分块坐标下降法

宋晓良

大连理工大学数学科学学院

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

提纲

1. 分块坐标下降法

2. 应用举例

问题形式

考虑具有如下形式的问题：

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

- $\mathcal{X}$是函数的可行域，自变量x拆分成s个变量块$x_1,x_2,\cdots,x_s$，每个变量块$x_i\in\mathbb{R}^{n_i}$.
- 函数f是关于x的可微函数，每个$r_i(x_i)$关于x_i是适当的闭凸函数，但不一定可微.
- 目标函数F的性质体现在f，每个r_i以及自变量的分块上. 通常情况下，f对于所有变量块x_i不可分，但单独考虑每一块自变量时，f有简单结构；r_i只和第i个自变量块有关，因此r_i在目标函数中是一个可分项.
- 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.

问题形式

- 分组LASSO 模型：参数$x=(x_1,x_2,\cdots,x_G)\in\mathbb{R}^p$ 可以分成G 组，且$\{x_i\}_{i=1}^G$中只有少数的非零向量.

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

- K-均值聚类问题的等价形式：

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

- 低秩矩阵恢复: 设$b\in\mathbb{R}^m$是已知的观测向量，$\mathcal{A}$是线性映射.

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

其中$\alpha,\beta>0$为正则化参数.

问题形式

- 非负矩阵分解：设$\mathcal{M}$是已知张量，考虑求解如下极小化问题：

$$\min_{A_1,A_2,\cdots,A_N\geq 0} \quad \frac{1}{2}\|\mathcal{M}-A_1\circ A_2\circ\cdots\circ A_N\|_F^2+\sum_{i=1}^{N}\lambda_i r_i(A_i),$$

其中“$\circ$”表示张量的外积运算.

- 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$m$，现在我们要从$A$中学习出一个字典$D\in\mathbb{R}^{m\times k}$和系数矩阵$X\in\mathbb{R}^{k\times n}$：

$$\begin{aligned}\min_{D,X} \quad & \frac{1}{2n}\|DX-A\|_F^2+\lambda\|X\|_1,\\ \text{s.t.} \quad & \|D\|_F\leq 1.\end{aligned}$$

在这里自变量有两块，分别为D和X，此外对D还存在球约束$\|D\|_F\leq 1$.

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

- **分块坐标下降法更新方式**：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

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

- 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.

变量更新方式

在每一步更新中，通常使用以下三种更新格式之一：

$$x_i^k = \underset{x_i \in \mathcal{X}_i^k}{\operatorname{argmin}} \left\{ f_i^k(x_i) + r_i(x_i) \right\}, \tag{1}$$

$$x_i^k = \underset{x_i \in \mathcal{X}_i^k}{\operatorname{argmin}} \left\{ f_i^k(x_i) + \frac{L_i^{k-1}}{2} \|x_i - x_i^{k-1}\|_2^2 + r_i(x_i) \right\}, \tag{2}$$

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

- $L_i^k > 0$为常数
- 在更新格式(3)中，$\hat{x}_i^{k-1}$采用外推定义：

$$\hat{x}_i^{k-1} = x_i^{k-1} + \omega_i^{k-1}(x_i^{k-1} - x_i^{k-2}), \tag{4}$$

其中$\omega_i^k \geq 0$为外推的**权重**，$\hat{g}_i^k \overset{\text{def}}{=\!=} \nabla f_i^k(\hat{x}_i^{k-1})$为外推点处的梯度．

算法格式

Algorithm 1 分块坐标下降法

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

- 三种格式都有其适用的问题，特别是子问题是否可写出显式解
- 在每一步更新中，三种迭代格式对不同自变量块可以混合使用，不必仅仅局限于一种.

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，坐标下降算法的数值表现也不相同.

- 格式(1)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(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.

- 格式(2) (3) 则是对格式(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．使用格式(2)可使得算法收敛性在函数F为非严格凸时有所改善.

- 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．此外，格式(3)的计算量很小，比较容易实现.

不收敛反例

值得注意的是，对于非凸函数$f(x)$，分块坐标下降法可能失效．Powell 在1973 年就给出了一个使用格式(1)但不收敛的例子:

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

其中$(x_i-1)_+^2$的含义为先对(x_i-1)取正部再平方．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

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

这个迭代序列有两个聚点$(-1,1,-1)$与$(1,-1,1)$，但这两个点都不是F的稳定点．

提纲

1. 分块坐标下降法

2. 应用举例

考虑最基本的非负矩阵分解问题

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

可以计算梯度

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

注意到在格式(3)中，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如下：

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

其中t_k^x, t_k^y是步长，