

第五章 复合优化算法

修贤超

<https://xianchaoxiu.github.io>

- 5.1 近似点梯度法
- 5.2 Nesterov 加速算法
- 5.3 分块坐标下降法
- 5.4 交替方向乘子法

■ 考虑如下复合优化问题

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

□ $f(x)$ 为可微函数 (可能非凸)

□ $h(x)$ 可能为不可微函数

■ 定义 对于一个凸函数 h , 定义邻近算子为

$$\text{prox}_h(x) = \arg \min_u \left\{ h(u) + \frac{1}{2} \|u - x\|_2^2 \right\}$$

■ 定理 如果 h 为闭凸函数, 则对任意 x 有 $\text{prox}_h(x)$ 存在且唯一

■ **定理** 若 h 是适当的闭凸函数, 则

$$u = \text{prox}_h(x) \quad \Leftrightarrow \quad x - u \in \partial h(u)$$

证明 若 $u = \text{prox}_h(x)$, 则由最优性条件得 $0 \in \partial h(u) + (u - x)$, 因此 $x - u \in \partial h(u)$. 反之, 若 $x - u \in \partial h(u)$ 则由**次梯度**的定义可得到

$$h(v) \geq h(u) + (x - u)^\top (v - u), \quad \forall v \in \text{dom } h$$

两边同时加 $\frac{1}{2}\|v - x\|^2$, 即有

$$\begin{aligned} h(v) + \frac{1}{2}\|v - x\|^2 &\geq h(u) + (x - u)^\top (v - u) + \frac{1}{2}\|(v - u) - (x - u)\|^2 \\ &\geq h(u) + \frac{1}{2}\|u - x\|^2, \quad \forall v \in \text{dom } h \end{aligned}$$

根据定义可得 $u = \text{prox}_h(x)$

■ 给定 ℓ_1 范数 $h(x) = t\|x\|_1$, 则 $\text{prox}_{th}(x) = \text{sign}(x) \max\{|x| - t, 0\}$

证明 邻近算子 $u = \text{prox}_{th}(x)$ 的最优性条件为

$$x - u \in t\partial\|u\|_1 = \begin{cases} \{t\}, & u > 0 \\ [-t, t], & u = 0 \\ \{-t\}, & u < 0 \end{cases}$$

\Downarrow

$$u = \begin{cases} x - t, & x > t \\ x + t, & x < -t \\ 0, & x \in [-t, t] \end{cases}$$

- 给定 ℓ_2 范数 $h(x) = t\|x\|_2$, 则 $\text{prox}_{th}(x) = \begin{cases} (1 - \frac{t}{\|x\|_2})x, & \|x\|_2 \geq t \\ 0, & \text{其他} \end{cases}$

证明 邻近算子 $u = \text{prox}_{th}(x)$ 的最优性条件为

$$x - u \in t\partial\|u\|_2 = \begin{cases} \{\frac{tu}{\|u\|_2}\}, & u \neq 0 \\ \{w : \|w\|_2 \leq t\}, & u = 0 \end{cases}$$

\Downarrow

$$u = \begin{cases} x - \frac{tx}{\|x\|_2}, & \|x\|_2 > t \\ 0, & \|x\|_2 \leq t \end{cases}$$

■ 邻近算子的计算规则

- ▣ 变量的常数倍放缩以及平移 ($\lambda \neq 0$)

$$h(x) = g(\lambda x + a), \quad \text{prox}_h(x) = \frac{1}{\lambda} \left(\text{prox}_{\lambda^2 g}(\lambda x + a) - a \right)$$

- ▣ 函数（及变量）的常数倍放缩 ($\lambda > 0$)

$$h(x) = \lambda g\left(\frac{x}{\lambda}\right), \quad \text{prox}_h(x) = \lambda \text{prox}_{\lambda^{-1}g}\left(\frac{x}{\lambda}\right)$$

- ▣ 加上线性函数

$$h(x) = g(x) + a^\top x, \quad \text{prox}_h(x) = \text{prox}_g(x - a)$$

- 加上二次项 ($u > 0$)

$$h(x) = g(x) + \frac{u}{2}\|x - a\|_2^2, \quad \text{prox}_h(x) = \text{prox}_{\theta g}(\theta x + (1 - \theta)a)$$

其中 $\theta = \frac{1}{1+u}$

- 向量函数

$$h\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \varphi_1(x) + \varphi_2(y), \quad \text{prox}_h\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \text{prox}_{\varphi_1}(x) \\ \text{prox}_{\varphi_2}(y) \end{bmatrix}$$

- 设 C 为闭凸集, 则示性函数 I_C 的邻近算子为点 x 到 C 的投影 $\mathcal{P}_C(x)$

$$\begin{aligned}\operatorname{prox}_{I_C}(x) &= \arg \min_u \left\{ I_C(u) + \frac{1}{2} \|u - x\|^2 \right\} \\ &= \arg \min_{u \in C} \|u - x\|^2 \\ &= \mathcal{P}_C(x)\end{aligned}$$

- 几何意义

$$u = \mathcal{P}_C(x) \quad \Leftrightarrow \quad (x - u)^\top (z - u) \leq 0, \quad \forall z \in C$$

近似点梯度法

■ 考虑复合优化问题

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

■ 对于光滑部分 f 做梯度下降, 对于非光滑部分 h 使用邻近算子

=====

算法 近似点梯度法

- 1 给定函数 $f(x), h(x)$, 初始点 x^0
- 2 **while** 未达到收敛准则 **do**
- 3 $x^{k+1} = \text{prox}_{t_k h}(x^k - t_k \nabla f(x^k))$
- 4 **end while**

对近似点梯度法的理解

■ 把迭代公式展开

$$x^{k+1} = \text{prox}_{t_k h}(x^k - t_k \nabla f(x^k))$$

\Downarrow

$$\begin{aligned} x^{k+1} &= \arg \min_u \left\{ h(u) + \frac{1}{2t_k} \|u - x^k + t_k \nabla f(x^k)\|^2 \right\} \\ &= \arg \min_u \left\{ h(u) + f(x^k) + \nabla f(x^k)^\top (u - x^k) + \frac{1}{2t_k} \|u - x^k\|^2 \right\} \end{aligned}$$

■ 根据邻近算子与次梯度的关系, 可改写为

$$x^{k+1} = x^k - t_k \nabla f(x^k) - t_k g^k, \quad g^k \in \partial h(x^{k+1})$$

■ 对光滑部分做显式的梯度下降, 对非光滑部分做隐式的梯度下降

步长选取

- 当 f 为梯度 L -利普希茨连续函数时, 可取固定步长 $t_k = t \leq \frac{1}{L}$
- 当 L 未知时可使用线搜索准则

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{1}{2t_k} \|x^{k+1} - x^k\|^2$$

- BB 步长
- 可构造如下适用于近似点梯度法的非单调线搜索准则

$$\psi(x^{k+1}) \leq C^k - \frac{c_1}{2t_k} \|x^{k+1} - x^k\|^2$$

应用举例: LASSO 问题

- 考虑用近似点梯度法求解 LASSO 问题

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

- 令 $f(x) = \frac{1}{2} \|Ax - b\|^2$, $h(x) = \mu \|x\|_1$, 则

$$\begin{aligned}\nabla f(x) &= A^\top (Ax - b) \\ \text{prox}_{t_k h}(x) &= \text{sign}(x) \max\{|x| - t_k \mu, 0\}\end{aligned}$$

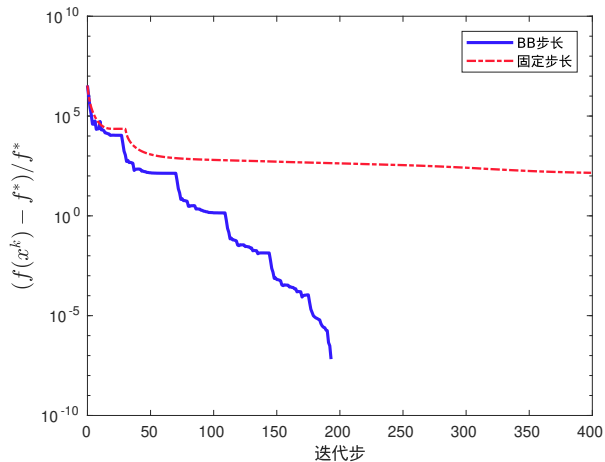
- 相应的迭代格式为

$$\begin{aligned}y^k &= x^k - t_k A^\top (Ax^k - b) \\ x^{k+1} &= \text{sign}(y^k) \max\{|y^k| - t_k \mu, 0\}\end{aligned}$$

即第一步做梯度下降, 第二步做收缩

应用举例: LASSO 问题

■ 使用 BB 步长加速收敛



应用举例：低秩矩阵恢复

■ 考虑低秩矩阵恢复模型

$$\min_{X \in \mathbb{R}^{m \times n}} \mu \|X\|_* + \frac{1}{2} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2$$

■ 令

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

■ 定义矩阵

$$P_{ij} = \begin{cases} 1, & (i,j) \in \Omega \\ 0, & \text{其他} \end{cases}$$

则

$$f(X) = \frac{1}{2} \|P \odot (X - M)\|_F^2$$

应用举例：低秩矩阵恢复

- 进一步可以得到

$$\begin{aligned}\nabla f(X) &= P \odot (X - M) \\ \text{prox}_{t_k h}(X) &= U \text{Diag}(\max\{|d| - t_k \mu, 0\}) V^\top\end{aligned}$$

- 得到近似点梯度法的迭代格式

$$\begin{aligned}Y^k &= X^k - t_k P \odot (X^k - M) \\ X^{k+1} &= \text{prox}_{t_k h}(Y^k)\end{aligned}$$

收敛性分析

■ **假设** 为了保证近似点梯度算法的收敛性

□ f 在 \mathbb{R}^n 上是凸的; ∇f 为 L -利普希茨连续, 即

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y$$

□ h 是适当的闭凸函数

□ 函数 $\psi(x) = f(x) + h(x)$ 的最小值 ψ^* 是有限的, 并且在点 x^* 处取到

■ **定理** 在假设下, 取定步长为 $t_k = t \in (0, \frac{1}{L}]$, 设 $\{x^k\}$ 为迭代产生序列, 则

$$\psi(x^k) - \psi^* \leq \frac{1}{2kt} \|x^0 - x^*\|^2$$

- 5.1 近似点梯度法
- 5.2 Nesterov 加速算法
- 5.3 分块坐标下降法
- 5.4 交替方向乘子法

典型问题形式

■ 考虑如下复合优化问题

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

□ $f(x)$ 是连续可微的凸函数, 且梯度是利普西茨连续的

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

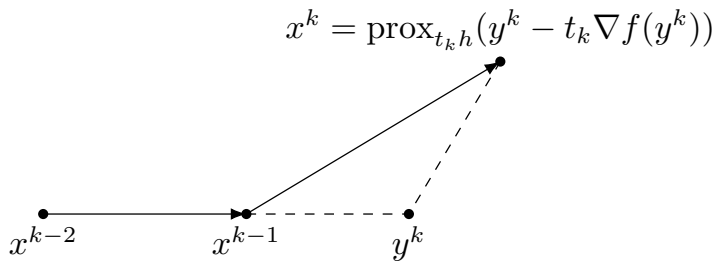
□ $h(x)$ 是适当的闭凸函数, 且邻近算子

$$\text{prox}_h(x) = \arg \min_{u \in \text{dom} h} \left\{ h(u) + \frac{1}{2} \|x - u\|^2 \right\}$$

■ 步长取常数 $t_k = 1/L$ 时, 近似点梯度法的收敛速度为 $\mathcal{O}(1/k)$

Nesterov 加速算法简史

- Nesterov 在 1983、1988、2005 提出了三种改进的一阶算法, 收敛速度 $\mathcal{O}\left(\frac{1}{k^2}\right)$
- Beck 和 Teboulle 在 2008 年提出了 FISTA 算法, 第一步沿着前两步的计算方向计算一个新点, 第二步在该新点处做一步近似点梯度迭代



算法 近似点梯度法

- 1 给定函数 $f(x), h(x)$, 初始点 x^0
- 2 **while** 未达到收敛准则 **do**
- 3 $x^{k+1} = \text{prox}_{t_k h}(x^k - t_k \nabla f(x^k))$
- 4 **end while**

=====

算法 7.2 FISTA 算法

- 1 输入 $x^0 = x^{-1} \in \mathbb{R}^n, k \leftarrow 1$
- 2 **while** 未达到收敛准则 **do**
- 3 计算 $y^k = x^{k-1} + \frac{k-2}{k+1}(x^{k-1} - x^{k-2})$
- 4 选取 $t_k = t \in (0, 1/L]$, 计算 $x^k = \text{prox}_{t_k h}(y^k - t_k \nabla f(y^k))$
- 5 $k \leftarrow k + 1$
- 6 **end while**

FISTA 的等价形式

算法 FISTA 算法的等价变形

- 1 输入 $v^0 = x^0 \in \mathbb{R}^n, k \leftarrow 1$
- 2 **while** 未达到收敛准则 **do**
- 3 计算 $y^k = (1 - \gamma_k)x^{k-1} + \gamma_k v^{k-1}$
- 4 选取 t_k , 计算 $x^k = \text{prox}_{t_k h}(y^k - t_k \nabla f(y^k))$
- 5 计算 $v^k = x^{k-1} + \frac{1}{\gamma_k}(x^k - x^{k-1})$
- 6 $k \leftarrow k + 1$
- 7 **end while**

第二类 Nesterov 加速算法

■ 第二类 Nesterov 加速算法

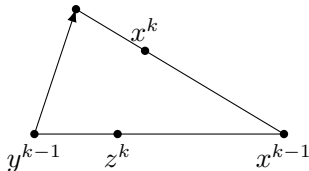
$$z^k = (1 - \gamma_k)x^{k-1} + \gamma_k y^{k-1}$$

$$y^k = \text{prox}_{(t_k/\gamma_k)h} \left(y^{k-1} - \frac{t_k}{\gamma_k} \nabla f(z^k) \right)$$

$$x^k = (1 - \gamma_k)x^{k-1} + \gamma_k y^k$$

■ 三个序列 $\{x^k\}$, $\{y^k\}$ 和 $\{z^k\}$ 都可以保证在定义域内

$$y^k = \text{prox}_{(t_k/\gamma_k)h} (y^{k-1} - (t_k/\gamma_k) \nabla f(z^k))$$



第三类 Nesterov 加速算法

■ 第三类 Nesterov 加速算法

$$z^k = (1 - \gamma_k)x^{k-1} + \gamma_k y^{k-1}$$

$$y^k = \text{prox}_{(t_k \sum_{i=1}^k 1/\gamma_i)h} \left(-t_k \sum_{i=1}^k \frac{1}{\gamma_i} \nabla f(z^i) \right)$$

$$x^k = (1 - \gamma_k)x^{k-1} + \gamma_k y^k$$

■ 计算 y^k 时需要利用全部已有的 $\{\nabla f(z^i)\}, i = 1, 2, \dots, k$

■ 取 $\gamma_k = \frac{2}{k+1}$, $t_k = \frac{1}{L}$ 时, 也有 $\mathcal{O}\left(\frac{1}{k^2}\right)$ 的收敛速度

针对非凸问题的 Nesterov 加速算法

- 考虑 $f(x)$ 是非凸函数, 但可微且梯度是利普希茨连续
- 非凸复合优化问题的加速梯度法框架

$$\begin{aligned}z^k &= \gamma_k y^{k-1} + (1 - \gamma_k) x^{k-1} \\y^k &= \text{prox}_{\lambda_k h}(y^{k-1} - \lambda_k \nabla f(z^k)) \\x^k &= \text{prox}_{t_k h}(z^k - t_k \nabla f(z^k))\end{aligned}$$

- 当 λ_k 和 t_k 取特定值时, 它等价于第二类 Nesterov 加速算法
- 当 f 为凸函数, 收敛速度为 $\mathcal{O}\left(\frac{1}{k^2}\right)$
- 当 f 为非凸函数, 收敛速度为 $\mathcal{O}\left(\frac{1}{k}\right)$

应用举例: LASSO 问题求解

- 考虑 LASSO 问题

$$\min_x \quad \frac{1}{2} \|Ax - b\|_2^2 + \mu \|x\|_1$$

- FISTA 算法可以由下面的迭代格式给出

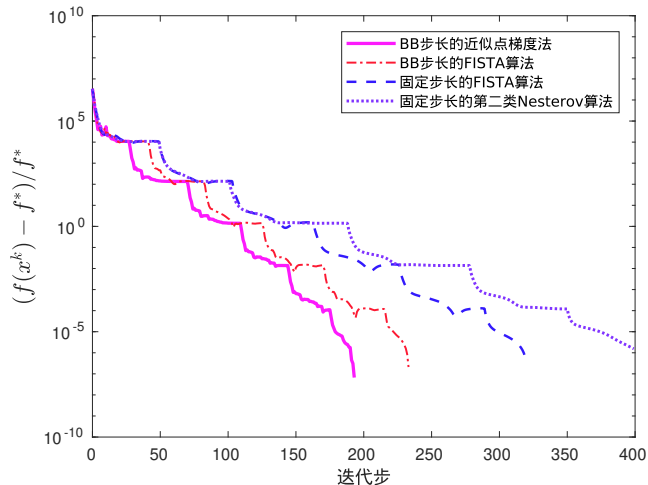
$$y^k = x^{k-1} + \frac{k-2}{k+1} (x^{k-1} - x^{k-2})$$

$$w^k = y^k - t_k A^\top (Ay^k - b)$$

$$x^k = \text{sign}(w^k) \max\{|w^k| - t_k \mu, 0\}$$

应用举例: LASSO 问题求解

- 取 $\mu = 10^{-3}$, 步长 $t = \frac{1}{L}$, 其中 $L = \lambda_{\max}(A^{\top}A)$



收敛性分析

- **定理** 在假设 7.1 下, 当用 FISTA 算法求解凸复合优化问题时, 若取固定步长 $t_k = 1/L$, 则

$$\psi(x^k) - \psi(x^*) \leq \frac{2L}{(k+1)^2} \|x^0 - x^*\|^2$$

- **推论** 在假设下, 当用 FISTA 算法求解凸复合优化问题时, 若迭代点 x^k, y^k , 步长 t_k 以及组合系数 γ_k 满足一定条件, 则

$$\psi(x^k) - \psi(x^*) \leq \frac{C}{k^2}$$

其中 C 仅与函数 f 和初始点 x^0 的选取有关

- 采用线搜索的 FISTA 算法具有 $\mathcal{O}\left(\frac{1}{k^2}\right)$ 的收敛速度

- 5.1 近似点梯度法
- 5.2 Nesterov 加速算法
- 5.3 分块坐标下降法
- 5.4 交替方向乘子法

问题形式

■ 考虑具有如下形式的问题

$$\min_{x \in \mathcal{X}} \quad F(x_1, x_2, \dots, x_s) = f(x_1, x_2, \dots, x_s) + \sum_{i=1}^s r_i(x_i)$$

- f 是关于 x 的可微函数, 但不一定凸
- $r_i(x_i)$ 关于 x_i 是适当的闭凸函数, 但不一定可微

■ 挑战和难点

- 在非凸问题上, 很多针对凸问题设计的算法通常会失效
- 目标函数的整体结构十分复杂, 变量的更新需要很大计算量

- 例 设参数 $x = (x_1, x_2, \dots, x_G) \in \mathbb{R}^p$, 分组 LASSO 模型

$$\min_x \quad \frac{1}{2n} \|b - Ax\|_2^2 + \lambda \sum_{i=1}^G \sqrt{p_i} \|x_i\|_2$$

- 例 设 $b \in \mathbb{R}^m$ 是已知的观测向量, 低秩矩阵恢复模型

$$\min_{X,Y} \quad \frac{1}{2} \|\mathcal{A}(XY) - b\|_2^2 + \alpha \|X\|_F^2 + \beta \|Y\|_F^2$$

- 例 设 $M \in \mathbb{R}^{m \times n}$ 是已知的矩阵, 非负矩阵分解模型

$$\min_{X,Y \geq 0} \quad \frac{1}{2} \|XY - M\|_F^2 + \alpha r_1(X) + \beta r_2(Y)$$

变量更新方式

- 按照 x_1, x_2, \dots, x_s 的次序依次固定其他 $(s - 1)$ 块变量极小化 F

- 辅助函数

$$f_i^k(\mathbf{x}_i) = f(x_1^k, \dots, x_{i-1}^k, \mathbf{x}_i, x_{i+1}^{k-1}, \dots, x_s^{k-1})$$

- 在每一步更新中, 通常使用以下三种更新格式之一

$$x_i^k = \arg \min_{x_i \in \mathcal{X}_i^k} \left\{ f_i^k(x_i) + r_i(x_i) \right\} \quad (1)$$

$$x_i^k = \arg \min_{x_i \in \mathcal{X}_i^k} \left\{ f_i^k(x_i) + \frac{L_i^{k-1}}{2} \|x_i - x_i^{k-1}\|_2^2 + r_i(x_i) \right\} \quad (2)$$

$$x_i^k = \arg \min_{x_i \in \mathcal{X}_i^k} \left\{ \langle \hat{g}_i^k, x_i - \hat{x}_i^{k-1} \rangle + \frac{L_i^{k-1}}{2} \|x_i - \hat{x}_i^{k-1}\|_2^2 + r_i(x_i) \right\} \quad (3)$$

算法 分块坐标下降法

```
1 选择两组初始点  $(x_1^{-1}, x_2^{-1}, \dots, x_s^{-1}) = (x_1^0, x_2^0, \dots, x_s^0)$ 
2 for  $k = 1, 2, \dots$  do
3   for  $i = 1, 2, \dots$  do
4     使用格式 (1)、(2)、(3) 更新  $x_i^k$ 
5   end for
6   if 满足停机条件 then
7     返回  $(x_1^k, x_2^k, \dots, x_s^k)$ , 算法终止
8   end if
9 end for
```

算法格式

- BCD 算法的子问题可采用三种不同的更新格式, 这三种格式可能会产生不同的迭代序列, 可能会收敛到不同的解, 数值表现也不相同
- 格式 (1) 是最直接的更新方式, 保证整个迭代过程的目标函数值是下降的. 然而由于 f 的形式复杂, 子问题求解难度较大
- 在收敛性方面, 格式 (1) 在强凸问题上可保证目标函数收敛到极小值, 但在非凸问题上不一定收敛

$$x_i^k = \arg \min_{x_i \in \mathcal{X}_i^k} \left\{ f_i^k(x_i) + r_i(x_i) \right\}$$

算法格式

- 格式 (2) (3) 是对格式 (1) 的修正, 不保证迭代过程目标函数的单调性, 但可以改善收敛性结果.
- 格式 (3) 实质上为目标函数的一阶泰勒展开近似, 在一些测试问题上有更好的表现, 可能的原因是使用一阶近似可以避开一些局部极小值点
- 格式 (3) 的计算量很小, 比较容易实现

$$x_i^k = \arg \min_{x_i \in \mathcal{X}_i^k} \left\{ f_i^k(x_i) + r_i(x_i) \right\}$$

$$x_i^k = \arg \min_{x_i \in \mathcal{X}_i^k} \left\{ f_i^k(x_i) + \frac{L_i^{k-1}}{2} \|x_i - x_i^{k-1}\|_2^2 + r_i(x_i) \right\}$$

$$x_i^k = \arg \min_{x_i \in \mathcal{X}_i^k} \left\{ \langle \hat{g}_i^k, x_i - \hat{x}_i^{k-1} \rangle + \frac{L_i^{k-1}}{2} \|x_i - \hat{x}_i^{k-1}\|_2^2 + r_i(x_i) \right\}$$

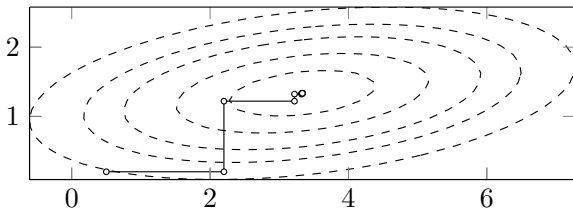
- 考虑二元二次函数的优化问题

$$\min f(x, y) = x^2 - 2xy + 10y^2 - 4x - 20y$$

- 采用格式 (1) 的分块坐标下降法

$$x^{k+1} = 2 + y^k \quad y^{k+1} = 1 + \frac{x^{k+1}}{10}$$

- 初始点为 $(x, y) = (0.5, 0.2)$ 时的迭代点轨迹



不收敛反例

■ 考虑

$$F(x_1, x_2, x_3) = -x_1x_2 - x_2x_3 - x_3x_1 + \sum_{i=1}^3 [(x_i - 1)_+^2 + (-x_i - 1)_+^2]$$

■ 设 $\varepsilon > 0$, 初始点取为

$$x^0 = \left(-1 - \varepsilon, 1 + \frac{\varepsilon}{2}, -1 - \frac{\varepsilon}{4}\right)$$

容易验证迭代序列满足

$$x^k = (-1)^k \cdot (-1, 1, -1) + \left(-\frac{1}{8}\right)^k \cdot \left(-\varepsilon, \frac{\varepsilon}{2}, -\frac{\varepsilon}{4}\right)$$

■ 迭代序列有两个聚点 $(-1, 1, -1)$ 与 $(1, -1, 1)$, 但都不是 F 的稳定点

应用举例: LASSO 问题求解

- 使用分块坐标下降法来求解 LASSO 问题

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

- 将自变量 x 记为 $x = [x_i \quad \bar{x}_i^\top]^\top$, 矩阵 A 记为 $A = [a_i \quad \bar{A}_i]$

- 应用格式 (1), 替换 $c_i = b - \bar{A}_i \bar{x}_i$, 原问题等价于

$$\min_{x_i} \quad f_i(x_i) = \mu |x_i| + \frac{1}{2} \|a_i\|^2 x_i^2 - a_i^\top c_i x_i$$

- 可直接写出最小值点

$$x_i^k = \arg \min_{x_i} f_i(x_i) = \begin{cases} \frac{a_i^\top c_i - \mu_i}{\|a_i\|^2}, & a_i^\top c_i > \mu \\ \frac{a_i^\top c_i + \mu_i}{\|a_i\|^2}, & a_i^\top c_i < -\mu \\ 0, & \text{其他} \end{cases}$$

应用举例：非负矩阵分解

- 考虑最基本的非负矩阵分解问题

$$\min_{X, Y \geq 0} f(X, Y) = \frac{1}{2} \|XY - M\|_F^2$$

- 计算梯度

$$\frac{\partial f}{\partial X} = (XY - M)Y^\top, \quad \frac{\partial f}{\partial Y} = X^\top(XY - M)$$

- 应用格式 (3), 当 $r_i(X)$ 为凸集示性函数时, 得到

$$\begin{aligned} X^{k+1} &= \max\{X^k - t_k^x(X^k Y^k - M)(Y^k)^\top, 0\} \\ Y^{k+1} &= \max\{Y^k - t_k^y(X^k)^\top(X^k Y^k - M), 0\} \end{aligned}$$

- 5.1 近似点梯度法
- 5.2 Nesterov 加速算法
- 5.3 分块坐标下降法
- 5.4 交替方向乘子法

典型问题形式

- 考虑如下凸问题

$$\begin{aligned} \min_{x_1, x_2} \quad & f_1(x_1) + f_2(x_2) \\ \text{s.t.} \quad & A_1 x_1 + A_2 x_2 = b \end{aligned} \tag{4}$$

- f_1, f_2 是适当的闭凸函数, 但不要求是光滑的
- 目标函数可以分成彼此分离的两块, 但是变量被线性约束结合在一起

问题形式举例

■ 例 可以分成两块人无约束优化问题

$$\min_x f_1(x) + f_2(x)$$

引入一个新的变量 z 并令 $x = z$, 将问题转化为

$$\begin{aligned} \min_{x,z} \quad & f_1(x) + f_2(z) \\ \text{s.t.} \quad & x - z = 0 \end{aligned}$$

■ 例 带线性变换的无约束优化问题

$$\min_x f_1(x) + f_2(Ax)$$

引入一个新的变量 z , 令 $z = Ax$, 则问题变为

$$\begin{aligned} \min_{x,z} \quad & f_1(x) + f_2(z) \\ \text{s.t.} \quad & Ax - z = 0 \end{aligned}$$

- 例 凸集 $C \subset \mathbb{R}^n$ 上的约束优化问题

$$\begin{array}{ll}\min_x & f(x) \\ \text{s.t.} & Ax \in C\end{array}$$

引入约束 $z = Ax$, 那么问题转化为

$$\begin{array}{ll}\min_{x,z} & f(x) + I_C(z) \\ \text{s.t.} & Ax - z = 0\end{array}$$

问题形式举例

■ 例 全局一致性问题

$$\min_x \sum_{i=1}^N \phi_i(x)$$

令 $x = z$, 并将 x 复制 N 份, 分别为 x_i , 那么问题转化为

$$\begin{aligned} \min_{x_i, z} \quad & \sum_{i=1}^N \phi_i(x_i) \\ \text{s.t.} \quad & x_i - z = 0, \quad i = 1, 2, \dots, N \end{aligned}$$

增广拉格朗日函数法

- 首先写出问题 (4) 的增广拉格朗日函数

$$L_{\rho}(x_1, x_2, y) = f_1(x_1) + f_2(x_2) + y^{\top}(A_1x_1 + A_2x_2 - b) + \frac{\rho}{2}\|A_1x_1 + A_2x_2 - b\|_2^2$$

- 增广拉格朗日函数法为如下更新

$$(x_1^{k+1}, x_2^{k+1}) = \arg \min_{x_1, x_2} L_{\rho}(x_1, x_2, y^k) \\ y^{k+1} = y^k + \tau \rho (A_1x_1^{k+1} + A_2x_2^{k+1} - b)$$

交替方向乘子法

- Alternating direction method of multipliers, ADMM
- 迭代格式如下

$$\begin{aligned}(x_1^{k+1}, x_2^{k+1}) &= \arg \min_{x_1, x_2} L_\rho(x_1, x_2, y^k) \\ y^{k+1} &= y^k + \tau \rho(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b) \\ &\Downarrow\end{aligned}$$

$$\begin{aligned}x_1^{k+1} &= \arg \min_{x_1} L_\rho(x_1, x_2^k, y^k) \\ x_2^{k+1} &= \arg \min_{x_2} L_\rho(x_1^{k+1}, x_2, y^k) \\ y^{k+1} &= y^k + \tau \rho(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b)\end{aligned}$$

原问题最优性条件

- 问题 (4) 的拉格朗日函数为

$$L(x_1, x_2, y) = f_1(x_1) + f_2(x_2) + y^\top (A_1 x_1 + A_2 x_2 - b)$$

- 若 x_1^*, x_2^* 为问题 (4) 的最优解, y^* 为对应的拉格朗日乘子, 则满足

$$0 \in \partial_{x_1} L(x_1^*, x_2^*, y^*) = \partial f_1(x_1^*) + A_1^\top y^* \quad (5a)$$

$$0 \in \partial_{x_2} L(x_1^*, x_2^*, y^*) = \partial f_2(x_2^*) + A_2^\top y^* \quad (5b)$$

$$A_1 x_1^* + A_2 x_2^* = b \quad (5c)$$

- 条件 (5c) 称为原始可行性条件
- 条件 (5a) 和条件 (5b) 称为对偶可行性条件

- 关于 x_2 的更新步骤

$$x_2^k = \arg \min_x \left\{ f_2(x) + \frac{\rho}{2} \|A_1 x_1^k + A_2 x - b + \frac{y^{k-1}}{\rho}\|^2 \right\}$$

- 根据最优性条件推出

$$0 \in \partial f_2(x_2^k) + A_2^\top [y^{k-1} + \rho(A_1 x_1^k + A_2 x_2^k - b)]$$

- 当 $\tau = 1$ 时知

$$0 \in \partial f_2(x_2^k) + A_2^\top y^k$$

■ 关于 x_1 的更新公式

$$x_1^k = \arg \min_x \left\{ f_1(x) + \frac{\rho}{2} \|A_1 x + A_2 x_2^{k-1} - b + \frac{y^{k-1}}{\rho}\|^2 \right\}$$

■ 假设子问题能精确求解, 根据最优性条件

$$0 \in \partial f_1(x_1^k) + A_1^\top [\rho(A_1 x_1^k + A_2 x_2^{k-1} - b) + y^{k-1}]$$

■ 当 $\tau = 1$ 时知

$$0 \in \partial f_1(x_1^k) + A_1^\top (y^k + \rho A_2 (x_2^{k-1} - x_2^k))$$

■ 对比条件 (5a)

$$0 \in \partial f_1(x_1^*) + A_1^\top y^*$$

$$\Downarrow$$

$$0 \in \partial f_1(x_1^k) + A_1^\top (y^k + \rho A_2(x_2^{k-1} - x_2^k))$$

■ 当 x_2 更新取到精确解且 $\tau = 1$ 时, 判断 ADMM 是否收敛只需要检测

□ 原始可行性

$$0 \approx \|r^k\| = \|A_1 x_1^k + A_2 x_2^k - b\|$$

□ 对偶可行性

$$0 \approx \|s^k\| = \|A_1^\top A_2(x_2^{k-1} - x_2^k)\|$$

- 考虑第一个子问题

$$\min_{x_1} f_1(x_1) + \frac{\rho}{2} \|A_1 x_1 - v^k\|^2$$

- 当子问题目标函数可微时, 线性化为

$$x_1^{k+1} = \arg \min_{x_1} \left\{ (\nabla f_1(x_1^k) + \rho A_1^\top (A_1 x_1^k - v^k))^\top x_1 + \frac{1}{2\eta_k} \|x_1 - x^k\|_2^2 \right\}$$

这等价于做一步**梯度下降**

- 当目标函数不可微时, 可以考虑只将二次项线性化

$$x_1^{k+1} = \arg \min_{x_1} \left\{ f_1(x_1) + \rho (A_1^\top (A_1 x_1^k - v^k))^\top x_1 + \frac{1}{2\eta_k} \|x_1 - x^k\|_2^2 \right\}$$

这等价于做一步**近似点梯度步**

- 考虑目标函数中含二次函数

$$f_1(x_1) = \frac{1}{2} \|Cx_1 - d\|_2^2$$

\Downarrow

$$(C^\top C + \rho A_1^\top A_1)x_1 = C^\top d + \rho A_1^\top v^k$$

- 首先对 $C^\top C + \rho A_1^\top A_1$ 进行 Cholesky 分解并缓存分解的结果, 在每步迭代中只需求解简单的三角形方程组
- 当 $C^\top C + \rho A_1^\top A_1$ 一部分容易求逆, 另一部分是低秩的情形时, 可以用 **SMW 公式** 来求逆

优化转移

- 为了方便求解子问题, 可以用一个性质好的矩阵 D 近似二次项 $A_1^\top A_1$, 即

$$x_1^{k+1} = \arg \min_{x_1} \left\{ f_1(x_1) + \frac{\rho}{2} \|A_1 x_1 - v^k\|_2^2 \right\}$$

\Downarrow

$$x_1^{k+1} = \arg \min_{x_1} \left\{ f_1(x_1) + \frac{\rho}{2} \|A_1 x_1 - v^k\|_2^2 + \frac{\rho}{2} (x_1 - x^k)^\top (D - A_1^\top A_1) (x_1 - x^k) \right\}$$

- 通过选取合适的 D , 优化转移简化子问题更容易计算
- 当 $D = \frac{\eta_k}{\rho} I$ 时, 优化转移等价于做单步的近似点梯度步

二次罚项系数的动态调节

- 二次罚项系数 ρ 太大会导致原始可行性 $\|r^k\|$ 下降很快, 但是对偶可行性 $\|s^k\|$ 下降很慢. 二次罚项系数 ρ 太小, 则会有相反的效果
- 动态调节惩罚系数 ρ 的大小, 使得原始可行性和对偶可行性能够以比较一致的速度下降到零

$$\rho^{k+1} = \begin{cases} \gamma_p \rho^k, & \|r^k\| > \mu \|s^k\| \\ \rho^k / \gamma_d, & \|s^k\| > \mu \|r^k\| \\ \rho^k, & \text{其他} \end{cases}$$

- 常见的选择为 $\mu = 10, \gamma_p = \gamma_d = 2$

多块问题的 ADMM

■ 考虑有多块变量的情形

$$\begin{aligned} \min_{x_1, x_2, \dots, x_N} \quad & f_1(x_1) + f_2(x_2) + \dots + f_N(x_N) \\ \text{s.t.} \quad & A_1 x_1 + A_2 x_2 + \dots + A_N x_N = b \end{aligned}$$

■ 多块 ADMM 迭代格式为

$$\begin{aligned} x_1^{k+1} &= \arg \min_x L_\rho(x, x_2^k, \dots, x_N^k, y^k) \\ x_2^{k+1} &= \arg \min_x L_\rho(x_1^{k+1}, x, \dots, x_N^k, y^k) \\ &\dots\dots\dots \\ x_N^{k+1} &= \arg \min_x L_\rho(x_1^{k+1}, x_2^{k+1}, \dots, x, y^k) \\ y^{k+1} &= y^k + \tau \rho(A_1 x_1^{k+1} + A_2 x_2^{k+1} + \dots + A_N x_N^{k+1} - b) \end{aligned}$$

其中 $\tau \in (0, (\sqrt{5} + 1)/2)$ 为步长参数

应用举例: LASSO 问题

- 考虑 LASSO 问题

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

- 转换为标准问题形式

$$\begin{aligned} \min_{x,z} \quad & \frac{1}{2} \|Ax - b\|^2 + \mu \|z\|_1 \\ \text{s.t.} \quad & x = z \end{aligned}$$

- 交替方向乘子法迭代格式为

$$\begin{aligned} x^{k+1} &= \arg \min_x \left\{ \frac{1}{2} \|Ax - b\|^2 + \frac{\rho}{2} \|x - z^k + y^k / \rho\|_2^2 \right\} \\ &= (A^\top A + \rho I)^{-1} (A^\top b + \rho z^k - y^k) \end{aligned}$$

- 交替方向乘子法迭代格式为

$$\begin{aligned} z^{k+1} &= \arg \min_z \left\{ \mu \|z\|_1 + \frac{\rho}{2} \|x^{k+1} - z + y^k / \rho\|^2 \right\} \\ &= \text{prox}_{(\mu/\rho)\|\cdot\|_1}(x^{k+1} + y^k / \rho) \\ y^{k+1} &= y^k + \tau \rho (x^{k+1} - z^{k+1}) \end{aligned}$$

- 在求解 x 迭代时, 可以使用固定的罚因子 ρ , 缓存矩阵 $A^\top A + \rho I$ 的初始分解
- 主要运算量来自更新 x 变量时求解线性方程组, 复杂度为 $O(n^3)$

应用举例: LASSO 问题

- 考虑 LASSO 问题的对偶问题

$$\begin{aligned} \min \quad & b^\top y + \frac{1}{2} \|y\|^2 \\ \text{s.t.} \quad & \|A^\top y\|_\infty \leq \mu \end{aligned}$$

- 引入约束 $A^\top y + z = 0$, 可以得到如下等价问题

$$\begin{aligned} \min \quad & \underbrace{b^\top y + \frac{1}{2} \|y\|^2}_{f(y)} + \underbrace{I_{\|z\|_\infty \leq \mu}(z)}_{h(z)} \\ \text{s.t.} \quad & A^\top y + z = 0 \end{aligned}$$

- 对约束 $A^\top y + z = 0$ 引入乘子 x , 对偶问题的增广拉格朗日函数

$$L_\rho(y, z, x) = b^\top y + \frac{1}{2} \|y\|^2 + I_{\|z\|_\infty \leq \mu}(z) - x^\top (A^\top y + z) + \frac{\rho}{2} \|A^\top y + z\|^2$$

应用举例: LASSO 问题

- 当固定 y, x 时, 对 z 的更新即向无穷范数球 $\{z \mid \|z\|_\infty \leq \mu\}$ 做欧几里得投影, 即将每个分量截断在区间 $[-\mu, \mu]$
- 当固定 z, x 时, 对 y 的更新即求解线性方程组

$$(I + \rho AA^\top)y = A(x^k - \rho z^{k+1}) - b$$

- ADMM 迭代格式为

$$\begin{aligned} z^{k+1} &= \mathcal{P}_{\|z\|_\infty \leq \mu}(x^k / \rho - A^\top y^k) \\ y^{k+1} &= (I + \rho AA^\top)^{-1}(A(x^k - \rho z^{k+1}) - b) \\ x^{k+1} &= x^k - \tau \rho (A^\top y^{k+1} + z^{k+1}) \end{aligned}$$

- 由于 $m \ll n$, 求解 y 更新的线性方程组需要的计算量是 $O(m^3)$

应用举例：矩阵分离问题

■ 考虑矩阵分离问题

$$\begin{aligned} \min_{X,S} \quad & \|X\|_* + \mu\|S\|_1 \\ \text{s.t.} \quad & X + S = M \end{aligned}$$

■ 引入乘子 Y 得到增广拉格朗日函数

$$L_\rho(X, S, Y) = \|X\|_* + \mu\|S\|_1 + \langle Y, X + S - M \rangle + \frac{\rho}{2}\|X + S - M\|_F^2$$

■ 对于 X 子问题

$$\begin{aligned} X^{k+1} &= \arg \min_X L_\rho(X, S^k, Y^k) \\ &= \arg \min_X \left\{ \|X\|_* + \frac{\rho}{2} \left\| X + S^k - M + \frac{Y^k}{\rho} \right\|_F^2 \right\} \\ &= \arg \min_X \left\{ \frac{1}{\rho} \|X\|_* + \frac{1}{2} \left\| X + S^k - M + \frac{Y^k}{\rho} \right\|_F^2 \right\} \\ &= U \text{Diag}(\text{prox}_{(1/\rho)\|\cdot\|_1}(\sigma(A))) V^\top \end{aligned}$$

其中 $A = M - S^k - \frac{Y^k}{\rho}$, $\sigma(A)$ 为 A 的所有非零奇异值构成的向量并且 $U \text{Diag}(\sigma(A)) V^\top$ 为 A 的约化奇异值分解

■ 对于 S 子问题

$$\begin{aligned} S^{k+1} &= \arg \min_S L_\rho(X^{k+1}, S, Y^k) \\ &= \arg \min_S \left\{ \mu \|S\|_1 + \frac{\rho}{2} \|X^{k+1} + S - M + \frac{Y^k}{\rho}\|_F^2 \right\} \\ &= \text{prox}_{(\mu/\rho)\|\cdot\|_1} \left(M - X^{k+1} - \frac{Y^k}{\rho} \right) \end{aligned}$$

■ 交替方向乘子法的迭代格式为

$$\begin{aligned} X^{k+1} &= U \text{Diag}(\text{prox}_{(1/\rho)\|\cdot\|_1}(\sigma(A))) V^\top \\ S^{k+1} &= \text{prox}_{(\mu/\rho)\|\cdot\|_1} \left(M - L^{k+1} - \frac{Y^k}{\rho} \right) \\ Y^{k+1} &= Y^k + \tau \rho (X^{k+1} + S^{k+1} - M) \end{aligned}$$

应用举例：全局一致性优化问题

■ 考虑全局一致性优化问题

$$\begin{aligned} \min_{x_i, z} \quad & \sum_{i=1}^N \phi_i(x_i) \\ \text{s.t.} \quad & x_i - z = 0, \quad i = 1, 2, \dots, N \end{aligned}$$

■ 增广拉格朗日函数为

$$L_\rho(x_1, \dots, x_N, z, y_1, \dots, y_N) = \sum_{i=1}^N \phi_i(x_i) + \sum_{i=1}^N y_i^\top (x_i - z) + \frac{\rho}{2} \sum_{i=1}^N \|x_i - z\|^2$$

■ 固定 z^k, y_i^k , 更新 x_i 的公式为

$$x_i^{k+1} = \arg \min_x \left\{ \phi_i(x) + \frac{\rho}{2} \|x - z^k + y_i^k / \rho\|^2 \right\}$$

应用举例：全局一致性优化问题

- 在一般情况下更新 x_i 的表达式为

$$x_i^{k+1} = \text{prox}_{\phi_i/\rho}(z^k - y_i^k/\rho)$$

- 固定 x_i^{k+1}, y_i^k , 关于 z 可以直接写出显式解

$$z^{k+1} = \frac{1}{N} \sum_{i=1}^N (x_i^{k+1} + y_i^k/\rho)$$

- 交替方向乘子法迭代格式为

$$x_i^{k+1} = \text{prox}_{\phi_i/\rho}(z^k - y_i^k/\rho), \quad i = 1, 2, \dots, N$$

$$z^{k+1} = \frac{1}{N} \sum_{i=1}^N (x_i^{k+1} + y_i^k/\rho)$$

$$y_i^{k+1} = y_i^k + \tau\rho(x_i^{k+1} - z^{k+1}), \quad i = 1, 2, \dots, N$$

Q&A

Thank you!

感谢您的聆听和反馈