机器学习中的优化算法

Lecture05: 罚函数法和增广拉格朗日函数法

张立平

清华大学数学科学系

办公室：理科楼#A302, Tel: 62798531

E-mail: lipingzhang@tsinghua.edu.cn

Contents and Acknowledgement

- 教材：**最优化：建模、算法与理论**

  http://bicmr.pku.edu.cn/ wenzw/bigdata2021.html

- 致谢：北京大学文再文教授

Outline of Penalty Method

- 约束优化的二次罚函数法
- 二次罚函数法应用
- ℓ_1精确罚函数法

等式约束优化的二次罚函数法

罚函数法的思想是将约束优化问题转化为无约束优化问题来进行求解．考虑等式约束的优化问题:

$$\begin{aligned} \min \quad & f(x) \\ \textbf{s.t.} \quad & c_i(x) = 0, \quad i \in \mathcal{E}, \end{aligned} \tag{1}$$

其中$x \in \mathbb{R}^n$, $\mathcal{E}$为等式约束的指标集, $f, c_i(x)$为连续可微函数.

定义(1)的二次罚函数为:

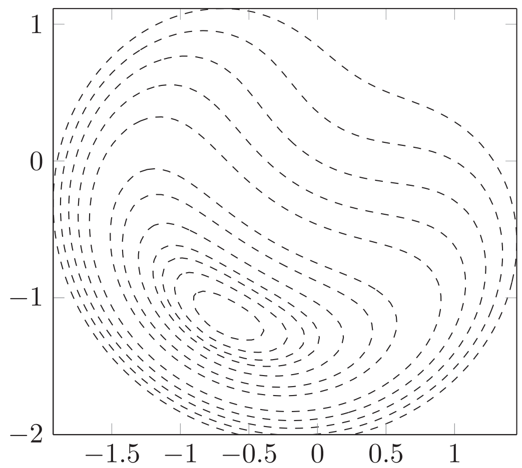
$$P_E(x, \sigma) = f(x) + \frac{1}{2}\sigma \sum_{i \in \mathcal{E}} c_i^2(x), \tag{2}$$

其中$\sigma > 0$称为**罚因子**.

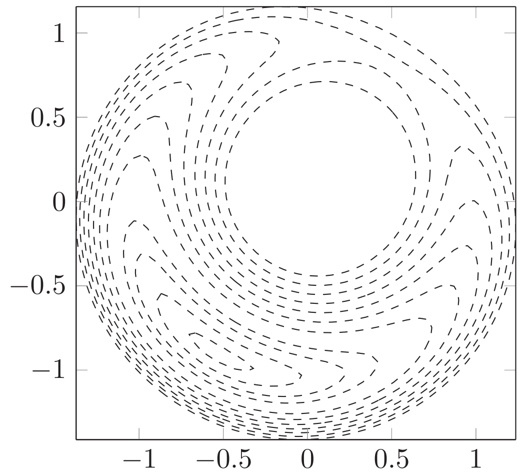
罚函数的作用

考虑优化问题: $\min x+\sqrt{3}y \quad$ **s.t.** $\; x^2+y^2=1$. 易知其最优解为$\left(-\frac{1}{2},-\frac{\sqrt{3}}{2}\right)^{\mathrm{T}}$, 二次罚函数为

$$P_E(x,y,\sigma)=x+\sqrt{3}y+\frac{\sigma}{2}\left(x^2+y^2-1\right)^2.$$



(a) $\sigma=1$



(b) $\sigma=10$

Figure 1: $P_E(x,y,\sigma)$的等高线

- 当σ选取过小时罚函数可能无下界.

考虑优化问题$\min\ -x^2+2y^2$, **s.t.** $x=1$. 显然最优解为$(1,0)^T$, 但二次罚函数$P_E(x,y,\sigma)=-x^2+2y^2+\frac{\sigma}{2}(x-1)^2$对任意的$\sigma\leq 2$无下界.

- 二次罚函数法: 在第k步迭代, 求解$x^{k+1}=\arg\min\limits_x P_E\left(x,\sigma_k\right)$. 选取$\sigma^{k+1}=\rho\sigma_k$, 其中$\rho>1$.

- 最优性条件及数值分析

  - (1)的KKT条件:

$$\begin{cases}\nabla f(x^*)-\sum_{i\in\mathcal{E}}\lambda_i^*\nabla c_i(x^*)=0,\\ c_i(x^*)=0,\quad \forall i\in\mathcal{E}.\end{cases}$$

  - (2)的一阶必要条件:

$$\nabla f(x)+\sum_{i\in\mathcal{E}}\sigma c_i(x)\nabla c_i(x)=0.$$

- 假设(1)和(2)有相同的解, 则应有:

$$\sigma c_i(x) \approx -\lambda_i^*, \quad \forall i \in \mathcal{E}.$$

因此, 为使约束$c_i(x) = 0$成立, 需要$\sigma \to \infty$.

- 考虑罚函数$P_E(x, \sigma)$的海瑟矩阵:

$$\begin{aligned}\nabla_{xx}^2 P_E(x,\sigma) &= \nabla^2 f(x) + \sum_{i\in\mathcal{E}} \sigma c_i(x)\nabla^2 c_i(x) + \sigma \nabla c(x) \nabla c(x)^{\mathrm{T}} \\ &\approx \nabla_{xx}^2 L(x, \lambda^*) + \sigma \nabla c(x) \nabla c(x)^{\mathrm{T}}.\end{aligned}$$

σ越来越大导致$\nabla_{xx}^2 P_E(x,\sigma)$ 条件数越来越大, 求解子问题的难度也会相应地增加. 因此在实际应用中, 不可能令罚因子趋于正无穷.

二次罚函数法的收敛性分析

Theorem 1. *设$\{x^k\}$是由二次罚函数法产生的迭代序列, σ_k单调上升趋于无穷, 则$\{x^k\}$的每个极限点都是(1)的全局最优解.*

Proof. 设 $\bar{x}$ 为(1)的全局最优解, 则有$P_E(x^{k+1},\sigma_k)\le P_E(\bar{x},\sigma_k)$,即

$$f(x^{k+1})+\frac{\sigma_k}{2}\sum_{i\in\mathcal{E}}c_i^2(x^{k+1})\le f(\bar{x})+\frac{\sigma_k}{2}\sum_{i\in\mathcal{E}}c_i^2(\bar{x})=f(\bar{x})$$

$$\Rightarrow\quad \sum_{i\in\mathcal{E}}c_i^2(x^{k+1})\le\frac{2}{\sigma_k}\left(f(\bar{x})-f(x^{k+1})\right).$$

设x^*是$\{x^k\}$的一个极限点, 不妨设$x^k\to x^*$, 则令$k\to\infty$, 得$\sum_{i\in\mathcal{E}}c_i^2(x^*)=0$, 故$x^*$为(1)的可行解. 又由$f(x^{k+1})\le f(\bar{x})$取极限得$f(x^*)\le f(\bar{x})$, 故$x^*$为(1)的全局最优解. □

Theorem 2. *设$\{x^k\}$是由二次罚函数法在子问题终止准则为$\|\nabla_x P_E(x^{k+1},\sigma_k)\| \le \varepsilon_k$时产生的迭代序列, 其中正数序列$\varepsilon_k \to 0$, $\sigma_k \to +\infty$. 若$\{x^k\}$的一个极限点x^*满足$\{\nabla c_i(x^*), i \in \mathcal{E}\}$线性无关, 则$x^*$是(1)的KKT点, 且其对应的拉格朗日乘子$\lambda^*$满足*

$$\lim_{k\to\infty}\left(-\sigma_k c_i(x^{k+1})\right) = \lambda_i^*, \quad \forall i \in \mathcal{E}.$$

Proof. 易知, $\nabla P_E(x,\sigma_k) = \nabla f(x) + \sum_{i\in\mathcal{E}} \sigma_k c_i(x)\nabla c_i(x)$. 由终止准则$\|\nabla_x P_E(x^{k+1},\sigma_k)\| \le \varepsilon_k$知,

$$\|\nabla f(x^{k+1}) + \sum_{i\in\mathcal{E}} \sigma_k c_i(x^{k+1})\nabla c_i(x^{k+1})\| \le \varepsilon_k$$

$$\Rightarrow \ \|\sum_{i\in\mathcal{E}} c_i(x^{k+1})\nabla c_i(x^{k+1})\| \le \frac{1}{\sigma_k}\left(\varepsilon_k + \|\nabla f(x^{k+1})\|\right).$$

不妨设$\{x^k\}$收敛于x^*, 则有$\sum_{i\in\mathcal{E}} c_i(x^*)\nabla c_i(x^*) = 0$. 因$\{\nabla c_i(x^*), i \in \mathcal{E}\}$线性无关, 故$c_i(x^*) = 0$, 从而$x^*$是(1)的可行解.

记$\nabla c(x) = [\nabla c_i(x)]_{i\in\mathcal{E}}$, $\lambda_i^k = \left(-\sigma_k c_i(x^{k+1})\right)_{i\in\mathcal{E}}$, 则

$$\nabla c(x^{k+1})\lambda^k = \nabla f(x^{k+1}) - \nabla P_E(x^{k+1}, \sigma_k). \qquad (3)$$

由条件知$\nabla c(x^*)$是列满秩矩阵且$x^k \to x^*$, 故由(3)知当k充分大时,

$$\lambda^k = \left(\nabla c(x^{k+1})^{\mathrm{T}} \nabla c(x^{k+1})\right)^{-1} \nabla c(x^{k+1})^{\mathrm{T}} \left(\nabla f(x^{k+1}) - \nabla_x P_E(x^{k+1}, \sigma_k)\right).$$

$$\Rightarrow \quad \lambda^* \overset{\text{def}}{=} \lim_{k\to\infty} \lambda^k = \left(\nabla c(x^*)^{\mathrm{T}} \nabla c(x^*)\right)^{-1} \nabla c(x^*)^{\mathrm{T}} \nabla f(x^*).$$

因此在$\left\|\nabla_x P_E(x^{k+1}, \sigma_k)\right\| \le \varepsilon_k$中令$k \to \infty$可得$(x^*, \lambda^*)$ 是(1)的KKT对. □

■ 注: 不管$\{\nabla c_i(x^*), i \in \mathcal{E}\}$ 是否线性无关, $\{x^k\}$的聚点总是$\|c(x)\|^2$ 的一个稳定点. 这说明即便没有找到可行解, 二次罚函数法也找到了使得约束$c(x) = 0$违反度相对较小的一个解.

■ 约束优化(4)的二次罚函数:

$$P(x, \sigma) = f(x) + \frac{\sigma}{2}\left(\sum_{i\in\mathcal{E}} c_i^2(x) + \sum_{i\in\mathcal{I}} (\max\{0, c_i(x)\})^2\right)$$

应用: 基追踪问题求解

考虑LASSO问题

$$\min \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1,$$

和基追踪(BP) 问题

$$\min \quad \|x\|_1 \quad \textbf{s.t.} \quad Ax = b.$$

BP问题的二次罚函数法:

$$\min \quad \|x\|_1 + \frac{\sigma}{2}\|Ax - b\|^2.$$

■ 注:

- 仅当μ趋于0时, LASSO问题的解收敛于BP问题的解.
- μ较小时LASSO问题病态, 收敛较慢, 可逐渐缩小μ的值求解子问题逼近.

LASSO 问题的二次罚函数法

1: 给定初值x_0, 最终参数μ, 初始参数μ_0, 因子$\gamma \in (0,1), k \leftarrow 0.$

2: 以x^k 为初值, 求解问题

$$x^{k+1} = \arg\min \left\{ \frac{1}{2}\|Ax-b\|^2 + \mu_k\|x\|_1 \right\}.$$

若$\mu_k = \mu$, 则停止迭代, 输出x^{k+1}.

3: 更新罚因子$\mu_{k+1} = \max\{\mu, \gamma\mu_k\}$.

4: 令$k \leftarrow k+1$, 转**2**

▶**矩阵补全问题:**

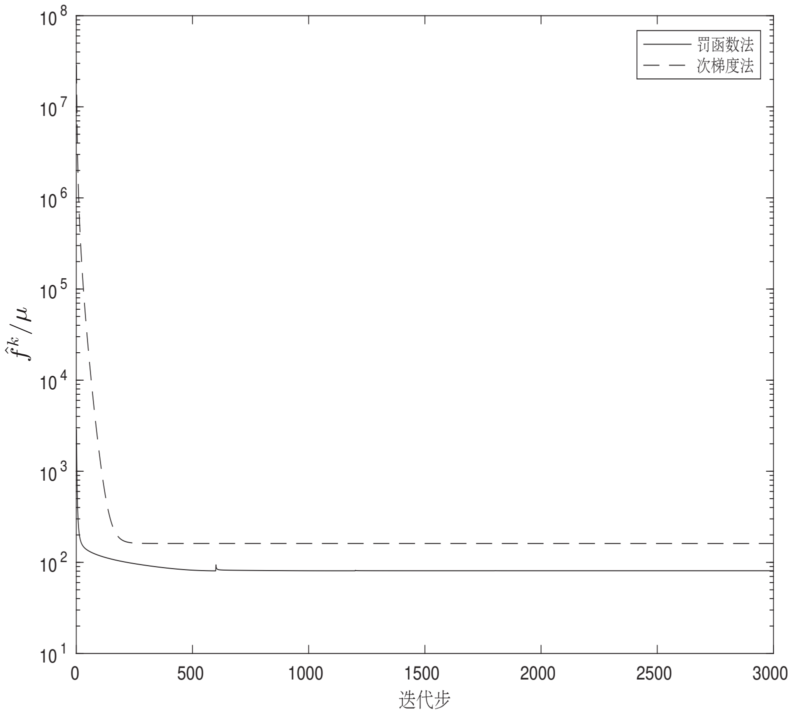
$$\begin{aligned} &\min \|X\|_* \\ &\text{s.t. } X_{ij} = M_{ij} \quad (i,j) \in \Omega \end{aligned}$$

其二次罚函数:

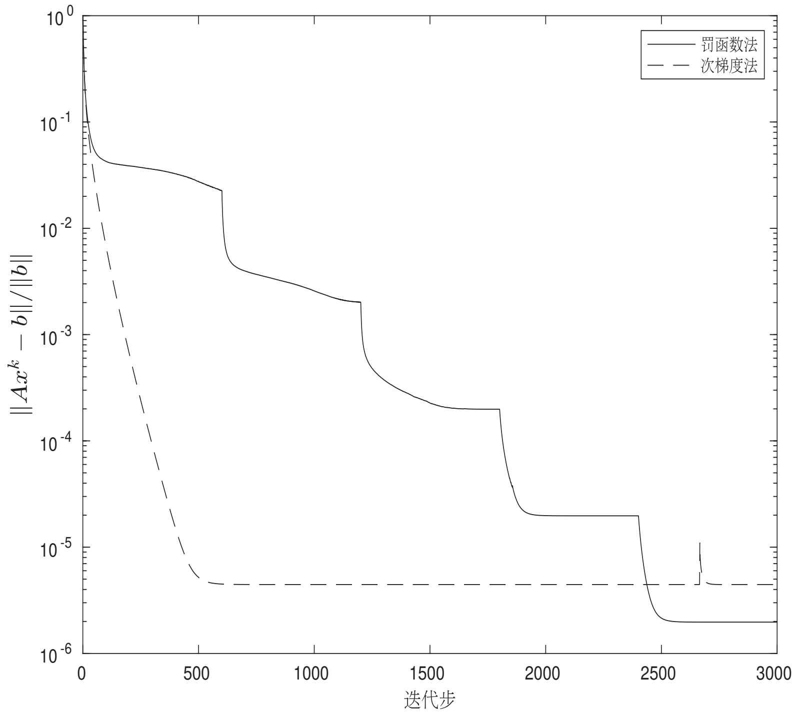
$$\min \quad \mu\|X\|_* + \frac{1}{2}\sum_{(i,j)\in\Omega}(X_{ij} - M_{ij})^2$$

求解LASSO 问题: 罚函数法和次梯度法

取LASSO 问题的正则化参数为$\mu = 10^{-3}$. 在罚函数法中, 令μ从10 开始, 因子$\gamma = 0.1$. 次梯度法选取固定步长$\alpha = 0.0002$.



(a) 函数值变化趋势



(b) 约束违反度变化趋势

精确罚函数法

精确罚函数是一种问题求解时不需要令罚因子趋于正无穷的罚函数. 常用的精确罚函数是ℓ_1罚函数.

考虑约束优化问题:

$$\begin{aligned} \min \quad & f(x) \\ \textbf{s.t.} \quad & c_i(x) = 0, \quad i \in \mathcal{E}, \\ & c_i(x) \le 0, \quad i \in \mathcal{I}. \end{aligned} \tag{4}$$

定义(4)的ℓ_1罚函数:

$$P(x,\sigma) = f(x) + \sigma\left(\sum_{i\in\mathcal{E}} |c_i(x)| + \sum_{i\in\mathcal{I}} \max\{0, c_i(x)\}\right). \tag{5}$$

ℓ_1精确罚函数法: 在第k步迭代, 以x^k为初始点, 求解

$$x^{k+1} = \arg\min_x P(x, \sigma_k).$$

选取$\sigma^{k+1} = \rho\sigma_k$, 其中$\rho > 1$.

ℓ_1精确罚函数法的收敛性分析

Lemma 1. *设x^*是(4)的一个可行解，且$c_i(x)$在x^*处是连续可微的，则下列结论等价:*

① *不存在均不为0的u_i $(i \in \mathcal{I}(x^*))$, v_i $(i \in \mathcal{E})$满足*

$$\sum_{i \in \mathcal{I}(x^*)} u_i \nabla c_i(x^*) + \sum_{i \in \mathcal{E}} v_i \nabla c_i(x^*) = 0, \quad u_i \geq 0,\ i \in \mathcal{I}(x^*).$$

② *存在$\epsilon > 0$, 对任意有界函数$b(x) : \mathcal{N}(x^*; \epsilon) \to \mathbb{R}^{|\mathcal{E}|}$, 均存在有界函数$d(x) : \mathcal{N}(x^*; \epsilon) \to \mathbb{R}^n$,使得对任意$x \in \mathcal{N}(x^*; \epsilon)$ 都有*

$$\nabla c_i(x)^T d(x) \leq -1,\ i \in \mathcal{I}(x^*); \quad \nabla c_i(x)^T d(x) = b_i(x),\ i \in \mathcal{E}.$$

Lemma 2. *设x^*是(4)的一个严格局部最优解，$c_i(x)(i \in \mathcal{E} \cup \mathcal{I})$在$x^*$处的邻域可微，则存在$\bar{\sigma}$, 使得对任意$\sigma \geq \bar{\sigma}$, 均存在正数$\epsilon(\sigma)$和向量$x(\sigma) \in \mathbb{R}^n$满足$x(\sigma) \in \mathcal{N}(x^*; \epsilon(\sigma))$且*

$$\lim_{\sigma \to \infty} \epsilon(\sigma) = 0, \quad P(x(\sigma), \sigma) \leq P(x, \sigma),\ \forall x \in \mathcal{N}(x^*; \epsilon(\sigma)).$$

ℓ_1精确罚函数法的收敛性定理

Theorem 3. 设x^*是(4)的一个严格局部最优解, $c_i(x)(i \in \mathcal{E} \cup \mathcal{I})$在$x^*$处的邻域可微. 若不存在均不为0的$u_i\ (i \in \mathcal{I}(x^*))$, $v_i\ (i \in \mathcal{E})$满足

$$\sum_{i \in \mathcal{I}(x^*)} u_i \nabla c_i(x^*) + \sum_{i \in \mathcal{E}} v_i \nabla c_i(x^*) = 0, \quad u_i \geq 0,\ i \in \mathcal{I}(x^*),$$

其中

$$\mathcal{I}(x^*) \overset{\text{def}}{=} \{i \in \mathcal{I} |\ c_i(x^*) = 0\},$$

则当罚因子σ 足够大时, x^* 也是$\min\ P(x, \sigma)$ 的一个局部最优解.

Theorem 4. 设x^*是(4)的一个严格局部最优解且满足KKT条件, 其对应的拉格朗日乘子为$\lambda_i^*(i \in \mathcal{E} \cup \mathcal{I})$, 则当罚因子$\sigma > \sigma^* = \|\lambda^*\|_\infty$ 时, x^* 也是$\min\ P(x, \sigma)$ 的一个局部最优解.

Outline of ALM

- 约束优化的增广拉格朗日函数法
- ALM应用: 基追踪问题

等式约束优化问题的增广拉格朗日函数法

等式约束优化问题(1)的增广拉格朗日函数:

$$L_\sigma(x,\lambda) = f(x) + \sum_{i\in\mathcal{E}} \lambda_i c_i(x) + \frac{\sigma}{2}\sum_{i\in\mathcal{E}} c_i^2(x).$$

ALM的子问题: 在第k步迭代, 给定罚因子σ_k和乘子λ^k, 以x^k为初始点, 求解

$$x^{k+1} \in \arg\min L_{\sigma_k}(x,\lambda^k).$$

x^{k+1}满足一阶最优性条件:

$$\nabla_x L_{\sigma_k}(x^{k+1},\lambda^k) = \nabla f(x^{k+1}) + \sum_{i\in\mathcal{E}} \left(\lambda_i^k + \sigma_k c_i(x^{k+1})\right) \nabla c_i(x^{k+1}) = 0.$$

(1)的KKT条件:

$$\nabla f(x^*) + \sum_{i\in\mathcal{E}} \lambda_i^* \nabla c_i(x^*) = 0, \quad c_i(x^*) = 0,\ i\in\mathcal{E}.$$

拉格朗日乘子的迭代: $\lambda_i^{k+1} = \lambda_i^k + \sigma_k c_i(x^{k+1})$.

约束违反度: $c_i(x^{k+1}) \approx \dfrac{\lambda_i^* - \lambda_i^k}{\sigma_k}$.

等式约束优化问题的增广拉格朗日函数法

1: 选取初始点x^0, 乘子λ^0, 罚因子$\sigma_0 > 0$, 罚因子更新常数$\rho > 0$, 约束违反度常数$\varepsilon > 0$ 和精度要求$\epsilon_k > 0$. 令$k = 0$.

2: 以x^k 为初始点, 求

$$x^{k+1} \in \arg\min \ L_{\sigma_k}(x, \lambda^k)$$

使其满足$\|\nabla_x L_{\sigma_k}(x^{k+1}, \lambda^k)\| \leq \epsilon_k$.

3: 若$\|c(x^{k+1})\| \leq \varepsilon$, 则输出近似解$(x^{k+1}, \lambda_k)$, 终止迭代.

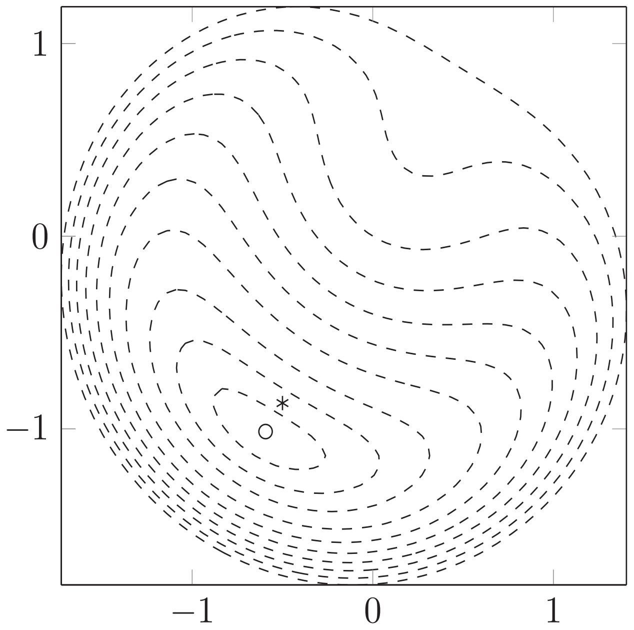
4: 更新乘子 $\lambda^{k+1} = \lambda^k + \sigma_k c(x^{k+1})$.

5: 更新罚因子 $\sigma_{k+1} = \rho\sigma_k$.

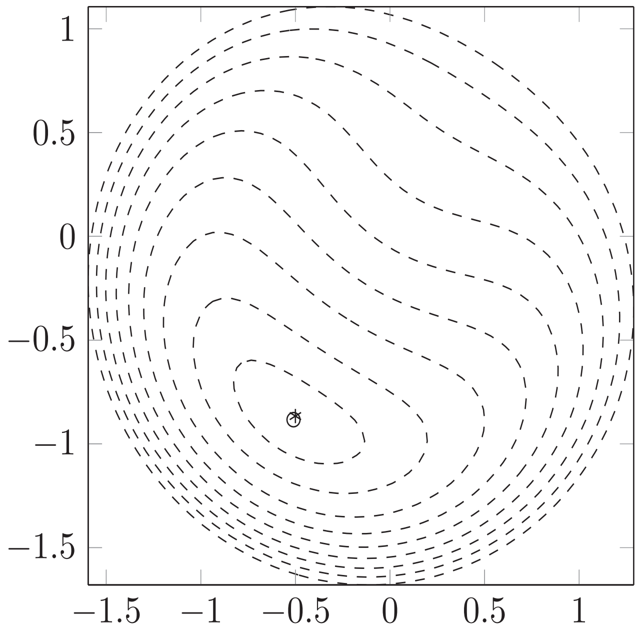
6: 令$k \leftarrow k + 1$, 转**2**.

【例】 $\min \ x+\sqrt{3}y, \quad$ **s.t.** $x^2+y^2=1.$

最优解为$x^* = \left(-\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)^{\mathrm{T}}$, 相应的拉格朗日乘子$\lambda^* = 1.$



(c) 二次罚函数法



(d) 增广拉格朗日函数法

- 二次罚函数法求出的最优解为$(-0.5957, -1.0319)$, 与最优解的欧氏距离约0.1915, 约束违反度为0.4197.
- 增广拉格朗日函数法求出的最优解为$(-0.5100, -0.8833)$, 与最优解的欧氏距离约0.02, 约束违反度为0.0403.
- 增广拉格朗日函数法具有比二次罚函数法更精确的寻优能力, 且约束违反度一般更低.
- 罚因子σ_k不应增长过快, 也不应增长过慢, 一般$\rho \in [2, 10]$或者设计更合理的自适应方法.

ALM的收敛性分析

Theorem 5. *设x^*, λ^*分别为问题(1)的局部最优解和相应的乘子, 且点x^*处LICQ和二阶充分条件成立, 则存在有限的常数$\bar{\sigma}$, 对任意的$\sigma \geq \bar{\sigma}$, x^*都是$\min\ L_\sigma(x, \lambda^*)$ 的严格局部最优解.*
反之, 若x^* 为$\min\ L_\sigma(x, \lambda^*)$ 的局部最优解且满足$c_i(x^*) = 0, i \in \mathcal{E}$, 则$x^*$为问题(1)的局部最优解.

Proof. 因x^* 为(1)的局部最优解且LICQ、二阶充分条件成立, 故

$$\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + \sum_{i \in \mathcal{E}} \lambda_i^* \nabla c_i(x^*) = 0, \quad c_i(x^*) = 0,\ i \in \mathcal{E}.$$
$$u^{\mathrm{T}} \nabla_{xx}^2 L(x^*, \lambda^*) u > 0, \quad \forall u:\ \nabla c(x^*) u = 0. \tag{6}$$

从而有$\nabla_x L_\sigma(x^*, \lambda^*) = 0$且

$$\nabla_{xx}^2 L_\sigma(x^*, \lambda^*) = \nabla_{xx}^2 L(x^*, \lambda^*) + \sigma \nabla c(x^*)^{\mathrm{T}} \nabla c(x^*).$$

为了证明x^*是$\min\ L_\sigma(x^*, \lambda^*)$的严格局部最优解, 只需证对于充分大的$\sigma$有$\nabla_{xx}^2 L_\sigma(x^*, \lambda^*) \succ 0$.

假设该结论不成立, 则对任意大的σ, 存在u_k 满足$\|u_k\|=1$, 且

$$u_k^{\mathrm{T}}\nabla_{xx}^2 L_\sigma(x^*,\lambda^*)u_k = u_k^{\mathrm{T}}\nabla_{xx}^2 L(x^*,\lambda^*)u_k + \sigma\|\nabla c(x^*)u_k\|^2 \le 0.$$

于是有

$$\|\nabla c(x^*)u_k\|^2 \le -\frac{1}{\sigma}u_k^{\mathrm{T}}\nabla_{xx}^2 L(x^*,\lambda^*)u_k \to 0 \ \ (\sigma\to\infty).$$

因$\{u_k\}$ 有界, 故必存在聚点, 设为 u. 则

$$\nabla c(x^*)u = 0, \quad u^{\mathrm{T}}\nabla_{xx}^2 L(x^*,\lambda^*)u \le 0.$$

这与(6)矛盾, 故结论成立.

反之, 若x^* 为$\min\ L_\sigma(x,\lambda^*)$ 的局部最优解且满足$c_i(x^*)=0, i\in\mathcal{E}$, 则对于任意与$x^*$ 充分接近的可行点x, 有

$$f(x^*) = L_\sigma(x^*,\lambda^*) \le L_\sigma(x,\lambda^*) = f(x).$$

因此, x^*为问题(1)的局部最优解. □

Theorem 6. *设乘子$\{\lambda^k\}$是有界的, 罚因子$\sigma_k \to +\infty$, $k \to \infty$, 精度$\eta_k \to 0$, 迭代点列$\{x^k\}$的一个子序列$\{x^{k_j+1}\}$收敛到x^*且在点x^*处LICQ成立, 则$c(x^*) = 0$且存在λ^*满足*

$$\lambda^{k_j+1} \to \lambda^* \text{ as } j \to \infty, \quad \nabla f(x^*) + \nabla c(x^*)\lambda^* = 0.$$

Proof. 对于增广拉格朗日函数$L_{\sigma_k}(x, \lambda^k)$,

$$\begin{aligned}\nabla_x L_{\sigma_k}(x^{k+1}, \lambda^k) &= \nabla f(x^{k+1}) + \nabla c(x^{k+1})\left(\lambda^k + \sigma_k c(x^{k+1})\right) \\ &= \nabla f(x^{k+1}) + \nabla c(x^{k+1})\lambda^{k+1} = \nabla_x L(x^{k+1}, \lambda^{k+1}).\end{aligned}$$

由于点x^*处LICQ成立, 故当x^{k_j+1}充分接近x^*时, **rank**$(\nabla c(x^{k_j+1})) = |\mathcal{E}|$, 从而有

$$\lambda^{k_j+1} = \left(\nabla c(x^{k_j+1})^{\mathrm{T}} \nabla c(x^{k_j+1})\right)^{-1} \nabla c(x^{k_j+1})^{\mathrm{T}} \left(\nabla_x L_{\sigma_k}(x^{k_j+1}, \lambda^{k_j}) - \nabla f(x^{k_j+1})\right).$$

因$\|\nabla_x L_{\sigma_k}(x^{k_j+1}, \lambda^{k_j})\| \le \eta_{k_j} \to 0$, 故

$$\lambda^{k_j+1} \to \lambda^* \overset{\text{def}}{=} -\left(\nabla c(x^*)^{\mathrm{T}} \nabla c(x^*)\right)^{-1} \nabla c(x^*)^{\mathrm{T}} \nabla f(x^*),$$

$$\nabla_x L(x^*, \lambda^*) = 0.$$

因$\{\lambda^k\}$是有界的, 且$\lambda^{k_j} + \sigma_{k_j} c\left(x^{k_j+1}\right) \to \lambda^*$, 故$\{\sigma_{k_j} c\left(x^{k_j+1}\right)\}$有界.
又$\sigma_k \to +\infty$, 则$c(x^*) = 0$. □

Theorem 7. *设x^*, λ^*分别是问题(1) 的严格局部最优解和相应的乘子, 则存在充分大的常数$\bar{\sigma} > 0$ 和充分小的常数$\delta > 0$, 若对某个k, 有*

$$\frac{1}{\sigma_k}\left\|\lambda^k - \lambda^*\right\| < \delta, \quad \sigma_k \geqslant \bar{\sigma},$$

则当$k \to \infty$有$\lambda^k \to \lambda^*$且$x^k \to x^*$. 而且,

① *若$\limsup \sigma_k < +\infty$ 且对任意的k有$\lambda^k \neq \lambda^*$, 则$\{\lambda^k\}$ 收敛的速度是Q-线性.*

② *若$\limsup \sigma_k = +\infty$ 且且对任意的k有$\lambda^k \neq \lambda^*$, 则$\{\lambda^k\}$收敛的速度是Q-超线性.*

凸优化的增广拉格朗日函数法

考虑凸优化问题:

$$\begin{aligned} &\min\ f(x), \\ &\textbf{s.t.}\ \ c_i(x) \le 0, i \in \mathcal{I}. \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} &\min_{x,s}\ f(x), \\ &\textbf{s.t.}\ \ c_i(x) + s_i = 0,\ \ s_i \ge 0,\ \ i \in \mathcal{I}. \end{aligned} \tag{7}$$

构造增广拉格朗日罚函数:

$$L_\sigma(x, s, \lambda) = f(x) + \sum_{i \in \mathcal{I}} \lambda_i \left(c_i(x) + s_i\right) + \frac{\sigma}{2} \sum_{i \in \mathcal{I}} \left(c_i(x) + s_i\right)^2.$$

固定x, 求解关于s的子问题:

$$\min_{s \ge 0}\ \ L_\sigma(x, s, \lambda) \quad \Leftrightarrow \quad s_i = \max\left\{-\frac{\lambda_i}{\sigma} - c_i(x), 0\right\},\ \ i \in \mathcal{I}.$$

于是凸优化(7)增广拉格朗日罚函数为:

$$L_\sigma(x, \lambda) = f(x) + \frac{\sigma}{2} \sum_{i \in \mathcal{I}} \left(\max\left\{\frac{\lambda_i}{\sigma} + c_i(x), 0\right\}^2 - \frac{\lambda_i^2}{\sigma^2}\right).$$

给定单调递增序列$\{\sigma_k\} \uparrow \sigma_\infty$, 初始乘子$\lambda^0$, 凸优化(7)的增广拉格朗日函数法为

$$\begin{cases} x^{k+1} \approx \underset{x\in\mathbb{R}^n}{\arg\min} L_{\sigma_k}(x,\lambda^k), \\ \lambda^{k+1} = \max\left\{0, \lambda^k + \sigma_k c(x^{k+1})\right\}. \end{cases} \tag{8}$$

► 不精确条件: 令$\phi_k(x) = L_{\sigma_k}(x,\lambda^k)$. (8)的显式最优解往往未知, 通常调用迭代算法求一个近似解. 为保证收敛性, 要求该近似解至少满足不精确条件:

$$\phi_k(x^{k+1}) - \inf \phi_k \leqslant \frac{\varepsilon_k^2}{2\sigma_k}, \quad \varepsilon_k \geqslant 0, \quad \sum_{k=1}^{\infty} \varepsilon_k < +\infty. \tag{9}$$

由于 $\inf \phi_k$ 是未知的, 直接验证(9) 式是数值上不可行的. 但是, 若ϕ_k 是α - 强凸函数, 则有

$$\phi_k(x) - \inf \phi_k \leqslant \frac{1}{2\alpha}\textbf{dist}^2\left(0, \partial\phi_k(x)\right).$$

可以进一步构造如下数值可验证的不精确条件:

$$\textbf{dist}\left(0, \partial\phi_k(x^{k+1})\right) \leq \sqrt{\frac{\alpha}{\sigma_k}}\varepsilon_k, \quad \varepsilon_k \geqslant 0, \quad \sum_{k=1}^{\infty} \varepsilon_k < +\infty. \tag{10}$$

▶ 凸优化增广拉格朗日函数法的收敛性:

Theorem 8. *假设$\{x^k\}$,$\{\lambda^k\}$为问题(7)通过(8)式生成的序列, x^{k+1}满足不精确条件(9). 如果问题(7)的Slater约束品性成立, 那么序列$\{\lambda^k\}$是有界序列且收敛到λ^∞, 并且λ^∞为对偶问题的一个最优解.*

如果存在一个γ, 使得下水平集$\{x \in \mathcal{X} \mid f(x) \leqslant \gamma\}$是非空有界的, 那么序列$\{x^k\}$也是有界的, 并且其所有的聚点都是问题(7)的最优解.

▶ 一般约束问题的增广拉格朗日函数法: 考虑一般的约束优化问题:

$$
\begin{aligned}
\min \quad & f(x), \\
\textbf{s.t.} \quad & c_i(x) = 0, i \in \mathcal{E}, \\
& c_i(x) \le 0, i \in \mathcal{I}.
\end{aligned}
$$

其增广拉格朗日函数如下:

$$
\begin{aligned}
L_\sigma(x, \lambda, \mu) =& f(x) + \sum_{i \in \mathcal{E}} \lambda_i c_i(x) + \frac{\sigma}{2} \sum_{i \in \mathcal{E}} c_i^2(x) \\
& + \frac{\sigma}{2} \sum_{i \in \mathcal{I}} \left(\max\left\{ \frac{\mu_i}{\sigma} + c_i(x), 0 \right\}^2 - \frac{\mu_i^2}{\sigma^2} \right).
\end{aligned}
$$

基追踪问题的ALM算法

考虑基追踪问题(BP)

$$\min_{x\in\mathbb{R}^n} \quad \|x\|_1, \quad \textbf{s.t.} \quad Ax=b, \tag{11}$$

其中 $A\in\mathbb{R}^{m\times n}(m\leqslant n),\ b\in\mathbb{R}^m$.

基追踪问题(11)的对偶问题为:

$$\begin{aligned}&\min_{y\in\mathbb{R}^m} b^{\mathrm{T}}y\\ &\textbf{s.t.}\ \ \|A^{\mathrm{T}}y\|_\infty\le 1.\end{aligned} \quad \Leftrightarrow \quad \begin{aligned}&\min_{y\in\mathbb{R}^m,\ s\in\mathbb{R}^n} b^{\mathrm{T}}y\\ &\textbf{s.t.}\quad A^{\mathrm{T}}y-s=0,\ \|s\|_\infty\le 1.\end{aligned} \tag{12}$$

其增广拉格朗日函数为

$$\begin{aligned}L_\sigma(x,\lambda) &= \|x\|_1+\lambda^{\mathrm{T}}(Ax-b)+\frac{\sigma}{2}\|Ax-b\|_2^2\\ &= \|x\|_1+\frac{\sigma}{2}\left\|Ax-b+\frac{\lambda}{\sigma}\right\|_2^2.\end{aligned}$$

▶ 基追踪问题(11)的ALM法: 固定σ, 第k步迭代更新格式为

$$\begin{cases} x^{k+1} = \underset{x\in\mathbb{R}^n}{\arg\min} \left\{ \|x\|_1 + \frac{\sigma}{2} \left\| Ax - b + \frac{\lambda^k}{\sigma} \right\|_2^2 \right\}, \\ \lambda^{k+1} = \lambda^k + \sigma \left(Ax^{k+1} - b \right). \end{cases} \tag{13}$$

设迭代初始点$x^0 = \lambda^0 = 0$, 由(13)得

$$0 \in \partial \left\| x^{k+1} \right\|_1 + \sigma A^{\mathrm{T}} \left(Ax^{k+1} - b + \frac{\lambda^k}{\sigma} \right)$$

因此

$$-A^{\mathrm{T}} \lambda^{k+1} \in \partial \|x^{k+1}\|_1. \tag{14}$$

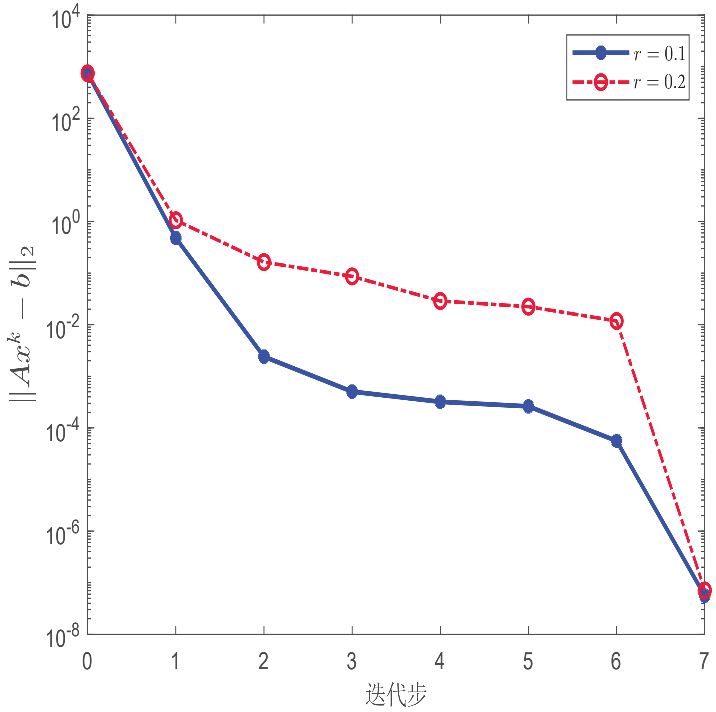
- BP问题的实例: 取A是512×1024规模的随机矩阵(每个元素从标准正态分布中抽样), b定义为

$$b = Au,$$

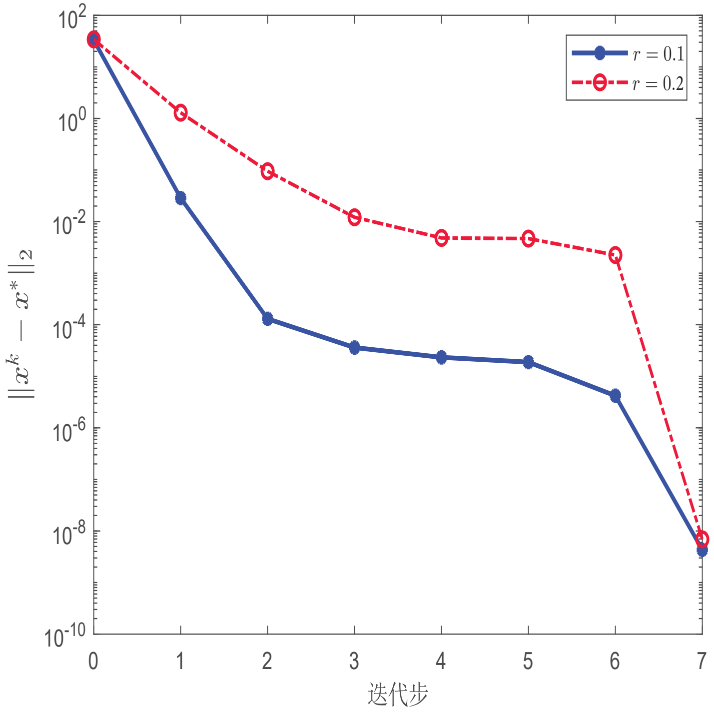
其中$u \in \mathbb{R}^{1024}$是服从正态分布随机稀疏向量, 设其稀疏度$r = 0.1$或0.2, 即分别约具有102/205 个服从正态分布的非零分量.

固定罚因子σ, 采用近似点梯度法不精确地求解(13)中关于x的子问题以得到x^{k+1}.

设置求解精度$\eta_k = 10^{-k}$, 并使用**BB步长**作为线搜索的初始步长. 下图展示了算法产生的迭代点与最优点的欧式距离的变化, 以及它们约束违反度的变化趋势. 由图可知: 固定的罚因子σ 也可以使增广拉格朗日函数法收敛.



(e) 约束违反度



(f) 与最优点的距离

▶ 对固定的二次罚项系数$\sigma = 1$, BP的ALM迭代格式(13)具有有限终止性.

基追踪问题ALM法的收敛性分析

Lemma 3. *设$\{x^k\}, \{\lambda^k\}$ 是ALM算法*(13)*从初始点$x^0 = \lambda^0 = 0$产生的迭代序列, 则$\|Ax^{k+1} - b\|_2 \le \|Ax^k - b\|_2$, 并且若存在$\tilde{x}$ 满足$A\tilde{x} = b$, 则*

$$\frac{\sigma}{2}\|Ax^k - b\|_2^2 \le \frac{1}{k}\|\tilde{x}\|_1. \tag{15}$$

Proof. 由(13)知,

$$\begin{aligned}&\|x^{k+1}\|_1 + (\lambda^k)^{\mathrm{T}}(Ax^{k+1} - b) + \frac{\sigma}{2}\|Ax^{k+1} - b\|_2^2 \\ \le &\|x^k\|_1 + (\lambda^k)^{\mathrm{T}}(Ax^k - b) + \frac{\sigma}{2}\|Ax^k - b\|_2^2.\end{aligned}$$

由$\|x\|_1$ 的凸性和(14)知,

$$\|x^{k+1}\|_1 \ge \|x^k\|_1 + \langle -A^{\mathrm{T}}\lambda^k, x^{k+1} - x^k\rangle.$$

因此,

$$\|Ax^{k+1} - b\|_2 \le \|Ax^k - b\|_2.$$

由(13)知,

$$A^{\mathrm{T}}(\lambda^{k+1}-\lambda^{k})=\sigma A^{\mathrm{T}}(Ax^{k+1}-b), \tag{16}$$

于是, 由$\|Ax-b\|_2^2$凸及(16), $\|x\|_1$凸及(14)可得

$$\begin{aligned}&\frac{\sigma}{2}\|Ax^{k+1}-b\|_2^2-\frac{\sigma}{2}\|Ax-b\|_2^2\le\langle A^{\mathrm{T}}(\lambda^{k+1}-\lambda^{k}),x^{k+1}-x\rangle\\ =&\langle A^{\mathrm{T}}\lambda^{k+1},x^{k+1}-x\rangle-\langle A^{\mathrm{T}}\lambda^{k},x^{k}-x\rangle-\langle A^{\mathrm{T}}\lambda^{k},x^{k+1}-x^{k}\rangle\\ \le&\langle A^{\mathrm{T}}\lambda^{k+1},x^{k+1}-x\rangle-\langle A^{\mathrm{T}}\lambda^{k},x^{k}-x\rangle+\|x^{k+1}\|_1-\|x^{k}\|_1.\end{aligned}$$

因此,

$$k\left(\frac{\sigma}{2}\|Ax^{k}-b\|_2^2-\frac{\sigma}{2}\|Ax-b\|_2^2\right)\le\sum_{j=1}^{k}\left(\frac{\sigma}{2}\|Ax^{j}-b\|_2^2-\frac{\sigma}{2}\|Ax-b\|_2^2\right)$$

$$\le\left\langle A^{\mathrm{T}}\lambda^{k},x^{k}-x\right\rangle+\|x^{k}\|_1-\left\langle A^{\mathrm{T}}\lambda^{0},x^{0}-x\right\rangle-\|x^{0}\|_1\le\|x\|_1,$$

取$x=\tilde{x}$, 我们即有

$$\frac{\sigma}{2}\left\|Ax^{k}-b\right\|_2^2\leqslant\frac{1}{k}\|\tilde{x}\|_1.$$

□

Lemma 4. *假设问题(11)的可行域非空, x^k 是由迭代格式(13)得到的满足$Ax^k = b$ 的迭代点, 则x^k 是BP 问题(11)的一个解.*

Proof. 对任意x, 由$\|x\|_1$ 的凸性和(14)式, 有

$$\begin{aligned} \|x^k\|_1 &\leqslant \|x\|_1 - \left\langle x - x^k, -A^{\mathrm{T}}\lambda^k \right\rangle \\ &= \|x\|_1 + \left\langle Ax - Ax^k, \lambda^k \right\rangle \\ &= \|x\|_1 + \left\langle Ax - b, \lambda^k \right\rangle, \end{aligned}$$

因此, 对任意的满足$Ax = b$ 的x, 都有$\left\|x^k\right\|_1 \leqslant \|x\|_1$. 故$x^k$ 是问题(11) 的最优解. □

Theorem 9 (BP的ALM收敛性定理)**.** *假设问题(11)的可行域非空, 迭代序列$\{(x^k, \lambda^k)\}$ 是由迭代格式(13) 从初始点 $x^0 = \lambda^0 = 0$ 产生的, 则存在正整数 K 使得任意的$x^k, k \geq K$ 是问题(11)的解.*

基追踪问题ALM法的收敛性定理证明

对指标集$\{1,2,\ldots,n\}$的任一划分$\left(I_+^j, I_-^j, E^j\right)$, 令

$$U^j \overset{\text{def}}{=} \left\{x \mid x_i \geqslant 0, i \in I_+^j; x_i \leqslant 0, i \in I_-^j; x_i = 0, i \in E^j\right\},$$

$$H^j \overset{\text{def}}{=} \min_{x\in\mathbb{R}^n} \left\{\frac{1}{2}\|Ax-b\|_2^2 \mid x \in U^j\right\}.$$

对于迭代点λ^k, 定义指标集$\{1,2,\ldots,n\}$的划分$\left(I_+^{j_k}, I_-^{j_k}, E^{j_k}\right)$为

$$I_+^{j_k} = \left\{i : \left(A^{\mathrm{T}}\lambda^k\right)_i = -1\right\},$$

$$I_-^{j_k} = \left\{i : \left(A^{\mathrm{T}}\lambda^k\right)_i = 1\right\},$$

$$E^{j_k} = \left\{i : \left(A^{\mathrm{T}}\lambda^k\right)_i \in (-1,1)\right\}.$$

并由U^j的定义和$-A^{\mathrm{T}}\lambda^k \in \partial\left\|x^k\right\|_1$知, $x^k \in U^{j_k}$.

因为问题(11)的可行域非空, 故存在$\tilde{x}$ 满足$\|A\tilde{x}-b\|=0$. 由引理3, 对任意满足$H^j>0$ 的j, 存在一个充分大的K_j 使得

$$x^k \notin U^j, \quad \forall k \geqslant K_j.$$

因此, 我们取$K=\max_j \left\{K_j \mid H^j>0\right\}$, 即有

$$H^{j_k}=0, \quad \forall k \geqslant K.$$

结合$\|x\|_1$ 的凸性和(14) 式, 对$k \geqslant K$ 有

$$\|x^k\|_1+(\lambda^k)^{\mathrm{T}} A x^k \leqslant \|x\|_1+(\lambda^k)^{\mathrm{T}} A x,$$

且等号成立当且仅当$x \in U^{j_k}$ (注意到$x^k \in U^{j_k}$). 由于$H^{j_k}=0$, 故可取$\tilde{x} \in U^{j_k}$ 且$\|A\tilde{x}-b\|=0$, 根据x^k 的最优性, 得到

$$\frac{\sigma}{2}\|Ax^k-b\|^2 \leqslant \|\tilde{x}\|_1-\|x^k\|_1+\left(\lambda^k\right)^{\mathrm{T}} A\left(\tilde{x}-x^k\right)+\frac{\sigma}{2}\|A\tilde{x}-b\|^2=0.$$

因此由引理4可知, $x^k(\forall k \geqslant K)$都是问题(11) 的最优解.

基追踪问题的Bregman算法

求解基追踪问题的一个通用方法是Bregman迭代算法. 对于凸函数$h(x) = \|x\|_1$, 定义其 Bregman距离:

$$D_h^g(x, y) = h(x) - h(y) - \langle g, x - y \rangle,$$

其中$g \in \partial h(y)$ 为函数 h 在点 y 处的一个次梯度.

▶ 基追踪问题的Bregman迭代算法:

$$\begin{cases} x^{k+1} = \underset{x \in \mathbb{R}^n}{\arg\min} \left\{ D_h^{g^k}\left(x, x^k\right) + \frac{1}{2}\|Ax - b\|_2^2 \right\}, \\ g^{k+1} = g^k - A^{\mathrm{T}}\left(Ax^{k+1} - b\right). \end{cases} \tag{17}$$

▶ BP的ALM法(13) 和Bregman迭代算法(17)具有如下的等价性质: 若(13)的初始点设置为$(x^0, -A^{\mathrm{T}}\lambda^0)$, 则有$g^k = -A^{\mathrm{T}}\lambda^k$. 在合理选取初始点时, 两个算法是等价的.