose α so that $D = \alpha I$ satisfies $Dy \approx s$:

$$\alpha = \arg\min_{\alpha} \|\alpha y - s\|^2 \Longrightarrow \alpha := \frac{s^\top y}{y^\top y}$$

$$\alpha = \arg\min_{\alpha} \|y - s/\alpha\|^2 \Longrightarrow \alpha := \frac{s^\top s}{s^\top y}$$

Globalization strategy for BB method

设d^k是点x^k处的下降方向，$M > 0$为给定的正整数：

$$f(x^k + \alpha d^k) \le \max_{0 \le j \le \min\{k,M\}} f(x^{k-j}) + c_1 \alpha \nabla f(x^k)^{\mathrm{T}} d^k,$$

其中$c_1 \in (0,1)$为给定的常数．

Algorithm 1: Raydan's method

```
Given x^0, set α > 0, M ≥ 0, σ, δ, ϵ ∈ (0,1), k = 0.
while ||g^k|| > ϵ do
    while f(x^k − αg^k) ≥ max_{0≤j≤min(k,M)} f_{k−j} − σα||g^k||^2 do
        set α = δα
    Set x^{k+1} := x^k − αg^k.
    Set α := max( min( −(α(g^k)^⊤ g^k)/((g^k)^⊤ y^k), α_M ), α_m ), k := k + 1.
```

Globalization strategy for BB method

设d^k是点x^k处的下降方向，$M > 0$为给定的正整数：

$$f(x^k + \alpha d^k) \leq C^k + c_1 \alpha \nabla f(x^k)^{\mathrm{T}} d^k,$$

其中C^k满足递推式$C^0 = f(x^0)$, $C^{k+1} = \frac{1}{Q^{k+1}}(\eta Q^k C^k + f(x^{k+1}))$，$Q^0 = 1$, $Q^{k+1} = \eta Q^k + 1$，参数$\eta, c_1 \in (0,1)$.

Algorithm 2: Hongchao and Hagger's method

Given x^0, set $\alpha > 0$, $\sigma, \delta, \eta, \epsilon \in (0,1)$, $k = 0$.
while $\|g^k\| > \epsilon$ **do**
  while $f(x^k - \alpha g^k) \geq C^k - \sigma \alpha \|g^k\|^2$ **do**
    set $\alpha = \delta \alpha$
  Set $x^{k+1} := x^k - \alpha g^k$, $Q^{k+1} = \eta Q^k + 1$ and $C^{k+1} = (\eta Q^k C^k + f(x^{k+1}))/Q^{k+1}$.
  Set $\alpha := \max\left(\min\left(-\frac{\alpha (g^k)^\top g^k}{(g^k)^\top y^k}, \alpha_M\right), \alpha_m\right)$, $k := k + 1$.

设二次函数$f(x,y)=x^2+10y^2$，并使用*BB*方法进行迭代，初始点为$(-10,-1)$，结果如下图所示．为了方便对比，我们也在此图中描绘了梯度法的迭代过程．可以很明显看出*BB*方法的收敛速度较快，在经历*15*次迭代后已经接近最优值点．从等高线也可观察到*BB*方法是非单调方法．

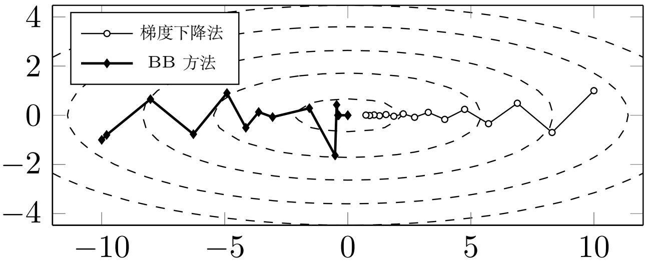


Figure: 梯度法与BB方法的前15次迭代

LASSO: 光滑化策略

LASSO 问题:

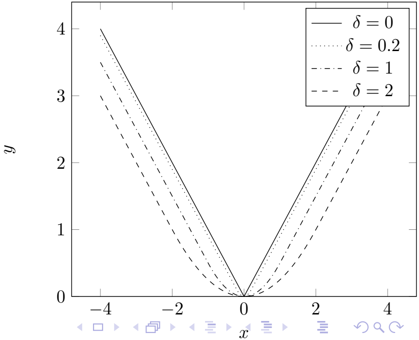
$$\min \quad f(x) = \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1.$$

求解光滑化问题

$$\min \quad f_\delta(x) = \frac{1}{2}\|Ax - b\|^2 + \mu\sum_{i=1}^{n} l_\delta(x_i),$$

一维光滑函数：

$$l_\delta(x) = \begin{cases} \frac{1}{2\delta}x^2, & |x| < \delta, \\ |x| - \frac{\delta}{2}, & \text{其他}. \end{cases}$$



采用连续化策略来从较大的正则化参数μ_0逐渐减小到μ．具体地，对于每一个μ_t，我们调用带BB步长的光滑化梯度法（这里光滑化参数$\delta_t = 10^{-3}\mu_t$）来求解对应的子问题．每个子问题的终止条件设为

$$|f_\delta(x^k) - f_\delta(x^{k-1})| < 10^{-4-t}, \text{ 或者 } \|\nabla f_\delta(x)\| < 10^{-1-t}.$$

当μ_t 的子问题求解完之后，设置

$$\mu_{t+1} = \max\{\mu_t\eta, \mu\},$$

其中η 为缩小因子，这里取为0.1．