BCD & ADMM 算法收敛理论

Lecture 5: Block Coordinate Descent (BCD) 算法

罗自炎

北京交通大学数统学院

E-mail: zyluo@bjtu.edu.cn

参考资料

- 教材与参考文献：
  - 最优化：建模、算法与理论
  - Bolte et al., MP 2014
  - Fazel et al., SIMAX 2013
  - Rockafellar & Wets, Variational Analysis
- 致谢： 北京大学文再文教授; 清华大学张立平教授

Outline of BCD

- 问题描述
- 分块坐标下降法
- 应用举例

BCD算法求解的优化模型

▶典型优化问题形式:

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

- 可行域：$\mathcal{X}$
- 决策变量: $x=(x_1,x_2,\cdots,x_s), x_i\in\mathbb{R}^{n_i}, \sum_{i=1}^{s}n_i=n$
- 目标函数：f是关于x的可微函数(coupled term)，$r_i(x_i)$关于x_i是适当的闭凸函数(separable terms)，但不一定可微.
- 求解该问题的难点在于如何利用分块结构处理不可分的函数f.

典型问题举例

- LASSO 模型 待估参数$x=(x_1,x_2,\cdots,x_p)^T\in\mathbb{R}^p$ 可以分成p 组

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

- 组LASSO 模型 待估参数$x=(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_G)^T\in\mathbb{R}^p$ 可以分成G 组, 且$\{\mathbf{x}_i\}_{i=1}^G$中只有少数的非零向量.

$$\min_x \quad \frac{1}{2n}\|b-Ax\|_2^2+\lambda\sum_{i=1}^G\|\mathbf{x}_i\|_2.$$

- *K*-means 聚类问题 给定p维空间中n个数据点$a_1, a_2, \cdots, a_n$，聚类问题就是要寻找k个不相交的非空集合$S_1, S_2, \ldots, S_k$，使得

$$\{a_1, a_2, \cdots, a_n\} = S_1 \cup S_2 \cup \cdots \cup S_k,$$

并且使得组内距离平方和最小，即

$$\begin{aligned} \min_{S_1, S_2, \cdots, S_k} \quad & \textstyle\sum_{i=1}^{k} \sum_{a \in S_i} \|a - c_i\|^2, \\ \text{s.t.} \quad & S_1 \cup S_2 \cup \cdots \cup S_k = \{a_1, a_2, \ldots, a_n\}, \\ & S_i \cap S_j = \emptyset, \quad \forall i \neq j \end{aligned}$$

- **等价转化：**

$$
\begin{aligned}
\min_{\Phi,H} \quad & \|A-\Phi H\|_F^2, \\
\text{s.t.} \quad & \Phi \in \mathbb{R}^{n\times k}, \text{每一行只有一个元素为1, 其余为0}, \\
& H \in \mathbb{R}^{k\times p}.
\end{aligned}
$$

其中：$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix} \in \mathbb{R}^{n\times p}$, $H = \begin{pmatrix} h_1^T \\ \vdots \\ h_k^T \end{pmatrix} \in \mathbb{R}^{k\times p}$,

$\Phi = (\phi_{ij}) \in \mathbb{R}^{n\times k}$, $\phi_{ij} = \begin{cases} 1, & a_i \in S_j; \\ 0, & a_i \notin S_j. \end{cases}$

典型问题举例

- 非负矩阵分解

$$\min_{X,Y\geq 0} \quad \frac{1}{2}\|M - XY\|_F^2$$

其中$M \in \mathbb{R}^{m\times n}$是已知矩阵, $X \in \mathbb{R}^{m\times k}$, $Y \in \mathbb{R}^{k\times n}$ 为待求非负矩阵.

- 非负张量分解

$$\min_{A_1,A_2,\cdots,A_N\geq 0} \quad \frac{1}{2}\|\mathcal{M} - [\![A_1, A_2, \ldots, A_N]\!]\|_F^2$$

其中$\mathcal{M} \in \mathbb{R}^{I_1\times\cdots\times I_N}$是已知张量,

$[\![A_1, A_2, \ldots, A_N]\!] = \sum_{i=1}^{k} a_i^{(1)} \circ \cdots \circ a_i^{(N)}$, “$\circ$” 表示张量的外积运算,

$A_j = [a_1^{(j)} \ \cdots \ a_k^{(j)}] \in \mathbb{R}^{I_j\times k}, j = 1, \ldots, N.$

典型问题举例

- 字典学习 设$A \in \mathbb{R}^{m\times n}$为$n$个观测, 每个观测的信号维数是$m$, 要从$A$中学习出一个字典$D \in \mathbb{R}^{m\times k}$和系数矩阵$X \in \mathbb{R}^{k\times n}$:

$$\min_{D,X} \quad \frac{1}{2n}\|DX - A\|_F^2 + \lambda\|X\|_1 + \frac{\mu}{2}\|D\|_F^2,$$

挑战和难点

- 函数f关于变量全体一般是非凸的，这使得问题求解具有挑战性
- 应用在非凸问题上的算法收敛性不易分析，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
- 目标: 发展一种更新方式简单且有全局收敛性（收敛到稳定点）的有效算法

变量划分

- 交替极小: 按照$x_1, x_2, \cdots, x_s$的次序依次固定其他$(s-1)$块变量极小化F, 完成一块变量的极小化后,它的值便立即被更新到变量空间中, 更新下一块变量时将使用每个变量最新的值.
- 变量划分

$$\mathcal{X}_i^k = \{x_i \in \mathbb{R}^{n_i} \mid (x_1^k, \cdots, x_{i-1}^k, x_i, x_{i+1}^{k-1}, \cdots, x_s^{k-1}) \in \mathcal{X}\}.$$

- 辅助函数

$$f_i^k(x_i) = f(x_1^k, \cdots, x_{i-1}^k, x_i, x_{i+1}^{k-1}, \cdots, x_s^{k-1}),$$

其中x_j^k表示在第k次迭代中第j块自变量的值, 函数f_i^k 表示在第k次迭代更新第i块变量时所需要考虑的目标函数的光滑部分. 考虑第i块变量时前$(i-1)$块变量已经完成更新, 因此上标为k; 而后面下标从$(i+1)$起的变量仍为旧的值, 因此上标为$(k-1)$.

三种变量更新方式

(i) 固定其他分量然后对单一变量求极小:

$$x_i^k = \arg\min_{x_i \in \mathcal{X}_i^k} \left\{ f_i^k(x_i) + r_i(x_i) \right\}. \tag{2}$$

(ii) 增加了一个邻近项 $\frac{L_i^{k-1}}{2}\|x_i - x_i^{k-1}\|_2^2$ 来限制下一步迭代不应该与当前位置相距过远, 增加邻近项的作用是使得算法能够收敛($L_i^k > 0$为常数):

$$x_i^k = \arg\min_{x_i \in \mathcal{X}_i^k} \left\{ f_i^k(x_i) + \frac{L_i^{k-1}}{2}\|x_i - x_i^{k-1}\|_2^2 + r_i(x_i) \right\}. \tag{3}$$

(iii) 对$f_i^k(x)$进行线性化以简化子问题的求解, 并引入了Nesterov 加速算法的技巧加快收敛:

$$x_i^k = \arg\min_{x_i \in \mathcal{X}_i^k} \left\{ \langle \hat{g}_i^k, x_i - \hat{x}_i^{k-1} \rangle + \frac{L_i^{k-1}}{2}\|x_i - \hat{x}_i^{k-1}\|_2^2 + r_i(x_i) \right\}, \tag{4}$$

其中$\hat{x}_i^{k-1}$采用外推定义:

$$\hat{x}_i^{k-1} = x_i^{k-1} + \omega_i^{k-1}(x_i^{k-1} - x_i^{k-2}), \tag{5}$$

$\omega_i^k \geq 0$为外推的**权重**, $\hat{g}_i^k \stackrel{\text{def}}{=\!=} \nabla f_i^k(\hat{x}_i^{k-1})$为外推点处的梯度. 取权重$\omega_i^k = 0$ 即可得到不带外推的更新格式, 此时等价于进行一次近端梯度法(Proximal Gradient)的更新.

BCD 算法框架

►分块坐标下降法(Block Coordinate Descent, BCD)

```
初始化：选择两组初始点(x_1^{-1}, x_2^{-1}, ···, x_s^{-1}) = (x_1^0, x_2^0, ···, x_s^0).
for k = 1, 2, ··· do
    for i = 1, 2, ··· do
        使用格式(2) 或(3) 或(4) 更新x_i^k.
    end for
    if 满足停机条件 then
        返回(x_1^k, x_2^k, ···, x_s^k)，算法终止.
    end if
end for
```

BCD算法格式解释

- 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，坐标下降算法的数值表现也不相同.
- 格式(2)是最直接的更新方式，它严格保证了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(2)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.
- 格式(3) (4) 则是对格式(2)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．使用格式(3)可使得算法收敛性在函数F为非严格凸时有所改善.
- 格式(4)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．此外，格式(4)的计算量很小，比较容易实现.

▶【例】 min $\quad f(x,y)=x^2-2xy+10y^2-4x-20y$

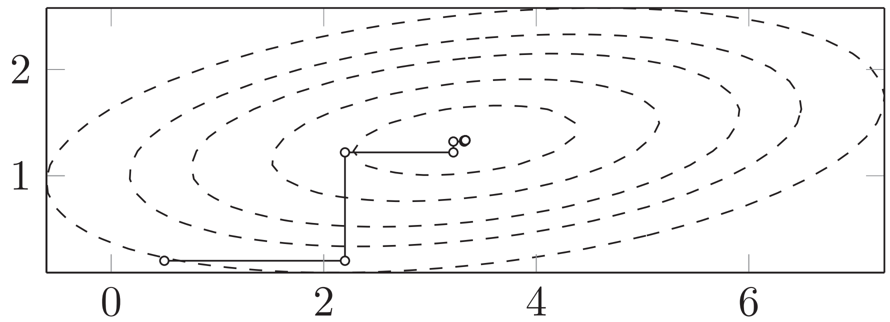
- 采用格式(2)的分块坐标下降法:

$$x^{k+1}=\arg\min_{x\in\mathbb{R}}\{x^2-2xy^k-4x\}=2+y^k,$$

$$y^{k+1}=\arg\min_{y\in\mathbb{R}}\{-2x^{k+1}y+10y^2-20y\}=1+\frac{x^{k+1}}{10}.$$

- 初始点: $(x,y)=(0.5,0.2)$时迭代点轨迹如图，在进行了7 次迭代后迭代点与最优解已充分接近.

- 对于比较病态的问题, 由于分块坐标下降法是对逐个分量处理, 它能较好地捕捉目标函数的各向异性, 而梯度法则会受到很大影响.



非凸$f(x)$, BCD可能失效!

▶不收敛反例 (Powell, 1973): (使用格式(2))

- 目标函数:

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

  $f(x_1, x_2, x_3) = -x_1x_2 - x_2x_3 - x_3x_1$ 非凸

- 梯度分量: $\nabla_{x_1} F(x_1, x_2, x_3) = -x_2 - x_3 + 2(x_1 - 1)_+ - 2(-x_1 - 1)_+$

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

  容易验证迭代序列满足

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

两个聚点: $u = (-1, 1, -1)^T, v = (1, -1, 1)^T$,

$$\nabla F(u) = (0, 2, 0)^T = -\nabla F(v)$$

因此u, v均不是F的稳定点

BCD应用举例

▶LASSO问题 $\min_x \mu\|x\|_1 + \dfrac{1}{2}\|Ax-b\|^2$

- 将自变量x记为$x=[x_i,\bar{x}_i^\top]^\top$, 其中$\bar{x}_i$为$x$去掉第$i$个分量而形成的列向量．相应地, 矩阵$A$在第$i$块的更新记为$A=[a_i\ \bar{A}_i]$, 其中$\bar{A}_i$为矩阵$A$去掉第$i$列而形成的矩阵．LASSO问题可以写为

$$\min_{x_i} \quad \mu|x_i| + \mu\|\bar{x}_i\|_1 + \frac{1}{2}\|a_i x_i - (b-\bar{A}_i\bar{x}_i)\|^2.$$

- 利用格式(2)更新第i块, 令$c_i = b-\bar{A}_i\bar{x}_i$, 则求解

$$\min_{x_i} \quad f_i(x_i) \xlongequal{\text{def}} \mu|x_i| + \frac{1}{2}\|a_i\|^2 {x_i}^2 - a_i^{\mathrm{T}} c_i x_i. \qquad (6)$$

$$= \|a_i\|^2 \min_{x_i}\{\tfrac{\mu}{\|a_i\|^2}|x_i| + \tfrac{1}{2}(x_i - a_i^{\mathrm{T}}c_i/\|a_i\|^2)^2 + const\}$$

- (6)的最小值点

$$x_i^k = \arg\min_{x_i} f_i(x_i) = \mathrm{Soft}_{\lambda_i}(\hat{u}_i) := \begin{cases} \hat{u}_i - \lambda_i, & \text{if } \hat{u}_i > \lambda_i, \\ \hat{u}_i + \lambda_i, & \text{if } \hat{u}_i < -\lambda_i, \\ 0, & \text{otherwise.} \end{cases}$$

其中$\lambda_i = \frac{\mu}{\|a_i\|^2}, \hat{u}_i = \frac{a_i^{\mathrm{T}} c_i}{\|a_i\|^2}$.

▶K-均值聚类问题

$$\begin{aligned} \min_{\Phi, H} \quad & \|A - \Phi H\|_F^2, \\ \text{s.t.} \quad & \Phi \in \mathbb{R}^{n\times k}, \text{每一行只有一个元素为1, 其余为0}, \\ & H \in \mathbb{R}^{k\times p}. \end{aligned}$$

- 当固定H时, 设Φ的每一行为ϕ_i^{T}, 则

$$A-\Phi H=\begin{pmatrix} a_1^{\mathrm{T}} \\ a_2^{\mathrm{T}} \\ \vdots \\ a_n^{\mathrm{T}} \end{pmatrix}-\begin{pmatrix} \phi_1^{\mathrm{T}} \\ \phi_2^{\mathrm{T}} \\ \vdots \\ \phi_n^{\mathrm{T}} \end{pmatrix} H=\begin{pmatrix} a_1^{\mathrm{T}}-\phi_1^{\mathrm{T}} H \\ a_2^{\mathrm{T}}-\phi_2^{\mathrm{T}} H \\ \vdots \\ a_n^{\mathrm{T}}-\phi_n^{\mathrm{T}} H \end{pmatrix}.$$

注意到ϕ_i只有一个分量为1, 其余分量为0, 不妨设其第j个分量为1, 此时$\phi_i^{\mathrm{T}} H$ 相当于将H 的第j行取出, 因此$\|a_i^{\mathrm{T}}-\phi_i^{\mathrm{T}} H\|$为$a_i^{\mathrm{T}}$与$H$的第$j$个行向量的距离. 我们的最终目的是极小化$\|A-\Phi H\|_F^2$, 所以$j$应该选矩阵$H$中距离$a_i^{\mathrm{T}}$最近的那一行:

$$\Phi_{ij}=\begin{cases} 1, & j=\arg\min_l \|a_i-h_l\|, \\ 0, & \text{otherwise.} \end{cases}$$

其中h_l^{T}表示矩阵H的第l行.

- 当固定Φ时, 考虑H的每一行h_j^{T}, 则有

$$\|A-\Phi H\|_F^2=\sum_{j=1}^{k}\sum_{a\in S_j}\|a-h_j\|^2,$$

因此只需对每个h_j求最小．设$\bar{a}_j$是目前第j类所有点的均值,
则$\sum\limits_{a\in S_j}\langle a-\bar{a}_j,\bar{a}_j-h_j\rangle=0$且

$$\begin{aligned}\sum_{a\in S_j}\|a-h_j\|^2&=\sum_{a\in S_j}\|a-\bar{a}_j+\bar{a}_j-h_j\|^2\\&=\sum_{a\in S_j}\left(\|a-\bar{a}_j\|^2+\|\bar{a}_j-h_j\|^2+2\langle a-\bar{a}_j,\bar{a}_j-h_j\rangle\right)\\&=\sum_{a\in S_j}\left(\|a-\bar{a}_j\|^2+\|\bar{a}_j-h_j\|^2\right).\end{aligned}$$

故$h_j=\bar{a}_j$可达到最小值.

▶**非负矩阵分解问题** $\min\limits_{X,Y\geq 0} \quad f(X,Y)=\frac{1}{2}\|XY-M\|_F^2$

- $f(X,Y)$的梯度

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

- 利用格式(4), 注意到当$r_i(X)$为凸集示性函数时即是求解到该集合的投影,因此得到分块坐标下降法:

$$X^{k+1}=\max\{X^k-t_k^x(X^kY^k-M)(Y^k)^{\mathrm{T}},0\},$$
$$Y^{k+1}=\max\{Y^k-t_k^y(X^k)^{\mathrm{T}}(X^kY^k-M),0\},$$

其中t_k^x, t_k^y是步长.

▶字典学习 $\min_{D,X} \quad \frac{1}{2n}\|DX-A\|_F^2+\lambda\|X\|_1+\frac{\mu}{2}\|D\|_F^2$

- 当固定变量D时, 考虑函数

$$f_D(X)=\frac{1}{2n}\|DX-A\|_F^2+\lambda\|X\|_1.$$

使用格式(4). 计算$f_D(X)$中光滑部分的梯度:

$$G=\frac{1}{n}D^{\mathrm{T}}(DX-A),$$

因此格式(4)的BCD:

$$X^{k+1}=\mathrm{Soft}_{t_k\lambda}\left(X^k-\frac{t_k}{n}(D^k)^{\mathrm{T}}(D^kX^k-A)\right),$$

其中t_k为步长.

- 当固定变量X时, 考虑函数

$$f_X(D)=\frac{1}{2n}\|DX-A\|_F^2+\frac{\mu}{2}\|D\|_F^2.$$

使用格式(2)．计算关于D^{T}的梯度:

$$\nabla_{D^{\mathrm{T}}} f_X(D) = \frac{1}{n} X(X^{\mathrm{T}} D^{\mathrm{T}} - A^{\mathrm{T}}) + \mu D^{\mathrm{T}}.$$

于是可得

$$D = AX^{\mathrm{T}}(XX^{\mathrm{T}} + n\mu I)^{-1}.$$

因为$X \in \mathbb{R}^{k\times n}$, $k \ll n$, 故可方便地求出XX^{T}的逆．故格式(2)的BCD:

$$D^{k+1} = A(X^{k+1})^{\mathrm{T}}(X^{k+1}(X^{k+1})^{\mathrm{T}} + n\mu I)^{-1}.$$

- 若先更新X再更新D, 则最终可以得到如下的分块坐标下降法:

$$X^{k+1} = \mathrm{Soft}_{t_k\lambda}\left(X^k - \frac{t_k}{n}(D^k)^{\mathrm{T}}(D^kX^k - A)\right),$$

$$D^{k+1} = A(X^{k+1})^{\mathrm{T}}(X^{k+1}(X^{k+1})^{\mathrm{T}} + n\mu I)^{-1}.$$