约束非光滑凸优化问题的光滑加速算法

献给吴从炘教授 85 华诞

边伟^{1,2}

1. 哈尔滨工业大学数学学院, 哈尔滨 150001;

2. 哈尔滨工业大学数学研究院, 哈尔滨 150001

E-mail: bianweilvse520@163.com

收稿日期: 2020-06-09; 接受日期: 2020-08-10; 网络出版日期: 2020-12-14

国家自然科学基金 (批准号: 11871178 和 61773136) 资助项目

摘要 本文考虑一类在信号处理、图像恢复和机器学习等多项科学和工程领域中具有广泛应用的约束非光滑凸优化问题. 近年来, 理论研究和数值实验均验证了外插项可有效提高算法的收敛速率, 带有外插的临近梯度算法在求解大规模优化问题中有显著优势. 因此, 本文利用光滑化技巧, 结合 Beck 和 Teboulle 提出的快速迭代收缩阈值算法, 对一类非光滑凸优化问题提出新的加速算法, 证明算法的任意聚点都是优化问题的最优解. 在算法分析中, 考虑光滑参数的不同更新准则, 给出目标函数值 $O(\ln k/k)$ 的全局收敛速率, 并证明迭代序列的变化趋势 $\lim_{k \rightarrow +\infty} \|x^{k+1} - x^k\| = 0$. 最后, 通过数值实验展示本文提出的算法对两类稀疏优化问题的良好求解能力和外插项对算法收敛速率的正面影响.

关键词 非光滑凸优化问题 加速算法 光滑化技巧 收敛速率

MSC (2010) 主题分类 90C30, 90C25, 65K05

1 引言

本文考虑如下约束凸优化问题:

$$\begin{aligned} \min f(x) &:= h(x) + g(x) \\ \text{s.t. } x &\in \mathcal{X}, \end{aligned} \quad (1.1)$$

其中 $h: \mathbb{R}^n \rightarrow \mathbb{R}$ 是连续凸函数 (未必可微), $g: \mathbb{R}^n \rightarrow \mathbb{R}$ 是一个真闭凸函数 (未必可微), $\mathcal{X}$ 是 $\mathbb{R}^n$ 中的非空闭凸集. 本文需要假设: 对任意的 $\alpha > 0$ 和 $y \in \mathbb{R}^n$, αg 关于集合 $\mathcal{X}$ 的临近算子 $\text{Prox}_{\alpha g}$ 可计算, 其中

$$\text{Prox}_{\alpha g}(y, \mathcal{X}) = \arg \min_{x \in \mathcal{X}} \left\{ g(x) + \frac{1}{2\alpha} \|x - y\|_2^2 \right\}.$$

英文引用格式: Bian W. Smoothing accelerated algorithm for constrained nonsmooth convex optimization problems (in Chinese). Sci Sin Math, 2020, 50: 1651–1666, doi: 10.1360/SSM-2020-0181

满足上述条件的优化模型 (1.1) 广泛存在于科学和工程应用中的许多领域, 如图像分解与还原、视觉编码、压缩感知、地质勘探、无线通信、人工智能、基因表达分析和风险管理等. 例如, ℓ_1 正则化稀疏优化模型

$$\min_{x \in \mathcal{X}} f(x) := h(x) + \lambda \|x\|_1, \quad (1.2)$$

其中 h 为控制数据拟合度的损失函数, $\|x\|_1$ 是控制解稀疏性的正则函数 (也称其为罚函数), λ 是协调数据拟合度与解稀疏性之间的平衡参数. 这里, ℓ_1 范数函数的临近算子具有如下显式结构:

$$\text{Prox}_{\alpha \|\cdot\|_1}(x) = \begin{cases} x_i - \alpha, & \text{若 } x_i > \alpha, \\ 0, & \text{若 } |x_i| \leq \alpha, \\ x_i + \alpha, & \text{若 } x_i < -\alpha. \end{cases}$$

关于临近算子的更多显式结构和性质可参见文献 [1].

现代应用领域中问题的规模越来越大, 有些问题涉及的变量和数据的数量已达到数百万. 众所周知, 一阶算法不依赖于问题的二阶信息, 计算成本低. 邻近向前 - 向后分裂算法是一类经典的一阶算法, 算法的每一步主要是计算可微函数的梯度与非可微函数的邻近算子, 该算法也常被称为临近梯度算法. 鉴于优化问题 (1.1) 中 g 的特殊性质, 临近算法成为求解此类问题的主要方法之一. 但现有算法和理论分析大多是在 h 为可微凸函数, 且其梯度全局 Lipschitz 连续的条件下建立起来的. 在此条件下, 利用邻近梯度算法求解该问题时, 生成的迭代序列的目标函数值为次线性全局收敛的, 即

$$f(x^k) - \min_{x \in \mathbb{R}^n} f(x) = O\left(\frac{1}{k}\right).$$

随后, 学者们提出了各种方法用于改善邻近梯度算法的收敛速率, 其中一个比较经典的方法是加入惯性项. 该方法与原始邻近梯度算法的不同之处是利用前两个迭代点的线性组合来更新下一个迭代点. 算法的主要框架如下:

$$\begin{cases} y^k = x^k + \beta_k(x^k - x^{k-1}), \\ x^{k+1} = \arg \min_{x \in \mathbb{R}^n} \left\{ h(y^k) + \langle \nabla h(y^k), x - y^k \rangle + \frac{1}{2\lambda_k} \|x - y^k\|^2 + g(x) \right\}, \end{cases} \quad (1.3)$$

其中 $\{\beta_k\} \subseteq [0, 1]$ 为惯性参数, 也称为外插系数, $\{\lambda_k\}$ 为步长参数. 在相同条件下, Beck 和 Teboulle [2] 提出了快速迭代阈值算法 (fast iterative shrinkage-thresholding algorithm, FISTA) 求解问题 (1.1), 该算法不仅保留了邻近梯度算法计算简便的优点, 而且将函数值序列的收敛速率改善为 $O(1/k)$. FISTA 框架与 (1.3) 相同, 但给出了惯性参数 $\{\beta_k\}$ 的更新格式. Nesterov [3] 独立地提出了一个与 FISTA 具有相同收敛速率的加速算法. O'Donoghue 和 Candès [4] 提出了带有重启策略的 FISTA 求解问题 (1.1), 并在 h 强凸、 $g = 0$ 时, 证明了该算法生成的函数值序列具有全局线性收敛性, 并通过数值实验展示了该算法的收敛速率快于 FISTA. 2016 年, Tao 等 [5] 讨论了利用 FISTA 求解 ℓ_1 正则化最小二乘问题, 在问题有唯一解且满足严格互补性条件时, 证明了 FISTA 生成的迭代序列是局部线性收敛的. 同年 Attouch 和 Peyrouquet [6] 研究了 Hilbert 空间上的凸优化问题, 分析了当 $\beta_k = \frac{k-1}{k+\alpha-1}$ ($\alpha > 3$) 时, 算法 (1.3) 生成的迭代序列与函数值序列的收敛速率分别为

$$\|x^{k+1} - x^k\| = o\left(\frac{1}{k}\right) \quad \text{和} \quad f(x^k) - \min f = o\left(\frac{1}{k^2}\right).$$

在适当条件下, Johnstone 和 Moulin^[7] 讨论了算法 (1.3) 生成的迭代序列的收敛性. 特别地, 当利用算法 (1.3) 求解 ℓ_1 正则化最小二乘问题且最优解唯一时, 算法生成的迭代序列具有局部线性收敛性. 文献 [8] 在误差界条件下, 证明了利用带外插的邻近梯度算法 (1.3) 求解问题 (1.1) 时, 生成的迭代序列及相应的函数值序列均为 R 线性收敛的. 随后, 文献 [9] 在目标函数满足一致 Lojasiewicz 条件下, 讨论了算法 (1.3) 的迭代序列的长度有限性, 即

$$\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| < \infty.$$

以上工作均是在 h 连续可微且其梯度 Lipschitz 连续的条件下给出的. 对于 h 可微但其梯度不具有 Lipschitz 性质的情形, Bauschke 等^[10] 提出了一类 Bregman 邻近梯度算法求解闭凸集 $\mathcal{X}$ 上的非光滑凸优化问题 (1.1), 其中 h 为可微凸函数 (梯度未必 Lipschitz 连续). Bregman 邻近梯度算法基本格式如下:

$$x^{k+1} = \arg \min_{x \in \mathcal{X}} \left\{ \langle \nabla h(x^k), x - x^k \rangle + g(x) + \frac{1}{\lambda_k} D_c(x, x^k) \right\}, \quad (1.4)$$

其中

$$D_c(x, y) = c(x) - c(y) - \langle \nabla c(y), x - y \rangle$$

为关于给定 Legendre 函数 $c: \mathbb{R}^n \rightarrow \mathbb{R}$ 的 Bregman 距离. 本文证明了当 $\lambda_k = \frac{1+\alpha}{2L}$ 时, 该算法生成的函数值序列的全局收敛速率为

$$f(x^k) - \min_{x \in \mathcal{X}} f \leq \frac{2L}{1+\alpha k} D_c(x^*, x^0),$$

其中 α 为 c 的对称系数, $L > 0$ 为使 $Lc - h$ 为凸函数的固定常数, x^* 可为问题 (1.1) 的任意最优解. 此外, 在 $x^k \rightarrow x \in \text{dom} c$ 等价于 $D_c(x, x^k) \rightarrow 0$ 的条件下, 文献 [10] 证明了迭代序列的收敛性.

对于 h 非光滑的情形, 加速算法的研究较少. 前期, Nesterov^[11] 利用非光滑凸函数 h 的一种光滑近似方法, 提出了可求解问题 (1.1) 的 ϵ 最优解的加速梯度算法. 这里, 若 $f(x^\epsilon) - \min f \leq \epsilon$, 则称 x^ϵ 为 (1.1) 的一个 ϵ 最优解. 同时, 在适当的 ϵ 取值下, Nesterov^[11] 将求得问题 (1.1) 的 ϵ 最优解的迭代次数改善为 $O(\frac{1}{\epsilon})$. 随后, 文献 [12] 将 Nesterov^[11] 的光滑化技巧应用于 Nash 均衡鞍点问题, 提出了一阶算法, 并证明了算法具有相同的最坏复杂性. 需要说明的是, 在上述两个工作中, ϵ 均是 h 近似函数定义中的一个不变参数, 所以, 算法的任意聚点仅可为原问题的 ϵ 最优解, 无法保证为最优解.

近期研究发现, 许多问题的目标函数并不能表达成形如 (1.1) 的 h 为可微函数的形式, 例如, 在 ℓ_1 正则化的稀疏模型 (1.2) 中, 很多损失函数都为非光滑凸函数. 光滑的 ℓ_2 损失函数是对线性回归问题在样本符合 Gauss 分布的条件下推导得到的, 对其的研究在过去几十年得到了广泛的关注和发展. 但是, 统计学家 Fan 和 Li^[13] 称, 现已有结果显示最小二乘损失函数虽然可解决一类线性回归问题, 但并非适用于所有类型的数据, 这时, 我们可选取对离群值具有较强抗干扰能力的损失函数, 如 ℓ_1 函数、分位数回归函数或更一般的 Huber 类函数. 在审查回归问题中, 为了寻找满足非线性系统 $\max\{Ax, 0\} = b$ 的稀疏解, 损失函数常定义为

$$h(x) = \|\max\{Ax, 0\} - b\|. \quad (1.5)$$

无论 ℓ_1 损失函数, 还是 (1.5) 中的损失函数, 均是非可微凸函数, 而且, 带有 $\max$ 算子的非可微损失函数, 如 Hinge 损失函数, 也常出现在分类问题中.

因此, 本文将主要关注优化问题 (1.1) 中 h 是非光滑凸函数的情形. 为克服 h 的非光滑性, 我们将对其引入光滑化技巧, 然后基于 FISTA^[2], 对优化问题 (1.1) 建立加速算法, 并给出其收敛性分析. 我们得到: 所建立算法的任意聚点均为优化问题 (1.1) 的最优解, 而且目标函数值具有 $O(\ln k/k)$ 的全局收敛速率.

符号约定 $\mathbb{R}^n$ 表示 n 维 Euclid 空间. $\mathbb{N}$ 表示非负整数全体. 对于 $x, y \in \mathbb{R}^n$,

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i, \quad \|x\| = \sqrt{\sum_{i=1}^n x_i^2},$$

$x \leq y$ 表示 $x_i \leq y_i, i = 1, \dots, n$. 给定 $\gamma \in \mathbb{R}$, γ_n 代表分量全为 γ 的 n 维列向量.

2 加速算法设计和分析

本节将针对优化模型 (1.1) 提出一种带有外插项的加速临近梯度算法, 并给出其收敛性分析. 由于此算法是在 FISTA^[2] 基础上通过对目标函数中的第一项采用光滑化而建立起来的, 所以, 称其为光滑快速迭代阈值算法 (smoothing FISTA), 简记为 FISTA_S.

在本文中, 我们需要如下假设条件:

假设 2.1 f 在 $\mathcal{X}$ 上水平集有界, 即对任意的 $\Gamma > 0$, 集合 $\{x \in \mathcal{X} : f(x) \leq \Gamma\}$ 有界.

显然, 若 $\mathcal{X}$ 有界, 则假设 2.1 成立. 在假设 2.1 成立的条件下, 由 f 的凸性知, 优化问题 (1.1) 的最优解集非空且有界, 记为 $\mathcal{M}$.

2.1 算法设计

鉴于问题 (1.1) 中 h 的非光滑性和 $\text{Prox}_{\alpha g}$ 的可计算性, 我们对 h 引入光滑化函数 $\tilde{h}$, 定义如下.

定义 2.1^[14] 对于非光滑凸函数 $h : \mathbb{R}^n \rightarrow \mathbb{R}$, 称 $\tilde{h} : \mathbb{R}^n \times (0, 1] \rightarrow \mathbb{R}$ 是 h 的一个光滑函数, 如果 $\tilde{h}$ 满足下述条件:

- (i) 对任意的 $\mu > 0$, $\tilde{h}(\cdot, \mu)$ 在 $\mathbb{R}^n$ 中连续可微;
- (ii) $\lim_{z \rightarrow x, \mu \downarrow 0} \tilde{h}(z, \mu) = h(x), \forall x \in \mathcal{X}$;
- (iii) (关于 x 的凸性) 对任意的 $\mu > 0$, $\tilde{h}(\cdot, \mu)$ 是 $\mathcal{X}$ 上的凸函数;
- (iv) (梯度一致性) $\{\lim_{z \rightarrow x, \mu \downarrow 0} \nabla_z \tilde{h}(z, \mu)\} \subseteq \partial h(x), \forall x \in \mathcal{X}$;
- (v) (关于 μ 的 Lipschitz 连续性) 存在 $\kappa > 0$ 使得

$$|\tilde{h}(x, \mu_2) - \tilde{h}(x, \mu_1)| \leq \kappa |\mu_1 - \mu_2|, \quad \forall x \in \mathcal{X}, \quad \mu_1, \mu_2 \in (0, 1];$$

(vi) (关于 x 的 Lipschitz 连续性) 存在 $L > 0$ 使得对于任意的 $\mu \in (0, 1]$, $\nabla_x \tilde{h}(\cdot, \mu)$ 在 $\mathcal{X}$ 上 Lipschitz 连续, 且 Lipschitz 常数为 $L\mu^{-1}$.

由定义 2.1(iv) 知, 对任意的 $x \in \mathbb{R}^n, \mu, \nu \in (0, 1]$, 有

$$|\tilde{h}(x, \mu) - \tilde{h}(x, \nu)| \leq \kappa |\mu - \nu|.$$

在上式中, 令 $\nu \rightarrow 0$, 由定义 2.1(ii) 得到

$$|\tilde{h}(x, \mu) - h(x)| \leq \kappa \mu, \quad \forall x \in \mathbb{R}^n, \quad \mu \in (0, 1]. \quad (2.1)$$

虽然定义 2.1 中条件较多, 但对很多常用的非光滑凸函数均可给出满足这些条件的光滑函数, 例如, ℓ_1 损失函数 $h(x) := \|Ax - b\|_1$ 的光滑函数可定义为

$$\tilde{h}(x, \mu) = \sum_{i=1}^m \tilde{\theta}(A_i x - b_i, \mu), \quad (2.2)$$

(1.5) 中损失函数 h 的光滑函数可定义为

$$\tilde{h}(x, \mu) = \sum_{i=1}^m \tilde{\theta}(\tilde{\phi}(A_i x, \mu) - b_i, \mu), \quad (2.3)$$

其中

$$\tilde{\theta}(s, \mu) = \begin{cases} |s|, & \text{若 } |s| > \mu, \\ \frac{s^2}{2\mu} + \frac{\mu}{2}, & \text{若 } |s| \leq \mu \end{cases}$$

为 $|s|$ 的一个光滑函数,

$$\tilde{\phi}(s, \mu) = \begin{cases} \max\{s, 0\}, & \text{若 } |s| > \mu, \\ \frac{(s + \mu)^2}{4\mu}, & \text{若 } |s| \leq \mu \end{cases}$$

为 $\max\{s, 0\}$ 的一个光滑函数. 关于光滑函数的更多构造方法和基本性质, 可参见文献 [14, 15].

基于光滑函数 $\tilde{h}$, 对于给定的 $\mu > 0$, 我们可将 FISTA 的算法框架应用于如下模型:

$$\begin{aligned} \min \quad & \tilde{f}(x, \mu) := \tilde{h}(x, \mu) + g(x) \\ \text{s.t.} \quad & x \in \mathcal{X}. \end{aligned} \quad (2.4)$$

如无特殊说明, 我们用 $\nabla \tilde{h}(x, \mu)$ 表示 $\tilde{h}(\cdot, \mu)$ 关于 x 的梯度. 由定义 2.1(vi), 定义 $\tilde{f}(\cdot, \mu)$ 在 (y, μ) 点的近似函数

$$Q(x, y, \mu) = \tilde{h}(y, \mu) + \langle \nabla \tilde{h}(y, \mu), x - y \rangle + \frac{1}{2} L \mu^{-1} \|x - y\|^2 + g(x).$$

由 g 的凸性知, $Q(x, y, \mu)$ 关于 x 强凸, 其强凸系数为 $L\mu^{-1}$, 即对任意的 $x, y, z \in \mathbb{R}^n$, $\mu \in (0, 1]$ 和 $\eta(z) \in \partial g(z)$, 成立

$$Q(x, y, \mu) \geq Q(z, y, \mu) + \langle \nabla \tilde{h}(y, \mu) + L\mu^{-1}(z - y) + \eta(z), x - z \rangle + \frac{1}{2} L \mu^{-1} \|x - z\|^2. \quad (2.5)$$

对给定的 $y \in \mathbb{R}^n$ 和 $\mu \in (0, 1]$,

$$\min_{x \in \mathcal{X}} Q(x, y, \mu) \quad (2.6)$$

存在唯一全局解, 记为 $p(y, \mu)$. 通过简单计算可知,

$$\begin{aligned} p(y, \mu) &= \arg \min_{x \in \mathcal{X}} \left\{ \frac{1}{2} L \mu^{-1} \|x - (y - L^{-1} \mu \nabla \tilde{h}(y, \mu))\|^2 + g(x) \right\} \\ &:= \text{Prox}_{L^{-1} \mu g}(y - L^{-1} \mu \nabla \tilde{h}(y, \mu), \mathcal{X}). \end{aligned} \quad (2.7)$$

利用凸优化问题 (2.6) 的最优性条件, 得到 $z = p(y, \mu)$ 是 (2.6) 的最优解, 当且仅当存在 $\bar{\eta}(z) \in \partial g(z)$, 使得对于任意的 $x \in \mathcal{X}$, 成立

$$\langle \nabla \tilde{h}(y, \mu) + L\mu^{-1}(z - y) + \bar{\eta}(z), x - z \rangle \geq 0. \quad (2.8)$$

基于以上分析, 在每一步将 FISTA^[2] 应用于优化问题 (2.4), 并逐步更新光滑参数 μ , 给出本文提出的 FISTA_S.

算法 1 FISTA_S

步 0: 给定 $y^0 = x^0 \in \mathcal{X}$, $t_0 = 1$, $\mu_0 \in (0, 1]$, $\sigma \in (0, 2)$. 令 $k = 0$.

步 1: 计算:

$$\begin{aligned} x^{k+1} &= p(y^k, \mu_k), \\ \mu_{k+1} &= \frac{\mu_0}{(k+1)^\sigma}, \\ t_{k+1} &= \frac{1 + \sqrt{1 + 4(\frac{\mu_k}{\mu_{k+1}})t_k^2}}{2}, \\ \theta_{k+1} &= \frac{t_k - 1}{t_{k+1}}, \\ y^{k+1} &= x^{k+1} + \theta_{k+1}(x^{k+1} - x^k). \end{aligned} \quad (2.9)$$

步 2: 置 $k := k + 1$, 转到步 1.

需要注意到, 与文献 [11] 中的算法不同, FISTA_S 中的光滑参数 μ_k 是逐步迭代且趋向于 0 的. 从后续的收敛性证明和数值实验中也可看出, μ_k 的迭代方法直接影响算法的收敛速度.

2.2 收敛性分析

以下, 记 $\{x^k\}$ 、 $\{y^k\}$ 、 $\{t_k\}$ 、 $\{\theta_k\}$ 和 $\{\mu_k\}$ 为 FISTA_S 生成的迭代序列. 为后续分析方便, 首先给出序列 $\{t_k\}$ 和 $\{\theta_k\}$ 满足的一些基本性质.

引理 2.1 对于 FISTA_S 生成的序列 $\{t_k\}$ 和 $\{\theta_k\}$, 下述性质成立:

- (i) $\frac{1}{2} + \sqrt{\frac{\mu_k}{\mu_{k+1}}}t_k \leq t_{k+1} \leq 1 + \sqrt{\frac{\mu_k}{\mu_{k+1}}}t_k$;
- (ii) $\mu_{k+1}t_{k+1}(t_{k+1} - 1) = \mu_k t_k^2$;
- (iii) $\frac{\sqrt{\mu_0}}{2-\sigma}((k+2)^{1-\frac{\sigma}{2}} - 1) \leq t_{k+1}\sqrt{\mu_{k+1}} \leq \frac{4\sqrt{\mu_0}}{2-\sigma}(k+2)^{1-\frac{\sigma}{2}}$;
- (iv) $t_{k+1}\mu_{k+1}^2 \leq \frac{4\mu_0^2}{2-\sigma}(k+1)^{1-2\sigma}$;
- (v) 对任意的 $k \in \mathbb{N}$, $\theta_{k+1}^2 \leq \frac{\mu_{k+1}}{\mu_k}$;
- (vi) 存在 $K \in \mathbb{N}$, 使得对任意的 $k \geq K$, 有

$$\theta_{k+1}^2 \leq \frac{\mu_{k+1}}{\mu_k} - \sqrt{\frac{\mu_{k+1}}{\mu_k}}t_{k+1}^{-1}.$$

证明 由 t_{k+1} 的定义, 结论 (i) 和 (ii) 显然成立. 下面依次证明 (iii)–(vi).

(iii) 由结论 (i), 有

$$t_{k+1}\sqrt{\mu_{k+1}} \leq \sqrt{\mu_{k+1}} + t_k\sqrt{\mu_k},$$

进而, 由 $t_0 = 1$, 对 $k \geq 0$, 有

$$t_{k+1}\sqrt{\mu_{k+1}} \leq \sum_{i=1}^{k+1} \sqrt{\mu_i} + t_0\sqrt{\mu_0} = \sum_{i=1}^{k+1} \sqrt{\mu_i} + \sqrt{\mu_0}. \quad (2.10)$$

根据 μ_k 在 (2.9) 中的定义, 可得

$$\sum_{i=1}^{k+1} \sqrt{\mu_i} = \sqrt{\mu_0} \sum_{i=1}^{k+1} i^{-\sigma/2} \leq \sqrt{\mu_0} \int_0^{k+1} t^{-\sigma/2} dt = \frac{2\sqrt{\mu_0}}{2-\sigma} (k+1)^{1-\sigma/2}.$$

将上式代入 (2.10), 得到

$$t_{k+1} \sqrt{\mu_{k+1}} \leq \frac{4\sqrt{\mu_0}}{2-\sigma} (k+1)^{1-\frac{\sigma}{2}}.$$

类似地, 成立

$$\begin{aligned} t_{k+1} \sqrt{\mu_{k+1}} &\geq \frac{1}{2} \sum_{i=1}^{k+1} \sqrt{\mu_i} + t_0 \sqrt{\mu_0} = \frac{1}{2} \sum_{i=1}^{k+1} \sqrt{\mu_i} + \sqrt{\mu_0} \geq \frac{1}{2} \sqrt{\mu_0} \int_1^{k+2} t^{-\sigma/2} dt + \sqrt{\mu_0} \\ &= \frac{\sqrt{\mu_0}}{2-\sigma} ((k+2)^{1-\frac{\sigma}{2}} - 1) + \sqrt{\mu_0} \geq \frac{\sqrt{\mu_0}}{2-\sigma} ((k+2)^{1-\frac{\sigma}{2}} - 1). \end{aligned}$$

(iv) 由结论 (iii), 得到

$$t_{k+1} \mu_{k+1}^2 = t_{k+1} \sqrt{\mu_{k+1}} \sqrt{\mu_{k+1}^3} \leq \frac{4\mu_0^2}{2-\sigma} (k+1)^{1-2\sigma}.$$

(v) 由结论 (i) 的第一个不等式, 可得 $t_{k+1}^2 \geq \frac{\mu_k}{\mu_{k+1}} t_k^2$, 所以,

$$\theta_{k+1}^2 \leq \frac{t_k^2}{t_{k+1}^2} \leq \frac{\mu_{k+1}}{\mu_k}.$$

(vi) 再次利用结论 (i) 的第一个不等式和 $\mu_{k+1} \leq \mu_k$, 有 $t_{k+1} \geq \frac{1}{2} + t_k \geq \frac{k+1}{2}$, 而且

$$t_k - 1 \leq \sqrt{\frac{\mu_{k+1}}{\mu_k}} \left(t_{k+1} - \frac{1}{2} \right) - 1 \leq \sqrt{\frac{\mu_{k+1}}{\mu_k}} t_{k+1} - 1.$$

进而,

$$(t_k - 1)^2 \leq \frac{\mu_{k+1}}{\mu_k} t_{k+1}^2 - 2\sqrt{\frac{\mu_{k+1}}{\mu_k}} t_{k+1} + 1.$$

由 $\lim_{k \rightarrow +\infty} \sqrt{\frac{\mu_{k+1}}{\mu_k}} t_{k+1} \geq \lim_{k \rightarrow +\infty} t_k = +\infty$, 结论 (vi) 成立. $\square$

引理 2.2 对于任意的 $x, y \in \mathbb{R}^n$ 和 $\mu \in (0, 1]$, 下式成立:

$$\tilde{f}(x, \mu) - \tilde{f}(p(y, \mu), \mu) \geq \frac{1}{2} L \mu^{-1} \|p(y, \mu) - y\|^2 + L \mu^{-1} \langle y - x, p(y, \mu) - y \rangle. \quad (2.11)$$

证明 由 $Q(x, y, \mu)$ 关于 x 的强凸性和 $p(y, \mu)$ 的定义, 即最优性条件 (2.8), 在 (2.5) 中, 令 $z := p(y, \mu)$, 则对任意的 $x \in \mathcal{X}$, 有

$$Q(p(y, \mu), y, \mu) \leq Q(x, y, \mu) - \frac{1}{2} L \mu^{-1} \|x - p_L(y, \mu)\|^2. \quad (2.12)$$

利用定义 2.1(vi), 有

$$\tilde{h}(p(y, \mu), \mu) \leq \tilde{h}(y, \mu) + \langle \nabla \tilde{h}(y, \mu), p(y, \mu) - y \rangle + \frac{1}{2} L \mu^{-1} \|p(y, \mu) - y\|^2.$$

在上式两边同时加上 $g(p(y, \mu))$, 得到

$$\tilde{f}(p(y, \mu), \mu) \leq Q(p(y, \mu), y, \mu). \quad (2.13)$$

把 (2.12) 和 (2.13) 相加得到

$$\begin{aligned} \tilde{f}(p(y, \mu), \mu) &\leq Q(x, y, \mu) - \frac{1}{2}L\mu^{-1}\|x - p_L(y, \mu)\|^2 \\ &= \tilde{h}(y, \mu) + \langle \nabla \tilde{h}(y, \mu), x - y \rangle + \frac{1}{2}L\mu^{-1}\|x - y\|^2 + g(x) - \frac{1}{2}L\mu^{-1}\|x - p(y, \mu)\|^2 \\ &\leq \tilde{f}(x, \mu) + \frac{1}{2}L\mu^{-1}\|x - y\|^2 - \frac{1}{2}L\mu^{-1}\|x - p(y, \mu)\|^2, \end{aligned} \quad (2.14)$$

其中, 在第二个不等式推导中使用了定义 2.1 中 $\tilde{h}(x, \mu)$ 关于 x 的凸性. 通过对上式简单整理, 即可得到此引理中的不等式 (2.11). $\square$

引理 2.3 若假设 2.1 成立, 则

$$\lim_{k \rightarrow \infty} \tilde{f}(x^k, \mu_k) + \kappa\mu_k + \frac{1}{2}L\mu_{k-1}^{-1}\|x^k - x^{k-1}\|^2$$

存在, 而且序列 $\{x^k\}$ 有界, 即存在 $R > 0$ 使得 $\|x^k\| \leq R, \forall k \in \mathbb{N}$.

证明 在 (2.14) 中, 令 $x := x^k, y := y^k, \mu := \mu_k$, 得到

$$\tilde{f}(p(y^k, \mu_k), \mu_k) \leq \tilde{f}(x^k, \mu_k) + \frac{1}{2}L\mu_k^{-1}\|x^k - y^k\|^2 - \frac{1}{2}L\mu_k^{-1}\|x^k - p(y^k, \mu_k)\|^2,$$

由 $x^{k+1} = p(y^k, \mu_k)$ 、 $y^k = x^k + \theta_k(x^k - x^{k-1})$ 、 $\mu_{k+1} \leq \mu_k$ 和定义 2.1(v), 上式可表示成

$$\tilde{f}(x^{k+1}, \mu_{k+1}) + \kappa\mu_{k+1} + \frac{1}{2}L\mu_k^{-1}\|x^{k+1} - x^k\|^2 \leq \tilde{f}(x^k, \mu_k) + \kappa\mu_k + \frac{1}{2}L\mu_k^{-1}\theta_k^2\|x^k - x^{k-1}\|^2. \quad (2.15)$$

将引理 2.1(v) 代入到不等式 (2.15), 得到

$$\tilde{f}(x^{k+1}, \mu_{k+1}) + \kappa\mu_{k+1} + \frac{1}{2}L\mu_k^{-1}\|x^{k+1} - x^k\|^2 \leq \tilde{f}(x^k, \mu_k) + \kappa\mu_k + \frac{1}{2}L\mu_{k-1}^{-1}\|x^k - x^{k-1}\|^2, \quad (2.16)$$

即

$$W_k := \tilde{f}(x^k, \mu_k) + \kappa\mu_k + \frac{1}{2}L\mu_{k-1}^{-1}\|x^k - x^{k-1}\|^2$$

是非增的. 由 $\{x^k\} \subseteq \mathcal{X}$, $f(x^k) \geq \min_{\mathcal{X}} f$, 进而由 (2.1) 可知,

$$W_k \geq \tilde{f}(x^k, \mu_k) + \kappa\mu_k \geq f(x^k) \geq \min_{\mathcal{X}} f,$$

即 W_k 下有界. 所以, 下述极限存在:

$$\lim_{k \rightarrow \infty} \tilde{f}(x^k, \mu_k) + \kappa\mu_k + \frac{1}{2}L\mu_{k-1}^{-1}\|x^k - x^{k-1}\|^2.$$

再由 (2.1) 和 (2.16) 可知,

$$\{x^k\} \subseteq \left\{x \in \mathcal{X} : f(x) \leq \tilde{f}(x^1, \mu_1) + \kappa\mu_1 + \frac{1}{2}L\mu_0^{-1}\|x^1 - x^0\|^2\right\}. \quad (2.17)$$

在 (2.14) 中, 令 $x := x^0, y := y^0, \mu := \mu_0$, 由 $x^0 = y^0$, 类似于 (2.15) 的分析, 得到

$$\tilde{f}(x^1, \mu_1) + \kappa\mu_1 + \frac{1}{2}L\mu_0^{-1}\|x^1 - x^0\|^2 \leq \tilde{f}(x^0, \mu_0) + \kappa\mu_0.$$

通过 (2.17) 和假设 2.1, 可知序列 $\{x^k\}$ 有界. $\square$

定理 2.1 若假设 2.1 成立, 则 FISTA_S 生成的序列 $\{x^k\}$ 的任意聚点都为优化问题 (1.1) 的最优解, 而且存在常数 $C > 0$, 使得

$$f(x^{k+1}) - \min_{\mathcal{X}} f \leq \begin{cases} Ck^{-\sigma}, & \text{若 } 0 < \sigma < 1, \\ C\frac{\ln k}{k}, & \text{若 } \sigma = 1, \\ Ck^{\sigma-2}, & \text{若 } 1 < \sigma < 2, \end{cases} \quad \forall k,$$

即对应于 $0 < \sigma < 1$, $\sigma = 1$ 和 $1 < \sigma < 2$, 目标函数值的收敛速率分别 $O(k^{-\sigma})$ 、 $O(\ln k/k)$ 和 $O(k^{\sigma-2})$.

证明 首先, 在 (2.11) 中, 令 $x := x^k$, $y := y^k$, $\mu := \mu_k$, 得到

$$\tilde{f}(x^k, \mu_k) - \tilde{f}(p(y^k, \mu_k), \mu_k) \geq \frac{1}{2}L\mu_k^{-1}\|p(y^k, \mu_k) - y^k\|^2 + L\mu_k^{-1}\langle y^k - x^k, p(y^k, \mu_k) - y^k \rangle.$$

记 $v^k = \tilde{f}(x^k, \mu_k) + \kappa\mu_k - f(x^*)$, 其中 $x^* \in \mathcal{M}$. 由 (2.9) 和定义 2.1(v), 上式可表示为

$$v^k - v^{k+1} \geq \frac{1}{2}L\mu_k^{-1}\|x^{k+1} - y^k\|^2 + L\mu_k^{-1}\langle y^k - x^k, x^{k+1} - y^k \rangle. \quad (2.18)$$

其次, 在 (2.11) 中, 令 $x := x^*$, $y := y^k$, $\mu := \mu_k$, 类似于 (2.18) 的分析, 得到

$$\tilde{f}(x^*, \mu_k) - \tilde{f}(x^{k+1}, \mu_k) \geq \frac{1}{2}L\mu_k^{-1}\|x^{k+1} - y^k\|^2 + L\mu_k^{-1}\langle y^k - x^*, x^{k+1} - y^k \rangle. \quad (2.19)$$

对于上式左边, 再次利用定义 2.1(v), 可得

$$\tilde{f}(x^*, \mu_k) - \tilde{f}(x^{k+1}, \mu_k) \leq f(x^*) + \kappa\mu_k - \tilde{f}(x^{k+1}, \mu_{k+1}) + \kappa(\mu_k - \mu_{k+1}) = -v^{k+1} + 2\kappa\mu_k.$$

所以, 由 (2.19) 推导得到

$$-v^{k+1} + 2\kappa\mu_k \geq \frac{1}{2}L\mu_k^{-1}\|x^{k+1} - y^k\|^2 + L\mu_k^{-1}\langle y^k - x^*, x^{k+1} - y^k \rangle. \quad (2.20)$$

令 $((2.18) \times (t_k - 1) + (2.20)) \times t_k$, 可得

$$\begin{aligned} & 2L^{-1}t_k(t_k - 1)\mu_kv^k - 2L^{-1}t_k^2\mu_kv^{k+1} \\ & \geq t_k^2\|x^{k+1} - y^k\|^2 + 2t_k\langle x^{k+1} - y^k, t_ky^k - (t_k - 1)x^k - x^* \rangle - 4\kappa L^{-1}t_k\mu_k^2. \end{aligned} \quad (2.21)$$

类似于文献 [2] 中的计算方法, 可得

$$t_k^2\|x^{k+1} - y^k\|^2 + 2t_k\langle x^{k+1} - y^k, t_ky^k - (t_k - 1)x^k - x^* \rangle = \|u^{k+1}\|^2 - \|u^k\|^2,$$

其中

$$u^k = t_{k-1}x^k - (t_{k-1} - 1)x^{k-1} - x^*.$$

由 (2.21) 和引理 2.1(ii) 可知,

$$2L^{-1}t_{k-1}^2\mu_{k-1}v^k - 2L^{-1}t_k^2\mu_kv^{k+1} \geq \|u^{k+1}\|^2 - \|u^k\|^2 - 4\kappa L^{-1}t_k\mu_k^2,$$

即

$$2L^{-1}t_k^2\mu_kv^{k+1} + \|u^{k+1}\|^2 \leq 2L^{-1}t_{k-1}^2\mu_{k-1}v^k + \|u^k\|^2 + 4\kappa L^{-1}t_k\mu_k^2.$$

因此,

$$2L^{-1}t_k^2\mu_k v^{k+1} + \|u^{k+1}\|^2 \leq 2L^{-1}t_0^2\mu_0 v^1 + \|u^1\|^2 + 4\kappa L^{-1} \sum_{i=1}^k t_i \mu_i^2. \quad (2.22)$$

下面估计上式中的右端项. 在 (2.14) 中, 令 $x := x^*$, $y := y^0$, $\mu := \mu_0$, 得到

$$\begin{aligned} 2L^{-1}t_0^2\mu_0 v^1 + \|u^1\|^2 &= 2L^{-1}\mu_0 v^1 + \|x^1 - x^*\|^2 \\ &\leq \|x^0 - x^*\|^2 + 4\kappa L^{-1}\mu_0^2. \end{aligned} \quad (2.23)$$

为给出目标函数值的收敛速率, 下面的工作是估计 $\sum_{i=1}^k t_i \mu_i^2$ 的上界和 $t_k^2 \mu_k$ 的下界. 利用引理 2.1(iv), 可知

$$\begin{aligned} \sum_{i=1}^k t_i \mu_i^2 &\leq \frac{4\mu_0^2}{2-\sigma} \sum_{i=1}^k i^{1-2\sigma} \\ &\leq \begin{cases} \frac{4\mu_0^2}{(2-\sigma)(2-2\sigma)} (k+1)^{2-2\sigma}, & \text{若 } 0 < \sigma < 1, \\ \frac{4\mu_0^2}{2-\sigma} (\ln k + 1), & \text{若 } \sigma = 1, \\ \frac{4\mu_0^2(2\sigma-1)}{(2-\sigma)(2\sigma-2)}, & \text{若 } 1 < \sigma < 2. \end{cases} \end{aligned} \quad (2.24)$$

利用引理 2.1(iii), 对于给定的 $\sigma \in (0, 2)$, 存在 $K \in \mathbb{N}$ 使得 $(k+1)^{1-\frac{\sigma}{2}} \geq 2, \forall k \geq K$, 所以,

$$t_k^2 \mu_k \geq \frac{\mu_0}{4(2-\sigma)^2} (k+1)^{2-\sigma}, \quad \forall k \geq K.$$

再次利用定义 2.1(v), 可得

$$v^{k+1} = \tilde{f}(x^{k+1}, \mu_{k+1}) + \kappa \mu_{k+1} - f(x^*) \geq f(x^{k+1}) - f(x^*).$$

将上述分析代入 (2.22), 可得证该定理. $\square$

由上述定理可知, 在理论分析上, 当 $\sigma = 1$ 时, 目标函数具有最优的收敛速率 $O(\ln k/k)$. 特别地, 当 $\mu_0 = 1$ 和 $\sigma = 1$ 时, 由以上证明可得到如下全局估计:

$$f(x^k) - f(x^*) \leq C(k) := \frac{2(\|x^0 - x^*\|^2 + 4\kappa L^{-1} + 4 \ln k + 4)}{k+1}. \quad (2.25)$$

注 2.1 根据定理 2.1 的证明, 其函数值收敛速率结果对于无穷维 Hilbert 空间下的优化问题 (1.1) 仍然成立. 本文所考虑的模型 (1.1) 是一类通用的泛化优化问题, 特定应用背景下的具体优化模型结构常可给算法的设计和收敛性分析带来新的启发和更好的结果, 尝试对几类具体应用背景下的特殊优化模型建立具有更快、更强收敛性的算法是一件更有意义的研究工作.

推论 2.1 若假设 2.1 成立, 则

- (i) $\sum_{k=1}^{+\infty} (k+1)^{\sigma-1} \|x^k - x^{k-1}\|^2 < +\infty$;
- (ii) 存在 $M > 0$, 使得对任意的 $k \in \mathbb{N}$, 有

$$\|x^{k+1} - x^k\| \leq \begin{cases} Mk^{-\sigma}, & \text{若 } 0 < \sigma < 1, \\ Mk^{-1} \sqrt{\ln k}, & \text{若 } \sigma = 1, \\ Mk^{-1}, & \text{若 } 1 < \sigma < 2. \end{cases}$$

证明 (i) 应用不等式 (2.15), 利用引理 2.1(vi), 对 $k \geq K$, 成立

$$\begin{aligned} & \frac{1}{2}L(t_k\mu_{k-1})^{-1}\|x^k - x^{k-1}\|^2 \\ & \leq \tilde{f}(x^k, \mu_k) + \kappa\mu_k + \frac{1}{2}L\mu_{k-1}^{-1}\|x^k - x^{k-1}\|^2 \\ & \quad - \left(\tilde{f}(x^{k+1}, \mu_{k+1}) + \kappa\mu_{k+1} + \frac{1}{2}L\mu_k^{-1}\|x^{k+1} - x^k\|^2 \right). \end{aligned}$$

对上式的 k 从 K 到 $+\infty$ 求和, 由引理 2.3, 得到

$$\sum_{k=K}^{+\infty} \frac{1}{2}L(t_k\mu_{k-1})^{-1}\|x^k - x^{k-1}\|^2 \leq W_K - \lim_{k \rightarrow +\infty} W_k < +\infty,$$

其中

$$W_k = \tilde{f}(x^k, \mu_k) + \kappa\mu_k + \frac{1}{2}L\mu_{k-1}^{-1}\|x^k - x^{k-1}\|^2.$$

再由 $\mu_{k-1}^{-1} \geq \frac{1}{2^\sigma}\mu_k^{-1}$, 得到

$$\sum_{k=1}^{+\infty} (t_k\mu_k)^{-1}\|x^k - x^{k-1}\|^2 < +\infty. \quad (2.26)$$

由引理 2.1(iii) 和 μ_k 的定义, 有

$$t_k\mu_k = t_k\sqrt{\mu_k}\sqrt{\mu_k} \leq \frac{4\mu_0}{2-\sigma}(k+1)^{1-\sigma} \left(\frac{k+1}{k} \right)^{\sigma/2} \leq \frac{8\mu_0}{2-\sigma}(k+1)^{1-\sigma}.$$

即得证结论 (i).

(ii) 通过 (2.22) 和 (2.23), 得到

$$\|u^{k+1}\|^2 \leq \|x^0 - x^*\|^2 + 4\kappa L^{-1} \sum_{i=0}^k t_i \mu_i^2.$$

由 (2.24) 知, 存在 $M > 0$, 使得

$$\|u^{k+1}\| \leq \begin{cases} Mk^{1-\sigma}, & \text{若 } 0 < \sigma < 1, \\ M\sqrt{\ln k}, & \text{若 } \sigma = 1, \\ M, & \text{若 } 1 < \sigma < 2. \end{cases}$$

再利用 $\|u^{k+1}\| = \|t_k x^{k+1} - (t_k - 1)x^k - x^*\| = \|t_k(x^{k+1} - x^k) + x^k - x^*\| \geq t_k\|x^{k+1} - x^k\| - \|x^k - x^*\|$, 以及 $\{x^k\}$ 的有界性和 $t_k \geq \frac{1}{2}k$, 得证结论 (ii). $\square$

对于任意的 $\sigma \in (0, 2)$, 推论 2.1(ii) 给出

$$\lim_{k \rightarrow +\infty} \|x^{k+1} - x^k\| = 0.$$

根据引理 2.2 和定理 2.1 的证明, 算法中的常数 L 仅需满足不等式 (2.13) 即可. 所以, 在无法或较难预估 L 具体取值的情形下, 我们可在 FISTA_S 中加入一个子循环以产生满足条件的 L , 进而给出算法 r-FISTA_S .

算法 2 r-FISTA_S

步 0: 给定 $y^0 = x^0 \in \mathcal{X}$, $t_0 = 1$, $\mu_0 \in (0, 1]$, $\sigma \in (0, 2)$, $L_0 \in [1, +\infty)$, $\eta > 1$. 令 $k = 0$.

步 1: 令 $L_k = L_0$.

(1a) 计算 $\hat{x}^{k+1} = p(y^k, \mu_k)$.

(1b) 如果

$$\tilde{h}(\hat{x}^{k+1}, \mu_k) \leq \tilde{h}(y^k, \mu_k) + \langle \nabla \tilde{h}(y^k, \mu_k), \hat{x}^{k+1} - y^k \rangle + \frac{1}{2} L_k \mu_k^{-1} \|\hat{x}^{k+1} - y^k\|^2, \quad (2.27)$$

则令 $x^{k+1} = \hat{x}^{k+1}$, 并转到步 2; 否则, 令 $L_k := \eta L_k$, 并回到步 (1a).

步 2: 计算

$$\begin{aligned} \mu_{k+1} &= \frac{\mu_0}{(k+1)^\sigma}, \\ t_{k+1} &= \frac{1 + \sqrt{1 + 4(\frac{\mu_k}{\mu_{k+1}})t_k^2}}{2}, \\ \theta_{k+1} &= \frac{t_k - 1}{t_{k+1}}, \\ y^{k+1} &= x^{k+1} + \theta_{k+1}(x^{k+1} - x^k). \end{aligned} \quad (2.28)$$

步 3: 置 $k := k + 1$, 转到步 1.

由定义 2.1(vi), 当 $L_k \geq L$ 时, (2.27) 显然成立. 所以, 对任意的 $k \in \mathbb{N}$, 有 $L_k \leq \eta L$, 而且步 1 内的子循环至多为 $\lfloor \log_\eta L \rfloor + 1$ 次, 其中 $\lfloor a \rfloor$ 表示不大于 a 的最大正整数. 类似于定理 2.1 的证明, 其结论对算法 r-FISTA_S 仍然成立. 另一方面, L 数值的大小直接影响算法的步长. 若 L 太小, 无法保证算法收敛于优化问题的最优解; 而 L 太大, 会大大增加所需的迭代步数, 降低计算效率. 因此, 通过线性搜索寻找适当的 L , 算法 r-FISTA_S 在一定程度上比 FISTA_S 具有更强的计算性能.

2.3 无外插情形

对于 FISTA_S, 当 $\theta_k = 0$, 即无外插项时, 算法为结合光滑函数的临近梯度 (smoothing proximal gradient) 算法, 记为 PG_S, 算法框架如下:

$$\begin{cases} x^{k+1} = p(x^k, \mu_k), \\ \mu_{k+1} = \frac{\mu_0}{(k+1)^\sigma}. \end{cases} \quad (2.29)$$

类似于对 FISTA_S 的分析方法, 取 $x_0 \in \mathcal{X}$, $\mu_0 = 1$, 算法 PG_S 生成的迭代序列具有下述性质.

推论 2.2 若假设 2.1 成立, 当 $\sigma \in (0, 1)$ 时, PG_S 算法生成的序列 $\{x^k\}$ 的任意聚点都为优化问题 (1.1) 的最优解, 而且

- (i) $\tilde{f}(x^k, \mu_k) + \kappa \mu_k$ 非增, 且极限存在;
- (ii) $\sum_{k=1}^{+\infty} \mu_k^{-1} \|x^{k+1} - x^k\|^2 < +\infty$;
- (iii) 存在常数 $C > 0$, 使得

$$f(x^{k+1}) - \min_{\mathcal{X}} f \leq C k^{-\min\{\sigma, 1-\sigma\}}, \quad \forall k.$$

证明 类似于引理 2.3 和定理 2.1 的证明方法, 可得结论 (i) 和 (ii). 下面证明结论 (iii).

在 (2.14) 中, 令 $x := x^*$, $y := x^k$, $\mu := \mu_k$, 则得到

$$\tilde{f}(x^{k+1}, \mu_k) \leq \tilde{f}(x^*, \mu_k) + \frac{1}{2} L \mu_k^{-1} \|x^* - x^k\|^2 - \frac{1}{2} L \mu_k^{-1} \|x^* - x^{k+1}\|^2.$$

利用定义 2.1(v), 得到

$$\tilde{f}(x^{k+1}, \mu_{k+1}) + \kappa\mu_{k+1} - f(x^*) \leq \frac{1}{2}L\mu_k^{-1}\|x^* - x^k\|^2 - \frac{1}{2}L\mu_k^{-1}\|x^* - x^{k+1}\|^2 + 2\kappa\mu_k.$$

在上式中取 $k = 0, 1, \dots, K$, 并求和, 利用结论 (i) 和 $\tilde{f}(x^{k+1}, \mu_{k+1}) + \kappa\mu_{k+1} \geq f(x^*)$, 得到

$$\begin{aligned} & (K+1)(\tilde{f}(x^{K+1}, \mu_{K+1}) + \kappa\mu_{K+1} - f(x^*)) \\ & \leq \frac{1}{2}L\mu_0^{-1}\|x^* - x^0\|^2 + \frac{1}{2}L \sum_{i=1}^K (\mu_i^{-1} - \mu_{i-1}^{-1})\|x^* - x^i\|^2 \\ & \quad - \frac{1}{2}L\mu_K^{-1}\|x^* - x^{K+1}\|^2 + 2\kappa \sum_{i=0}^K \mu_i. \end{aligned}$$

由假设 2.1 和结论 (i) 知, 序列 $\{x^k\}$ 有界, 记 $\|x^k - x^*\|^2 \leq \bar{R}$. 再由

$$\tilde{f}(x^{K+1}, \mu_{K+1}) + \kappa\mu_{K+1} \geq f(x^{K+1}),$$

得到

$$\begin{aligned} (K+1)(f(x^{K+1}) - f(x^*)) & \leq \frac{1}{2}L\mu_0^{-1}\|x^* - x^0\|^2 + \frac{1}{2}L\bar{R} \sum_{i=1}^K (\mu_i^{-1} - \mu_{i-1}^{-1}) + 2\kappa \sum_{i=0}^K \mu_i \\ & \leq \frac{1}{2}L\mu_0^{-1}\|x^* - x^0\|^2 + \frac{1}{2}L\bar{R}\mu_K^{-1} + 2\kappa \sum_{i=0}^K \mu_i \\ & \leq \frac{1}{2}L\mu_0^{-1}\|x^* - x^0\|^2 + \frac{1}{2}L\bar{R}\mu_0^{-1}K^\sigma + 2\kappa\mu_0 \left(\frac{1}{1-\sigma}K^{1-\sigma} + 1 \right). \end{aligned}$$

因此,

$$f(x^k) - f(x^*) \leq \frac{L\mu_0^{-1}\|x^* - x^0\|^2 + 4\kappa\mu_0}{2k} + \left(\frac{1}{2}L\mu_0^{-1}\bar{R} \right) k^{-1+\sigma} + \frac{2\kappa\mu_0}{1-\sigma} k^{-\sigma}.$$

结论 (iii) 得证. $\square$

3 数值实验

本节通过考虑两类具有代表性的 ℓ_1 正则化优化模型, 来展示 FISTA_S 和 r-FISTA_S 对优化问题 (1.1) 的数值求解效果.

例 3.1 (ℓ_1 损失函数与 ℓ_1 正则函数模型) 考虑如下不仅以 ℓ_1 为正则函数而且以 ℓ_1 为损失函数的稀疏优化模型:

$$\min_{0_n \leq x \leq 5_n} f(x) := \|Ax - b\|_1 + 0.01\|x\|_1, \quad (3.1)$$

其中 $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

令 $h(x) = \|Ax - b\|_1$, $g(x) = 0.01\|x\|_1$, 则 (3.1) 可写成 (1.1) 的形式, 且 $\text{Prox}_{\alpha g}(y, \mathcal{X})$ 具有显式解. 参照 (2.2) 构造 h 的光滑函数, 则定义 2.1(vi) 中的常数 L 可取为 $\|AA^T\|$. 令

$$(m, n, s) = (500, 5000, 50),$$

按照下述命令生成一组数据 (A, b, x_s) :

```
index = randperm(n); index = index(1:s);
x_s = zeros(n,1); B = randn(n,m); A = orth(B)^T;
x_s(index) = unifrnd(0,1,[s,1]); b = A*x_s.
```

记 $f_{\min} = f(x_s)$. 令 $x^0 = \mathbf{0}_n$, $\mu_0 = 1$, $L = \|AA^T\|$. 取 $\sigma = 1$, 图 1 展示了目标函数值 $f(x^k) - f_{\min}$ 和迭代序列 $\|x^k - x_s\|$ 的收敛情形, 其中实线和点线分别对应由 FISTA_S 和不带外插的 PG_S 算法生成的迭代序列的情形. 通过计算可知, 对于此例中的函数 h 和光滑函数 $\tilde{h}$, 定义 2.1 中的 $\kappa = \frac{1}{2}m$. (2.25) 给出目标函数收敛的一个上界估计. 虚线描述 (2.25) 中的 $C(k)$ 走势, 其中 $x^* = x_s$. 从图像可知, 实验效果与理论结果吻合, 而且相比不带外插的情形, 带有外插的 FISTA_S 生成的迭代序列具有更好的数值表现. 图 2 给出了 FISTA_S 中参数 σ 选取不同数值时目标函数值和迭代序列的收敛情形. 从此图像可得到与定理 2.1 吻合的结论. 特别地, 当 $\sigma > 2$ 时, 迭代序列 x^k 不收敛于优化问题 (3.1) 的最优解.

例 3.2 (max 损失函数与 ℓ_1 正则函数模型) 考虑如下稀疏优化模型:

$$\min_{-1_n \leq x \leq 1_n} f(x) := \|\max\{Ax, 0\} - b\|_1 + 0.01\|x\|_1, \quad (3.2)$$

其中 A 和 b 如例 3.1 中所定义. 类似地, 令 $h(x) = \|\max\{Ax, 0\} - b\|_1$, $g(x) = 0.01\|x\|_1$, 此问题仍可写成模型 (1.1) 的形式. 对于给定的 $(m, n, s) = (500, 5000, 50)$, 按照下述命令生成一组数据 (A, b) :

```
index = randperm(n); index = index(1:s);
B = randn(n,m); A=orth(B)^T; x_s = zeros(n,1);
x_s(index) = unifrnd(-1,1,[s,1]); y = max(A*x_s+0.01*randn(n,1),0).
```

利用 (2.3) 给出 h 的光滑函数 $\tilde{h}$. 鉴于对此 $\tilde{h}$, 不易给出定义 2.1(vi) 中常数 L 的取值, 我们使用算法 r-FISTA_S 求解问题 (3.2). 将算法 r-FISTA_S 中 L_k 的更新方法应用于算法 PG_S , 并记此算法为 r-PG_S . 令 $x^0 = \mathbf{0}_n$, $\mu_0 = 1$, $L_0 = 1$. 图 3(a) 给出了在 $\sigma = 1$ 情形下, 算法 r-FISTA_S 和 r-PG_S 中目标函数值的收敛情形, 以及 (2.25) 中 $C(k)$ 的走势, 其中的 L 取为 $\max\{L_k\}$; 图 3(b) 给出了 σ 取不同值时, 目标函数的收敛情形. 图 3 显示的数值效果与例 3.1 中的一致.

图 1 (网络版彩图) 当 $\sigma = 1$ 时, 算法 FISTA_S 和 PG_S 对问题 (3.1) 的求解效果. (a) 目标函数值 $f(x^k) - f_{\min}$ 的收敛情形; (b) 迭代序列 $\|x^k - x_s\|$ 的收敛情形

图 2 针对不同 σ , FISTA_S 对问题 (3.1) 的求解情形. (a) 目标函数值 $f(x^k) - f_{\min}$ 的收敛情形; (b) 迭代序列 $\|x^k - x_s\|$ 的收敛情形

图 3 算法对问题 (3.2) 的求解情形. (a) 算法 $r\text{-FISTA}_S$ 和 $r\text{-PG}_S$ 中, 目标函数值 $f(x^k)$ 的收敛情形; (b) 针对不同 σ , 目标函数值 $f(x^k)$ 的收敛情形

致谢 本文是在哈尔滨工业大学薛小平教授的推荐和鼓励下完成的, 在此向薛教授表示衷心的感谢. 本文所涉及的光滑函数的定义、性质是在香港理工大学陈小君教授的指导下建立的. 作者也借此机会向陈小君教授致以诚挚的谢意.

参考文献

- 1 Parikh N, Boyd S. Proximal algorithms. *Found Trends Optim*, 2014, 1: 127–239
- 2 Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J Imag Sci*, 2009, 2: 183–202
- 3 Nesterov Y. Gradient methods for minimizing composite functions. *Math Program*, 2013, 140: 125–161
- 4 O'Donoghue B, Candès E. Adaptive restart for accelerated gradient schemes. *Found Comput Math*, 2015, 15: 715–732
- 5 Tao S, Boley D, Zhang S. Local linear convergence of ISTA and FISTA on the LASSO problem. *SIAM J Optim*, 2016, 26: 313–336
- 6 Attouch H, Peypouquet J. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM J Optim*, 2016, 26: 1824–1834
- 7 Johnstone P R, Moulin P. Local and global convergence of a general inertial proximal splitting scheme for minimizing composite functions. *Comput Optim Appl*, 2017, 67: 259–292
- 8 Wen B, Chen X, Pong T K. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM J Optim*, 2017, 27: 124–145
- 9 Wen B, Xue X. On the convergence of the iterates of proximal gradient algorithm with extrapolation for convex nonsmooth minimization problems. *J Global Optim*, 2019, 75: 767–787

- 10 Bauschke H H, Bolte J, Teboulle M. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Math Oper Res*, 2017, 42: 330–348
- 11 Nesterov Y. Smooth minimization of non-smooth functions. *Math Program*, 2005, 103: 127–152
- 12 Hoda S, Gilpin A, Peña J, et al. Smoothing techniques for computing Nash equilibria of sequential games. *Math Oper Res*, 2010, 35: 494–512
- 13 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, 2001, 96: 1348–1360
- 14 Bian W, Chen X. A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. *SIAM J Numer Anal*, 2020, 58: 858–883
- 15 Chen X. Smoothing methods for nonsmooth, nonconvex minimization. *Math Program*, 2012, 134: 71–99

Smoothing accelerated algorithm for constrained nonsmooth convex optimization problems

Wei Bian

Abstract In this paper, we consider a class of constrained nonsmooth convex optimization problems, which have wide applications in signal processing, imaging recovery, machine learning, etc. Recently, theoretical analysis and numerical experiments show that the extrapolation can improve the convergence rate of the algorithms efficiently, which let the proximal gradient algorithm with extrapolation have significant advantages in solving large scale optimization problems. Motivated by the smoothing techniques, and the fast iterative shrinkage-thresholding algorithm (FISTA) proposed by Beck and Teboulle, we bring forward a new accelerated algorithm for solving the considered constrained nonsmooth convex optimization problems. We prove that any accumulated point of the algorithm is an optimal solution of the problem. In the analysis of the algorithm, we consider several different update methods for the smoothing parameter, and give the global convergence rate $O(\ln k/k)$ on the objective function values. Moreover, we analyze the convergence rate of the difference on the iterates $\lim_{k \rightarrow \infty} \|x^{k+1} - x^k\| = 0$. Finally, our numerical experiments illustrate the good performance of the proposed algorithm in solving two classes of sparse optimization problems and the positive influence of the extrapolation on the convergence rate.

Keywords nonsmooth convex optimization problem, accelerated algorithm, smoothing technique, convergence rate

MSC(2010) 90C30, 90C25, 65K05

doi: 10.1360/SSM-2020-0181