第七章 复合优化算法

修贤超

https://xianchaoxiu.github.io

目录

近似点算法

- 考虑一般形式的优化问题

$$\min_{x} \quad \psi(x)$$

- ψ 是一个适当的闭凸函数，并不要求连续或可微
- 次梯度法求解收敛较慢，且收敛条件苛刻
- 近似点梯度法做隐性的梯度下降

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

  - $\psi(x)$ 的邻近算子一般需要通过迭代求解
  - 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

- 用 FISTA 算法对近似点算法进行加速，其迭代格式为

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

- 第二类 Nesterov 加速算法的迭代格式可以写成

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

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

与增广拉格朗日函数法的关系

- 考虑具有如下形式的优化问题

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

- 例 7.4 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

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

  - 当 h 是凸集 C 上的示性函数时，等价于凸集约束问题

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

  - 当 $h(y) = \|y - b\|$ 时，等价于正则优化问题

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

对偶问题

- **原问题的增广拉格朗日函数法**

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

- **对偶问题**

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

近似点算法

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

- **原问题用增广拉格朗日函数法 $\Leftrightarrow$ 对偶问题用近似点算法**

应用举例: LASSO 问题

- 考虑 LASSO 问题

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

- 引入变量 $y = Ax - b$，等价地转化为

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

- 采用近似点算法进行求解，第 k 步迭代为

$$(x^{k+1}, y^{k+1}) \approx \arg\min_{(x,y)\in\mathbb{D}} \quad \left\{ f(x,y) + \frac{1}{2t_k}(\|x-x^k\|_2^2 + \|y-y^k\|_2^2) \right\}$$

其中 $\mathbb{D} = \{(x,y) \mid Ax - y = b\}$ 为可行域，t_k 为步长

应用举例: LASSO 问题

- 除了直接求解，还可以通过对偶问题求解
- 引入拉格朗日乘子 z，对偶函数为

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

其中

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

应用举例: LASSO 问题

- 记函数 $q_{\mu t_k}:\mathbb{R}\to\mathbb{R}$ 为

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

- 易知 $\Gamma_{\mu t_k}(u)$ 是关于 u 的连续可微函数且导数为

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

- 对偶问题为

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

应用举例: LASSO 问题

- 设对偶问题的逼近最优解为 z^{k+1}，根据最优性条件有

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

- 在第 k 步迭代，LASSO 问题的近似点算法的迭代格式

$$
\begin{cases}
z^{k+1} \approx \arg\max\limits_{z} \quad \Phi_k(z) \\
x^{k+1} = \mathrm{prox}_{\mu t_k \|x\|_1}(x^k - t_k A^\top z^{k+1}) \\
y^{k+1} = \dfrac{1}{t_k+1}(y^k + t_k z^{k+1})
\end{cases}
$$

- 根据 $\Phi_k(z)$ 的连续可微性，可以调用梯度法进行求解

收敛性分析

- 定理 7.6 设 ψ 是闭凸函数，最优值 ψ^* 有限且在 x^* 取到，则对近似点算法有

$$\psi(x^{(k)}) - \psi^* \le \frac{||x^{(0)} - x^*||_2^2}{2\sum_{i=1}^{k} t_i}, \qquad \forall\ k \ge 1$$

- 若 $\sum_i t_i \to \infty$，则算法收敛
- 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

加速版本的近似点算法

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

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

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

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

- 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 $k > 1$，θ_k 满足

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

收敛性分析

- **定理 7.7** 设 ψ 是闭凸函数, 最优值 ψ^* 有限且在 x^* 处取到. 假设参数 t_k, γ_k 按照加速策略选取，那么

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

- 若 $\sum_i \sqrt{t_i} \to \infty$，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 $\mathcal{O}\left(\frac{1}{k^2}\right)$

目录

问题形式

- 考虑具有如下形式的问题

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

  - 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

问题形式

- **例 7.5** 设参数 $x=(x_1,x_2,\cdots,x_G)\in\mathbb{R}^p$，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 7.6** 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 7.7** 设 M 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 $x_1, x_2, \cdots, x_s$ 的次序依次固定其他 $(s-1)$ 块变量极小化 F
- 辅助函数

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

- 在每一步更新中，通常使用以下三种更新格式之一

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

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

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

算法格式

算法 7.9 分块坐标下降法

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)、(2)、(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

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，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
- 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．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

例 7.8

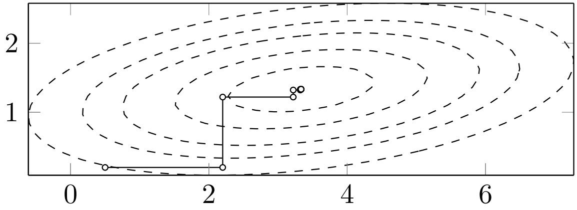
- 考虑二元二次函数的优化问题

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

- 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

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

- 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 $x = [x_i, \bar{x}_i^\top]^\top$, 矩阵 A 在第 i 块的更新记为 $A = [a_i \bar{A}_i]$
- 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

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

应用举例: 非负矩阵分解

- 考虑最基本的非负矩阵分解问题

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

- 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

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

Q&A

Thank you!

感谢您的聆听和反馈