2020 年 11 月 计 算 数 学 第 42 卷第 4 期
Nov., 2020 MATHEMATICA NUMERICA SINICA Vol.42, No.4

青年评述

复合凸优化的快速邻近点算法*1)

酈旭东2)
(复旦大学大数据学院, 上海数学中心, 上海 200433)

摘 要

在大数据时代, 随着数据采集手段的不断提升, 大规模复合凸优化问题大量的出现在包括统计数据分析, 机器与统计学习以及信号与图像处理等应用中. 本文针对大规模复合凸优化问题介绍了一类快速邻近点算法. 在易计算的近似准则和较弱的平稳性条件下, 本文给出了该算法的全局收敛与局部渐近超线性收敛结果. 同时, 我们设计了基于对偶原理的半光滑牛顿法来高效稳定求解邻近点算法所涉及的重要子问题. 最后, 本文还讨论了如何通过深入挖掘并利用复合凸优化问题中由非光滑正则函数所诱导的非光滑二阶信息来极大减少半光滑牛顿算法中求解牛顿线性系统所需的工作量, 从而进一步加速邻近点算法.

关键词: 复合优化; 邻近点算法; 半光滑牛顿算法.
MR (2010) 主题分类: 65F10, 90C06, 90C25, 90C31.

1. 前 言

记 $\mathcal{X},\mathcal{Y}$ 为有限维欧式空间, 每个空间都有内积 $\langle\cdot,\cdot\rangle$, 和它的诱导范数 $\|\cdot\|$. 给定线性算子 $\mathcal{A}:\mathcal{X}\to\mathcal{Y}$, 正常闭 (proper closed) 凸函数 $h:\mathcal{X}\to\Re$ 和 $p:\mathcal{X}\to(-\infty,+\infty]$, 数据向量 $c\in\mathcal{X}$, 本文考虑下面的复合凸优化问题:

$$\min_{x\in\mathcal{X}}\Big\{h(\mathcal{A}x)-\langle c,x\rangle+p(x)\Big\}. \tag{1.1}$$

文献中 h 通常为光滑损失函数, 用来减少数据拟合误差; p 通常为非光滑正则函数, 用来诱导最优解的某种性质, 如稀疏性、平整性、低秩性等.

当前, 由于数据采集技术的进步, 大数据时代对我们传统的数据分析手段带来了新的挑战. 为了应对大样本量以及高维数据问题带来的挑战, 模型 (1.1) 被广泛的应用于统计数据分析、机器学习、信号与图像处理等方面. 例如, 为了处理高维线性回归模型, 提高回归预测的准确性与鲁棒性, Tibshirani 在 [55] 中考虑了如下 Lasso 模型以期获得稀疏回归解

$$\min_{x\in\Re^n}\left\{\frac{1}{2}\|Ax-b\|^2+\lambda_1\|x\|_1\right\},$$

* 2020 年 9 月 28 日收到.
1) 基金项目: 国家自然科学基金 (11901107), 中国科协青年人才托举工程 (2019QNRC001), 上海市扬帆计划 (19YF1402600), 上海市科委项目 (19511120700) 资助.
2) 作者简介: 酈旭东, 复旦大学大数据学院青年研究员, 上海数学中心青年研究员. 2010 年本科毕业于中国科学技术大学数学系, 2015 年在新加坡国立大学数学系获博士学位. 博士毕业后曾在新加坡国立大学数学系与美国普林斯顿大学运筹与金融工程系任博士后研究员, 2018 年 9 月入职复旦大学, 于 2019 年获得由国际数学优化协会 (Mathematical Optimization Society) 所颁发的连续优化青年学者最佳论文奖, 2020 年入选第五届中国科协青年人才托举工程, 现任期刊《Mathematical Programming Computation》编委.

其中矩阵 $A \in \Re^{m\times n}$ 和向量 $b \in \Re^m$ 为给定数据, $\lambda_1 > 0$ 为给定参数. 显然该问题是模型 (1.1) 的一个特例, 特别的, 该问题对应于在 (1.1) 中取 $\mathcal{X} = \Re^n$, $\mathcal{Y} = \Re^m$, $h(x) = \frac{1}{2}\|x-b\|^2$, $c = 0$, $p(x) = \lambda_1\|x\|_1$. 随后, 为更好的拟合应用中数据出现的其它特殊结构, 如结构化稀疏性、一致性、聚集性等, 研究人员设计了更多的正则化函数, 提出了如 Fused Lasso [56], Clustered Lasso [51], OSCAR [6], Sparse Group Lasso [62] 和 Elastic Net [67] 等多种复合凸优化问题. 另外, 模型 (1.1) 也被广泛应用于矩阵优化 (matrix optimization) 中, 例如当 $\mathcal{X}$ 取成矩阵空间, p 取成矩阵核范数 (nuclear norm) $\|\cdot\|_*$ 时, (1.1) 就变成了图像与信号处理领域受到广泛关注的低秩矩阵优化问题 [2]. 其它更多的关于复合凸矩阵优化问题可以参见 [10, 12, 13].

这些应用吸引了优化领域众多研究人员为模型 (1.1) 或者它的各种等价形式设计开发各种高效算法. 这里, 我们做一个简要的总结. 首先, 在一些假设下, 流行的一阶算法如邻近梯度算法 (proximal gradient method) [3], 加速邻近梯度算法 (accelerated proximal gradient method) [1, 42], 一些算子分裂方法 (operator splitting method) [7] 以及交替方向乘子法 (alternating direction method of multipliers method) [19, 20] 都可以用来求解问题 (1.1) 或者它的等价变形. 然而一阶算法通常收敛速度较慢, 在很多大规模问题上数值表现差强人意. 因此, 很多研究人员开始考虑探索利用二阶信息来设计算法求解问题 (1.1), 如邻近牛顿算法 (proximal Newton method) [25, 63], 正则牛顿算法 (regularized Newton method) [59], 以及本文将要介绍的基于对偶半光滑牛顿的邻近点算法 (dual semismooth Newton based proximal point algorithm, PPDNA) [29, 30, 34, 38, 65]. 对于满足一些正则性条件的非退化问题, 邻近牛顿算法与正则牛顿算法的数值表现较好, 通常可以获得比上述一阶算法质量更好的近似最优解. 同时, 邻近牛顿算法的表现还与其所涉及的子问题求解方式密切相关. 而本文主要关注的 PPDNA 则不需要额外的非退化假设. 该算法的核心思想在于精细地对复合凸优化问题的内蕴非光滑结构进行分析, 并通过利用这些结构设计并求解条件数较好的对称正定非光滑牛顿线性系统. 事实上, 现有文献中大量关于 PPDNA 的数值实验表明, 对于很多困难大规模复合凸优化问题, 无论是寻找低精度还是高精度近似解, 该算法的效率都远超前文提到的流行一阶算法. 这一表现与邻近牛顿和正则牛顿算法的数值表现截然不同, 这两种算法的效率通常只有在寻求高精度解的时候才能超越一阶算法.

本文将要介绍的 PPDNA 主要包含两层迭代算法. 其中, 外层算法为邻近点算法 (proximal point algorithm, PPA), 内层基于对偶原理 (duality theory) 使用半光滑牛顿算法 (semismooth Newton method) 求解 PPA 中的子问题. 此算法框架设计灵感来源于半正定规划 (semidefinite programming) 领域的文献, 如 [28, 60, 64]. 特别的, 我们根据复合凸优化问题对偶问题的特质为 PPA 设计了易计算的近似准则并得到了它的全局收敛性. 另外, 基于较弱的平稳性 (calmness) 假设, 我们证明了外层 PPA 具有渐近超线性收敛速度. 这种局部快速收敛性质是 PPDNA 获得数值成功的基本保障. 同时, 通过利用 PPA 子问题的对偶问题目标函数光滑, 强凸等有利性质, 一类高效的半光滑牛顿算法被用来高效求解 PPA 子问题. 另外, 对于这些对偶问题的目标函数, 我们还深入探索与挖掘它们的广义 Hessian 矩阵中由非光滑正则函数 p 所诱导的特殊结构, 并利用这些结构极大减少求解牛顿线性系统所需工作量. 从文献 [29,30,34,38,65] 中大量关于 PPDNA 的数值实验可以观察到, 对很多大规模复合凸优化问题, PPDNA 中半光滑牛顿算法每步迭代的工作量与当前流行一阶算法每步迭代工作量相当, 甚至更少. 这一数值表现与关于二阶算法的传统观点 (即二阶算法通常每步迭代工作量巨大) 相反. PPDNA 取得这一进展的重要原因之一即为对原问题所蕴含的非光滑二阶信息的深入探索与利用. 因此, 我们认为由 p 所带来的非光滑二阶信息是高效稳定求解大规模困难复合

凸优化问题 (1.1) 的关键, 目前来看 PPDNA 框架可以较为完美的利用这一信息. PPDNA 的成功说明二阶信息, 尤其是非光滑二阶信息, 可以并且应该被精巧地利用到大规模优化问题求解算法的设计与开发中去.

本文将介绍如何为大规模复合凸优化问题 (1.1) 设计适合的邻近点算法, 并介绍如何基于对偶原理利用半光滑牛顿算法求解邻近点算法的子问题, 最后介绍如何挖掘并利用非光滑二阶信息加速邻近点算法.

2. 预备知识

为了更好的展示、讨论我们的算法, 我们在本节介绍一些必要的预备知识, 包括多值集合映射 (set-valued mapping) 的度量次正则性 (metrically subregularity) 以及邻近点算法.

对 $\mathcal{X}$ 上给定的增广实值 (extended-real valued) 正常闭凸函数 $p:\mathcal{X}\to(-\infty,+\infty]$, 我们定义该函数的 Moreau-Yosida 正则函数为

$$\psi_p(x):=\min_{y\in\mathcal{X}}\Big\{p(y)+\frac{1}{2}\|y-x\|^2\Big\},\quad x\in\mathcal{X},$$

其关联的邻近算子为

$$\mathrm{Prox}_p(x):=\operatorname*{argmin}_{y\in\mathcal{X}}\Big\{p(y)+\frac{1}{2}\|y-x\|^2\Big\},\quad x\in\mathcal{X}.$$

我们由 [47, Theorem 31.5] 可知 ψ_p 是一个连续可微的凸函数, Prox_p 是一个 Lipschitz 常数为 1 的全局 Lipschitz 连续函数, 并且有

$$\nabla\psi_p(x)=x-\mathrm{Prox}_p(x),\quad\forall\,x\in\mathcal{X}. \tag{2.1}$$

本文还会用到如下关于 Moreau-Yosida 正则函数的著名等式 (Moreau's identities [47, Theorem 31.5]): 对于任意的 $x\in\mathcal{X}$,

$$\begin{cases}\mathrm{Prox}_p(x)+\mathrm{Prox}_{p^*}(x)=x,\\ \min\limits_{y\in\mathcal{X}}\Big\{p(y)+\frac{1}{2}\|y-x\|^2\Big\}+\min\limits_{z\in\mathcal{X}}\Big\{p^*(z)+\frac{1}{2}\|z-x\|^2\Big\}=\frac{1}{2}\|x\|^2,\end{cases} \tag{2.2}$$

这里 p^* 是 p 的 Fenchel 共轭 (Fenchel's conjugate) 函数 [47, Section 12].

令 $G:\mathcal{X}\rightrightarrows\mathcal{Y}$ 为一集合映射, 定义 G 的图为 $\mathrm{gph}\,G:=\{(u,v)\in\mathcal{X}\times\mathcal{Y}\mid v\in G(u)\}$ 以及 G 的逆映射 $G^{-1}(v)=\{u\in\mathcal{X}\mid v\in G(u)\}$, $v\in\mathcal{Y}$. 对于任给的 $u\in\mathcal{X}$ 和 $\rho>0$, 我们记 $\mathbb{B}_\rho(u):=\{s\in\mathcal{X}\mid\|s-u\|\le\rho\}$. 下面关于 G 的度量正则性的定义可以在 [14, Section 3.8(3H)] 找到.

定义 1. 集合映射 $G:\mathcal{X}\rightrightarrows\mathcal{Y}$ 在 $\bar u$ 关于 $\bar v$ 是次度量正则的, 如果 $(\bar u,\bar v)\in\mathrm{gph}\,G$, 并且存在正常数 $\delta,\varepsilon,\kappa>0$ 使得

$$\mathrm{dist}(u,G^{-1}(\bar v))\le\kappa\,\mathrm{dist}(\bar v,G(u)\cap\mathbb{B}_\delta(\bar v)),\quad\forall\,u\in\mathbb{B}_\varepsilon(\bar u).$$

这里常数 κ 被称为次度量正则模.

由 [14, Theorem 3H.3], 我们知道对于任何的集合映射 $G:\mathcal{X}\rightrightarrows\mathcal{Y}$ 和 G 图上的点 $(\bar u,\bar v)\in\mathrm{gph}\,G$, 映射 G 在 $\bar u$ 点关于 $\bar v$ 是次度量正则的当且仅当映射 G^{-1} 在 $\bar v$ 关于 $\bar u$ 是平稳的 (calm),

即存在正常数 $\delta', \varepsilon', \kappa' > 0$ 使得

$$G^{-1}(v) \cap \mathbb{B}_{\delta'}(\bar{u}) \subseteq G^{-1}(\bar{v}) + \kappa' \|v - \bar{v}\| \mathbb{B}_{\mathcal{X}}, \quad \forall\, v \in \mathbb{B}_{\varepsilon'}(\bar{v}). \tag{2.3}$$

这里 $\mathbb{B}_{\mathcal{X}}$ 是 $\mathcal{X}$ 中的单位球.

集合映射的次度量正则性和平稳性是现代变分分析理论中重要的概念, 研究人员从包括非光滑分析与扰动性分析等多种角度对它们进行大量的研究. 感兴趣的读者可以从 Dontchev 和 Rockafellar 的书 [14] 中了解关于这些概念更多更全面的研究. 当集合映射的次度量正则性和平稳性与优化问题联系起来时, 我们可以利用这些条件分析很多知名算法, 如邻近梯度算法 [37, 58], 邻近点算法 [26, 33, 39, 48] 以及广义牛顿算法 [15, 16, 41] 的收敛速度. 本文主要的算法框架为邻近点算法, 下面我们做一些简要介绍.

给定 $\mathcal{X}$ 上的极大单调算子 $T : \mathcal{X} \rightrightarrows \mathcal{X}$, 邻近点算法 (PPA) 被用来求解 $z \in \mathcal{X}$ 使得如下包含关系成立

$$0 \in T(z). \tag{2.4}$$

给定正项单调数列 $\{\sigma_k\}$ 使得 $\sigma_k \uparrow \sigma_\infty \leq \infty$[1] 和初始点 $z^0 \in \mathcal{X}$, PPA 第 $k+1$ 步的迭代公式可以写成:

$$z^{k+1} \approx P_k(z^k) := (I + \sigma_k T)^{-1}(z^k), \quad \forall\, k \geq 0, \tag{2.5}$$

这里 I 为 $\mathcal{X}$ 中的单位算子. 在 [48] 中, 为了保证 PPA 的全局收敛, Rockafellar 提出使用如下近似准则来计算 z^{k+1}, 即 $P_k(z^k)$ 的近似值:

(A) $\|P_k(z^k) - z^{k+1}\| \leq \varepsilon_k, \quad \varepsilon_k \geq 0, \quad \sum_{k=1}^{\infty} \varepsilon_k < \infty,$

(B) $\|P_k(z^k) - z^{k+1}\| \leq \eta_k \|z^{k+1} - z^k\|, \quad 0 \leq \eta_k < 1, \quad \sum_{k=1}^{\infty} \eta_k < \infty.$

仔细观察可以发现, 上面的近似准则在实际计算中无法直接使用, 因为 $P_k(z^k)$ 一般情况下无法被精确计算. 因此在 PPA 的实际使用中, 一个重要的问题即为设计易于计算的近似准则. 本文将在后面的小节中讨论如何在求解模型 (1.1) 时为 PPA 设计适合计算的近似准则. 下面我们首先给出 PPA 的全局收敛结果. 该结果及其证明可以在文献 [48] 中找到.

命题 1. 假设集合 $T^{-1}(0)$ 非空. 令 $\{z^k\}$ 为 PPA (2.5) 执行近似准则 (A) 所产生的无穷点列. 那么序列 $\{z^k\}$ 收敛到问题 (2.4) 的某一个解.

下面, 我们讨论 PPA 的局部收敛速度. 在 Rockafellar 的经典工作 [48] 中, 他证明了当映射 T^{-1} 关于原点是 Lipschitz 连续时, 序列 $\{z^k\}$ 有渐近超线性收敛速度. 这里我们说集合映射 $\Gamma : \mathcal{X} \rightrightarrows \mathcal{Y}$ 是关于某点 $\bar{u} \in \mathcal{X}$ 是 Lipschitz 连续的, 如果 $\Gamma(\bar{u}) = \{\bar{v}\}$ 且存在正常数 κ, ε 使得

$$\|v - \bar{v}\| \leq \kappa \|u - \bar{u}\|, \quad \forall\, v \in \Gamma(u), \quad \forall\, u \in \mathbb{B}_{\varepsilon}(\bar{u}).$$

因此, 由 T^{-1} 关于原点是 Lipschitz 连续的, 我们可以推出 $T^{-1}(0)$ 是一个单点集, 即问题 (2.4) 有唯一解. 这一唯一性假设在很多情况下都显得异常严格, 因此该结论无法被应用到更多的实际问题中去. Luque 在 [39] 中放松了上面的条件, 他松弛了上面提到的集合映射 T^{-1} 的 Lipschitz 连续性的假设, 并研究了 PPA 的局部收敛速度. 具体来说, 在收敛速度分析中, 他利用如下假设代替了 Lipschitz 连续假设: 存在正常数 ε, κ 使得

$$\operatorname{dist}(z, T^{-1}(0)) \leq \kappa \|u\|, \quad \forall\, z \in T^{-1}(u), \quad \forall\, u \in \mathbb{B}_{\varepsilon}(0).$$

[1] 事实上, 如果仅需证明 PPA 的收敛性, 我们不需要假设 σ_k 的单调性, 而只需要确保 σ_k 的下界大于 0.

显然该条件弱于 T^{-1} 在原点处有 Lipschitz 连续性的条件. 事实上, 这个条件即为 Robionson 在 [45] 中提出的 T^{-1} 在原点是局部上 Lipschitz 连续的条件. 关于此条件, 在文献 [46] 中, Robionson 得到了关于多面体多值映射 (polyhedral multifunction) 的一个重要定理, 即任意的多面体多值映射总是局部上 Lipschitz 连续的. 另外, Sun 在其博士论文 [53] 中得到的重要结果之一表明一个正常闭凸函数是分片线性二次的 (piecewise linear-quadratic), 当且仅当它的次梯度映射 (subgradient mapping) 是一个多面体多值映射. 由此, 我们知道前面提到的 Lasso 问题以及 Fused Lasso 问题的目标函数的次梯度映射是局部上 Lipschitz 连续的. 当然, 对于其它很多非多面体的多值映射, 如很多复合矩阵优化问题目标函数的次梯度 [10–12, 12, 27, 28], 局部上 Lipschitz 连续条件依然太过局限. 因此, 本文考虑使用前面提到的平稳条件 (2.3) 来替代局部上 Lipschitz 连续条件. 显然, 由平稳性的定义 (2.3), 我们不难看出平稳性条件弱于局部上 Lipschitz 连续条件. 事实上, Cui et al. 在 [10] 中的研究表明对于很大一类矩阵优化问题, 目标函数次梯度映射的平稳性是可以在一些较弱的假设下保证的.

下面, 我们给出 PPA 在平稳性条件下的局部渐近超线性收敛的结果. 该结果可以看成是 [48, Theorem 2] 和 [39, Theorem 2.1] 的推广. 同时, 我们注意到在 PPA 的子问题被精确求解的情况下 (即 $z^{k+1} = P_k(z^k), \forall\, k \geq 0$) 类似的结论已经在 [26] 中讨论过. 这个结果及其证明可以在文献 [11, Proposition 1] 处找到.

命题 2. 假设 $T^{-1}(0)$ 非空, 令 $\{z^k\}$ 为 PPA (2.5) 执行近似准则 (A) 以及 (B) 所产生的无穷点列. 那么序列 $\{z^k\}$ 收敛到 $z^\infty \in T^{-1}(0)$, 即问题 (2.4) 的某一个解. 同时, 如果 T^{-1} 在原点关于 z^∞ 是平稳的且模为 κ, 那么存在 $\bar{k} \geq 0$ 使得对于任意的 $k \geq \bar{k}$,

$$\operatorname{dist}(z^{k+1}, T^{-1}(0)) \leq \mu_k \operatorname{dist}(z^k, T^{-1}(0)),$$

其中 $\mu_k := \left[\eta_k + (\eta_k+1)\kappa/\sqrt{\kappa^2+\sigma_k^2}\right]/(1-\eta_k) \to \mu_\infty := \kappa/\sqrt{\kappa^2+\sigma_\infty^2}$ $(\mu_\infty = 0 \text{ if } \sigma_\infty = +\infty)$.

注 1. 近年来, 新的研究表明为得到 PPA 的渐近超线性收敛速度, 集合映射 T^{-1} 的平稳性假设还可以进一步放松为如下的正则性条件: 对于任意的正数 $r > 0$, 存在 $\kappa > 0$ 使得,

$$\operatorname{dist}(x, T^{-1}(0)) \leq \kappa \operatorname{dist}(0, T(x)) \quad \forall\, x \in \mathcal{X} \text{ 满足 } \operatorname{dist}(x, T^{-1}(0)) \leq r. \tag{2.6}$$

文献 [32, 65] 考虑了条件 (2.6), 并证明了条件 (2.6) 弱于 T^{-1} 在原点处平稳这一条件, 同时在该条件下证明了 PPA 有渐近超线性的收敛速度. 特别的, [32] 还研究了一类预处理 PPA 的收敛性质.

作为一个简单的应用, PPA 可以用来求解优化问题如

$$\min_{x \in \mathcal{X}} g(x),$$

这里 $g : \mathcal{X} \to (-\infty, +\infty]$ 为任一给定的正常闭凸函数. 该问题的最优性条件可以写为 $0 \in \partial g(x)$. 由 [47, Corollary 31.5.2], 我们知道次梯度映射 ∂g 是极大单调算子, 因此 PPA 可以用来求解这一最优性包含系统并写出如下的迭代公式: 给定正项单调数列 $\{\sigma_k\}$ 及初始点 $x^0 \in \mathcal{X}$, 对于任意的 $k \geq 0$,

$$x^{k+1} \approx (I + \sigma_k \partial g)^{-1}(x^k) = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ g(x) + \frac{1}{2\sigma_k}\|x - x^k\|^2 \right\} = \operatorname{Prox}_{\sigma_k g}(x^k). \tag{2.7}$$

其中第一个等式由此处优化问题的最优性条件得到, 第二个等式由邻近算子的定义得到. 显然, 迭代公式 (2.7) 的收敛性质可以由命题 1 和 2 得到. 在优化领域, PPA 还有很多其它的应

用方式比如将 PPA 应用到对偶目标函数的次梯度映射, 我们会得到著名的增广 Lagrangian 算法 (augmented Lagrangian method). 感兴趣的读者可以参见 Rockafellar 的经典文献 [48, 49].

3. PPA 求解问题 (1.1)

本节我们考虑利用上节讨论的 PPA 来求解问题 (1.1). 在描述具体算法之前, 为简化讨论, 我们对损失函数 h 做如下假设:

假设 1.

1. $h:\mathcal{Y}\to\Re$ 是连续可微函数且存在 $\alpha_h>0$ 使得梯度 ∇h 满足
$$\|\nabla h(y')-\nabla h(y)\|\le(1/\alpha_h)\|y'-y\|,\quad \forall\, y',y\in\mathcal{Y};$$
2. h 是本质局部强凸 (essentially locally strongly convex) 函数 [22], 即对于任意给定的紧凸集 $K\in\mathcal{Y}$, 存在正数 $\beta_K>0$ 使得对于任意的 $\lambda\in[0,1]$,
$$(1-\lambda)h(y')+\lambda h(y)\ge h((1-\lambda)y'+\lambda y)+\frac{1}{2}\beta_K\lambda(1-\lambda)\|y'-y\|^2,\quad \forall\, y',y\in K.$$

注意到假设 1 是研究复合凸优化问题时文献中常见的假设, 例如线性回归与逻辑回归模型中的损失函数 h 都满足上面的假设, 因此我们的假设限定性并不强. 事实上, 我们后面所设计的算法并不完全依赖于假设 1. 当该假设不成立时, 如 h 为非光滑损失函数时, 只需少许改动本文所提算法即可处理该情况. 感兴趣的读者可以参见 [35,54]. 这里, 为了简化讨论并保持本文的可读性与专注性, 我们将在假设 1 下讨论算法的设计. 在假设 1 条件下, 我们发现 h^*, 即函数 h 的共轭函数, 有一些特别的性质. 我们将这些性质总结在下面的命题中. 该命题的证明可以从 [50, Proposition 12.60] 和 [22, Corollary 4.4] 得到.

命题 3. 设假设 1 成立, 下列结论成立

1. h^* 是强凸 (strongly convex) 函数且其强凸系数 (modulus) 为 α_h;
2. h^* 是本质可微的 (essentially differentiable [22])[1] , 即开集 $C=\mathrm{int}(\mathrm{dom}\,h^*)$ 非空, h^* 在开集 C 上可微且对于 C 中收敛序列 $\{y^k\}$, 如果 $\{y^k\}$ 收敛到 C 的某一边界点 y 那么 $\lim_{k\to\infty}\|\nabla h^*(y^k)\|=+\infty$;
3. ∇h^* 是局部 Lipschtiz 连续的.

在这些准备工作的基础上, 下面我们给出快速高效的 PPA 框架来求解问题 (1.1). 具体来说给定正项单调数列 $\{\sigma_k\}$ 及初始点 $x^0\in\mathcal{X}$, 我们考虑如下的迭代步骤
$$x^{k+1}\approx P_k(x^k):=\mathop{\mathrm{argmin}}_{x\in\mathcal{X}}\left\{f_k(x):=h(\mathcal{A}x)-\langle c,\,x\rangle+p(x)+\frac{1}{2\sigma_k}\|x-x^k\|^2\right\},\quad \forall\, k\ge0. \tag{3.1}$$
这里我们采用 PPA 的一个重要原因是如命题 1 和 2 所示, 在较弱的假设条件下, 该算法有很好的全局收敛性与局部渐近超线性收敛性质. 我们注意到 PPA 所展示出来的理论收敛性质远好于其它被用于求解问题 (1.1) 的流行的一阶算法如邻近梯度算法、加速邻近梯度算法与交

[1] 此概念在 [47, Section 26] 也称为本质光滑 (essentially smooth).

替方向乘子算法等. 同时, 不难发现 PPA 的效率可以说完全被子问题 (3.1) 的求解效率所决定. 因此, 我们需要挖掘利用这些子问题的特殊结构, 设计高效的子问题求解算法. 下面, 本文将阐述一种基于对偶原理的高效算法来求解 PPA 所遇到的子问题. 特别的, 我们将通过求解子问题 (3.1) 的对偶问题来获取子问题 (3.1) 自身的高质量近似解. 本文采取对偶手段有以下两点原因: (1) 在很多高维机器与统计学习问题中, 我们通常有 $n = \dim(\mathcal{X}) \gg m = \dim(\mathcal{Y})$, 即特征数量远大于样本数量. 因此, 对偶问题变量的维数 m 远小于原问题 (3.1) 变量的维数 n, 更有利于算法的设计与开发; (2) 由命题 3 中 h^* 的性质, 我们可以推出对偶问题的光滑性与强凸性, 即对偶问题理论性质较好. 事实上, 这些理论性质还有利于我们为 PPA 设计易计算的近似准则来代替准则 (A) 和 (B).

3.1. 对偶算法求解子问题 (3.1)

本小节将给出求解子问题 (3.1) 的对偶算法并给出针对模型 (1.1) 的特殊 PPA 以及其收敛结果. 我们首先给出子问题 (3.1) 的对偶问题. 由于该子问题被写成了无约束优化问题的形式[1)] , 推导对偶问题的常用方法不能很方便的处理这类问题. 这里我们采用 [50, Section 11.H] 中提到的基于扰动性分析 (perturbation analysis) 的框架写出对偶问题. 为此, 我们定义如下扰动函数

$$\widetilde{f}(x,u,\eta;\sigma) = h(\mathcal{A}x+u) - \langle c,\, x\rangle + p(x) + \frac{1}{2\sigma}\|x-\eta\|^2, \quad (x,u,\eta,\sigma) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{X}\times R.$$

注意到 $\widetilde{f}$ 关于变量 x, u 是联合凸函数 (jointly convex). 另外, 问题 (3.1) 等价于 $\min_{x\in\mathcal{X}} \widetilde{f}\,(x,\, 0,\, x^k;\, \sigma_k)$.

通过利用部分对偶 (partial dualization) 的技术, 我们可以写出如下 Lagrangian 函数 l:

$$l(x,\xi,\eta;\sigma) = \inf_{u\in\mathcal{Y}}\left\{\widetilde{f}(x,u,\eta;\sigma) - \langle u,\, \xi\rangle\right\} = -h^*(\xi) + p(x) - \langle c - \mathcal{A}^*\xi,\, x\rangle + \frac{1}{2\sigma}\|x-\eta\|^2,$$

对所有 $(x,\xi,\eta,\sigma) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{X}\times\Re$. 根据这一函数, 我们现在可以写出问题 (3.1) 如下形式的对偶问题:

$$\min_{\xi\in\mathcal{Y}} \Psi_k(\xi, x^k), \tag{3.2}$$

这里

$$\begin{aligned}\Psi_k(\xi,x^k) :=\ & -\inf_{x\in\mathcal{X}} l(x,\xi,x^k;\sigma_k) = h^*(\xi) - \inf_{x\in\mathcal{X}}\left\{p(x) + \langle\mathcal{A}^*\xi - c,\, x\rangle + \frac{1}{2\sigma_k}\|x-x^k\|^2\right\}\\ =\ & h^*(\xi) + \inf_{s\in\mathcal{X}}\left\{p^*(s) - \langle x^k,\, \mathcal{A}^*\xi + s - c\rangle + \frac{\sigma_k}{2}\|\mathcal{A}^*\xi + s - c\|^2\right\}, \quad \forall\,\xi\in\mathcal{Y}.\end{aligned} \tag{3.3}$$

其中 h^* 和 p^* 分别是 h and p 的 Fenchel 共轭函数, 第三个等式由 Moreau 等式 (2.2) 得来. 同时, 由命题 3 中 h^* 的性质以及 (2.1) 可知, $\Psi_k(\xi, x^k)$ 关于 ξ 是一个强凸函数并且是本质可微的且有

$$\nabla\Psi_k(\xi,x^k) \ =\ \nabla h^*(\xi) - \mathcal{A}\mathrm{Prox}_{\sigma_k p}(x^k - \sigma_k(\mathcal{A}^*\xi - c)), \quad \forall\,\xi\in\mathrm{int}(\mathrm{dom}h^*). \tag{3.4}$$

[1)] 事实上, 由于非光滑函数 p 的存在, 本质上 (3.1) 还是一个约束优化问题.

在此框架下, 我们可以写出下面的弱对偶 (weak duality) 结果:

$$\begin{aligned}\inf_{\xi\in\mathcal{Y}}\Psi_k(\xi,x^k)=\inf_{\xi\in\mathcal{Y}}\big(-\inf_{x\in\mathcal{X}}l(x,\xi,x^k;\sigma_k)\big)&=\inf_{\xi\in\mathcal{Y}}\sup_{x\in\mathcal{X}}-l(x,\xi,x^k;\sigma_k)\\&\geq\sup_{x\in\mathcal{X}}\inf_{\xi\in\mathcal{Y}}-l(x,\xi,x^k;\sigma_k)=\sup_{x\in\mathcal{X}}-f_k(x).\end{aligned}\tag{3.5}$$

同时, 由 [47, Theorem 37.3] 可知这里 $\inf_{\xi\in\mathcal{Y}}$ 和 $\sup_{x\in\mathcal{X}}$ 是可交换的, 因此 (3.5) 中不等式可以写成等式, 即得到如下强对偶 (strong duality) 结果:

$$\inf_{\xi\in\mathcal{Y}}\Psi_k(\xi,x^k)=\sup_{x\in\mathcal{X}}-f_k(x).\tag{3.6}$$

由于 $\Psi_k(\xi,x^k)$ 关于 ξ 是一个强凸本质可微函数, 我们知道问题 (3.2) 有唯一解记为 $\hat{\xi}^k$ 且 $\hat{\xi}^k\in\mathrm{int}(\mathrm{dom}\,h^*)$. 令 $\hat{x}^k=\mathrm{Prox}_{\sigma_k p}\big(x^k-\sigma_k(\mathcal{A}^*\hat{\xi}^k-c)\big)$, 由 (3.4) 可知,

$$\mathcal{A}\hat{x}^k=\nabla h^*(\hat{\xi}^k).\tag{3.7}$$

利用 (3.7) 和 [47, Theorem 23.5], 我们知道

$$h^*(\hat{\xi}^k)+h(\mathcal{A}\hat{x}^k)=\langle\mathcal{A}\hat{x}^k,\,\hat{\xi}^k\rangle$$

进一步, 由强对偶结果 (3.6), 有

$$\begin{aligned}\inf_{\xi\in\mathcal{Y}}\Psi_k(\xi,x^k)=\Psi_k(\hat{\xi}^k,x^k)&=h^*(\hat{\xi}^k)-p(\hat{x}^k)-\langle\mathcal{A}^*\hat{\xi}^k-c,\,\hat{x}^k\rangle-\frac{1}{2\sigma_k}\|\hat{x}^k-x^k\|^2\\&=-h(\mathcal{A}\hat{x}^k)-p(\hat{x}^k)+\langle c,\,\hat{x}^k\rangle\\&=-f_k(\hat{x}^k)=\sup_{x\in\mathcal{X}}-f_k(x).\end{aligned}\tag{3.8}$$

由上式, 我们可知由 $\hat{\xi}^k$ 所得到的 $\hat{x}^k$ 是原问题 (3.1) 的最优解, 即 $P_k(x^k)=\hat{x}^k=\mathrm{Prox}_{\sigma_k p}\big(x^k-\sigma_k(\mathcal{A}^*\hat{\xi}^k-c)\big)$. 这样, 我们就可将原问题 (3.1) 的求解转换为对偶问题 (3.3) 的求解. 注意到问题 (3.2) 的最优解 $\hat{\xi}^k$ 一般没有显式解 (closed form solution), 因此我们需要利用迭代算法去得到 $\hat{\xi}^k$ 的近似解记为 ξ^{k+1}. 这样一来, 通过上面讨论的对偶手段所得到的 $x^{k+1}=\mathrm{Prox}_{\sigma_k p}\big(x^k-\sigma_k(\mathcal{A}^*\xi^{k+1}-c)\big)$ 也是 $P_k(x^k)$ 的一个近似解.

上面的讨论启发我们写出基于对偶手段针对模型 (1.1) 特别设计的 PPA 框架.

算法 1: 基于对偶手段的 PPA 求解 (1.1)

输入 x^0, σ_0, $\{\epsilon_k\}$, $\{\delta_k\}$, $\sigma_\infty\leq\infty$. 对 $k=0,1,\ldots,$ 迭代

1: 近似求解问题 (3.2)

$$\xi^{k+1}\approx\operatorname*{argmin}_{\xi\in\mathcal{Y}}\left\{\Psi_k(\xi,x^k)\right\}.$$

2: 计算 $x^{k+1}=\mathrm{Prox}_{\sigma_k p}\big(x^k-\sigma_k(\mathcal{A}^*\xi^{k+1}-c)\big)$.

3: 更新 $\sigma_{k+1}\uparrow\sigma_\infty$.

为了保证算法 3.1 优越的收敛性质, 如预备知识小节讨论的一样, 我们需要保证迭代过程中 x^{k+1} 与 $P_k(x^k)$ 足够靠近, 即保证近似准则 (A) 和 (B) 成立. 这里我们基于函数 Ψ_k 的光

滑性与强凸性来设计易计算的近似准则来代一般情况下无法计算的近似准则 (A) 和 (B). 下面的引理表明, 对偶手段计算 $P_k(x^k)$ 时产生的误差可以被近似求解对偶问题 (3.2) 带来的误差控制住.

引理 1. 给定 $x^k \in \Re^n$ 和 $\sigma_k > 0$, 下面的公式成立

$$\|\mathrm{Prox}_{\sigma_k p}(\tilde{x}^k(\xi)) - P_k(x^k)\|^2 \le 2\sigma_k\big(\Psi_k(\xi, x^k) - \inf_{\xi\in\mathcal{Y}} \Psi_k(\xi, x^k)\big), \quad \forall\, \xi \in \mathcal{Y},$$

这里 $\tilde{x}^k(\xi) = \mathrm{Prox}_{\sigma_k p}\big(x^k - \sigma_k(\mathcal{A}^*\xi - c)\big)$.

证明. 由 Ψ_k 的定义 (3.3) 和 (2.1) 可以观察到, 作为 η 的函数, Ψ_k 可微且梯度为

$$\nabla_\eta \Psi_k(\xi, x^k) = \frac{1}{\sigma_k}\big(\tilde{x}^k(\xi) - x^k\big), \quad \forall\, \xi \in \mathcal{Y}.$$

并且, 不难发现函数 Ψ_k 关于 ξ 为凸函数, 关于 η 为凹函数. 因此, 我们可以对任意的 $w \in \mathcal{X}$ 得到如下不等式

$$\begin{aligned}
&\Psi_k(\xi, x^k) + \langle w - x^k,\, \nabla_\eta \Psi(\xi, x^k)\rangle \ge \Psi_k(\xi; w) \ge \inf_{\xi\in\mathcal{Y}} \Psi_k(\xi; w)\\
&\quad = \inf_{\xi\in\mathcal{Y}}\big(-\inf_{x\in\mathcal{X}} l(x, \xi, w; \sigma_k)\big) = \inf_{\xi\in\mathcal{Y}} \sup_{x\in\mathcal{X}} -l(x, \xi, w; \sigma_k)\\
&\quad = \sup_{x\in\mathcal{X}} \inf_{\xi\in\mathcal{Y}} -l(x, \xi, w; \sigma_k) = \sup_{x\in\mathcal{X}} \left\{\inf_{\xi\in\mathcal{Y}} \{h^*(\xi) - \langle \mathcal{A}x,\, \xi\rangle\} - p(x) + \langle c,\, x\rangle - \frac{1}{2\sigma_k}\|x - w\|^2\right\}\\
&\quad = \sup_{x\in\mathcal{X}} \left\{-h(\mathcal{A}x) - p(x) + \langle c,\, x\rangle - \frac{1}{2\sigma_k}\|x - w\|^2\right\}\\
&\qquad \ge -p(P_k(x^k)) - h(\mathcal{A}P_k(x^k)) + \langle c,\, P_k(x^k)\rangle - \frac{1}{2\sigma_k}\|P_k(x^k) - w\|^2.
\end{aligned}$$

其中, $\inf_{\xi\in\mathcal{Y}}$ 和 $\sup_{x\in\mathcal{X}}$ 的可交换性可以从 [47, Theorem 37.3] 得到. 同时, 由 (3.8), 我们有

$$\inf_{\xi\in\mathcal{Y}} \Psi_k(\xi, x^k) = -p(P_k(x^k)) - h(\mathcal{A}P_k(x^k)) + \langle c,\, P_k(x^k)\rangle - \frac{1}{2\sigma_k}\|P_k(x^k) - x^k\|^2.$$

因此, 对于任意的 $w \in \mathcal{X}$, 我们有

$$\Psi_k(\xi, x^k) - \inf_{\xi\in\mathcal{Y}} \Psi_k(\xi, x^k) \ge \frac{1}{2\sigma_k}\big(\|P_k(x^k) - x^k\|^2 - \|P_k(x^k) - w\|^2 - 2\langle w - x^k,\, \sigma_k\nabla_\eta\Psi(\xi, x^k)\rangle\big).$$

注意到上面不等式的右边是关于 w 的二次函数, 它的最大值在 $w^* = P_k(x^k) - \sigma_k\nabla_\eta\Psi(\xi, x^k) = P_k(x^k) + x^k - \tilde{x}^k(\xi)$ 处取到, 且最大值为 $\frac{1}{2\sigma_k}\|P_k(x^k) - \tilde{x}^k(\xi)\|^2$. 这样我们就证明了本引理.

由命题 3 中 h^* 的强凸性, 我们不难得到下面的不等式

$$\Psi_k(\xi, x^k) - \inf_{\xi\in\mathcal{Y}} \Psi_k(\xi, x^k) \le (1/2\alpha_h)\|\nabla\Psi_k(\xi, x^k)\|^2, \quad \forall\, \xi \in \mathcal{Y}. \tag{3.9}$$

现在, 结合引理 1 与 (3.9), 我们有

$$\|x^{k+1} - P_k(x^k)\|^2 \le (\sigma_k/\alpha_h)\|\nabla\Psi_k(\xi^{k+1}, x^k)\|^2,$$

并可提出如下易计算的近似准则来代替 (A) 和 (B):

(A') $\|\nabla\Psi_k(\xi^{k+1}, x^k)\| \le \sqrt{\alpha_h/\sigma_k}\,\varepsilon_k, \quad \varepsilon_k \ge 0, \quad \sum_{k=1}^\infty \varepsilon_k < \infty,$

(B') $\|\nabla\Psi_k(\xi^{k+1}, x^k)\| \le \sqrt{\alpha_h/\sigma_k}\,\eta_k\|\mathrm{Prox}_{\sigma_k p}\big(x^k - \sigma_k(\mathcal{A}^*\xi^{k+1} - c)\big) - x^k\|,$
$0 \le \eta_k < 1, \quad \sum_{k=1}^\infty \eta_k < \infty.$

不难看出近似准则 (A') 和 (B') 成立时近似准则 (A) 和 (B) 也能被满足. 同时, 满足这些准则只需梯度 $\nabla\Psi_k(\xi^{k+1}, x^k)$ 的范数足够小, 即问题 (3.2) 解得足够精确. 现在我们可以基于命题 1 和 2 给出算法 3.1 的收敛结果. 为了叙述的简练, 我们将问题 (1.1) 中的目标函数记为 f, 即 $f(x) := h(\mathcal{A}x) - \langle c, x\rangle + p(x)$, $x \in \mathcal{X}$.

定理 1. 设问题 (1.1) 有非空解集 Ω. 令 $\{x^k\}$ 为算法 3.1 在近似准则 (A') 下产生的无穷序列. 那么 $\{x^k\}$ 收敛到问题 (1.1) 的某个最优解 $x^* \in \Omega$.

如果, 近似准则 (B') 也同时被满足而且 ∂f^{-1} 在原点关于 x^* 是平稳的且模为 κ, 那么存在 $\bar{k} \geq 0$ 使得对任意的 $k \geq \bar{k}$,

$$\mathrm{dist}(x^{k+1}, \Omega) \leq \mu_k \mathrm{dist}(x^k, \Omega),$$

这里 $\mu_k = \left[\eta_k + (\eta_k + 1)\kappa/\sqrt{\kappa^2 + \sigma_k^2}\right]/(1 - \eta_k) \to \mu_\infty = \kappa/\sqrt{\kappa^2 + \sigma_\infty^2} < 1$ $(\mu_\infty = 0 \text{ if } \sigma_\infty = +\infty)$.

注 2. 正如预备知识小节所讨论的, 这里关于 ∂f^{-1} 平稳性的假设是比较弱的. 统计于机器学习里很多问题都满足这个假设. 例如, 在 Lasso, Fused Lasso 以及 Elastic net 回归问题中, 由于它们的目标函数 f 都是分片线性二次的, 所以由 Sun 的博士论文 [53](同时参见 [50, Proposition 12.30]) 与 Robionson 关于多面体多值映射误差界的结论 [46], 我们知道这些问题目标函数的 ∂f^{-1} 在原点是平稳的. 对于其它非分片线性二次问题, ∂f^{-1} 平稳性的研究也是优化理论研究的重要方向. 例如, 在 [36, 57] 中, 作者研究了带 ℓ_1 或 Elastic net 正则项的逻辑回归问题的目标函数的 ∂f^{-1} 的平稳性; 文献 [66] 基于有界正则性 (bounded regularity) 条件给出了 p 为核范数的情况下 ∂f^{-1} 平稳性的充分条件; 文献 [10, 11] 基于二次增长条件 (quadratic growth condition) 给出了一大类矩阵优化问题对应的 ∂f^{-1} 平稳性的充分条件.

3.2. 半光滑牛顿法求解问题 (3.2)

本小节我们考虑利用半光滑牛顿 (semismooth Newton) 算法求解算法 3.1 中的子问题 (3.2). 半光滑牛顿算法的局部快速收敛性质使得子问题求解效率极大提高进而提升了算法 3.1 的整体效率.

给定 $\tilde{x} \in \mathcal{X}$ 和 $\sigma > 0$, 根据问题 (3.2) 的形式, 我们定义下面的函数 $\psi : \mathcal{Y} \to (-\infty, +\infty]$:

$$\psi(\xi) := -\inf_{x \in \mathcal{X}} l(x, \xi, \tilde{x}; \sigma) = h^*(\xi) + \inf_{s \in \mathcal{X}} \left\{ p^*(s) - \langle \tilde{x}, \mathcal{A}\xi + s - c\rangle + \frac{\sigma}{2}\|\mathcal{A}\xi + s - c\|^2 \right\}.$$

本小节考虑如下的优化问题

$$\min_{\xi \in \mathcal{Y}} \psi(\xi). \tag{3.10}$$

由假设 1, 我们从 (2.1) 和命题 3 知道 ψ 是 $\mathcal{Y}$ 上的强凸本质可微函数. 特别的 ψ 在 $\mathrm{int}(\mathrm{dom} h^*)$ 可微并且梯度是局部 Lipschitz 连续的:

$$\nabla\psi(\xi) \;=\; \nabla h^*(\xi) - \mathcal{A}\mathrm{Prox}_{\sigma p}(-\sigma\mathcal{A}^*\xi + \tilde{x} + \sigma c), \quad \forall\, \xi \in \mathrm{int}(\mathrm{dom} h^*).$$

命题 3 进一步说明问题 (3.10) 有唯一解 $\xi^* \in \mathrm{int}(\mathrm{dom} h^*)$, 且 ξ^* 满足下面的最优条件方程:

$$\nabla\psi(\xi) = 0, \quad \xi \in \mathrm{int}(\mathrm{dom} h^*). \tag{3.11}$$

所以优化问题 (3.10) 的求解可以转换为求解上面的最优条件方程. 我们注意到 $\mathrm{Prox}_{\sigma p}(\nabla h^*)$ 仅为 (局部) Lipschitz 连续函数, 因此 $\nabla\psi$ 一般不再是可微函数, 所以无法利用传统的牛顿算

法求解. 由于有广泛的应用, 此类非光滑方程的求解一直以来受到优化领域研究人员的关注. 特别的, 半光滑牛顿算法 [24, 40, 44, 52] 被认为是求解此类非光滑方程一个非常有效的算法. 文献 [29, 30, 32, 34, 38, 43, 60, 64, 65] 中很多计算结果也展示了半光滑牛顿算法的数值优越性. 下面, 我们给出半光滑函数的定义, 这些定义是从文献 [24, 40, 44, 52] 中总结而来.

定义 2. 令 $\mathcal{O} \subseteq \mathcal{X}$ 为开集, $F : \mathcal{O} \subseteq \mathcal{X} \to \mathcal{Y}$ 为局部 Lipschitz 连续函数, $\mathcal{K} : \mathcal{O} \subseteq \Re^n \rightrightarrows \Re^{m\times n}$ 为非空紧值 (compact valued) 上半连续 (upper-semicontinous) 的集合映射. F 在 $x \in \mathcal{O}$ 关于多值映射 $\mathcal{K}$ 半光滑的, 如果 F 在 x 是方向可微的 (directional differentiable) 且对于任意的$V \in \mathcal{K}(x + \Delta x)$ 满足 $\Delta x \to 0$,

$$F(x + \Delta x) - F(x) - V\Delta x = o(\|\Delta x\|). \tag{3.12}$$

如果 (3.12) 被以下等式代替

$$F(x + \Delta x) - F(x) - V\Delta x = O(\|\Delta x\|^{1+\gamma}),$$

这里 $\gamma > 0$ 是正常数, 那么 F 在 x 关于 $\mathcal{K}$ 是 γ- 阶半光滑的[1)] . 另外, 我们称 F 是 $\mathcal{O}$ 上关于 $\mathcal{K}$ 的半光滑函数是指 F 在 $\mathcal{O}$ 上任意点关于 $\mathcal{K}$ 都是半光滑的. 默认的, 如果 $\mathcal{O}$ 为全空间或某约定的开集, $\mathcal{K}$ 为 F 的 Clarke 广义 Jacobian (Clarke's generalized Jacobian), 我们在叙述中无歧义的省略 $\mathcal{O}$ 和 $\mathcal{K}$, 称 F 是半光滑函数.

注 3. 我们注意到很多常用的函数都是半光滑函数, 如凸函数与光滑函数. 特别的, 连续分片仿射函数 (continuous piecewise affine functions) 和二次连续可微函数 (twice continuously differentiable functions) 都是强半光滑函数. 例如, ℓ_1 范数的邻近映射 $\mathrm{Prox}_{\|\cdot\|_1}$ 是强半光滑函数. 更多关于半光滑函数与强半光滑函数的讨论参见文献 [17].

由于 ∇h^* 和 $\mathrm{Prox}_{\sigma p}$ 都是局部 Lipschitz 连续的, 我们可以分别定义它们的 Clarke 广义 Jacobian [8], 记为 $\partial(\nabla h^*)$ 与 $\partial(\mathrm{Prox}_{\sigma p})$. 进一步, 我们可以定义目标函数 ψ 的广义 Hessian, 即 $\nabla\psi$ 的 Clarke 广义 Jacobian, 记为 $\partial^2\psi$. 一般来说 $\partial^2\psi$ 没有可计算的显式表达式, 因此我们定义如下替代映射

$$\hat{\partial}^2\psi(\xi) := \partial(\nabla h^*)(\xi) + \sigma\mathcal{A}\partial\mathrm{Prox}_{\sigma p}(-\sigma\mathcal{A}^*\xi + \tilde{x} + \sigma c)\mathcal{A}^*, \quad \forall\,\xi \in \mathcal{Y}. \tag{3.13}$$

对于任意给定的 $\xi \in \mathcal{Y}$, 由 [23, Example 2.5], 我们知道

$$\partial^2\psi(\xi)(d) = \hat{\partial}^2\psi(\xi)(d), \quad \forall d \in \mathcal{Y}.$$

由此, $\hat{\partial}^2\psi$ 被认为是 $\partial^2\psi$ 一个较好的替代映射. 给定 $\xi \in \mathcal{Y}$, 令 $V = H + \sigma\mathcal{A}U\mathcal{A}^*$ 其中 $H \in \partial(\nabla h^*)(\xi)$, $U \in \partial\mathrm{Prox}_{\sigma p}(-\sigma\mathcal{A}^*\xi + \tilde{x} + \sigma c)$, 则 $V \in \hat{\partial}^2\psi(\xi)$. 并且, 由于 h^* 为强凸函数, 我们有 H 为对称正定 (symmetric positive definite) 矩阵, 进一步有 V 也是对称正定矩阵. 现在我们给出半光滑牛顿算法求解系统 (3.11) 的基本框架.

在一些较弱的假设下, 我们可以证明算法 3.2 的全局和局部超线性收敛性质. 具体定理如下.

定理 2. 设 ∇h^* 及 $\mathrm{Prox}_{\sigma p}$ 为强半光滑函数, 令 $\{\xi^j\}$ 为算法 3.2 产生的无穷序列. 那么, $\{\xi^j\}$ 收敛到问题 (3.10) 的唯一解 ξ^* 且

$$\|\xi^{j+1} - \hat{\xi}\| = O(\|\xi^j - \hat{\xi}\|^{1+\tau}),$$

[1)] 特别的, 如果 $\gamma = 1$, 我们称之为强半光滑.

算法 2：半光滑牛顿算法求解系统 (3.11)

输入 $\mu \in (0,1/2)$, $\bar{\eta} \in (0,1)$, $\tau \in (0,1]$, $\delta \in (0,1)$, $\xi^0, \tilde{x}, \sigma$. 对 $j = 0,1,\ldots,$ 迭代

1: 选取 $H_j \in \partial(\nabla h^*)(\xi^j)$, $U_j \in \partial \mathrm{Prox}_{\sigma p}(-\sigma \mathcal{A}^* \xi^j + \tilde{x} + \sigma c)$, 令 $V_j = H_j + \sigma \mathcal{A} U_j \mathcal{A}^*$, 精确求解线性系统

$$V_j h = -\nabla \psi(\xi^j) \tag{3.14}$$

或使用共轭梯度算法 (conjugate gradient) 得到 h^j 使得

$$\|V_j h^j + \nabla \psi(\xi^j)\| \le \min(\bar{\eta}, \|\nabla \psi(\xi^j)\|^{1+\tau}).$$

2: 设定 $\alpha_j = \delta^{m_j}$, 其中 m_j 为第一个非负整数 m 使得

$$\xi^j + \delta^m h^j \in \mathrm{int}(\mathrm{dom}\, h^*), \quad \text{以及} \quad \psi(\xi^j + \delta^m h^j) \le \psi(\xi^j) + \mu \delta^m \langle \nabla \psi(\xi^j), h^j \rangle.$$

3: 计算 $\xi^{j+1} = \xi^j + \alpha_j h^j$.

其中 $\tau \in (0,1]$ 为算法 3.2 中给定的参数.

证明. 在 ∇h^* 及 $\mathrm{Prox}_{\sigma p}$ 强半光滑的假设下, 不难得到 $\nabla \psi$ 在 $\mathcal{X}$ 上关于 $\hat{\partial}^2 \psi$ (定义式 (3.13)) 是强半光滑的. 注意到 V_j 总是对称正定矩阵, 定理结论可以从文献 [29,30,64] 中相应证明步骤得到.

注 4. 在很多应用中, 前面提到的 $\partial(\nabla h^*)$ 与 $\partial \mathrm{Prox}_{\sigma p}$ 也可能没有可计算的显式表达式. 这时我们可以采取同样的思路寻找替代映射. 具体来说, 如果我们可以找到 $\partial(\nabla h^*)$ 与 $\partial \mathrm{Prox}_{\sigma p}$ 的较易计算的多值替代映射 $\mathcal{P}$ 与 $\mathcal{U}$, 且满足 ∇h^* 和 $\mathrm{Prox}_{\sigma p}$ 分别关于 $\mathcal{P}$ 和 $\mathcal{U}$ 是强半光滑的. 那么, 对任意的 $\xi \in \mathcal{Y}$, 我们定义

$$\mathcal{V}(\xi) = \{V \mid V = P + \sigma \mathcal{A} U \mathcal{A}^*,\ P \in \mathcal{P}(\xi),\ U \in \mathcal{U}(-\sigma \mathcal{A}^* \xi + \tilde{x} + \sigma c)\}.$$

不难证明, ψ 关于 $\mathcal{V}$ 是强半光滑的. 这样, 我们可以相对容易地实现算法 3.2 并证明如同定理 2 一样的收敛性结果. 事实上, 文献 [30,34,65] 已经采用了这里所描述的技术.

4. 半光滑牛顿算法的高效实现

本节我们考虑半光滑牛顿算法 3.2 的具体实现方法. 从算法 3.2 的描述可以看到, 我们主要需要考虑三个方面: (a) 非光滑正则函数 p 的邻近算子 Prox_p 的计算; (b) 线性系统 (3.14) 中 V_j 的构造; (c) 线性系统 (3.14) 的高效求解. 其中, 如同注 4 所讨论的, (b) 中 V_j 的构造需要根据函数 h 和 p 具体的形式进行讨论. 这些讨论偏离了本文主要目标, 故我们在本小节仅关注 (a) 和 (c). 对 ∇h^* 特别是 Prox_p 的广义 Jacobian 的计算及其替代映射的设计感兴趣的读者可以参考文献 [30,31,34,38,65].

4.1. 邻近算子 Prox_p 的计算

这里, 我们主要考虑定义在 $\Re^n$ 上的非光滑正则函数 p 的邻近算子 Prox_p 的计算. 关于定义在矩阵空间上的非光滑正则函数 p 的邻近算子的相关计算则更为复杂, 由于篇幅及作者

知识的限制, 本文就不作讨论了, 感兴趣的读者可以参见 [12,13]. 我们首先给出机器与统计学习中常见的正则函数并给出它们对应的邻近算子. 特别的, 我们将一些常见的正则函数总结在表 1 中.

表 1 正则函数 p 参数 ($\lambda_1 > 0$, $\lambda_2 > 0$)

	$p(x)$	文献
Lasso	$\lambda_1\|x\|_1$	[55]
Fused Lasso	$\lambda_1\|x\|_1 + \lambda_2\sum_{i=2}^{n}\|x_i - x_{i-1}\|$	[56]
Clustered Lasso	$\lambda_1\|x\|_1 + \lambda_2\sum_{i<j}\|x_i - x_j\|$	[51]
OSCAR	$\lambda_1\|x\|_1 + \lambda_2\sum_{i<j}\max\{\|x_i\|, \|x_j\|\}$	[6]
Sparse group Lasso	$\lambda_1\|x\|_1 + \lambda_2\sum_{l=1}^{g} w_l\|x_{G_l}\|$	[62]
Elastic net	$\lambda_1\|x\|_1 + \lambda_2\|x\|_2^2$	[67]

表 1 中最简单的正则函数为 ℓ_1 正则函数 $\lambda_1\|\cdot\|_1$, 它的邻近算子有如下显式表达式:

$$\mathrm{Prox}_{\lambda_1\|\cdot\|_1}(x) = \mathrm{sgn}(x)\circ\max\{\mathrm{abs}(x) - \lambda_1, 0\}, \quad \forall\, x \in \Re^n,$$

其中 "sgn" 是符号函数, "abs(x)" 表示 x 的绝对值, 而 "$\circ$" 代表向量分量相乘. 同时, 简要考察可以发现, Elastic net 正则函数所对应的邻近算子的计算也可以简化为 ℓ_1 范数的邻近算子的计算. 而表 1 中其它邻近算子的计算则需要更多的考察. 我们注意到表 1 中其它四种正则函数都有相同的相加结构, 即它们都可以表示为 ℓ_1 范数与另一个凸函数之和. 这种特殊的求和结构给邻近算子的计算带来了巨大的帮助, 由此我们可以得到这些邻近算子的半显式表达式 (semi-closed-form representation). 特别的, 由于 Fussed Lasso 正则函数可以表示为 ℓ_1 范数与总变差 (total variation) 范数之和, Friedman et al. 在 [18] 中得到 Fussed Lasso 正则函数的邻近算子可以由 ℓ_1 范数的邻近算子与总变差范数的邻近算子的复合得到. 随后, Yu 在文献 [61] 中给出了更一般的结果. 特别的, 他指出如果函数 p 可以表示为两个函数 p_1 与 p_2 之和, 即 $p(x) = p_1(x) + p_2(x)$, $\forall\, x \in \Re^n$, 那么在一些关于 p, p_1, p_2 的正则性条件下有

$$\mathrm{Prox}_p(x) = \mathrm{Prox}_{p_1}(\mathrm{Prox}_{p_2}(x)), \quad \forall\, x \in \Re^n.$$

幸运的是, 表 1 中的正则函数都满足 [61] 中的正则条件. 因此基于上面的分解公式, 同 Fused Lasso 情形一样, 我们可以得到 Clustered Lasso [34] 和 Sparse group Lasso [65] 所对应的邻近算子的半显式表达式. 表 1 中比较特别的是 OSCAR 正则函数, 我们采用一种比分解公式更快速的计算方式. 注意到 OSCAR 正则函数可等价变换成 [38]:

$$\lambda_1\|x\|_1 + \lambda_2\sum_{i<j}\max\{|x_i|, |x_j|\} = \sum_{i=1}^{n} r_i|x|_i^{\downarrow}, \quad \forall\, x \in \Re^n,$$

这里 $|x|_1^{\downarrow} \geq |x|_2^{\downarrow} \geq \cdots |x|_n^{\downarrow}$ 且 $r_i = \lambda_1 + \lambda_2(n-i)$, $i = 1, \cdots, n$. 这样我们得到新函数 $\kappa_r(x) := \sum_{i=1}^{n} r_i|x|_i^{\downarrow}$ 且 $r_1 \geq r_2 \geq \cdots r_n \geq 0$. κ_r 在文献中又被称为排序加权 ℓ_1 范数且它的邻近算子可以用 pool-adjacent-violators 算法 [4,5] 高效计算. 这样我们也随之得到了 OSCAR 正则函数的邻近算子. 基于这些讨论, 我们将表 1 中正则函数的邻近算子总结在下表中.

表 2 p **的邻近算子**

	$\mathrm{Prox}_p(x)$	文献
Lasso	$\mathrm{Prox}_{\lambda_1\|\cdot\|_1}(x)$	[55]
Fused Lasso	$\mathrm{Prox}_{\lambda_1\|\cdot\|_1}(\mathrm{Prox}_{TV}(x))$	[18]
Clustered Lasso	$\mathrm{Prox}_{\lambda_1\|\cdot\|_1}(\mathrm{Prox}_{Cl}(x))$	[34]
OSCAR	$\mathrm{Prox}_{\kappa_r}(x)$	[38]
Sparse group Lasso	$\mathrm{Prox}_{\lambda_2 w_l\|\cdot\|}(\mathrm{Prox}_{\lambda_1\|\cdot\|_1}(x_l)),\ x_l\in\Re^{\|G_l\|},\ l=1,2,\cdots,g$	[65]
Elastic net	$\mathrm{Prox}_{\lambda_1/(1+2\lambda_2)\|\cdot\|_1}(x/(1+2\lambda_2))$	[67]

在表 2 中,

$$\mathrm{Prox}_{TV}(x) := \underset{z\in\Re^n}{\mathrm{argmin}}\Big\{\frac{1}{2}\|z-x\|^2+\lambda_2\sum_{i=2}^{n}|z_i-z_{i-1}|\Big\},$$
$$\mathrm{Prox}_{Cl}(x) := \underset{z\in\Re^n}{\mathrm{argmin}}\Big\{\frac{1}{2}\|z-x\|^2+\lambda_2\sum_{i<j}|z_i-z_j|\Big\}.$$

这里 $\mathrm{Prox}_{TV}(x)$ 可以用 Condat 的直接算法得到 [9]; $\mathrm{Prox}_{Cl}(x)$ 的快速计算可以在文献 [34] 中找到. 另外, $\mathrm{Prox}_{\lambda_2 w_l}$ 可以由如下公式得到:

$$\mathrm{Prox}_{\lambda_2 w_l}(u)=\begin{cases}\dfrac{u}{\|u\|}\max\{\|u\|-\lambda_2 w_l,0\}, & \text{如果}u\neq 0,\\ 0, & \text{如果}u=0.\end{cases}$$

4.2. 线性系统 (3.14) 的高效求解

下面我们考虑线性系统 (3.14) 的. 与上节类似, 为了简化叙述, 我们将问题 (1.1) 的空间 $\mathcal{X}$ 和 $\mathcal{Y}$ 特别地固定为 $\mathcal{X}=\Re^n$, $\mathcal{Y}=\Re^m$. 由此, 我们可以得到线性映射 $\mathcal{A}$ 在 $\Re^n$ 和 $\Re^m$ 的标准基下的矩阵表示 $A\in\Re^{m\times n}$. 给定 $(\xi,\tilde{x})\in\Re^m\times\Re^n$, $\sigma>0$, 我们考虑如下的线性系统:

$$(H+\sigma AUA^T)d=-\nabla\psi(\xi), \tag{4.1}$$

这里 $H\in\partial(\nabla h^*)(\xi)$, $U\in\partial\mathrm{Prox}_{\sigma p}(\tilde{x}-\sigma(A^T\xi-c))$. 由于 H 是对称正定矩阵, 线性系统 (4.1) 可以等价的写成

$$\big(I_m+\sigma(L^{-1}A)U(L^{-1}A)^T\big)(L^Td)=-L^{-1}\nabla\psi(\xi),$$

其中 L 是由 H 的 (稀疏) Cholesky 分解 (Cholesky decomposition) 所得到的可逆矩阵且满足 $H=LL^T$. 在很多应用中, H 是稀疏矩阵. 事实上, 如果损失函数 h 取成二次或者逻辑损失函数, 则 H 为对角矩阵. 所以, 计算矩阵 L 和它的逆矩阵的时间在大多数情况下可以被忽略. 因此, 在这里, 我们不失一般性地考虑系统 (4.1) 的一个简化版本:

$$(I_m+\sigma AUA^T)d=-\nabla\psi(\xi), \tag{4.2}$$

这个系统恰好是损失函数 h 取为二次损失函数的情况. 为了进一步体现本小节的核心思想, 我们考虑特殊的非光滑正则函数 p 来进一步简化系统 (4.2). 特别的, 给定正参数 $\lambda_1>0$, 我们考虑 ℓ_1 范数为正则函数, 即 $p(\cdot)=\lambda_1\|\cdot\|_1$. 这样, (4.2) 恰好是如下 Lasso 问题所对应的半光滑牛顿系统:

$$\min_{x\in\Re^n}\left\{\frac{1}{2}\|Ax-b\|^2+\lambda_1\|x\|_1\right\},$$

其中 $b \in \Re^m$ 是给定数据向量. 在当前的设定下, $U \in \Re^{n\times n}$ 是对角矩阵. 粗略的考察可知, 这时计算矩阵 AUA^T 和对给定向量 $d \in \Re^m$ 的矩阵向量乘积 AUA^Td 的工作量分别是 $\mathcal{O}(m^2n)$ 和 $\mathcal{O}(mn)$. 当矩阵 A 的维数巨大时, 这些操作的巨大工作量使得常用的算法如 Cholesky 分解方法与共轭梯度法不适合用来求解大规模的系统 (4.2). 幸运的是, 在 Lasso 问题中, 如果我们可以探索并利用矩阵 U 的稀疏结构, 我们可以极大的减少这些对实际计算不利的工作量. 接下来, 我们将详细讨论如何去探索利用 U 的稀疏结构, 这里的 U 的稀疏性即为前言中提到的非光滑二阶信息, 在文献 [29] 中也被称为二阶稀疏性.

记 $x = \tilde{x} - \sigma(A^T\xi - c)$, 我们总是在实现中按照如下方式选取矩阵对角矩阵 $U = \mathrm{Diag}(u)$, 这里对角元素 u 的定义是

$$u_i = \begin{cases} 0, & \text{if} \quad |x_i| \le \sigma\lambda_1, \\ 1, & \text{otherwise}, \end{cases} \qquad i = 1, \ldots, n.$$

由于 $\mathrm{Prox}_{\sigma\lambda_1\|\cdot\|_1}(x) = \mathrm{sign}(x) \circ \max\{|x| - \sigma\lambda_1, 0\}$, 我们不难证明 $U \in \partial\mathrm{Prox}_{\sigma\lambda_1\|\cdot\|_1}(x)$. 定义索引指数集合 $\mathcal{J} := \{j \mid |x_j| > \sigma\lambda_1, j = 1, \ldots, n\}$, 记集合 $\mathcal{J}$ 中的元素个数为 r, 即 $r = |\mathcal{J}|$. 利用 U 特别的 0-1 结构, 我们可以将矩阵 AUA^T 进行如下等价改写

$$AUA^T = (AU)(AU)^T = A_{\mathcal{J}}A_{\mathcal{J}}^T, \tag{4.3}$$

其中 $A_{\mathcal{J}} \in \Re^{m\times r}$ 为 A 的子矩阵, 由 A 中下标在 $\mathcal{J}$ 内的列向量组成. 这样, 由 (4.3) 可知, 计算矩阵 AUA^T 和矩阵与向量乘积 AUA^Td 的工作量分别减少到 $\mathcal{O}(m^2r)$ 和 $\mathcal{O}(mr)$. 由于 ℓ_1 范数正则函数 p 具有提升最优解稀疏度的性质, 通常情况下 r 会远小于特征数量 n. 因此, 通过探索并利用 U 的稀疏结构即非光滑二次信息, 利用 Cholesky 分解求解线性系统 (4.2) 时的工作量可以被极大的减少. 特别的, 该求解过程的总工作量由 $\mathcal{O}(m^2(m+n))$ 减少到了 $\mathcal{O}(m^2(m+r))$. 图 1 形象地展示了工作量的减少. 图中红色的部分相当于矩阵 A 中那些占绝大部分的不在集合 $\mathcal{J}$ 中的列. 由此, 我们可以看到尽管特征数量 n 特别巨大 (如 $n \approx 10^7$),

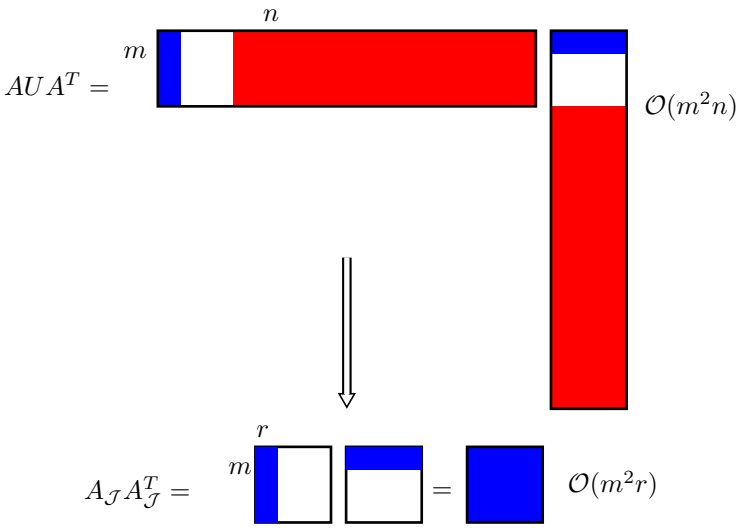

图 1　计算工作量由 $\mathcal{O}(m^2n)$ 减少到 $\mathcal{O}(m^2r)$

只要 m 和 r 大小合适 (如 10^4 左右), 牛顿线性系统 (4.2) 仍然可以使用 Cholesky 分解方法高效地进行求解.

如果同时样本数量 m 也远大于 $\mathrm{Prox}_{\sigma\lambda_1\|\cdot\|_1}(x)$ 的非零元素个数 r, 即 $(r \ll m)$, 那么我们还可以进一步缩减计算工作量. 与直接分解一个 $m \times m$ 维的矩阵不同, 这里我们可以借助于 Sherman-Morrison-Woodbury 公式 [21] 来通过求一个 $r \times r$ 维小矩阵的逆来得到 $m \times m$ 维大矩阵 $I_m + \sigma AUA^T$ 的逆:

$$(I_m + \sigma AUA^T)^{-1} = (I_m + \sigma A_{\mathcal{J}} A_{\mathcal{J}}^T)^{-1} = I_m - A_{\mathcal{J}}(\sigma^{-1} I_r + A_{\mathcal{J}}^T A_{\mathcal{J}})^{-1} A_{\mathcal{J}}^T.$$

图 2 展示了 $r \times r$ 维矩阵 $A_{\mathcal{J}}^T A_{\mathcal{J}}$ 的计算工作量. 利用这种方法, 求解牛顿线性系统 (4.2) 的工作量由 $\mathcal{O}(m^2(m+r))$ 进一步缩减到 $\mathcal{O}(r^2(m+r))$. 可见对优化模型内蕴非光滑二阶信息的探索, 以及将这些信息与数值线性代数技巧精巧的结合可以使我们极大的减少求解牛顿线性系统的工作量, 从而对求解原始优化问题的算法 3.1 进行加速. 当然, 如果共轭梯度法被用来求解系统 (4.2), 我们也可以用类似的方式做到算法工作量的减少与加速, 这里就不再赘述了.

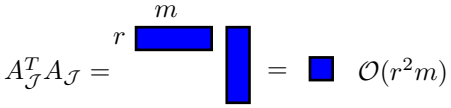

图 2 计算工作量进一步缩减到 $\mathcal{O}(r^2 m)$

对于损失函数 h 不是二次函数的一般情况, 从以上的讨论可知, 只要 $\mathrm{Prox}_{\sigma\lambda_1\|\cdot\|_1}(x)$ 中非零元个数 r 足够小 (如小于 $\sqrt{n}$), 且 $H \in \partial(\nabla h^*)(\xi)$ 维稀疏矩阵 (如对角矩阵), 我们总是可以以较低的工作量求解线性系统 (4.1). 这里一个常见的问题是, 尽管 Lasso 问题只有稀疏最优解, 在算法进行阶段 (尤其是开始阶段), 很有可能 $\mathrm{Prox}_{\sigma\lambda_1\|\cdot\|_1}(x)$ 中非零元的个数很大, 甚至与 n 同阶. 该问题中的困难可以从两个层次进行解决: 首先, 这种现象在实际情况中较为少见, 因为在实现中, 算法总是用稀疏可行解为初始点, 例如零向量; 其次, 就算这种现象出现了, 我们仅需要使用较少步的共轭梯度迭代就可以很好地处理这个线性系统, 因为这个时候参数 σ 通常比较小, 线性系统的条件数会比较好, 且当前迭代点距离最优解较远, 线性系统不需要精确求解.

本小节对正则函数 $p(\cdot)$ 取为 $\lambda_1\|\cdot\|_1$ 时半光滑牛顿算法的高效实现进行了细致的讨论. 对于其它的正则函数, 矩阵 U 的结构更为复杂, 因此线性系统 (4.2) 的求解需要更精细的考量. 本文由于篇幅受限, 这里就不一一叙述了, 感兴趣的读者可以参见文献 [30, 34, 38, 65]. 这些文献中的数值算例表明, 如果可以高质量地利用问题内蕴的二阶特殊结构, 并与数值线性代数技术精巧结合, 半光滑牛顿算法每步迭代的工作量将会与当前流行一阶算法每步迭代工作量相当, 甚至更少. 注意到这与二阶算法每步迭代工作量巨大的传统观点相反, 而我们所设计的算法取得这一进展的重要原因之一即为对原问题所蕴含的非光滑性的深入探索与利用. 从这个角度来看, 优化模型中的非光滑特性不应该让我们恐惧. 恰恰相反, 它们的存在是我们求解超大规模问题的希望, 它们是我们在算法设计过程中应该积极拥抱的重要元素.

5. 总 结

本文介绍了求解复合凸优化问题的一类快速邻近点算法, 重点阐述了算法的全局收敛与

超线性局部收敛性质. 同时, 本文描述了如何结合对偶技巧与半光滑牛顿算法高效稳定求解邻近点算法的子问题, 并结合实际稀疏优化问题, 描述了如何探索利用问题的非光滑内蕴结构, 高效实现半光滑牛顿算法. 借此作者希望向读者介绍并推荐非光滑场景下的二阶算法, 并希望本文介绍的算法与思路可以被用来求解更多大规模优化问题.

致谢. 本文作者感谢两位匿名审稿人的建议和指正, 并特别感谢新加坡国立大学林媚霞博士对本文内容的帮助.

参 考 文 献

[1] Beck A and Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems[J]. SIAM Journal on Imaging Sciences, 2009, 2: 183–202.

[2] Benjamin R, Fazel M and Parrilo P A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization[J]. SIAM Review, 2010, 52: 471–501.

[3] Bertsekas D P. Nonlinear Programming[M], Athena Scientific, 1999.

[4] Best M J and Chakravarti N. Active set algorithms for isotonic regression; a unifying framework[J]. Mathematical Programming, 1990, 47: 425–439.

[5] Bogdan M, van den Berg E, Sabatti C, Su W and Candès E J. SLOPE-adaptive variable selection via convex optimization[J]. Annals of Applied Statistics, 2015, 9: 1103–1140.

[6] Bondell H D and Reich B J. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR[J]. Biometrics, 2008, 64: 115–123.

[7] Bauschke H H and Combettes P L. Convex Analysis and Monotone Operator Theory in Hilbert Spaces[M]. Springer, New York, 2011.

[8] Clarke F H. Optimization and Nonsmooth Analysis[M]. SIAM, 1990.

[9] Condat L. A direct algorithm for 1-D total variation denoising[J]. IEEE Signal Processing Letters, 2013, 20: 1054–1057.

[10] Cui Y, Ding C and Zhao X Y. Quadratic growth conditions for convex matrix optimization problems associated with spectral functions[J]. SIAM Journal on Optimization, 2017, 27: 2332–2355.

[11] Cui Y, Sun D F and Toh K C. On the R-superlinear convergence of the KKT residuals generated by the augmented Lagrangian method for convex composite conic programming[J]. Mathematical Programming, 2019, 178: 381–415.

[12] Ding C. An Introduction to a Class of Matrix Optimization Problems[D]. Ph.D Thesis, Department of Mathematics, National University of Singapore, 2012.

[13] Ding C, Sun D F and Toh K C. An introduction to a class of matrix cone programming[J]. Mathematical Programming, 2014, 144: 141–179.

[14] Dontchev A L and Rockafellar R T. Implicit Functions and Solution Mappings[M]. Springer, New York, 2009.

[15] Fischer A. Local behavior of an iterative framework for generalized equations with nonisolated solutions[J]. Mathematical Programming, 2002, 94: 91–124.

[16] Facchinei F, Fischer A and Herrich M. An LP-Newton method: nonsmooth equations, KKT systems, and nonisolated solutions[J]. Mathematical Programming, 2014, 146: 1–36.

[17] Facchinei F and Pang J S. Finite-Dimensional Variational Inequalities and Complementarity Problems[M]. Springer, New York, 2003.

[18] Friedman J, Hastie T, Hofling H and Tibshirani R. Pathwise coordinate optimization[J]. The annals of applied statistics, 2007, 1: 302–332.

[19] Gabay D and Mercier B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation[J]. Computers & Mathematics with Applications, 1976, 2: 17–40.

[20] Glowinski R and Marroco A. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires[J]. Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique, 1975, 9: 41–76.

[21] Golub G and Van Loan C F. Matrix Computations[M]. 3nd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[22] Goebel R and Rockafellar R T. Local strong convexity and local Lipschitz continuity of the gradient of convex functions[J]. Journal of Convex Analysis, 2008, 15: 263–270.

[23] Hiriart-Urruty J B, Strodiot J J and Nguyen V H. Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data[J]. Applied Mathematics and Optimization, 1984, 11: 43–56.

[24] Kummer B, Newton's method for non-differentiable functions[J]. Advances in Mathematical Optimization, 1988, 45: 114–125.

[25] Lee J D, Sun Y and Saunders M A. Proximal Newton-type methods for minimizing composite functions[J]. SIAM Journal on Optimization, 2014, 24: 1420–1443.

[26] Leventhal D. Metric subregularity and the proximal point method[J]. Journal of Mathematical Analysis and Applications, 2009, 360: 681–688.

[27] Li X D, Sun D F and Toh K C. A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions[J]. Mathematical Programming, 2016, 155: 333–373.

[28] Li X D, Sun D F and Toh K C. QSDPNAL: A two-phase augmented Lagrangian method for convex quadratic semidefinite programming[J]. Mathematical Programming Computation, 2018, 10: 703–743.

[29] Li X D, Sun D F and Toh K C. A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems[J]. SIAM Journal on Optimization, 2018, 28: 433–458.

[30] Li X D, Sun D F and Toh K C. On efficiently solving the subproblems of a level-set method for fused Lasso problems[J]. SIAM Journal on Optimization, 2018, 28: 1842–1866.

[31] Li X D, Sun D F and Toh K C. On the efficient computation of a generalized Jacobian of the projector over the Birkhoff polytope[J]. Mathematical Programming, 2020, 178: 419–446.

[32] Li X D, Sun D F and Toh K C, An asymptotically superlinearly convergent semismooth Newton augmented Lagrangian method for Linear Programming[J]. SIAM Journal on Optimization, 2020, 30: 2410–2440.

[33] Li G Y and Mordukhovich B S. Hölder metric subregularity with applications to proximal point method[J]. SIAM Journal on Optimization, 2012, 22: 1655–1684.

[34] Lin M, Liu Y J, Sun D and Toh K C. Efficient sparse hessian based algorithms for the clustered lasso problem[J]. SIAM Journal on Optimization, 2019, 29: 2026–2052.

[35] Lin M, Sun D F, Toh K C and Yuan Y. A dual Newton based preconditioned proximal point algorithm for exclusive lasso models, arXiv:1902.00151, 2019.

[36] Luo Z Q and Tseng P. On the linear convergence of descent methods for convex essentially smooth minimization[J]. SIAM J. Control and Optimization, 1992, 30: 408–425.

[37] Luo Z Q and Tseng P. Error bounds and convergence analysis of feasible descent methods: a general approach[J]. Annals of Operations Research, 1993, 46: 157–178.

[38] Luo Z, Sun D F, Toh K C and Xiu N. Solving the OSCAR and SLOPE models using a semismooth Newton-based augmented Lagrangian method[J]. Journal of Machine Learning Research, 2019, 20: 1–25.

[39] Luque F J. Asymptotic convergence analysis of the proximal point algorithm[J]. SIAM Journal on Control and Optimization, 1984, 22: 277–293.

[40] Mifflin R. Semismooth and semiconvex functions in constrained optimization[J]. SIAM Journal on Control and Optimization, 1977, 15: 959–972.

[41] Mordukhovich B S and Ouyang W. Higher-order metric subregularity and its applications[J]. Journal of Global Optimization, 2015, 63: 777–795.

[42] Nesterov Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$[J]. Soviet Mathematics Doklady, 1983, 27: 372–376.

[43] Qi H and Sun D F. A quadratically convergent Newton method for computing the nearest correlation matrix[J]. SIAM Journal on Matrix Analysis and Applications, 2006, 28: 360–385.

[44] Qi L and Sun J. A nonsmooth version of Newton's method[J]. Mathematical Programming, 1993, 58: 353–367.

[45] Robinson S M. An implicit-function theorem for generalized variational inequalties. Technical Summary Report No. 1672, Mathematics Research Center, University of Wisconsin-Madison, 1976; available from National Technical Information Service under Accession No. ADA031952.

[46] Robinson S M. Some continuity properties of polyhedral multifunctions, In Mathematical Programming at Oberwolfach, vol. 14 of Mathematical Programming Studies, Springer, Berlin, Heidelberg, 1981, 206–214.

[47] Rockafellar R T. Convex Analysis[M]. Princeton University Press, 1970.

[48] Rockafellar R T. Monotone operators and the proximal point algorithm[J]. SIAM Journal on Control and Optimization, 1976, 14: 877–898.

[49] Rockafellar R T. Augmented Lagrangians and applications of the proximal point algorithm in convex programming[J]. Mathematics of Operations Research, 1976, 1: 97–116.

[50] Rockafellar R T and Wets R J B. Variational Analysis[M]. Springer, New York, 1998.

[51] She Y. Sparse regression with exact clustering[J]. Electronic Journal of Statistics, 2010, 4: 1055–1096.

[52] Sun D F and Sun J. Semismooth matrix-valued functions[J]. Mathematics of Operations Research, 2002, 27: 150–169.

[53] Sun J. On Monotropic Piecewise Qudratic Programming[D]. Ph.D Thesis, Department of Mathematics, University of Washington, Seattle, 1986.

[54] Tang P, Wang C, Sun D F and Toh K C. A sparse semismooth Newton based proximal majorization-minimization algorithm for nonconvex square-root-loss regression problems. Journal of Machine Learning Research, in print, 2020.

[55] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B, 1996, 58: 267–288.

[56] Tibshirani R, Saunders M, Rosset S, Zhu J and Knight K. Sparsity and smoothness via the fused lasso[M]. Journal of the Royal Statistical Society: Series B, 2005, 67: 91–108.

[57] Tseng P and Yun S. A coordinate gradient descent method for nonsmooth separable minimization[J]. Mathematical Programming, 2010, 125: 387–423.

[58] Tseng P. Approximation accuracy, gradient methods, and error bound for structured convex optimization[J]. Mathematical Programming, 2010, 125: 263–295.

[59] Xiao X, Li Y, Wen Z and Zhang L. A regularized semi-smooth Newton method with projection steps for composite convex programs[J]. Journal of Scientific Computing, 2018, 76: 364–389.

[60] Yang L Q, Sun D F and Toh K C. SDPNAL+: A majorized semismooth Newton-CG augmented Lagrangian method for semidefinite programming with nonnegative constraints[J]. Mathematical Programming Computation, 2015, 7: 331–366.

[61] Yu Y. On decomposing the proximal map, in Advances in Neural Information Processing Systems, 2013, 91–99.

[62] Yuan M and Lin Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society: Series B, 2006, 68: 49–67.

[63] Yue M X, Zhou Z and So A M C. A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo– Tseng error bound property[J]. Mathematical Programming, 2019, 174: 327–358.

[64] Zhao X Y, Sun D F and Toh K C. A Newton-CG augmented Lagrangian method for semidefinite programming[J]. SIAM Journal on Optimization, 2010, 20: 1737–1765.

[65] Zhang Y J, Zhang N, Sun D F and Toh K C. An efficient Hessian based algorithm for solving large-scale sparse group Lasso problems[J]. Mathematical Programming, 2020, 179, 223–263.

[66] Zhou Z R and So A M C. A unified approach to error bounds for structured convex optimization problems[J]. Mathematical Programming, 2017, 165: 689–728.

[67] Zou H and Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B, 2005, 67: 301–320.

EFFICIENT PROXIMAL POINT ALGORITHM FOR CONVEX COMPOSITE OPTIMIZATION

Li Xudong

(*School of Data Science, and Shanghai Center for Mathematical Sciences, Fudan University, Shanghai 200433, China*)

Abstract

In the Big Data era, with the advent of convenient automated data collection technologies, large-scale composite convex optimization problems are ubiquitous in many applications, such as massive data analysis, machine and statistical learning, image and signal processing. In this paper, we review a class of efficient proximal point algorithms for solving the large-scale composite convex optimization problems. Under the easy-to-implement stopping criteria and mild calmness conditions, we show the proximal point algorithm enjoys global and local asymptotic superlinear convergence. Meanwhile, based on the duality theory, we propose an efficient semismooth Newton method for handling the subproblems in the proximal point algorithm. Lastly, to further accelerate the proximal point algorithm, we fully exploit the nonsmooth second order information induced by the nonsmooth regularizer in the problem to achieve a dramatic reduction of the computational costs of solving the involved semismooth Newton linear systems.

Keywords: composite optimization; proximal point algorithm; semismooth Newton method.

2010 Mathematics Subject Classification: 65F10, 90C06, 90C25, 90C31.