

# 近似点算法

文再文

北京大学北京国际数学研究中心

教材《最优化：建模、算法与理论》配套电子教案

<http://bicmr.pku.edu.cn/~wenzw/optbook.html>

致谢：本教案由陈乐恒、邓展望协助准备

# 提纲

- 1 近似点算法
- 2 与增广拉格朗日函数法的关系
- 3 应用举例: LASSO问题
- 4 收敛性分析
- 5 Moreau-Yosida 正则化

# 近似点算法

考虑一般形式的优化问题：

$$\min_x \psi(x) \quad (1)$$

其中 $\psi$ 是一个适当的闭凸函数，并不要求连续或可微。

对于不可微的情形，可以使用次梯度法求解，但这种方式收敛较慢，且收敛条件苛刻。我们考虑用近似点梯度法做隐性的梯度下降：

$$\begin{aligned} x^{k+1} &= \text{prox}_{t_k \psi}(x^k) \\ &= \arg \min_u \left\{ \psi(u) + \frac{1}{2t_k} \|u - x^k\|_2^2 \right\} \end{aligned} \quad (2)$$

- $\psi(x)$ 的邻近算子一般需要通过迭代求解
- 迭代格式(2)的目标函数强凸，相比原问题更利于迭代法的求解

# FISTA算法加速

可以用FISTA算法对近似点算法进行加速，其迭代格式为：

$$x^k = \text{prox}_{t_k \psi} \left( x^{k-1} + \gamma_k \frac{1 - \gamma_{k-1}}{\gamma_{k-1}} (x^{k-1} - x^{k-2}) \right)$$

第二类Nesterov加速算法的迭代格式可以写成：

$$v^k = \text{prox}_{(t_k/\gamma_k)\psi} (v^{k-1}), \quad x^k = (1 - \gamma_k) x^{k-1} + \gamma_k v^k$$

关于算法参数的选择有两种策略：

- 固定步长  $t_k = t$  以及  $\gamma_k = \frac{2}{k+1}$ ;
- 可变步长  $t_k$ , 当  $k = 1$  时取  $\gamma_1 = 1$ ; 当  $k > 1$  时,  $\gamma_k$  来自下面的方程

$$\frac{(1 - \gamma_k) t_k}{\gamma_k^2} = \frac{t_{k-1}}{\gamma_{k-1}^2}$$

# 提纲

- 1 近似点算法
- 2 与增广拉格朗日函数法的关系
- 3 应用举例: LASSO问题
- 4 收敛性分析
- 5 Moreau-Yosida 正则化

考虑具有如下形式的优化问题：

$$\min_{x \in \mathbb{R}^n} f(x) + h(Ax) \quad (3)$$

其中  $f, h$  为适当的闭凸函数,  $A \in \mathbb{R}^{m \times n}$

下面给出了优化问题(3)的一些常见例子：

- 当  $h$  是单点集  $\{b\}$  的示性函数时，等价于线性等式约束优化问题

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{s.t. } Ax = b$$

- 当  $h$  是凸集  $C$  上的示性函数时，等价于凸集约束问题

$$\min f(x) \quad \text{s.t. } Ax \in C$$

- 当  $h(y) = \|y - b\|$  时，等价于正则优化问题

$$\min f(x) + \|Ax - b\|$$

# 对偶问题

写出上述优化问题的拉格朗日函数：

$$L(x, y, z) = f(x) + h(y) + z^T(Ax - y)$$

则对偶问题为：

$$\max \quad \psi(z) = \inf_{x,y} L(x, y, z) = -f^*(-A^T z) - h^*(z) \quad (4)$$

我们有以下结论：

对对偶问题用近似点算法  $\iff$  对原问题用增广拉格朗日函数法

## 对偶函数的邻近算子

对于对偶问题(4)，近似点算法迭代格式如下

$$z^{k+1} = \text{prox}_{t\psi}(z^k) = \arg \min_z \left\{ f^*(-A^T z) + h^*(z) + \frac{1}{2t_k} \|z - z^k\|_2^2 \right\}$$

事实上也可以写成： $z^{k+1} = \text{prox}_{t\psi}(z^k) = z^k + t^k(A\hat{x}^k - \hat{y})$ ，其中

$$(\hat{x}, \hat{y}) = \underset{x, y}{\operatorname{argmin}} \left( f(x) + g(y) + z^T(Ax - y) + \frac{t}{2} \|Ax - y\|_2^2 \right)$$

也就是说， $\hat{x}, \hat{y}$  最小化增广拉格朗日函数，近似点算法迭代格式对应增广拉格朗日函数法中的乘子更新



# 定理证明

## 引理

设  $f(x)$  是适当的闭凸函数,  $f^*(y)$  是其共轭函数, 则对任意的  $y \in \text{dom} f^*$  和  $x \in \text{dom} f$  有

$$y \in \partial f(x) \Leftrightarrow x \in \partial f^*(y) \quad (5)$$

## Proof.

由于  $f$  是适当闭函数, 因此有  $f^{**} = f$ .

根据  $f^*(y)$  的定义,  $y \in \partial f(x)$  即表明  $x$  满足最优性条件, 因此

$$x^T y - f(x) = f^*(y)$$

由自共轭性得

$$f^{**}(x) = f(x) = x^T y - f^*(y)$$

这说明  $y$  满足最优性条件, 即  $x \in \partial f^*(y)$ .

另一方向结论同理可得.



# 定理证明

## Proof.

首先写出最小化增广拉格朗日函数的优化问题:

$$\begin{aligned} \min_{x,y,w} \quad & f(x) + h(y) + \frac{t}{2} \|w\|_2^2 \\ \text{s.t.} \quad & Ax - y + z/t = w \end{aligned}$$

对约束  $Ax - y + \frac{z}{t} = w$  引入乘子  $u$ , 由最优性条件有:

$$A\hat{x} - \hat{y} + \frac{z}{t} = w, \quad -A^T u \in \partial f(\hat{x}), \quad u \in \partial h(\hat{y}), \quad tw = u$$

消去  $w$  得  $u = z + t(Ax - y)$ . 又根据引理(5)得:

$$\hat{x} \in \partial f^*(-A^T u), \quad \hat{y} \in \partial h^*(u)$$

代入  $u$  的表达式可得  $0 \in -A\partial f^*(-A^T u) + \partial h^*(u) + \frac{1}{t}(u - z)$ , 这正是  $u = \text{prox}_{t\psi}(z)$  的最优性条件.

另一方面, 若有  $u = \text{prox}_{t\psi}(z)$ , 则选取  $\hat{x} \in \partial f^*(-A^T u)$  及  $\hat{y} \in \partial h^*(u)$ , 即可恢复出增广拉格朗日函数法中的变量 □

# 增广拉格朗日函数法

选择初始点  $z^{(0)}$  并迭代以下步骤:

## 1 最小化增广拉格朗日函数

$$(\hat{x}, \hat{y}) = \operatorname{argmin}_{x, y} \left( f(x) + h(y) + \frac{t_k}{2} \|Ax - y + \frac{1}{t_k} z^k\|_2^2 \right)$$

## 2 乘子更新

$$z^{k+1} = z^k + t_k (A\hat{x} - \hat{y})$$

- 这等价于对对偶问题使用近似点算法
- (对偶问题)可以使用加速版本的近似点算法来加快收敛速度
- 通常第一步先求解一个较不精确的最值

## 例子

$$\min f(x) + h(Ax)$$

- 等式约束 ( $h$  是  $\{b\}$  点处的示性函数)

$$\begin{aligned}\hat{x} &= \operatorname{argmin}_x \left( f(x) + z^T(Ax - b) + \frac{t}{2} \|Ax - b\|_2^2 \right) \\ u &= z + t(A\hat{x} - b)\end{aligned}$$

- 凸集约束 ( $h$  是凸集  $C$  上示性函数):

$$\begin{aligned}\hat{x} &= \operatorname{argmin}_x \left( f(x) + \frac{t}{2} d(Ax + z/t)^2 \right) \\ u &= z + t(A\hat{x} - P(A\hat{x} + z/t))\end{aligned}$$

其中,  $P(u)$  是  $u$  在  $C$  上的投影,  $d(u) = \|u - P(u)\|_2$  是 Euclid 距离

# 提纲

- 1 近似点算法
- 2 与增广拉格朗日函数法的关系
- 3 应用举例: LASSO问题
- 4 收敛性分析
- 5 Moreau-Yosida 正则化

# LASSO

考虑LASSO问题：

$$\min_{x \in \mathbb{R}^n} \psi(x) = \mu \|x\|_1 + \frac{1}{2} \|Ax - b\|_2^2. \quad (6)$$

引入变量 $y = Ax - b$ ，问题(6)可以等价地转化为

$$\min_{x, y} f(x, y) = \mu \|x\|_1 + \frac{1}{2} \|y\|_2^2, \quad \text{s.t. } Ax - y - b = 0. \quad (7)$$

对于问题(7)，我们采用近似点算法进行求解，其第 $k$ 步迭代为

$$(x^{k+1}, y^{k+1}) \approx \operatorname{argmin}_{(x, y) \in \mathcal{D}} \left\{ f(x, y) + \frac{1}{2t_k} (\|x - x^k\|_2^2 + \|y - y^k\|_2^2) \right\}, \quad (8)$$

其中 $\mathcal{D} = \{(x, y) \mid Ax - y = b\}$ 为可行域， $t_k$ 为步长。由于问题(8)没有显式解，我们需要采用迭代算法来进行求解，比如罚函数法，增广拉格朗日方法等等。

除了直接求解问题(8)，一种比较实用的方式是通过偶问题的解来构造 $(x^{k+1}, y^{k+1})$ 。引入拉格朗日乘子 $z$ ，问题(8)的对偶函数为：

$$\begin{aligned}\Phi_k(z) &= \inf_x \left\{ \mu \|x\|_1 + z^T A x + \frac{1}{2t_k} \|x - x^k\|_2^2 \right\} \\ &\quad + \inf_y \left\{ \frac{1}{2} \|y\|_2^2 - z^T y + \frac{1}{2t_k} \|y - y^k\|_2^2 \right\} - b^T z \\ &= \mu \Gamma_{\mu t_k}(x^k - t_k A^T z) - \frac{1}{2t_k} (\|x_k - t_k A^T z\|_2^2 - \|x_k\|_2^2) \\ &\quad - \frac{1}{2(t_k + 1)} (\|z\|_2^2 + 2(y^k)^T z - \|y^k\|_2^2) - b^T z.\end{aligned}$$

这里，

$$\Gamma_{\mu t_k}(u) = \inf_x \left\{ \|x\|_1 + \frac{1}{2\mu t_k} \|x - u\|_2^2 \right\}.$$

# LASSO

通过简单地计算，并记函数  $q_{\mu t_k} : \mathbb{R} \rightarrow \mathbb{R}$  为

$$q_{\mu t_k}(v) = \begin{cases} \frac{v^2}{2\mu t_k}, & |v| \leq t, \\ |v| - \frac{\mu t_k}{2}, & |v| > t, \end{cases}$$

我们有  $\Gamma_{\mu t_k}(u) = \sum_{i=1}^n q_{\mu t_k}(u_i)$ ，其为极小点  $x = \text{prox}_{\mu t_k \|x\|_1}(u)$  处的目标函数值。易知  $\Gamma_{\mu t_k}(u)$  是关于  $u$  的连续可微函数且导数为：

$$\nabla_u \Gamma_{\mu t_k}(u) = u - \text{prox}_{\mu t_k \|x\|_1}(u).$$

那么，问题(8)的对偶问题为

$$\min_z \quad \Phi_k(z).$$

设对偶问题的逼近最优解为  $z^{k+1}$ ，那么根据问题(8)的最优性条件，我们有

$$\begin{cases} x^{k+1} = \text{prox}_{\mu t_k \|x\|_1}(x^k - t_k A^T z^{k+1}), \\ y^{k+1} = \frac{1}{t_k + 1}(y^k + t_k z^{k+1}). \end{cases}$$



# LASSO

在第 $k$ 步迭代, LASSO (6) 问题的近似点算法的迭代格式写为:

$$\begin{cases} z^{k+1} \approx \operatorname{argmax}_z \Phi_k(z), \\ x^{k+1} = \operatorname{prox}_{\mu t_k \|x\|_1} (x^k - t_k A^T z^{k+1}), \\ y^{k+1} = \frac{1}{t_k + 1} (y^k + t_k z^{k+1}). \end{cases} \quad (9)$$

根据 $\Phi_k(z)$ 的连续可微性, 我们可以调用梯度法进行求解。另外可以证明 $\Phi_k(z)$ 是半光滑的, 从而调用半光滑牛顿法来更有效地求解。为了保证算法(9)的收敛性, 我们采用以下 $z^{k+1}$ 满足以下不精确收敛准则:

$$\begin{aligned} \|\nabla \Phi_k(z^{k+1})\|_2 &\leq \sqrt{\alpha/t_k} \epsilon_k, \quad \epsilon_k \geq 0, \quad \sum_k \epsilon_k < \infty, \\ \|\nabla \Phi_k(z^{k+1})\|_2 &\leq \sqrt{\alpha/t_k} \delta_k \|(x^{k+1}, y^{k+1}) - (x^k, y^k)\|^2, \quad \delta_k \geq 0, \quad \sum_k \delta_k < \infty, \end{aligned} \quad (10)$$

其中 $\epsilon_k, \delta_k$ 是人为设定的参数,  $\alpha$ 为 $\Phi_k$ 的强凹参数(即 $-\Phi_k$ 的强凸参数)。

# 提纲

- 1 近似点算法
- 2 与增广拉格朗日函数法的关系
- 3 应用举例: LASSO问题
- 4 收敛性分析
- 5 Moreau-Yosida 正则化

# 收敛性分析

基本假设：

- $\psi$  是闭凸函数 (因此,  $\text{prox}_{t\psi}(x)$  唯一确定,  $\forall x$ )
- 最优值  $\psi^*$  有限且在  $x^*$  取到

结论：

$$\psi(x^{(k)}) - \psi^* \leq \frac{\|x^{(0)} - x^*\|_2^2}{2 \sum_{i=1}^k t_i} \quad \forall k \geq 1$$

- 这表明,  $\sum_i t_i \rightarrow \infty$  时收敛
- 若  $t_i$  固定或在一个正下界以上变化, 则收敛速率为  $1/k$
- $t_i$  可以任意选取, 然而邻近算子的计算代价依赖于  $t_i$

# 收敛性分析

## Proof.

对原问题应用近似点梯度法(即  $f(x) = 0$  的情形), 则下式在  $t > 0$  的情形下自然满足:

$$f(x - tG_t(x)) \leq f(x) + t\nabla f(x)^T G_t(x) + \frac{t}{2} \|G_t(x)\|_2^2$$

根据近似点梯度法中定理1的证明, 我们有

$$t_i (\psi(x^i) - \psi^*) \leq \frac{1}{2} (\|x^{i-1} - x^*\|_2^2 - \|x^i - x^*\|_2^2)$$

且  $\{\psi(x^i)\}$  是单调下降的序列. 因此

$$\left( \sum_{i=1}^k t_i \right) (\psi(x^k) - \psi^*) \leq \sum_{i=1}^k t_i (\psi(x^i) - \psi^*) \leq \frac{1}{2} \|x^0 - x^*\|_2^2$$



# 加速版本的近似点算法

**FISTA** (令  $f(x) = 0$ ): 取  $x^{(0)} = x^{(-1)}$  且对于  $k > 1$  有

$$x^{(k)} = \text{prox}_{t_k f} \left( x^{(k-1)} + \theta_k \frac{1 - \theta_{k-1}}{\theta_{k-1}} (x^{(k-1)} - x^{(k-2)}) \right)$$

第二类 **Nesterov** 加速算法: 取  $x^{(0)} = v^{(0)}$  且对于  $k \geq 1$

$$v^{(k)} = \text{prox}_{(t_k/\theta_k)f}(v^{(k-1)}), \quad x^{(k)} = (1 - \theta_k)x^{(k-1)} + \theta_k v^{(k)}$$

参数选择策略

- 固定步长:  $t_k = t$  以及  $\theta_k = 2/(k+1)$
- 变化步长: 选择任意的  $t_k > 0, \theta_1 = 1$ , 对于任意  $k > 1$ , 从下面的方程中解出  $\theta_k$

$$\frac{(1 - \theta_k)t_k}{\theta_k^2} = \frac{t_{k-1}}{\theta_{k-1}^2}$$

# 收敛性分析

基本假设：

- $\psi$  是闭凸函数 (因此,  $\text{prox}_{t\psi}(x)$  对于所有的  $x$  存在且唯一)
- 最优值  $\psi^*$  有限且在  $x^*$  取到

结论：

$$\psi(x^{(k)}) - \psi^* \leq \frac{2\|x^{(0)} - x^*\|_2^2}{(2\sqrt{t_1} + \sum_{i=2}^k \sqrt{t_i})^2}, \quad k \geq 1$$

- 这表明, 若  $\sum_i \sqrt{t_i} \rightarrow \infty$ , 则保证收敛
- 步长  $t_i$  取固定值或有正下界时, 其收敛速度可达到  $\mathcal{O}(\frac{1}{k^2})$  (实际上也需控制  $t_i$  上界以便子问题可快速求解)

# 收敛性分析

## Proof.

在  $f(x) = 0$  的情况下使用Nesterov加速算法中的证明:  
由于  $f(x) = 0$ , 对任意的  $t > 0$ ,

$$f(x) \leq f(y) + \nabla f(y)^T(x - y) + \frac{1}{2t} \|x - y\|_2^2, \quad \forall x, y$$

于是  $\psi(x^k) - \psi^* \leq \frac{\gamma_k^2}{2t_k} \|x^0 - x^*\|_2^2$  成立, 取固定步长  $t_k = t$  和  $\gamma_k = \frac{2}{k+1}$ ,

$$\frac{\gamma_k^2}{2t_k} = \frac{2}{(k+1)^2 t}$$

对于变步长,

$$\frac{\gamma_k^2}{2t_k} \leq \frac{2}{\left(2\sqrt{t_1} + \sum_{i=2}^k \sqrt{t_i}\right)^2}$$

分别将对应不等式代入 (23) 式就得到定理中的结论.



# 提纲

- 1 近似点算法
- 2 与增广拉格朗日函数法的关系
- 3 应用举例: LASSO问题
- 4 收敛性分析
- 5 Moreau-Yosida 正则化



# Moreau-Yosida 正则化

闭凸函数  $f$  的 Moreau-Yosida 正则化 (又称为 Moreau envelope) 定义为

$$\begin{aligned} f_{(t)}(x) &= \inf_u \left( f(u) + \frac{1}{2t} \|u - x\|_2^2 \right) \quad (t > 0) \\ &= f(\text{prox}_{tf}(x)) + \frac{1}{2t} \|\text{prox}_{tf}(x) - x\|_2^2 \end{aligned} \quad (11)$$

根据定义我们可以立刻得到：

- $f_{(t)}$  是凸函数 (一个以  $x, u$  为自变量的凸函数对  $u$  取下确界)
- $f_{(t)}$  的定义域为  $\mathbb{R}^n$  (回忆  $\text{prox}_{tf}(x)$  对于任意的  $x$  存在且唯一)

## 例子

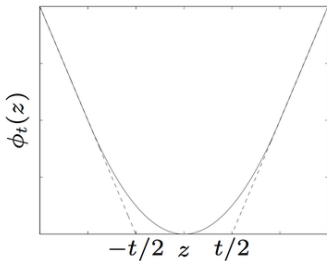
- 示性函数: 光滑化后的 $f$  是Euclid距离的平方

$$f(x) = I_C(x), \quad f_{(t)}(x) = \frac{1}{2t}d(x)^2$$

- $l_1$  范数: 光滑化后的函数为Huber损失函数

$$f(x) = \|x\|_1, \quad f_{(t)}(x) = \sum_{k=1}^n \phi_t(x_k)$$

$$\phi_t(z) = \begin{cases} z^2/(2t) & |z| \leq t \\ |z| - t/2 & |z| \geq t \end{cases}$$



# Moreau envelope的共轭函数

$$f_{(t)}(x) = \inf_u \left( f(u) + \frac{1}{2t} \|u - x\|_2^2 \right)$$

- $f_{(t)}$  是  $f(u)$  与  $\|v\|_2^2/(2t)$  的卷积下确界:

$$f_{(t)}(x) = \inf_{u+v=x} \left( f(u) + \frac{1}{2t} \|v\|_2^2 \right)$$

- $f_{(t)}$  的共轭是  $f(u)$  的共轭与  $\|v\|_2^2/(2t)$  的和:

$$(f_{(t)})^*(y) = f^*(y) + \frac{t}{2} \|y\|_2^2$$

- 因此,  $f_{(t)}$  的共轭是  $t$ -强凸的

# Moreau envelope的梯度

$$f_{(t)}(x) = \sup_y (x^T y - (f_{(t)})^*(y)) = \sup_y (x^T y - f^*(y) - \frac{t}{2} \|y\|_2^2)$$

- 最大值点  $y$  唯一且满足：

$$\frac{x}{t} - \partial \frac{f^*(y)}{t} - y = 0 \iff y = \operatorname{argmin}_u \left( \frac{f^*(u)}{t} + \frac{1}{2} \|u - \frac{x}{t}\|_2^2 \right)$$

- 同时有  $x \in \partial(f_{(t)})^*(y) \iff y \in \partial f_{(t)}(x)$ , 根据唯一性, 最大值点  $y$  即是  $f_{(t)}$  的梯度:

$$\nabla f_{(t)}(x) = \operatorname{prox}_{(1/t)f^*}(x/t) = \frac{1}{t}(x - \operatorname{prox}_{tf}(x))$$

- 梯度  $\nabla f_{(t)}$  为  $(1/t)$ -利普希茨连续  
(根据邻近算子性质:  $\|\operatorname{prox}_h(\mathbf{x}_1) - \operatorname{prox}_h(\mathbf{x}_2)\|_2 \leq \|\mathbf{x}_1 - \mathbf{x}_2\|_2$ )

# 近似点算法的解释

使用梯度法来最小化Moreau envelope

$$\min_x f_{(t)}(x) = \inf_u \left( f(u) + \frac{1}{2t} \|u - x\|_2^2 \right) \quad (12)$$

这是最小化  $f(x)$  优化问题的光滑化形式，且满足：

- 问题(12)的最优值点  $x$  同时也是原问题  $f$  的最小值点
- $f_{(t)}$  可微且梯度利普希茨连续 ( $L = 1/t$ )

我们可以使用固定步长： $t_k = 1/L = t$

$$x^{(k)} = x^{(k-1)} - t \nabla f_{(t)}(x^{(k-1)}) = \text{prox}_{tf}(x^{(k-1)})$$

这就是固定步长近似点算法的迭代格式

## 增广拉格朗日函数法的解释

$$\begin{aligned}(\hat{x}, \hat{y}) &= \operatorname{argmin}_{x, y} \left( f(x) + g(y) + \frac{t}{2} \|Ax - y + (1/t)z\|_2^2 \right) \\ z &= z + t(A\hat{x} - \hat{y})\end{aligned}$$

- 固定步长  $t$ ，对偶乘子更新即为对光滑化的对偶问题使用梯度下降
- 如果我们消去  $y$ ，关键变量  $x$  的更新可以理解为光滑化  $g$ ：

$$\hat{x} = \operatorname{argmin}_x \left( f(x) + g_{(1/t)}(Ax + (1/t)z) \right)$$

例子：最小化  $f(x) + \|Ax - b\|_1$

$$\hat{x} = \operatorname{argmin}_x \left( f(x) + \phi_{1/t}(Ax - b + (1/t)z) \right)$$

其中  $\phi_{1/t}$  表示将Huber损失函数应用到各分量

## proximal point algorithm and fast proximal point algorithm

- O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control and Optimization (1991)
- O. Güler, *New proximal point algorithms for convex minimization*, SIOPT (1992)
- O. Güler, *Augmented Lagrangian algorithm for linear programming*, JOTA (1992)

## augmented Lagrangian algorithm

- D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods* (1982)