



结构稀疏优化模型的理论及算法

胡耀华^{1*}, 李昱帆², 刘艳艳¹, 覃静³

1. 深圳大学数学科学学院, 深圳 518060;

2. 中山大学理学院, 深圳 518107;

3. 中山大学药学院 (深圳), 深圳 518107

E-mail: mayhhu@szu.edu.cn, liyufan@mail.sysu.edu.cn, liuyanyan2023@email.szu.edu.cn, qinj29@mail.sysu.edu.cn

收稿日期: 2024-02-07; 接受日期: 2024-04-17; 网络出版日期: 2024-06-12; * 通信作者

国家自然科学基金 (批准号: 12222112, 12071306 和 32170655)、广东省教育厅重点领域专项 (批准号: 2023ZDZX1017)、深圳市科技计划 (批准号: RCJC20221008092753082, 202206193000001 和 20220817122906001) 和深圳大学科研团队培育项目 (批准号: 2023QNT011) 资助项目

摘要 结构稀疏优化在压缩感知、信号与图像处理、机器学习和生物医学等领域应用广泛, 是优化领域非常热门的研究课题. 本文首先介绍稀疏优化与组稀疏优化的相关模型, 并概述混合稀疏优化与联合稀疏优化的相关模型及近年来的进展; 其次从理论与算法两个方面阐述关于稀疏优化与组稀疏优化模型的研究进展, 包括相合性理论与求解算法; 最后提出一些结构稀疏优化领域值得深入研究的方向. 总体来说, 结构稀疏优化的研究虽已取得一些成果, 但仍存在理论问题亟待解决.

关键词 稀疏优化 组稀疏优化 相合性理论 一阶算法 二阶算法 收敛性理论

MSC (2020) 主题分类 65K05, 49M37, 90C26

1 引言

在当今的信息爆炸时代, 人们对信息的需求量暴增, 所面临的数据量与数据规模越来越大, 运用传统的香农采样定理处理数据的局限性愈发明显. 为了不失真地恢复信号, 香农采样定理要求采样频率大于或等于信号最高带宽的 2 倍 (参见文献 [132]). 随着信号带宽的增大以及采样频率的增高, 对信号处理的要求变得愈加苛刻, 因为需要针对预先收集的数据, 去除其中的冗余信息, 只保留和存储关键信息. 然而较高的采样频率会使得采样数据量过大, 进而导致信息的冗余度增高, 因此不得不消耗较高的存储空间和计算成本.

在实际应用中, 高维数据的分量之间往往会存在一定程度的相关性, 进而可以保证数据在某些变换下是稀疏的或是可压缩的. 稀疏性指数据的分量中只有极少数的元素不为零, 绝大部分元素为零. Donoho^[44]、Candès 和 Tao^[29] 基于信号的稀疏性和可压缩性提出的压缩感知技术引起了业界学者的关注, 在信号处理^[29, 44, 58]、机器学习^[71, 169]、基因组学数据分析^[121, 152]、统计建模^[51] 和磁共振成

英文引用格式: Hu Y H, Li Y F, Liu Y Y, et al. Theory and algorithms of structured sparse optimization problems (in Chinese). Sci Sin Math, 2024, 54: 1045–1070, doi: 10.1360/SSM-2024-0032

像^[52,60,104]等相关领域成为了研究热点. 压缩感知理论的诞生突破了传统采样技术的局限性, 即如果原始信号(图像)满足稀疏性, 则可以通过极少的采样数据来精确恢复原始信号(图像). 这不仅缩小了采样数据的规模, 而且大大降低了处理冗余信息的难度.

结构稀疏优化指决策变量具有结构稀疏性的优化问题, 旨在寻找欠定线性系统的结构稀疏解. 第 1.1–1.4 小节分别介绍稀疏优化模型、组稀疏优化模型、混合稀疏优化模型和联合稀疏优化模型.

1.1 稀疏优化模型

在很多实际问题中, 高维信号经过压缩处理通常可以表示为稀疏或近似稀疏的向量, 进而可以由少量观测值恢复, 即可表示为求解如下线性方程组的解:

$$b = Ax + e, \quad (1.1)$$

其中, $b \in \mathbb{R}^m$ 为观测向量, $A \in \mathbb{R}^{m \times n}$ 是感知矩阵 ($m \ll n$), $x \in \mathbb{R}^n$ 是原始信号, $e \in \mathbb{R}^m$ 是噪声向量. 由于 $m \ll n$, 欠定线性方程组 (1.1) 是病态的, 存在无穷多个解. 此时, 若在解空间中加上稀疏性的先验条件, 则 (1.1) 存在有效解. 根据测量过程是否含噪声, 可表示为下述 ℓ_0 范数极小化问题:

$$\min_{x \in \mathbb{R}^n} \{ \|x\|_0 : Ax = b \text{ 或 } \|Ax - b\|_2 \leq \varepsilon \}, \quad (1.2)$$

其中, $\varepsilon > 0$ 是噪声水平, $\|x\|_0$ 表示 x 中非零元素的个数, 通常被称为 ℓ_0 范数¹⁾, 它可以刻画向量的稀疏度, 特别地, 若 $\|x\|_0 \leq s$, 则称 x 为 s -稀疏信号. 当解的稀疏先验信息或噪声水平未知时, 可以通过如下 ℓ_0 范数正则化问题来恢复稀疏信号:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_0, \quad (1.3)$$

其中 $\lambda > 0$ 为正则化参数, 用于平衡系统数据拟合的准确性与解的稀疏性. ℓ_0 范数极小化问题 (1.2) 与 ℓ_0 范数正则化问题 (1.3) 均是 NP (non-deterministic polynomial)- 难的 (参见文献 [1,111]), 直接求解困难. 为求解这两类问题, 一般对 ℓ_0 范数进行松弛, 将 (1.2) 和 (1.3) 转化为下述易于求解的松弛问题:

$$\min_{x \in \mathbb{R}^n} \{ \Phi(x) : Ax = b \text{ 或 } \|Ax - b\|_2 \leq \varepsilon \}, \quad (1.4a)$$

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \Phi(x), \quad (1.4b)$$

其中稀疏诱导函数 $\Phi: \mathbb{R}^n \rightarrow \mathbb{R}$ 是 ℓ_0 范数的松弛, 也被称为罚函数.

由于 ℓ_1 范数是 ℓ_0 范数在单位球内的最佳凸近似, 若用 ℓ_1 范数松弛 ℓ_0 范数, 即 $\Phi(x) = \|x\|_1 := \sum_{i=1}^n |x_i|$, 此时 (1.4a) 和 (1.4b) 分别变为凸的 ℓ_1 范数极小化问题与 ℓ_1 范数正则化问题, ℓ_1 范数正则化问题也被称为 LASSO (least absolute shrinkage and selection operator) 模型^[139]. 得益于 ℓ_1 范数的凸性, 快速求解 ℓ_1 范数松弛问题变得可行. 基于 ℓ_1 范数的稀疏优化研究可追溯到 20 世纪 90 年代, 其相关的理论性质与算法研究在近年来蓬勃发展, 已经成功应用于信号与图像处理^[8,146]、生物信息学^[133,166] 和计算机视觉^[160].

基于 ℓ_1 范数的稀疏优化研究虽然取得了很大的成功, 但仍有理论和实证研究表明 ℓ_1 范数容易过度惩罚稀疏向量中的元素值, 使得解产生偏差, 导致 ℓ_1 范数的稀疏诱导性不佳 (参见文献 [31,51,57,108]),

1) ℓ_0 范数与后面的 ℓ_p 范数实际上并非范数, 只是一种叫法.

在现实应用中通常会导致解的次优稀疏性^[175]. 研究发现, 使用非凸松弛函数近似 ℓ_0 范数可以得到比 ℓ_1 范数更稀疏的解. 例如, 若用 ℓ_p ($0 < p < 1$) 范数松弛 ℓ_0 范数, 即 $\Phi(x) = \|x\|_p^p := \sum_{i=1}^n |x_i|^p$, 直观来看, p 值越小, ℓ_p 范数越能刻画稀疏度, 并且 $\lim_{p \rightarrow 0} \|x\|_p^p = \|x\|_0$, $\lim_{p \rightarrow 1} \|x\|_p^p = \|x\|_1$. 从理论上来看, ℓ_p 范数比 ℓ_1 范数具有低阶惩罚的显著优点, 它在较弱的条件下也能保证精确恢复 (参见文献 [76, 103]). 从数值上来看, ℓ_p 范数正则化相比 ℓ_1 范数正则化模型也具有更强的稀疏诱导性, 它能保证从更少的观测值中恢复数据 (参见文献 [36, 70, 71, 94, 151, 163]). 特别地, 当 $p = 1/2$ 时, 基于基因调控网络的研究发现 $\ell_{1/2}$ 范数正则化模型在生物学意义上能获得比 ℓ_1 范数正则化模型更可靠的解 (参见文献 [69, 71, 121, 122, 152]).

$\Phi(x)$ 还可取对数和、变换 ℓ_1 范数、分式 ℓ_1 范数、光滑截断绝对偏差 (smoothly clipped absolute deviation, SCAD) 和极小极大凹罚 (minimax concave penalty, MCP). 类似 ℓ_p 范数, 对数和、变换 ℓ_1 范数、分式 ℓ_1 范数比 ℓ_1 范数具有更强的稀疏诱导性^[30, 105, 113], SCAD 与 MCP 在选择重要变量时相较于 ℓ_1 范数不会产生过度偏差, 可以同时用于变量选择和参数估计, 具备更优越的统计性质 (参见文献 [51, 171]). 以上所有松弛函数均可以表示为 $\Phi(x) = \sum_{i=1}^n \varphi(x_i)$, 其中 $\varphi(t) : \mathbb{R} \rightarrow \mathbb{R}$ 的具体数学形式如表 1 所示.

除上述松弛函数外, 基于排序思想的罚函数也有广泛的研究, 基于不同松弛函数 Φ 的排序罚函数^[19, 20, 117, 138] 在表 2²⁾ 给出. 此外, 基于截断思想的罚函数也引起了很多关注, 基于不同松弛函数 Φ 的部分罚函数和盖型罚函数详见表 2. 特别地, 当 Φ 取 ℓ_1 范数时, 部分 ℓ_1 罚函数也称为部分 LASSO, 它在求解压缩感知和逻辑回归时所产生的解优于 ℓ_1 范数^[102], 现多被用于深度学习领域^[170]; 当 Φ 取不同的松弛函数时, 相应的盖型罚函数被应用于字典学习^[80]、图像恢复与重建^[114, 118].

表 1 常见的松弛函数

松弛函数	$\varphi(t)$
ℓ_1 范数 ^[84, 139]	$ t $
ℓ_p 范数 ^[31, 57, 163]	$ t ^p$
对数和 ^[30]	$\log(\frac{ t }{\tau} + 1)$
变换 ℓ_1 范数 ^[113]	$\frac{\tau t }{1+\tau t }$
分式 ℓ_1 范数 ^[105]	$\frac{(\tau+1) t }{\tau+ t }$
SCAD ^[51]	$\begin{cases} \lambda t, & \text{当 } t \leq \lambda \text{ 时,} \\ \frac{2\tau\lambda t - (t^2 + \lambda^2)}{2(\tau-1)}, & \text{当 } \lambda < t \leq \tau\lambda \text{ 时,} \\ \frac{\lambda^2(\tau^2-1)}{2(\tau-1)}, & \text{当 } t > \tau\lambda \text{ 时} \end{cases}$
MCP ^[171]	$\begin{cases} \lambda t - \frac{t^2}{2\tau}, & \text{当 } t \leq \tau\lambda \text{ 时,} \\ \frac{1}{2}\tau\lambda^2, & \text{当 } t > \tau\lambda \text{ 时} \end{cases}$

表 2 特殊的罚函数

罚函数	$\varphi(t)$
排序罚函数 ^[19, 20, 75, 117, 138]	$\lambda_i \varphi _{[i]}$
部分罚函数 ^[10, 102, 170]	$\begin{cases} 0, & \text{当 } 1 \leq i \leq s \text{ 时,} \\ \varphi _{[i]}, & \text{当 } i > s \text{ 时} \end{cases}$
盖型罚函数 ^[80, 114, 118, 175]	$\min(\tau, \varphi _{[i]})$

2) 表中的 $|\varphi|_{[i]}$ 表示 $\{|\varphi(t_i)| : i = 1, \dots, n\}$ 中第 i 大的元素.

一般而言, 表 1 和 2 中的罚函数具有可分离结构, 但要注意的是, 对于排序罚函数, 当罚参数序列 $\{\lambda_i\}$ 按升序排列时, 对应的排序罚函数不可分离. 此时最小的权重分配给最强的信号, 相应的排序罚相比 LASSO 恢复性能更好^[75]. 而当罚参数序列 $\{\lambda_i\}$ 按降序排列, 即最大的权重分配给最强的信号时^[19, 20, 117, 138], 排序 ℓ_1 范数罚函数的计算复杂度与 LASSO 大体相同. 此外, 近年来流行的 $\ell_1 - \ell_2$ 罚函数 (即 $\|x\|_1 - \|x\|_2$)^[50, 100, 167] 和 ℓ_1/ℓ_2 罚函数 (即 $\|x\|_1/\|x\|_2$)^[68, 77, 124, 148] 均具有不可分离结构. $\ell_1 - \ell_2$ 罚最早由 Esser 等^[50] 提出, 系列研究表明, 它相比 ℓ_1 范数具有更强的稀疏诱导性, 已被成功应用于图像去噪、去模糊和磁共振成像等领域 (参见文献^[96, 100, 167]). ℓ_1/ℓ_2 罚函数最早由 Hoyer^[68] 提出, 具有尺度不变性^[77], 现如今在盲反卷积^[86, 128] 和稀疏滤波^[78, 120] 等领域应用颇多.

1.2 组稀疏优化模型

稀疏优化作为一类重要的结构优化问题, 已经得到了广泛的应用. 但要注意的是, 实际问题中的数据除了具有整体的稀疏性, 往往还存在组稀疏性. 组稀疏指向量的分量具有分组结构, 即每组的分量要么全为零要么全部非零, 且非零分量一般集中在某些组内. 利用组稀疏结构建模可以降低数据精确恢复所需要的样本数, 进而达到更好的恢复性能. 近年来, 组稀疏优化方法已被成功应用于统计学习^[73, 181]、机器学习^[131, 175]、图像处理^[101, 154]、生物信息学^[115, 189] 和神经影像学^[65] 等领域.

令 $x = (x_{\mathcal{G}_1}^T, \dots, x_{\mathcal{G}_r}^T)^T \in \mathbb{R}^n$, 其中 $\{\mathcal{G}_1, \dots, \mathcal{G}_r\}$ 为指标集 $[n] := \{1, 2, \dots, n\}$ 的一个分割, 即 $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$ 且 $\bigcup_{i=1}^r \mathcal{G}_i = [n]$. 当 $i = 1, 2, \dots, r$ 时, 记 $x_{\mathcal{G}_i} \in \mathbb{R}^{n_i}$ 为 x 对应下标在 $\mathcal{G}_i$ 中元素构成的子向量. 若规定 $0^0 = 0$, 则 x 的组稀疏性可由 $\ell_{p,q}$ ($p > 0, q \geq 0$) 范数衡量:

$$\|x\|_{p,q} := \begin{cases} \left(\sum_{i=1}^r \|x_{\mathcal{G}_i}\|_p^q \right)^{\frac{1}{q}}, & q > 0, \\ \sum_{i=1}^r \|x_{\mathcal{G}_i}\|_p^0, & q = 0. \end{cases}$$

特别地, $\|x\|_{p,0} = \|x\|_{2,0}$ 表示非零组数. 当 $p = q$ 时, $\|x\|_{p,p} = \|x\|_p$. 另外, 当 $r = n$ 时, 对任意的 $p > 0, q > 0$, $\ell_{p,q}$ 范数即为 ℓ_q 范数. 类似第 1.1 小节用 ℓ_0 范数刻画稀疏性, 可以利用 $\ell_{2,0}$ 范数来刻画组稀疏性, 进而考虑下述 $\ell_{2,0}$ 范数极小化问题与 $\ell_{2,0}$ 范数正则化问题:

$$\min_{x \in \mathbb{R}^n} \{ \|x\|_{2,0} : Ax = b \text{ 或 } \|Ax - b\|_2 \leq \varepsilon \}, \quad (1.5a)$$

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda \|x\|_{2,0}, \quad (1.5b)$$

受限于 $\|\cdot\|_{2,0}$ 的组合性质, 直接求解组稀疏优化问题 (1.5a) 与 (1.5b) 较为困难. 类似稀疏优化的松弛策略, 考虑下述松弛模型:

$$\min_{x \in \mathbb{R}^n} \{ \Phi(\|x_{\mathcal{G}}\|_p) : Ax = b \text{ 或 } \|Ax - b\|_2 \leq \varepsilon \}, \quad (1.6a)$$

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda \Phi(\|x_{\mathcal{G}}\|_p), \quad (1.6b)$$

其中 $\Phi(\|x_{\mathcal{G}}\|_p)$ 为组稀疏诱导函数, 它用 ℓ_p 范数刻画组内结构信息, 是组稀疏性 $\|x\|_{p,0}$ 的松弛. 当 Φ 取 ℓ_1 范数且 $p = 2$ 时, 相应的凸 $\ell_{2,1}$ 范数正则化模型由 Yuan 和 Lin^[168] 在研究统计学中的分组变量选择时被提出, 它是 ℓ_1 范数正则化模型的组稀疏推广, 也被称为组 LASSO 模型. 相较于 LASSO 模型, Huang 和 Zhang^[74] 发现组 LASSO 模型在处理具备强组稀疏结构的信号时性能表现优于 LASSO,

可以显著减少精确恢复所需要的样本数. 当 Φ 取 ℓ_q 范数 ($0 \leq q < 1$) 且 $p \geq 1$ 时, 相应的非凸 $\ell_{p,q}$ 范数正则化模型被 Hu 等^[71] 应用于基因调控网络, 研究发现, $\ell_{p,1/2}$ 范数正则化模型的恢复精度与稳健性均明显优于组 LASSO.

鉴于非凸罚函数在稀疏优化中良好的性能表现, 除 $\ell_{p,q}$ 范数外, 其他诱导组稀疏的非凸组罚函数也被提出, 如组 SCAD^[23, 72, 92, 153, 158]、组 MCP^[23, 72, 92, 158] 与组盖型罚函数^[118, 178] 等. 相较于组 LASSO, 基于基因表达和遗传关联研究的实例证明了组 SCAD 和组 MCP 罚函数具有更强的预测准确性^[23], 基于图像重建的应用研究也表明组盖型罚函数可以获得更高质量的解 (参见文献^[118, 178]).

1.3 混合稀疏优化模型

虽然稀疏优化与组稀疏优化均取得了重大成功, 但实际问题中的数据往往兼具整体稀疏性和组稀疏性, 例如, 基因推断网络问题中往往涉及组内稀疏性^[71, 134], 高维多元分类问题往往涉及组间稀疏性^[147]. 为了有效刻画组内稀疏性和组间稀疏性, 可考虑下述混合稀疏正则化模型:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda_1 \|x\|_0 + \lambda_2 \|x\|_{2,0}, \quad (1.7)$$

其中 $\lambda_1, \lambda_2 \geq 0$ 为惩罚参数. 当 $\lambda_1 = 0$ 时, 模型 (1.7) 退化为组稀疏正则化模型 (1.5b); 当 $\lambda_2 = 0$ 时, 模型 (1.7) 退化为稀疏正则化模型 (1.3). 针对模型 (1.7), Yap 等^[165] 设计了非单调迭代硬阈值算法, 并将其应用于扩散磁共振成像恢复. Li 等^[93] 将模型 (1.7) 中的数据拟合项推广为更一般的凸函数, 并设计了两类凸差算法进行求解.

类似于稀疏优化与组稀疏优化的凸松弛策略, Friedman 等^[59] 和 Simon 等^[134] 分别用 ℓ_1 范数和 $\ell_{2,1}$ 范数进行松弛, 考虑下述 ℓ_1 - $\ell_{2,1}$ 混合稀疏正则化模型:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda_1 \|x\|_1 + \lambda_2 \|x\|_{2,1}, \quad (1.8)$$

它也被称为混合稀疏 LASSO 模型. 相较 LASSO 模型, 混合稀疏 LASSO 模型可以有效克服对迭代变量的每个分量惩罚力度相同这一局限性, 且相较组 LASSO 模型能产生组内稀疏的解. 混合稀疏 LASSO 模型自提出以来得到了广泛的研究. Simon 等^[134] 在基因表达研究中通过比较 LASSO、组 LASSO 与混合稀疏 LASSO 的预测准确性, 验证了混合稀疏 LASSO 模型的高效性. Vincent 和 Hansen^[147] 将混合稀疏 LASSO 模型运用至基于多元回归模型分类问题, 发现多元混合稀疏 LASSO 分类器的性能明显优于多元 LASSO. 此外, 模型 (1.8) 还被应用于气候变化预测^[33]、不确定数据的特征选择^[161] 与目标追踪^[191] 等领域. 目前, 求解模型 (1.8) 的算法包括坐标下降算法^[59, 88]、可分离稀疏重构法^[136]、半光滑 Newton 增广 Lagrange 方法^[180] 与线性化交替方向乘法^[95] 等.

相比非凸松弛在稀疏优化与组稀疏优化中的广泛研究与成功应用, 目前关于混合稀疏优化非凸松弛的理论及算法研究还偏少, 希望未来在这一方面能有更深入的研究.

1.4 联合稀疏优化模型

回顾第 1.1–1.3 小节, 稀疏优化模型、组稀疏优化模型和混合稀疏优化模型均基于单一观测值进行数据恢复 (参见文献^[29, 44]), 而在实际应用中, 通常会利用矩阵 $A \in \mathbb{R}^{m \times n}$ 针对多个信号 $x_1, \dots, x_l$ 同时收集测量值 $b_1 = Ax_1, \dots, b_l = Ax_l$, 其中 $x_1, \dots, x_l \in \mathbb{R}^n$ 共享相同的支撑集, 被称为联合稀疏向量. 多测量向量的稀疏优化问题也被称为联合稀疏优化问题, 指利用测量向量的同步性同时恢复 l 个信号 $x_1, \dots, x_l$, 进而提升稀疏恢复能力和估计精度. 联合稀疏优化问题自提出以来, 已被广泛应用于

各类重要问题的建模, 如多传感器信号处理^[34, 37, 110, 145]、多任务机器学习^[4]、多通道信号处理^[49]和分布式压缩感知^[126]等.

联合稀疏优化问题可以表示为 $B = AX$, 其中, $B = (b_1, b_2, \dots, b_l) \in \mathbb{R}^{m \times l}$ 为测量向量, $X = (x_1, x_2, \dots, x_l) \in \mathbb{R}^{n \times l}$. 一般地, 假设矩阵 X 满足 s -行稀疏性, 即 X 的非零行数不超过 s . 类似于第 1.1–1.3 小节, 可利用松弛的思想, 考虑下述联合稀疏松弛模型:

$$\min_{X \in \mathbb{R}^{n \times l}} \{ \|X\|_{p,q}^q : B = AX \text{ 或 } \|B - AX\|_F \leq \varepsilon \}, \quad (1.9a)$$

$$\min_{X \in \mathbb{R}^{n \times l}} \|B - AX\|_F^2 + \lambda \|X\|_{p,q}^q, \quad (1.9b)$$

其中, $\|X\|_F := (\sum_{i,j=1}^{m,n} X_{ij}^2)^{1/2}$ 表示 Frobenius 范数; $\|X\|_{p,q}^q := \sum_{i=1}^m \|X_{i \cdot}\|_p^q$ 表示 X 的混合 $\ell_{p,q}$ 范数 ($p \geq 1, q \geq 0$), 这里 $X_{i \cdot}$ 为 X 的第 i 行且 $\|X_{i \cdot}\|_p = (\sum_{j=1}^n |X_{ij}|^p)^{1/p}$.

目前, 联合稀疏优化相关的理论与算法研究主要集中在凸松弛方面. 理论上, Chen 和 Huo^[34] 基于最小线性相关子集 (spark) 条件在无噪声情形下证明混合 $\ell_{1,1}$ 范数松弛模型 (1.9a) 可精确恢复联合稀疏优化问题的真实解. Tropp 等^[141, 143] 基于相干性条件, 在含噪情形下分别建立了混合 $\ell_{\infty,1}$ 范数松弛模型 (1.9a) 及 (1.9b) 的解与联合稀疏优化问题真实解的误差界估计. 此外, 学者们还基于零空间性质^[145]、限制等距性^[99, 106]、互不相干性^[47]与秩感知^[41]等正则性条件建立了混合 $\ell_{2,1}$ 范数松弛模型 (1.9b) 保障精确恢复的理论. 算法上, 求解联合稀疏优化的方法相继被提出, 包括贪婪算法^[13, 34, 141]、路径跟踪算法^[116]和快速阈值 Landweber 算法^[54]. 研究表明, 相比稀疏优化, 联合稀疏优化能显著减少精确恢复原始信号所需要的样本数.

在非凸松弛方面, Cotter 等^[38]提出了求解混合 $\ell_{2,q}$ 范数松弛模型 (1.9b) 的多焦点欠定系统求解器算法 ($0 \leq q \leq 1$), 并证明了其局部收敛性. Chartrand 和 Wohlberg^[32] 基于广义的阈值压缩算子建立了联合稀疏优化的非凸松弛模型, 并设计了交替方向乘子法对其求解. Ling 等^[98]利用对数和罚函数松弛联合稀疏正则化问题, 并设计了迭代重加权 ℓ_1 与 ℓ_2 范数极小化方法对其求解. 目前, 联合稀疏优化的非凸松弛研究仍然处于起步阶段, 期待未来在这一方面有更多的研究成果.

下文将重点围绕稀疏优化模型和组稀疏优化模型展开综述, 主要介绍相关模型的相合性理论与求解算法. 第 2 节分别总结稀疏优化模型与组稀疏优化模型的相合性理论, 包含还原界理论与精确恢复理论. 第 3 节总结概述结构稀疏优化领域的著名求解算法. 第 4 节是对本文的总结及对未来工作的展望.

2 相合性理论

第 1.1 和 1.2 小节介绍各类松弛模型虽然可用于求解 ℓ_0 和 $\ell_{2,0}$ 范数极小化 (或正则化) 问题, 但松弛模型的科学性即松弛模型何时能寻找到稀疏解, 以及松弛模型的准确性, 即通过松弛模型获得的最优解与真实解之间的误差界均是需要关注的问题. 第 2.1 和 2.2 小节分别介绍稀疏优化模型中经典的正则性条件及其在各类条件下的相合性理论. 第 2.3 和 2.4 小节分别介绍组稀疏优化模型中常见的正则性条件及其在各类条件下的相合性理论.

2.1 稀疏优化模型的正则性条件

针对稀疏优化模型, 正则性条件不仅可用于分析模型的理论性质, 保障模型最优解与真实信号之间的一致性, 而且可用来保障优化算法的收敛性. 目前流行的正则性条件包括互不相干性 (mutual

incoherence property, MIP)^[46]、限制等距性质 (restricted isometry property, RIP)^[29]、稀疏特征值条件 (sparse eigenvalue condition, SEC)^[43]、限制特征值条件 (restricted eigenvalue condition, REC)^[11]、 p -限制特征值条件 (p -restricted eigenvalue condition, p -REC)^[71]、零空间性质 (null space property, NSP)^[58]、值空间性质 (range space property, RSP)^[182] 与限制可逆因子 (restricted invertibility factor, RIF)^[172]. 接下来, 给出具体介绍.

定义 2.1 (MIP^[46]) 设 $A = [A_{\cdot 1}, A_{\cdot 2}, \dots, A_{\cdot n}] \in \mathbb{R}^{m \times n}$ 满足列正交化, 即 $\|A_{\cdot j}\|_2 = 1, \forall j \in [n]$. MIP 条件通常假设 A 的相干性常数 μ 很小, 其中 μ 定义如下:

$$\mu = \max_{i \neq j} |A_{\cdot i}^T A_{\cdot j}|.$$

根据定义可知, 相干性常数易于计算.

定义 2.2 (RIP^[29]) 设 $A \in \mathbb{R}^{m \times n}$, A 的 s -限制等距常数 $\delta_s \in (0, 1)$ 定义为

$$\delta_s := \min\{\delta : (1 - \delta)\|x\|^2 \leq \|Ax\|^2 \leq (1 + \delta)\|x\|^2, \forall x \in \mathbb{R}^n \text{ 且 } \|x\|_0 \leq s\},$$

此时, 称矩阵 A 满足常数为 δ_s 的 s -RIP.

RIP 通常难以计算 (参见文献 [6]), 但它可由相干性常数 μ 得到 (参见文献 [26, 性质 4.1]). 此外, s -RIP 还可等价表示为 (参见文献 [58, 定义 2])

$$\delta_s = \max_{\mathcal{T} \subseteq [n], |\mathcal{T}| \leq s} \|A_{\mathcal{T}}^T A_{\mathcal{T}} - I\|_2^2.$$

特别地, 研究发现 Bernoulli 矩阵、Gauss 矩阵和次 Gauss 矩阵以高概率满足 RIP^[28, 29, 58]. 由于 $A_{\mathcal{T}}^T A_{\mathcal{T}}$ 的期望特征值与单位特征值的不对称偏差, Blanchard 等^[12] 提出了非对称的限制等距性质.

定义 2.3 (非对称 RIP^[12]) 设 $A \in \mathbb{R}^{m \times n}$, 定义 A 的非对称限制等距常数 α_s 与 β_s 分别为

$$\alpha_s := \min_{c \geq 0} \{c : (1 - c)\|x\|_2^2 \leq \|Ax\|_2^2, \forall x \in \mathbb{R}^n \text{ 且 } \|x\|_0 \leq s\},$$

$$\beta_s := \min_{c \geq 0} \{c : \|Ax\|_2^2 \leq (1 + c)\|x\|_2^2, \forall x \in \mathbb{R}^n \text{ 且 } \|x\|_0 \leq s\}.$$

类似地, 非对称的限制等距常数 α_s 与 β_s 可分别等价表示为 (参见文献 [5])

$$\alpha_s = 1 - \min_{\mathcal{T} \subseteq [n], |\mathcal{T}| \leq s} \lambda_{\min}(A_{\mathcal{T}}^T A_{\mathcal{T}}), \quad \beta_s = \max_{\mathcal{T} \subseteq [n], |\mathcal{T}| \leq s} \lambda_{\max}(A_{\mathcal{T}}^T A_{\mathcal{T}}) - 1.$$

定义 2.4 (SEC^[43]) 设 $A \in \mathbb{R}^{m \times n}$, 若 $A^T A$ 的 s -最小稀疏特征值大于 0, 即

$$\phi_{\min}(s) := \min_{\|x\|_0 \leq s} \frac{x^T A^T A x}{x^T x} > 0,$$

则称 A 满足 s -稀疏特征值条件.

如果 $\phi_{\min}(2s) > 0$ 成立, 则线性系统 $Ax = b$ 的 s -稀疏解是唯一的. 如果 $A^T A$ 的 $2s$ -最小稀疏特征值为零 (即 $\phi_{\min}(2s) = 0$), 则难以从含噪的观测值中恢复出真正的 s -稀疏解.

定义 2.5 (REC^[11]) 设 $A \in \mathbb{R}^{m \times n}$, $s \leq t \leq n$, $s + t \leq n$, $\mathcal{I} \subseteq [n]$, $\mathcal{I}(x; t)$ 表示 x 在 $\mathcal{I}^c$ 中绝对值最大的 t 个元素的指标集. 如果

$$\phi_1(s, t) := \min \left\{ \frac{\|Ax\|_2}{\|x_{\mathcal{T}}\|_2} : |\mathcal{I}| \leq s, \|x_{\mathcal{I}^c}\|_1 \leq \|x_{\mathcal{I}}\|_1, \mathcal{T} = \mathcal{I}(x; t) \cup \mathcal{I} \right\} > 0,$$

则称 A 满足关于 (s, t) 的限制特征值条件.

在 SEC 的基础上提出的 REC 可以由 RIP 保证^[11], 故 Bernoulli 矩阵、Gauss 矩阵和次 Gauss 矩阵仍以高概率满足 REC (参见文献 [129]). 而广泛的相关 Gauss 设计矩阵虽满足 REC 却高概率不满足 RIP^[127].

定义 2.6 (p -REC^[71]) 设 $A \in \mathbb{R}^{m \times n}$, $0 \leq p \leq 1$, $s \leq t \leq n$, $s+t \leq n$, $\mathcal{I} \subseteq [n]$, $\mathcal{I}(x; t)$ 表示 x 在 $\mathcal{I}^c$ 中绝对值最大的 t 个元素的指标集, 如果

$$\phi_p(s, t) := \min \left\{ \frac{\|Ax\|_2}{\|x_{\mathcal{T}}\|_2} : |\mathcal{I}| \leq s, \|x_{\mathcal{I}^c}\|_p \leq \|x_{\mathcal{I}}\|_p, \mathcal{T} = \mathcal{I}(x; t) \cup \mathcal{I} \right\} > 0,$$

则称 A 满足关于 (s, t) 的 p -限制特征值条件.

p -REC 刻画了 $A^T A$ 限制在满足 ℓ_p 范数不等式的向量集上的正定性. 此外, p 越小, p -REC 越弱. 特别地,

$$1\text{-REC} \Rightarrow \frac{1}{2}\text{-REC} \Rightarrow 0\text{-REC}.$$

定义 2.7 (NSP^[58]) 设 $\mathcal{N}(A) := \{x \in \mathbb{R}^n : Ax = 0\}$ 为 $A \in \mathbb{R}^{m \times n}$ 的零空间.

(1) 对任意的 $\mathcal{S} \subseteq [n]$ 且 $|\mathcal{S}| \leq s$, 如果对任意的 $v \in \mathcal{N}(A) \setminus \{0\}$, 都有 $\|x_{\mathcal{S}}\|_1 < \rho \|x_{\mathcal{S}^c}\|_1$ 成立, 则当 $\rho = 1$ 时, 称 A 满足 s -零空间性质; 当 $\rho \in (0, 1)$ 时, 称 A 满足 s -稳定的零空间性质.

(2) 对任意的 $\mathcal{S} \subseteq [n]$ 且 $|\mathcal{S}| \leq s$, 如果对任意的 $v \in \mathbb{R}^n$, 都有 $\|x_{\mathcal{S}}\|_1 < \rho \|x_{\mathcal{S}^c}\|_1 + \tau \|Ax\|_2$ 成立, 其中 $\rho \in (0, 1)$, $\tau > 0$, 则称 A 满足 s -鲁棒零空间性质.

结合 NSP 的定义, 会发现 NSP 严格弱于 RIP, 详见文献 [24, 56]. 特别地, 当定义 2.7(1) 中的 ℓ_1 范数为 ℓ_p 范数时, 相应的正则性条件被称为 p -NSP^[179].

定义 2.8 (RSP^[182]) 如果对任意不相交的指标集 $\mathcal{I}, \mathcal{J} \subseteq [n]$ 且 $|\mathcal{I}| + |\mathcal{J}| \leq s$, 存在 $\eta \in \mathcal{R}(A^T)$ 使得

$$\eta_i = 1, \quad \forall i \in \mathcal{I}; \quad \eta_i = -1, \quad \forall i \in \mathcal{J}; \quad |\eta_i| < 1, \quad \forall i \notin \mathcal{I} \cup \mathcal{J}$$

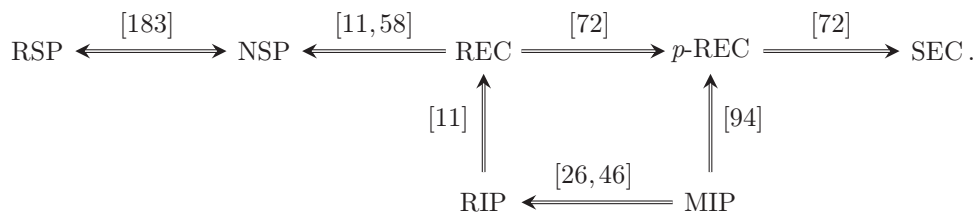
成立, 则称 A^T 满足 s -RSP, 其中 $\mathcal{R}(A^T)$ 为 A^T 的像空间.

注意到, s -RSP 等价于 s -NSP^[182]. 在 RSP 的基础上, 文献 [183, 185, 188] 还提出了弱 RSP 与限制弱 RSP, 它们均弱于 RIP 与 NSP.

定义 2.9 (RIF^[172]) 设 $q \geq 1$, $\xi > 0$, $\mathcal{S} \subseteq [n]$, 限制可逆因子定义如下:

$$\text{RIF}_q(\xi, \mathcal{S}) = \inf \left\{ \frac{|\mathcal{S}|^{1/q} \|A^T A u\|_{\infty}}{m \|u\|_q} : \Phi(u_{\mathcal{S}^c}) \leq \xi \Phi(u_{\mathcal{S}}) \right\}.$$

以上为一些常用的正则性条件, 它们之间的关系如下所示:



2.2 稀疏优化模型的相合性理论

基于第 2.1 小节的各类正则性条件, 可以针对不同的松弛模型阐述相应的相合性理论. 为方便起见, 不妨设 $\bar{x}$ 满足 $b = A\bar{x} + e$ 且 $\mathcal{S} = \text{supp}(\bar{x})$, x_0^* 和 $x_{0,\varepsilon}^*$ 分别为 (1.2) 在无噪和含噪情形下的最优解, x_Φ^* 和 $x_{\Phi,\varepsilon}^*$ 分别为 (1.4a) 在无噪和含噪情形下的全局最优解, $x_{0,\lambda}^*$ 和 $x_{\Phi,\lambda}^*$ 分别为 (1.3) 和 (1.4b) 的全局最优解. 特别地, 在含噪情形下, 当 $\Phi(x) = \|x\|_1$ 时, 相应的松弛问题 (1.4a) 与 (1.4b) 的全局最优解分别记为 $x_{1,\varepsilon}^*$ 和 $x_{1,\lambda}^*$. 在不引起混淆的情形下, 其他的罚函数类似处理. 另外, 记 $\sigma_s(\bar{x})_p := \inf_{\|z\|_0 \leq s} \|\bar{x} - z\|_p$ 为 $\bar{x}$ 在 ℓ_p 范数下的最佳 s -稀疏逼近.

(I) 凸松弛情形. 基于 MIP 条件, Donoho 等^[45] 证明了当 $\|\bar{x}\|_0 \leq (1/\mu + 1)/4$ 时, ℓ_1 范数极小化问题具有如下的精确恢复结果:

$$\|x_{1,\varepsilon}^* - \bar{x}\|_2 \leq \frac{4\varepsilon^2}{\sqrt{1 - \mu(4\|\bar{x}\|_0 - 1)}}. \quad (2.1)$$

基于 RIP 条件, Candès 和 Tao^[29] 在无噪情形下, 建立了 ℓ_1 范数极小化问题与 ℓ_0 范数极小化问题的等价性, 证明了当 $\delta_s + \delta_{2s} + \delta_{3s} < 1$ 时, $\bar{x}$ 为 ℓ_1 范数极小化问题的唯一最优解; 在含噪情形下, 证明当 $\delta_{3s} + 3\delta_{4s} < 2$ 时, 有

$$\|x_{1,\varepsilon}^* - \bar{x}\|_2 \leq O(\varepsilon).$$

Cai 和 Zhang^[27] 证明了当 $\delta_{ts} < \sqrt{(t-1)/t}$ ($t \geq 4/3$) 时, 求解 ℓ_1 范数极小化问题可以精确恢复所有 s -稀疏信号. 特别地, 当 $\delta_{2s} < 1/\sqrt{2}$ 时, $x_{1,\varepsilon}^*$ 满足

$$\|x_{1,\varepsilon}^* - \bar{x}\|_2 \leq O(\varepsilon) + O\left(\frac{\sigma_s(\bar{x})_1}{\sqrt{s}}\right).$$

基于 REC 条件, Bickel 等^[11] 分别就 LASSO 和 Dantzig 估计, 证明了

$$\|x_{1,\lambda}^* - \bar{x}\|_2^2 \leq O(s\lambda^2)$$

以不低于 $1 - n^{1-\zeta^2/8}$ 的概率成立 ($\zeta > 2\sqrt{2}$). van de Geer 和 Bühlmann^[144] 针对 ℓ_1 范数正则化问题证明了

$$\begin{aligned} \|Ax_{1,\lambda}^* - b\|_2^2 + \lambda\|(x_{1,\lambda}^*)_{\mathcal{S}^c}\|_1 &\leq s\lambda^2\phi_1^2(s, s), \\ \|x_{1,\lambda}^* - \bar{x}\|_2^2 &\leq 2s\lambda^2\phi_1^{-4}(s, s). \end{aligned}$$

(II) 非凸松弛情形. 基于 MIP 条件, Dong 等^[42] 针对 ℓ_p 范数极小化问题, 证明了当 $\|\bar{x}\|_0 \leq \gamma^{1/p-2}(\mu+1)/(4^{1/p}\mu)$ 时, $x_{p,\varepsilon}^*$ 满足

$$\|x_{p,\varepsilon}^* - \bar{x}\|_2 \leq \frac{2\varepsilon}{\sqrt{1 - \mu(\frac{4^{1/p}}{\gamma^{2/p-2}}\|\bar{x}\|_0 - 1)}}, \quad (2.2)$$

其中 $\gamma \in [1/\|\bar{x}\|_0, n/\|\bar{x}\|_0]$. 当 $\gamma > 2$ 时, (2.2) 中还原界的上界比 (2.1) 中的界更紧, 这意味着 ℓ_p 范数极小化模型比 LASSO 模型的预测准确性更高, 且对真实信号的稀疏度要求更低.

基于 RIP 条件, Song 和 Xia^[135] 证明了当 $\delta_{(tp+1)s} < 1/\sqrt{tp-2+1}$ ($t > 0$) 时, ℓ_p 范数极小化问题在无噪声情形下可以精确恢复 s -稀疏信号; 并在噪声水平满足 $\|e\|_2 \leq \varepsilon$ 或 $\|A^T e\|_\infty \leq \varepsilon$ 时, 分别建立了

ℓ_p 范数极小化恢复稀疏信号的还原界. 另外, 当 $\delta_{2s} \in [\frac{\sqrt{2}}{2}, 1)$ 且 p 在一定阈值范围内时, Wen 等^[159] 证明了

$$\|x_{p,\varepsilon}^* - \bar{x}\|_2^p \leq O(s^{\frac{p}{2}-1} \sigma_s(\bar{x})_p^p) + O(\varepsilon^p),$$

特别地, 当 $\varepsilon = 0$ 且 $\bar{x}$ 恰好为 s - 稀疏信号时, ℓ_p 范数极小化问题可实现精确恢复.

基于非对称的 RIP 条件, Foucart 和 Lai^[57] 证明了如果对任意的整数 $t \geq s$, $\gamma_{2t} := \beta_{2t}^2 / \alpha_{2t}^2$ 满足 $\gamma_{2t} - 1 < 4(\sqrt{2} - 1)(t/s)^{1/p-1/2}$, 则 ℓ_p 范数极小化在无噪情形下可以精确恢复 s - 稀疏向量, 且在含噪情形下, 有

$$\|x_{p,\varepsilon}^* - \bar{x}\|_2 \leq O\left(\frac{\sigma_s(\bar{x})_p}{t^{\frac{1}{p}-\frac{1}{2}}}\right) + O\left(\frac{\varepsilon}{\beta_{2s}}\right).$$

基于 p -REC 条件, Hu 等^[71] 针对 ℓ_p 范数正则化问题, 证明了 $x_{p,\lambda}^*$ 满足

$$\begin{aligned} \|Ax_{p,\lambda}^* - b\|_2^2 + \lambda \|(x_{p,\lambda}^*)_{S^c}\|_1 &\leq s\lambda^{\frac{2}{2-p}} \phi_p^{\frac{2-p}{2p}}(s, s), \\ \|x_{p,\lambda}^* - \bar{x}\|_2^2 &\leq 2s\lambda^{\frac{2}{2-p}} \phi_p^{\frac{2-p}{4}}(s, s). \end{aligned}$$

基于 RIF 正则性条件, Zhang 和 Zhang^[172] 针对一般的稀疏正则化优化模型 (1.4b), 证明了 x_Φ^* 满足

$$\|x_\Phi^* - \bar{x}\|_2 \leq (1 + \eta)\lambda^* |\mathcal{S}|^{\frac{1}{2}} \text{RIF}_q(\xi, \mathcal{S}),$$

其中, $\eta \in (0, 1)$, $\xi = (\eta + 1)/(1 - \eta)$, $\lambda^* := \inf_{t>0} \{t/2 + \rho(t; \lambda)/t\}$ 为罚函数的阈值水平, 这里 Φ 包括 ℓ_p 范数、SCAD、MCP 和盖型 ℓ_1 范数罚函数.

2.3 组稀疏优化模型的正则性条件

在求解组稀疏优化问题时, 通常考虑相应的组罚函数松弛模型, 但罚函数 (或正则项) 的选取会影响松弛模型的性能, 进而导致与原模型的真实解产生偏差. 因此估计松弛模型的解与真实解之间的误差界至关重要, 这能保证松弛模型的准确性. 鉴于组稀疏优化模型与稀疏优化模型的密切联系, 稀疏优化模型的正则性条件大多可延伸至组稀疏优化, 如组相干性 (group coherent, GC)^[47]、组限制等距条件 (group RIP, GRIP)^[48]、组零空间性质 (group NSP, GNSP)^[61, 85, 137] 与组限制特征值条件 (group REC, GREC)^[71]. 下面回顾这些正则性条件.

定义 2.10 (GC^[47]) 对于分组 $\{\mathcal{G}_i\}_{i=1}^r$, 其每组维数大小相同, 即 $n_i = d, \forall i \in [r]$, 定义 $A \in \mathbb{R}^{m \times n}$ 的组相干性常数 μ_G 与次相干性常数 μ_S 分别为

$$\mu_G = \max_{i \neq j} \frac{1}{d} \|A_{\mathcal{G}_i}^* A_{\mathcal{G}_j}\|_2, \quad \mu_S = \max_{1 \leq l \leq r} \max_{(l-1)d \leq i \neq j \leq ld} |A_{\cdot i}^* A_{\cdot j}|,$$

其中 A^* 表示 A 的伴随矩阵.

当 $d = 1$ 时, 组相干性常数 μ_G 恰好退化为稀疏优化的相干性常数 μ (定义 2.1). 此外, 基于 MIP 的另一种组相干性常数的推广形式可参见文献 [81], 其上界可由相干性常数 μ 控制.

定义 2.11 (GRIP^[48]) 设 $A \in \mathbb{R}^{m \times n}$, $S \in \mathbb{N}$ 满足 $S \leq n$, 定义 A 在分组结构 $G = \{\mathcal{G}_1, \dots, \mathcal{G}_r\}$ 上的组限制等距性常数 $\Delta_S \in (0, 1)$ 为

$$\Delta_S := \min\{\Delta : (1 - \Delta)\|x\|^2 \leq \|Ax\|^2 \leq (1 + \Delta)\|x\|^2, \forall x \in \mathbb{R}^n \text{ 满足 } \|x\|_{2,0} \leq S\}.$$

容易发现, GRIP 是 RIP 的延伸, 但它比 RIP 更弱一些 (参见文献 [7, 48]).

定义 2.12 (GREC^[71]) 设 $A \in \mathbb{R}^{m \times n}$, $0 \leq q \leq p \leq 2$, 称 A 满足关于 (S, N) 的 (p, q) -组限制特征值条件, 如果

$$\phi_{p,q}(S, N) := \min \left\{ \frac{\|Ax\|_2}{\|x_{\mathcal{G}_N}\|_{p,2}} : |\mathcal{J}| \leq S, \|x_{\mathcal{G}_{\mathcal{J}^c}}\|_{p,q} \leq \|x_{\mathcal{G}_{\mathcal{J}}}\|_{p,q}, \mathcal{N} = \mathcal{J}(x; N) \cup \mathcal{J} \right\} > 0,$$

其中, $S \leq N \ll r$, $x \in \mathbb{R}^n$, $\mathcal{J} \subseteq [r]$, $\mathcal{J}(x; N)$ 表示 $\|x_{\mathcal{G}_j}\|_p$ 在 $\{\|x_{\mathcal{G}_j}\|_p : j \in \mathcal{J}^c\}$ 上最大的 N 个组所对应的指标集.

由定义可知, (p, q) -GREC 是 p -REC 的延伸, 它比 p -REC 更弱. 因为相比 p -REC, (p, q) -GREC 将每个组视为一个元素, 在相关约束上的自由度约为 $s/n_{\max}$, 这显然比 s 小.

定义 2.13 设 $\mathcal{N}(A) := \{x \in \mathbb{R}^n : Ax = 0\}$ 为 $A \in \mathbb{R}^{m \times n}$ 的零空间.

(1) ($\ell_{2,1}$ -GNSP^[137]) 对任意的 $\mathcal{J} \subseteq [r]$ 且 $|\mathcal{J}| \leq S$, 如果对任意的 $x \in \mathcal{N}(A) \setminus \{0\}$, 都有 $\|x_{\mathcal{J}}\|_{2,1} \leq \|x_{\mathcal{J}^c}\|_{2,1}$ 成立, 则称 A 满足 $\ell_{2,1}$ -组零空间性质;

(2) ($\ell_{2,1}$ -鲁棒 GNSP^[61, 85]) 对任意的 $\mathcal{J} \subseteq [r]$ 且 $|\mathcal{J}| \leq S$, 如果存在 $\rho \in (0, 1)$, $\tau > 0$, 使得 $\|x_{\mathcal{J}}\|_2 \leq \frac{\rho}{\sqrt{S}} \|x_{\mathcal{J}^c}\|_{2,1} + \tau \|Ax\|_2$ 对任意的 $x \in \mathbb{R}^n$ 成立, 则称 A 满足 $\ell_{2,1}$ -鲁棒组零空间性质.

2.4 组稀疏优化模型的相合性理论

类似地, 基于第 2.3 小节的正则性条件, 学者们研究了各类组稀疏优化松弛模型的相合性理论. 在不引起混淆的情形下, 与第 2.2 小节类似, 当 $\Phi(\|x_{\mathcal{G}_i}\|_p) = \|x\|_{p,q}$ 时, 记 $\ell_{p,q}$ 范数极小化问题 (1.6a) 与正则化问题 (1.6b) 在含噪情形下的最优解分别为 $x_{p,q,\varepsilon}^*$ 与 $x_{p,q,\lambda}^*$. 另外, 记 $\sigma_S(\bar{x})_q := \inf_{\|z\|_{2,0} \leq S} \|\bar{x} - z\|_{2,q}$ 为 $\bar{x}$ 在 $\ell_{2,q}$ 范数下的最佳 S 组稀疏逼近.

(I) 凸松弛情形. 基于 GRIP 条件, Eldar 和 Mishali^[48] 证明了当 $\Delta_{2S} < 0.414$ 时, $\ell_{2,1}$ 范数极小化问题可以精确地恢复任意 S -组稀疏信号, 且

$$\|x_{2,1,\varepsilon}^* - \bar{x}\|_2 \leq O\left(\frac{\sigma_S(\bar{x})_1}{\sqrt{S}}\right) + O(\varepsilon). \quad (2.3)$$

Lin 和 Li^[97] 将 GRIP 的条件改进为 $\Delta_{2S} < 0.4931$, 并给出了精确恢复的另一个充分条件 $\Delta_S < 0.307$. Gao 和 Ma^[61] 进一步将相合性理论 (2.3) 成立的 GRIP 条件改进到 $\Delta_{2S} < 0.6426$. 此外, 基于 $\ell_{2,1}$ 鲁棒 GNSP 条件, Gao 和 Ma^[61] 及 Koep 等^[85] 针对 $\ell_{2,1}$ 范数极小化问题, 也建立了类似于 (2.3) 的还原界.

(II) 非凸松弛情形. 基于 GRIP 条件, Wang 等^[156] 证明了当 $\Delta_{2S} < 1/2$ 时, 存在与 Δ_{2S} 相关的常数 q_0 , 对任意的 $q < q_0$, $\ell_{2,q}$ ($0 < q \leq 1$) 范数极小化问题可以稳健地恢复组稀疏信号, 即

$$\|x_{2,q,\varepsilon}^* - \bar{x}\|_2 \leq O\left(\frac{\|\bar{x}_{T_0^c}\|_{2,q}}{S^{\frac{1}{q}-\frac{1}{2}}}\right) + O(\varepsilon),$$

其中 T_0 是 $\bar{x}$ 关于 ℓ_2 范数最大的 S 个分组所对应的指标集.

基于 GRIP 条件, Feng 等^[53] 在含噪情形下, 建立了 $\ell_{2,q}$ ($0 < q < 1$) 范数正则化问题的相合性理论, 即当 $\Delta_{2S} \in (0, 1)$ 且 (q, Δ_{2S}) 满足一定条件时, 有

$$\|x_{2,q,\lambda}^* - \bar{x}\|_{2,q}^2 \leq O\left(\frac{\varepsilon^2}{\lambda}\right) + O(\sigma_S(\bar{x})_q^q) + O(S^{1-\frac{q}{2}}\varepsilon^q) + O(S^{1-\frac{q}{2}}). \quad (2.4)$$

ε 和 S 越小, (2.4)中的还原界估计越紧. 此外, 当 q 接近 1 时, $\ell_{2,q}$ 范数正则化问题的还原界接近组 LASSO 在 GRIP 下建立的还原界 $O(\lambda^2)$ (参见文献 [107]).

基于 (p, q) -GREC 条件, Hu 等^[71] 建立了 $\ell_{p,q}$ ($0 < q < 1 \leq p$) 范数正则化问题的全局还原界:

$$\|x_{p,q,\lambda}^* - \bar{x}\|_2^2 \leq 2\lambda^{\frac{2}{2-q}} S \phi_{p,q}^{\frac{2-q}{4}}(S, S).$$

当 λ 足够小时, 真实稀疏解 $\bar{x}$ 都可以由 $x_{p,q,\lambda}^*$ 精确恢复.

第 2.2 小节及上述内容均为基于各类正则性条件针对稀疏优化与组稀疏松弛优化模型建立的相合性理论. 事实上, 在含噪情形下, 若直接求解 ℓ_0 范数极小化问题 (1.2) 与 $\ell_{2,0}$ 范数约束优化问题

$$\min_{x \in \mathbb{R}^n} \{\|Ax - b\|_2^2 : \|x\|_{2,0} \leq S\}, \quad (2.5)$$

则也存在相应的相合性理论. 例如, 当 $\|x\|_0 \leq (1 + 1/\mu)/2$ 时, Donoho 等^[45] 证明了 (1.2) 可稳定地恢复原始信号, 且 $x_{0,\varepsilon}^*$ 满足

$$\|x_{0,\varepsilon}^* - \bar{x}\|_2 \leq \frac{4\varepsilon^2}{\sqrt{1 - \mu(2\|\bar{x}\|_0 - 1)}},$$

当噪声向量 e 中的每个元素独立服从于 Gauss 分布 $N(0, \sigma^2)$ 时, Hazimeh 等^[67] 证明了 (2.5) 的最优解 $x_{2,0,S}^*$ 以不低于 $1 - (S/r)^S$ 的概率满足

$$\|x_{2,0,S}^* - \bar{x}\|_{2,1} \leq \sigma S \left[\frac{\bar{T}_S + \log(r/S)}{m} \right]^{1/2} [\gamma_{2S}]^{-1},$$

其中, $\bar{T}_S = \max_{|J| \leq S} \sum_{i \in J} T_i / S$, T_i 为每个分组 $i \in [r]$ 中的特征数, $\gamma_S := \min_{|\text{Gsupp}(x)| \leq S} \frac{\sqrt{S}\|Ax\|_2}{\sqrt{m}\|x\|_{2,1}}$.

3 求解算法

鉴于结构稀疏优化在业界和工程上的显著应用, 学者们相继提出了诸多求解结构稀疏优化模型的高效算法. 接下来, 本文将根据不同的问题模型分别从松弛 ℓ_0 范数与直接处理 ℓ_0 范数两个角度, 将求解算法分为松弛类算法与 ℓ_0 范数极小化算法两大类, 再分别介绍各自的一阶与二阶求解算法.

3.1 松弛类算法

本小节主要介绍求解松弛问题 (1.4a) 和 (1.4b) 的一阶与二阶求解算法.

3.1.1 一阶松弛算法

(I) 迭代收缩阈值算法 (iterative shrinkage thresholding algorithm, ISTA)^[39] 是一类非常受关注的梯度类算法, 也称为软阈值算法. 求解松弛问题 (1.4b) 的 ISTA 迭代过程为

$$x^{k+1} = \mathbb{S}_{\lambda v}(x^k - vA^T(Ax^k - b)), \quad (3.1)$$

其中, $v > 0$ 是步长, 一般通过回溯步长规则获取; $\mathbb{S}_\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 为软阈值收缩算子, 对给定的 $x \in \mathbb{R}^n$, $\mathbb{S}_\alpha(x)$ 的第 i 个分量为

$$\mathbb{S}_\alpha(x)_i := (|x_i| - \alpha)_+ \text{sgn}(x_i), \quad i \in [n],$$

这里 $t_+ := \max(0, t)$, $\text{sgn}(\cdot)$ 是符号函数. S_α 将比 α 小的分量值收缩为 0, 将比 α 大的分量值变小. 从 (3.1) 可以看出, 在每步迭代中, ISTA 先对数据拟和项作梯度下降, 再通过软阈值收缩算子更新 x^{k+1} , 算法结构简单且易于实现. 记 (1.4b) 的目标函数为 $F(x) := f(x) + \lambda\|x\|_1$, 其中 $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, 文献 [8, 定理 3.1] 证明了由 ISTA 生成序列 $\{x^k\}$ 的收敛速度为 $O(1/k)$, 满足

$$F(x^k) - F(x^*) \leq \frac{\alpha L(f)}{2k} \|x^0 - x^*\|_2^2,$$

其中, x^* 为 (1.4b) 的最优解, $L(f)$ 表示 ∇f 的 Lipschitz 常数. 基于 ISTA 还衍生了许多高效的加速算法, 例如, Beck 和 Teboulle^[8] 采用 Nesterov 加速技巧, 提出了复杂度为 $O(1/k^2)$ 的快速迭代收缩阈值算法 (fast iterative shrinkage thresholding algorithm, FISTA); Jiao 等^[82] 采用延拓技巧, 提出了带有延拓的迭代收缩阈值算法, 并基于 MIP 条件给出了算法的误差界估计.

注意到, ISTA 其实是邻近梯度算法 (proximal gradient algorithm, PGA) 应用于 LASSO 模型的一个特例. 对于复合优化问题

$$\min_{x \in \mathbb{R}^n} F(x) := f(x) + g(x), \quad (3.2)$$

其中 f 为损失函数, g 可能是凸函数也可能非凸, 求解它的 PGA 迭代格式为

$$x^{k+1} = \text{Prox}_{vg}(x^k - v\nabla f(x^k)),$$

其中 $\text{Prox}_{vg} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 为邻近算子, 定义如下:

$$\text{Prox}_{vg}(x) := \underset{y \in \mathbb{R}^n}{\text{argmin}} g(y) + \frac{1}{2v} \|y - x\|_2^2, \quad \forall x \in \mathbb{R}^n.$$

当 $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ 且 $g(x) = \lambda\|x\|_1$ 时, PGA 即退化为 ISTA. 特别地, 不动点连续 (fixed point continuation, FPC) 算法^[66] 实际上也是一类邻近梯度方法. 当 $g(x) = \lambda\|x\|_p^p$ 时, 文献 [63, 173, 192] 基于邻近算子设计了相应的 PGA, 并借助 Kurdyka-Łojasiewicz (KL) 框架^[2, 3] 给出了算法的线性或次线性收敛速度. 另外, PGA 也适用于组稀疏优化问题. 当 $g(x) = \lambda\|x\|_{p,q}^q$ 时, Hu 等^[71] 应用 PGA 求解 (3.2), 并基于 KL 框架建立了 PGA 的全局收敛性与局部线性收敛速度.

(II) 交替方向乘子法 (alternating direction method of multipliers, ADMM) 作为一类广泛用于求解大规模凸优化问题的算法, 已被用于求解结构稀疏优化问题 (参见文献 [21, 150, 164]). 对于约束优化问题

$$\min_{x \in \mathbb{R}^n, y \in \mathbb{R}^n} \{f(x) + g(y) : Ax + By = b\}, \quad (3.3)$$

其增广 Lagrange 函数为

$$\mathcal{L}(x, y, w) = f(x) + g(y) - w^T(Ax + By - b) + \frac{\beta}{2} \|Ax + By - b\|_2^2,$$

其中, w 是乘子向量, $\beta > 0$ 是罚参数. ADMM 通常将优化问题 (3.3) 分解为两个子问题来交替求解, 通过固定某个变量来更新另一个变量, 即分别对 y 和 x 交替求极小, 其数学形式如下:

$$\begin{aligned} y^{k+1} &\in \underset{y}{\text{argmin}} \mathcal{L}(x^k, y, w^k), \\ x^{k+1} &\in \underset{x}{\text{argmin}} \mathcal{L}(x, y^{k+1}, w^k), \\ w^{k+1} &= w^k - \gamma\beta(Ax^{k+1} + By^{k+1} - b). \end{aligned}$$

当 $f(x) = \frac{1}{2\lambda} \|Ax - b\|_2^2$ 且 $g(x) = \|x\|_1$ 时, (3.2) 即为 (1.4b), 此时引入辅助变量 $y = Ax - b$, (3.2) 可等价于 (3.3). 求解 (1.4b) 的线性 ADMM 迭代策略由如下梯度下降步、软阈值操作和乘子更新规则构成:

$$y^{k+1} = \frac{\lambda\beta}{1 + \lambda\beta} \left(\frac{w^k}{\beta} - (Ax^k - b) \right), \quad (3.4a)$$

$$x^{k+1} = \mathbb{S}_{\frac{\tau}{\beta}} \left(x^k - \tau A^T \left(Ax^k + y^{k+1} - b - \frac{w^k}{\beta} \right) \right), \quad (3.4b)$$

$$w^{k+1} = w^k - \gamma\beta(Ax^{k+1} + y^{k+1} - b). \quad (3.4c)$$

Yang 和 Zhang^[164] 证明了由线性 ADMM (3.4a)–(3.4c) 生成的序列 $\{y^k, x^k, w^k\}$ 收敛到 (y^*, x^*, w^*) , 其中 x^* 为 (1.4b) 的解. Wang 等^[149] 在研究分布式机器学习任务时, 提出用组 ADMM 求解优化问题, 并证明了其收敛速度为 $O(1/k)$. 当噪声分布不均匀或厚尾时, 最小绝对偏差方法在线性回归模型中具有更高效稳健的表现. 特别地, 当 $f(x) = \|Ax - b\|_1$ 且 $g(x) = \lambda_1 \|x\|_1 + \lambda_2 \|x\|_{2,1}$ 时, Wang 等^[150] 采用了线性化的 ADMM 求解混合稀疏优化问题 (3.2), 并证明了对任意给定的初始点算法均能收敛到 (1.8) 的解. 另外, ADMM 作为增广 Lagrange 方法的变体, 当目标函数非凸非光滑时, Wang 等^[157] 建立了其全局收敛性; 当目标函数非凸时, Jia 等^[79] 基于误差界条件, 建立了其局部线性收敛性.

(III) 块坐标下降 (block coordinate descent, BCD) 方法^[22, 123, 162] 是一类求解具有多块变量优化问题的高效方法, 被广泛用于处理大规模问题. 若将变量 x 分成 r 块 $x_{\mathcal{G}_1}, \dots, x_{\mathcal{G}_r}$, (3.2) 可记为

$$\min_{x \in \mathbb{R}^n} f(x_{\mathcal{G}_1}, \dots, x_{\mathcal{G}_r}) + \sum_{i=1}^r g_i(x_{\mathcal{G}_i}), \quad (3.5)$$

BCD 方法在每一步迭代时, 逐次更新所有变量块 $x_{\mathcal{G}_i}$. 具体地, 在第 k 步迭代中, BCD 方法的第 i 块更新方式如下:

$$x_{\mathcal{G}_i}^k \in \operatorname{argmin}_{x_{\mathcal{G}_i}} f(x_{\mathcal{G}_1}^k, \dots, x_{\mathcal{G}_{i-1}}^k, x_{\mathcal{G}_i}, x_{\mathcal{G}_{i+1}}^{k-1}, \dots, x_{\mathcal{G}_n}^{k-1}) + g_i(x_{\mathcal{G}_i}).$$

BCD 方法已被用于求解结构稀疏优化问题, 例如, 在 $f(x) = \frac{1}{2} \|Ax - b\|_2^2$ 情形下, 若 $g(x) = \lambda \|x\|_1$, 求解 ℓ_1 范数正则化模型 (1.4b) 的 BCD 方法为

$$x_i^k \in \operatorname{argmin}_{x_i} \lambda |x_i| + \frac{1}{2} \|A_i\|_2^2 x_i^2 - A_i^T (b - \bar{A}_i \bar{x}_i) x_i, \quad \forall i \in [r],$$

其中, $\bar{A}_i$ 为 A 除去第 i 列后构成的矩阵, $\bar{x}_i$ 为 x 除去第 i 个分量后构成的向量. 当 $g(x) = \lambda \|x\|_{2,1}$ 时, Qin 等^[123] 应用 BCD 方法求解组 LASSO 并建立了全局收敛性; 当 g 为 SCAD 或 MCP 罚函数时, Breheny 和 Huang^[22] 采用 BCD 方法求解相应的 SCAD 和 MCP 正则化松弛模型并建立了相应的收敛性, 证明了由 BCD 方法生成的序列收敛到既是局部极小值也是全局坐标方向极小值点. 另外, Xu 和 Yin^[162] 应用 BCD 方法求解一般的非凸正则化问题, 并基于 KL 理论证明了 BCD 方法全局收敛到问题的临界点, 且建立了其渐近收敛性.

(IV) 凸差 (difference of convex functions, DC) 算法是由 Le Thi 和 Tao^[90]、Pham Dinh 和 Le Thi^[119] 以及 Le Thi 等^[89] 提出的一种基于局部最优性条件和 DC 对偶理论的优化算法, 多用于求解无约束的 DC 规划问题. 以 (3.2) 为例, 若其目标函数 $F(x) = f(x) + g(x)$ 可表示为 2 个凸函数 F_1 与 F_2 之差, 即 $F(x) = F_1(x) - F_2(x)$, 则称 F 为 DC 函数. 求解 (3.2) 的 DC 算法具体为: 在第 k 步迭代中, 结合 DC 规划的局部最优性条件与对偶理论, 用 F_2 的一阶逼近来近似 F_2 , 再对近似后的凸优化

问题进行求解, 迭代如下所示:

$$\begin{aligned} y^k &\in \partial F_2(x^k), \\ x^{k+1} &\in \operatorname{argmin}_{x \in \mathbb{R}^n} \{F_1(x) - (F_2(x^k) + \langle x - x^k, y^k \rangle)\}. \end{aligned}$$

注意到, DC 算法是一种无需线搜索但具有全局收敛性的下降方法, 并且对于一般的 DC 规划, DC 算法具有线性收敛性; 对于多面体 DC 规划, 则具备有限步收敛性 (参见文献 [89, 119]).

目前, DC 算法已被用于结构稀疏优化问题. 当 $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ 且 $g(x) = \lambda(\|x\|_1 - \|x\|_2)$ 时, Yin 等^[167]应用 DC 算法求解压缩感知中的 $\ell_1 - \ell_2$ 范数正则化问题, 并建立了算法的全局收敛性. 当 $g(x) = \lambda\|x\|_0$ 时, Le Thi 等^[89]针对 ℓ_0 范数正则化问题 (3.2), 采用 DC 近似方法, 在一定条件下证明了近似问题与原问题的等价性, 及其全局 (或局部) 最优解之间的一致性. 当 $g(x) = \lambda_1\|x\|_0 + \lambda_2\|x\|_{2,0}$ 时, Li 等^[93]给出了混合稀疏正则化问题 (3.2) 的 DC 松弛模型, 并建立了松弛问题与原问题在全局最优解下的等价性. Gotoh 等^[64]针对 (3.9a), 给出了稀疏约束 $\{\|x\|_0 \leq s\}$ 的精确 DC 表示, 并采用 DC 算法求解.

(V) 迭代重加权类算法 (iteratively reweighted algorithms) 是一类适用于求解优化问题 (3.2) 的高效算法, 一般用于处理 g 非凸的情形, 下面以 $g(x) = \lambda\|x\|_p^p$ 为例来展开介绍 ($0 < p < 1$).

迭代重加权 ℓ_1 范数极小化 (iteratively reweighted ℓ_1 minimization, IRL1) 算法^[30, 35, 36]利用重新加权的思想, 动态保证对向量中较大分量的惩罚力度小于较小分量, 从而有效避免 ℓ_1 范数的过度惩罚性, 更科学地惩罚迭代向量中的元素值. 注意到, 加权方式不同, IRL1 的迭代更新策略也有所不同. 对 $g(x)$ 作一阶光滑化处理, 有

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \sum_{i=1}^n (|x_i| + \epsilon)^p, \quad (3.6)$$

其中 $\epsilon > 0$. 求解 (3.2) 的 IRL1 的一种迭代策略如下所述 (参见文献 [36]):

$$x^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^n} \{f(x) + \lambda\|W^k x\|_1\}, \quad (3.7)$$

其中权重矩阵 $W^k = \operatorname{diag}(w_1^k, \dots, w_n^k)$ 且 $w_i^k = p/(|x_i^k| + \epsilon)^{1-p}$ ($i = 1, \dots, n$).

注意到, 权重 w_i^k 可以较好地连续估计非零分量的位置. 最大的非零分量通常最有可能被识别为非零项, 一旦这些位置被识别, 相应的权重就会降低, 减少其惩罚力度, 进而更灵敏更准确地识别出剩余较小但非零的分量位置. 在 IRL1 的每次迭代中, (3.7) 实际上都在求解一个凸的加权 ℓ_1 范数正则化问题.

文献 [36, 定理 1 和 4] 证明了由 IRL1 生成的序列 $\{x^k\}$ 是有界的, 且该序列的任意聚点都是优化问题 (3.6) 的稳定点. 在特定区域内, (3.6) 的稳定点是其全局最小值解, 且在一定条件下, IRL1 具有近似线性收敛性.

迭代重加权最小二乘 (iteratively reweighted least squares, IRLS) 算法^[40, 87]也是一种经典的迭代重加权类算法. 在 $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ 情形下, 对 $g(x)$ 作二阶光滑 $g(x) \approx \sum_{i=1}^n (x_i^2 + \epsilon^2)^{\frac{p}{2}}$, Lai 等^[87]提出了 IRLS 算法, 其主要迭代过程如下:

$$\begin{aligned} x^{k+1} &= \left\{ x \in \mathbb{R}^n : \lambda \left[\frac{px_i}{(\epsilon_k^2 + (x_i^k)^2)^{1-p/2}} \right]_{1 \leq i \leq n} + A^T(Ax - b) = 0 \right\}, \\ \epsilon_{k+1} &= \min\{\epsilon_k, \lambda r(x^{k+1})_{s+1}\}, \end{aligned}$$

其中, s 为预估计的稀疏水平, $r(z)$ 表示 $z \in \mathbb{R}^n$ 的分量按其绝对值降序排列后构成的向量. 基于 RIP 条件, 文献 [87, 定理 2.2] 证明了当 $\delta_{2s} < 1$ 且 $\epsilon_k \rightarrow \epsilon_*$ ($k \rightarrow \infty$) 时, 由 IRLS 生成的序列 $\{x^k\}$ 存在收敛子列. 特别地, 当 $\epsilon_* > 0$ 时, 任意收敛子列的极限点 $x_{\epsilon_*, \lambda}$ 为 (3.2) 在 $\epsilon = \epsilon_*$ 时的临界点, 且 $x_{\epsilon_*, \lambda}$ 与真实 s -稀疏解之间的误差界为 $O(\sqrt{\lambda}) + O(\sigma_s(x_{\epsilon_*, \lambda})_2)^3$.

3.1.2 二阶松弛算法

邻近 Newton 型算法 (proximal Newton-type algorithm, PNA) 是一类将 Newton 法与广义邻近算法结合的方法, 可看作 PGA 的二阶延伸, 最早由 Lee 等 [91] 在求解凸复合优化问题 (3.2) 时提出. 具体地, PNA 先利用广义邻近算子确定搜索方向 d^k , 进而利用回溯法选取步长 v_k , 迭代过程如下:

$$\begin{aligned} d^k &= \text{Prox}_g^{H^k}(x^k - H_k^{-1} \nabla f(x^k)) - x^k, \\ x^{k+1} &= x^k + v_k d^k, \end{aligned} \quad (3.8)$$

其中 $H_k = \nabla^2 f(x^k)$ 为 f 在 x^k 处的 Hesse 矩阵, 且广义邻近算子定义如下:

$$\text{Prox}_g^H(x) := \underset{y \in \mathbb{R}^n}{\operatorname{argmin}} g(y) + \frac{1}{2} \|y - x\|_H^2, \quad \forall y \in \mathbb{R}^n,$$

这里 $\|x\|_H^2 := x^T H x$, $H \in \mathbb{S}_{++}^n$ 对称正定. 与 Newton 法类似, PNA 具有二次收敛性. 若取 H^k 为 $\nabla^2 f(x^k)$ 的近似, (3.8) 被称为邻近拟 Newton 算法 (proximal quasi-Newton algorithm, PQNA). 此时, 若 H^k 的奇异值有界, PQNA 具有超线性收敛性.

目前, PNA 已被用于求解结构稀疏优化问题. 例如, 当 $g(x) = \lambda \|x\|_1$ 且 f 取最小二乘损失或逻辑回归损失时, 类似于 FISTA, Ghanbari 和 Scheinberg [62] 提出了加速的 PQNA 求解凸正则化问题, 并证明了若 H^k 满足某些假设, 算法的收敛速率可达到 $O(1/k^2)$. 另外, 当 $g(x)$ 非凸时, Rakotomamonjy 等 [125] 将 DC 框架与 Newton 法相结合, 提出了 DC 近端 Newton 法求解 (3.2), 并证明了由 DC 近端 Newton 法生成的序列的任意极限点都是 (3.2) 的稳定点. Kanzow 和 Lechner [83] 提出了非精确的近端 Newton 型方法, 并证明了该方法可全局收敛到 (3.2) 的稳定点.

3.2 ℓ_0 范数极小化算法

本小节针对目标函数 (或约束) 涉及 ℓ_0 范数的结构稀疏优化问题

$$\min_{x \in \mathbb{R}^n} \{f(x) : \|x\|_0 \leq s\}, \quad (3.9a)$$

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \|x\|_0, \quad (3.9b)$$

介绍其一阶与二阶求解算法. 特别地, 第 3.2.1 小节以 f 为最小二乘损失为例介绍具体算法.

3.2.1 一阶贪婪算法

(I) 正交匹配追踪 (orthogonal matching pursuit, OMP) 算法 [130, 140, 142] 是一类著名的贪婪算法. 具体地, OMP 算法在每一步迭代中, 首先找到感知矩阵 A 中与当前残差 $r^{k+1} = b - Ax^k$ 最相关的一

3) $\sigma_s(\cdot)_2$ 为第 2.2 小节中介绍的最佳 s -稀疏逼近.

列; 其次将该列对应的指标并入支撑集 $\mathcal{S}_k$, 进而得到新的支撑集 $\mathcal{S}_{k+1}$; 最后在 $\mathcal{S}_{k+1}$ 上应用最小二乘更新 x^{k+1} , 迭代过程如下:

$$i = \operatorname{argmax}_{i \notin \mathcal{S}_k} |\langle A_i, b - Ax^k \rangle|, \quad (3.10a)$$

$$\mathcal{S}_{k+1} = \{i\} \cup \mathcal{S}_k, \quad (3.10b)$$

$$x^{k+1} = \operatorname{argmin}_z \{\|Az - b\|_2, \operatorname{supp}(z) \subseteq \mathcal{S}_{k+1}\}. \quad (3.10c)$$

注意到, OMP 算法简洁高效, 且由于残差与已选择的列是正交的, 迭代过程中不会两次选择同一列, 因此迭代次数至多为 A 的列数; 但由于 $\mathcal{S}_{k+1}$ 随迭代逐次增大, OMP 算法无法自动矫正迭代中可能出现的指标错误. 目前, 基于 MIP 或 REC 条件, Cai 和 Wang^[25] 证明了 OMP 算法以较高概率精确识别稀疏信号的支撑集. 基于 RIP 条件, 文献 [177, 定理 2.1] 证明了由 OMP 算法生成的迭代序列 $\{x^{k+1}\}$ 满足

$$\|x^{k+1} - \bar{x}\|_2 \leq \frac{\sqrt{6}\epsilon_t(\bar{x})}{\rho_-(t)},$$

其中, $t \in \mathbb{N}$ 与 $\mathcal{S}$ 相关, $\epsilon_t(\bar{x})$ 是限制梯度最优常数, $\rho_-(t)$ 是限制强凸常数. 特别地, OMP 算法也可用于求解组稀疏优化问题. Ben-Haim 和 Eldar^[9] 应用 OMP 算法求解组稀疏优化问题, 并基于 GC 条件给出了组 OMP 算法生成的迭代序列的误差界估计. 压缩采样匹配追踪 (compressive sampling matching pursuit, CoSaMP) 算法^[112] 通过将采样技巧与 OMP 算法结合, 实现了对 OMP 的改进, 其主要迭代过程如下:

$$\begin{aligned} y &= A^T(b - Ax^k), \\ \mathcal{T} &= \operatorname{supp}(x^k) \cup \operatorname{supp}(y_{2s}), \\ z &= \operatorname{argmin}_x \{\|Ax - b\|_2 : \operatorname{supp}(x) \subseteq \mathcal{T}\}, \\ x^{k+1} &= \operatorname{argmin}_x \{\|x - z\| : \|x\|_0 \leq s\}. \end{aligned}$$

注意到, 在每步迭代中, 与 OMP 算法依次将下标并入 $\operatorname{supp}(x^k)$ 不同, CoSaMP 算法先放大了支撑集的搜索范围 (新的支撑集 $\mathcal{T}$ 的基数不超过 $3s$), 再利用最小二乘获得近似解, 最后保留近似解的最大 s 个分量. 这样能有效避免 OMP 算法无法自动矫正迭代过程可能出现的错误指标, 因此, CoSaMP 具备一定的支撑集自我矫正纠错能力. 目前, 基于 RIP 条件, 文献 [112, 定理 4.1] 证明了由 CoSaMP 生成的序列能收敛到 (3.9a) 的最优解 $\bar{x}$, 并给出了二者的误差界估计. 具体地, 若 $\delta_{4s} \leq 0.1$, 则由 CoSaMP 生成的序列 $\{x^k\}$ 是 s -稀疏的, 且 $\|x^k - \bar{x}\|_2 \leq 2^{-k}\|\bar{x}\|_2 + 15\|e\|_2$.

前向后向 (forward-backward, FoBa) 自适应贪婪算法是由 Zhang^[176] 提出的一种具有自我矫正能力的算法, 它结合前向贪婪算法和后向贪婪算法, 进一步改进了 OMP 算法. 具体地, FoBa 先执行前向步, 即 OMP 迭代步 (3.10a)–(3.10c) 来添加新的指标集. 若

$$\delta_k := \|Ax^{k-1} - b\|_2^2 - \|Ax^k - b\|_2^2 \leq \epsilon,$$

则算法终止; 否则, 令 $k = k + 1$, 执行后向步. 具体地, 计算 $j_k = \operatorname{argmin}_{j \in \mathcal{S}_k} \|A(x^k - x_j^k e^j) - b\|_2^2$, 如果

$$\|A(x^k - x_j^k e^j) - b\|_2^2 - \|Ax^{k-1} - b\|_2^2 \leq \frac{1}{2}\delta_k, \quad (3.11)$$

则令 $k = k - 1$, 接受后向步来删除错误指标, 并在新的支撑集上更新迭代:

$$\begin{aligned}\mathcal{S}_k &= \mathcal{S}_{k+1} \setminus \{j_{k+1}\}, \\ x^k &= \underset{z}{\operatorname{argmin}} \{\|Az - b\|_2, \operatorname{supp}(z) \subseteq \mathcal{S}_k\}.\end{aligned}$$

FoBa 通过自适应规则步 (3.11), 删除前向步搜索的错误指标, 进行自我矫正; 弥补 OMP 算法的不足的同时, 也继承了 OMP 算法、前向贪婪算法与后向贪婪算法的快速高效性. 此外, 基于 RIP 或 SEC 条件, Zhang^[176] 证明了 FoBa 以高概率收敛到 (3.9a) 的解.

(II) 迭代硬阈值 (iterative hard thresholding, IHT) 算法是由 Blumensath 和 Davies^[16, 17] 提出的一类用于求解 ℓ_0 范数约束优化问题 (3.9a) 的阈值迭代算法, 适用于求解大规模问题, 其迭代格式为

$$x^{k+1} = \mathbb{H}_s(x^k - vA^T(Ax^k - b)),$$

其中 $\mathbb{H}_s: \mathbb{R}^n \rightarrow \mathbb{R}^n$ 为硬阈值算子, 对给定的 $x \in \mathbb{R}^n$, $\mathbb{H}_s(x)$ 仅保留 x 中绝对值最大的 s 项分量, 其余分量为 0.

IHT 与 ISTA 思想类似, 在每步迭代中均先对数据拟合项作梯度下降, 再通过阈值算子更新 x^{k+1} . 二者的不同在于, ISTA 中的阈值算子为软阈值算子, 而 IHT 中的则是硬阈值算子. 相较 ISTA, IHT 在迭代过程中不会减少和改变 x^k 中相对较大的分量值. 文献 [16, 定理 4] 证明了, 当 $\|A\|_2 \leq 1$ 时, 由 IHT 算法生成的序列 $\{x^k\}$ 收敛到 (3.9a) 的局部极小值; 文献 [17, 定理 4] 证明了, 当 $\delta_{3s} < 1/\sqrt{32}$ 时,

$$\|\bar{x} - x^k\|_2 \leq 2^{-k} \|\sigma_s(\bar{x})_2\|_2 + 6\tilde{\varepsilon}_s, \quad (3.12)$$

其中, $\bar{x}$ 为 (3.9a) 的最优解, $\tilde{\varepsilon}_s = \|\bar{x} - x^k\|_2 + \frac{1}{\sqrt{s}} \|\bar{x} - \sigma_s(\bar{x})_2\|_1 + \|e\|_2$.

类似 ISTA, 基于 IHT 算法也衍生了许多加速的硬阈值类算法, 例如, Blumensath 和 Davies^[18] 采用自适应步长策略, 提出了规范的迭代硬阈值 (normalized iterative hard thresholding, NIHT) 算法, 并基于非对称的 RIP 条件建立了 NIHT 的相合性理论; Blanchard 等^[14, 15] 用共轭梯度代替 IHT 中的梯度, 提出了共轭梯度迭代硬阈值 (conjugate gradient iterative hard thresholding, CGIHT) 算法, 兼具梯度下降法的低复杂度性与共轭梯度法的渐近收敛性.

硬阈值追踪 (hard thresholding pursuit, HTP) 算法由 Foucart^[55] 提出, 是 IHT 和 CoSaMP 算法的组合. 具体地, HTP 算法先将中间迭代点 y^{k+1} 的前 s 大分量对应的下标确定为支撑集, 再在支撑集上应用最小二乘更新 x^{k+1} , 其迭代过程如下:

$$y^{k+1} = x^k - vA^T(Ax^k - b), \quad \mathcal{S}_{k+1} = \operatorname{supp}(\mathbb{H}_s(y^{k+1})), \quad x^{k+1} = \underset{z}{\operatorname{argmin}} \{\|Az - b\|_2, \operatorname{supp}(z) \subseteq \mathcal{S}_{k+1}\}.$$

基于 RIP 条件, 文献 [55, 定理 3.5] 建立了 HTP 算法的误差界估计, 证明了当 $\delta_{3s} < 1/\sqrt{3}$ 时 HTP 算法生成的序列 $\{x^k\}$ 与 (3.9a) 的最优解 $\bar{x}$ 有如下关系:

$$\|x^k - \bar{x}_S\|_2 \leq \rho^k \|x^0 - \bar{x}_S\|_2 + \tau \frac{1 - \rho^k}{1 - \rho} \|A\bar{x}_{S^c} + e\|_2, \quad (3.13)$$

其中 $\rho \in (0, 1)$ 和 $\tau > 0$ 为与 δ_{2s} 相关的常数. 比较 (3.12) 与 (3.13), HTP 算法不但无需假设 $\|A\|_2 \leq 1$, 而且对 RIP 的要求也放宽了.

近几年, Zhao^[184]、Zhao 和 Luo^[186] 提出了一类最优 k 阈值算法, 它在保证算法执行阈值压缩的同时, 还能保证目标函数下降. 基于最优 k 阈值正则化模型的一阶近似, Zhao 和 Luo^[187] 还提出了一类自然阈值算法, 其计算成本显著低于最优 k 阈值算法.

3.2.2 二阶贪婪算法

(I) 基于 Newton 步的迭代硬阈值 (Newton-step-based IHT, NSIHT) 算法^[109] 由 Meng 和 Zhao 提出, 该算法将 Newton 步与 IHT 算法结合, 可以求解稀疏约束的最小二乘问题, 即 (3.9a) 中取 $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, 其迭代格式如下:

$$x^{k+1} = \mathbb{H}_s(x^k + v(A^T A + \epsilon I)^{-1} A^T (b - Ax^k)).$$

当 A 满足 $\delta_{3s} < 1/\sqrt{3}$ 时且其最大奇异值 σ_1 满足特定关系时, 由 NSIHT 算法生成的序列 $\{x^{k+1}\}$ 满足

$$\|x^{k+1} - \bar{x}_S\|_2 \leq \rho^k \|x^0 - \bar{x}_S\|_2 + \tau \|A\bar{x}_{S^c} + e\|_2,$$

其中, $\rho \in (0, 1)$, τ 是与 ρ 、 σ_1 、 ϵ 和 v 相关的正常数 (参见文献 [109, 定理 4.1]). 此外, Meng 和 Zhao^[109] 还将 Newton 步与 HTP 算法相结合, 提出了基于 Newton 步的硬阈值追踪算法 (Newton-step-based HTP, NSHTP), 并建立了与 NSIHT 类似的收敛性结果.

(II) 贪婪投影梯度 Newton (greedy projection gradient Newton, GPGN) 算法是由 Wang 等^[155] 提出的一类二阶算法, 它将投影梯度法和 Newton 法相结合, 可用于求解稀疏约束的逻辑回归问题, 即 (3.9a) 中取 f 为逻辑回归损失. 具体地, GPGN 算法先作梯度下降, 再作稀疏投影, 迭代格式如下:

$$\tilde{x}^{k+1} = \mathcal{P}_S(x^k - v_k \nabla f(x^k)).$$

若 $\text{supp}(\tilde{x}^{k+1}) = \text{supp}(x^k)$, 则在 $\tilde{x}^{k+1}$ 的支撑集上计算 Newton 步, 记 $\tilde{\mathcal{S}}_{k+1} := \text{supp}(\tilde{x}^{k+1})$, 且令

$$\hat{x}_{\tilde{\mathcal{S}}_{k+1}}^{k+1} = \tilde{x}_{\tilde{\mathcal{S}}_{k+1}}^{k+1} - (\nabla_{\tilde{\mathcal{S}}_{k+1}}^2 f(\tilde{x}^{k+1}))^{-1} \nabla_{\tilde{\mathcal{S}}_{k+1}} f(\tilde{x}^{k+1});$$

否则, $x^{k+1} = \tilde{x}^{k+1}$. 最后, 如果 $f(\hat{x}^{k+1}) \leq f(\tilde{x}^{k+1}) - \frac{1}{2}\|\hat{x}^{k+1} - \tilde{x}^{k+1}\|_2^2$, 则 $x^{k+1} := \hat{x}^{k+1}$, 否则, $x^{k+1} := \tilde{x}^{k+1}$. Wang 等^[155] 证明了 GPGN 算法具有全局收敛性和局部二次/次线性收敛性. 不同于前面介绍的二阶算法, GPGN 算法在支撑集上计算 Newton 步, 可有效降低计算成本.

针对模型 (3.9b), 当 f 为逻辑回归损失且正则项为 $\lambda\|x\|_{p,0}$ ($p \geq 1$) 时, Zhang 等^[174] 在多维的稀疏逻辑回归背景下, 提出了 $\ell_{p,0}$ -近端 Newton 算法 ($\ell_{p,0}$ -PNA), 引入了强 α -稳定点刻画其与全局最小值 (或局部极小值) 的等价性, 并证明了 $\ell_{p,0}$ -PNA 的局部二次收敛性.

(III) Newton 硬阈值追踪 (Newton HTP, NHTP) 算法是由 Zhou 等^[190] 提出的一类二阶算法, 它将限制 Newton 步迭代与 HTP 相结合, 可用于求解一般的 ℓ_0 约束优化问题 (3.9a). 具体地, 通过引入 η -稳定点, 即满足 $x \in \mathcal{P}_s(x - \eta \nabla f(x))$ 的点, 将 (3.9a) 转化为求解下述非线性方程:

$$F_\eta(x; \eta) := \begin{bmatrix} \nabla_T f(x) \\ x_{T^c} \end{bmatrix} = 0, \quad \exists x \in \mathcal{T}(x; \eta), \quad (3.14)$$

其中 $\mathcal{T}(x; \eta) := \{T \subset [n] : |T| = s, T \supseteq \text{supp}(z), \exists z \in \mathcal{P}_s(x - \eta \nabla f(x))\}$. 在每步迭代中, NHTP 首先选取 T_k , 接着对非线性方程 (3.14) 应用 Newton 法并在 T_k 上执行 Newton 步, 若由此得到的 Newton 方向 d_N^k 满足下降不等式, 则选择搜索方向 d^k 为 d_N^k , 否则选择 d^k 为限制梯度方向 d_g^k ; 最后通过在 T_k 上执行 Armijo 线搜索, 进行下一次迭代

$$x^{k+1} = \begin{bmatrix} x_{T_k}^k + \alpha d_{T_k}^k \\ 0 \end{bmatrix}, \quad \alpha > 0.$$

基于限制强凸性和限制强光滑条件, 文献 [190, 定理 10] 证明了 NHTP 具有全局收敛性与二次收敛性.

4 结论与展望

结构稀疏优化问题与数据处理密切相关, 在信号处理、图像科学、机器学习和生物医学等应用领域取得了很大的成功, 引起了业界的广泛关注. 本文针对稀疏优化与组稀疏优化的基本模型、相合性理论与求解算法进行了概述总结, 对混合稀疏优化和联合稀疏优化的研究作了简要介绍, 以便读者快速了解结构稀疏优化的相关内容, 进一步开展结构稀疏优化的相关研究.

结构稀疏优化发展至今, 虽然积累了丰富的研究成果, 但仍有很多问题亟待解决. 现有的算法大多是一阶算法, 如何设计高效快速的二阶算法, 如何针对含约束的组稀疏优化问题进行算法设计及建立收敛性理论, 如何针对混合稀疏优化问题进行算法设计与求解, 非凸稀疏优化算法收敛到全局最优解的公开问题, 如何提出更弱的正则性条件以保证精确恢复与相合性理论, 如何针对低秩矩阵与低秩张量优化问题研究相合性理论, 设计高效算法等问题亟待解决.

参考文献

- 1 Arora S, Barak B. Computational Complexity: A Modern Approach. Cambridge: Cambridge Univ Press, 2010
- 2 Attouch H, Bolte J, Redont P, et al. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality. *Math Oper Res*, 2010, 35: 438–457
- 3 Attouch H, Bolte J, Svaiter B F. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math Program*, 2013, 137: 91–129
- 4 Bach F R. Consistency of the group Lasso and multiple kernel learning. *J Mach Learn Res*, 2008, 9: 1179–1225
- 5 Bah B, Tanner J. Improved bounds on restricted isometry constants for Gaussian matrices. *SIAM J Matrix Anal Appl*, 2010, 31: 2882–2898
- 6 Bandeira A S, Dobriban E, Mixon D G, et al. Certifying the restricted isometry property is hard. *IEEE Trans Inform Theory*, 2013, 59: 3448–3450
- 7 Baraniuk R G, Cevher V, Duarte M F, et al. Model-based compressive sensing. *IEEE Trans Inform Theory*, 2010, 56: 1982–2001
- 8 Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J Imag Sci*, 2009, 2: 183–202
- 9 Ben-Haim Z, Eldar Y C. Near-oracle performance of greedy block-sparse estimation techniques from noisy measurements. *IEEE J Sel Top Signal Process*, 2011, 5: 1032–1047
- 10 Bertsimas D, Copenhaver M S, Mazumder R. The trimmed Lasso: Sparsity and robustness. *arXiv:1708.04527*, 2017
- 11 Bickel P J, Ritov Y, Tsybakov A B. Simultaneous analysis of Lasso and Dantzig selector. *Ann Statist*, 2009, 37: 1705–1732
- 12 Blanchard J D, Cartis C, Tanner J. Compressed sensing: How sharp is the restricted isometry property? *SIAM Rev*, 2011, 53: 105–125
- 13 Blanchard J D, Cermak M, Hanle D, et al. Greedy algorithms for joint sparse recovery. *IEEE Trans Signal Process*, 2014, 62: 1694–1704
- 14 Blanchard J D, Tanner J, Wei K. Conjugate gradient iterative hard thresholding: Observed noise stability for compressed sensing. *IEEE Trans Signal Process*, 2015, 63: 528–537
- 15 Blanchard J D, Tanner J, Wei K. CGIHT: Conjugate gradient iterative hard thresholding for compressed sensing and matrix completion. *Inf Inference*, 2015, 4: 289–327
- 16 Blumensath T, Davies M E. Iterative thresholding for sparse approximations. *J Fourier Anal Appl*, 2008, 14: 629–654
- 17 Blumensath T, Davies M E. Iterative hard thresholding for compressed sensing. *Appl Comput Harmon Anal*, 2009, 27: 265–274
- 18 Blumensath T, Davies M E. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE J Sel Top Signal Process*, 2010, 4: 298–309
- 19 Bogdan M, van den Berg E, Sabatti C, et al. SLOPE—adaptive variable selection via convex optimization. *Ann Appl Stat*, 2015, 9: 1103–1140
- 20 Bogdan M, van den Berg E, Su W J, et al. Statistical estimation and testing via the sorted L1 norm. *arXiv:1310.1969*, 2013
- 21 Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Machine Learn*, 2010, 3: 1–122
- 22 Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann Appl Stat*, 2011, 5: 232–253

- 23 Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat Comput*, 2015, 25: 173–187
- 24 Cahill J, Chen X M, Wang R R. The gap between the null space property and the restricted isometry property. *Linear Algebra Appl*, 2016, 501: 363–375
- 25 Cai T T, Wang L. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans Inform Theory*, 2011, 57: 4680–4688
- 26 Cai T T, Xu G W, Zhang J. On recovery of sparse signals via ℓ_1 minimization. *IEEE Trans Inform Theory*, 2009, 55: 3388–3397
- 27 Cai T T, Zhang A. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Trans Inform Theory*, 2014, 60: 122–132
- 28 Candès E J, Romberg J, Tao T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans Inform Theory*, 2006, 52: 489–509
- 29 Candès E J, Tao T. Decoding by linear programming. *IEEE Trans Inform Theory*, 2005, 51: 4203–4215
- 30 Candès E J, Wakin M B, Boyd S P. Enhancing sparsity by reweighted L1 minimization. *J Fourier Anal Appl*, 2008, 14: 877–905
- 31 Chartrand R, Staneva V. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 2008, 24: 657–682
- 32 Chartrand R, Wohlberg B. A nonconvex ADMM algorithm for group sparsity with sparse groups. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Piscataway: IEEE, 2013, 6009–6013
- 33 Chatterjee S, Banerjee A, Chatterjee S, et al. Sparse group Lasso for regression on land climate variables. In: *Proceedings of the IEEE 11th International Conference on Data Mining Workshops*. Piscataway: IEEE, 2011, 1–8
- 34 Chen J, Huo X M. Theoretical results on sparse representations of multiple-measurement vectors. *IEEE Trans Signal Process*, 2006, 54: 4634–4643
- 35 Chen X J, Xu F M, Ye Y Y. Lower bound theory of nonzero entries in solutions of ℓ_2 - ℓ_p minimization. *SIAM J Sci Comput*, 2010, 32: 2832–2852
- 36 Chen X J, Zhou W J. Convergence of the reweighted ℓ_1 minimization algorithm for ℓ_2 - ℓ_p minimization. *Comput Optim Appl*, 2014, 59: 47–61
- 37 Choi J W, Shim B, Ding Y C, et al. Compressed sensing for wireless communications: Useful tips and tricks. *IEEE Commun Surv Tutor*, 2017, 19: 1527–1550
- 38 Cotter S F, Rao B D, Engan K, et al. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans Signal Process*, 2005, 53: 2477–2488
- 39 Daubechies I, Defrise M, De Mol C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm Pure Appl Math*, 2004, 57: 1413–1457
- 40 Daubechies I, DeVore R, Fornasier M, et al. Iteratively reweighted least squares minimization for sparse recovery. *Comm Pure Appl Math*, 2010, 63: 1–38
- 41 Davies M E, Eldar Y C. Rank awareness in joint sparse recovery. *IEEE Trans Inform Theory*, 2012, 58: 1135–1146
- 42 Dong Z L, Yang X Q, Dai Y H. A unified recovery bound estimation for noise-aware ℓ_q optimization model in compressed sensing. *arXiv:1609.01531*, 2016
- 43 Donoho D L. For most large underdetermined systems of linear equations the minimal ℓ_1 -norm solution is also the sparsest solution. *Comm Pure Appl Math*, 2006, 59: 797–829
- 44 Donoho D L. Compressed sensing. *IEEE Trans Inform Theory*, 2006, 52: 1289–1306
- 45 Donoho D L, Elad M, Temlyakov V N. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans Inform Theory*, 2006, 52: 6–18
- 46 Donoho D L, Huo X M. Uncertainty principles and ideal atomic decomposition. *IEEE Trans Inform Theory*, 2001, 47: 2845–2862
- 47 Eldar Y C, Kuppinger P, Bolcskei H. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans Signal Process*, 2010, 58: 3042–3054
- 48 Eldar Y C, Mishali M. Robust recovery of signals from a structured union of subspaces. *IEEE Trans Inform Theory*, 2009, 55: 5302–5316
- 49 Eldar Y C, Rauhut H. Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE Trans Inform Theory*, 2009, 56: 505–519
- 50 Esser E, Lou Y F, Xin J. A method for finding structured sparse solutions to nonnegative least squares problems with applications. *SIAM J Imag Sci*, 2013, 6: 2010–2046
- 51 Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, 2001, 96: 1348–1360
- 52 Feng L, Benkert T, Block K T, et al. Compressed sensing for body MRI. *Magn Reson Imag*, 2017, 45: 966–987
- 53 Feng X, Yan S, Wu C. The $\ell_{2,q}$ regularized group sparse optimization: Lower bound theory, recovery bound and algorithms. *Appl Comput Harmon Anal*, 2020, 49: 381–414
- 54 Fornasier M, Rauhut H. Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM J Numer*

- Anal, 2008, 46: 577–613
- 55 Foucart S. Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM J Numer Anal*, 2011, 49: 2543–2563
- 56 Foucart S. Stability and robustness of ℓ_1 -minimizations with Weibull matrices and redundant dictionaries. *Linear Algebra Appl*, 2014, 441: 4–21
- 57 Foucart S, Lai M J. Sparsest solutions of underdetermined linear systems via ℓ_q -minimization for $0 < q \leq 1$. *Appl Comput Harmon Anal*, 2009, 26: 395–407
- 58 Foucart S, Rauhut H. *A Mathematical Introduction to Compressive Sensing*. New York: Springer, 2013
- 59 Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso. arXiv:1001.0736, 2010
- 60 Gamper U, Boesiger P, Kozerke S. Compressed sensing in dynamic MRI. *Magn Reson Med*, 2008, 59: 365–373
- 61 Gao Y, Ma M D. A new bound on the block restricted isometry constant in compressed sensing. *J Inequal Appl*, 2017, 2017: 174
- 62 Ghanbari H, Scheinberg K. Proximal quasi-Newton methods for regularized convex optimization with linear and accelerated sublinear convergence rates. *Comput Optim Appl*, 2018, 69: 597–627
- 63 Gong P H, Zhang C S, Lu Z S, et al. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In: *Proceedings of the 30th International Conference on Machine Learning*, vol. 28. Brookline: Microtome Publishing, 2013, 37–45
- 64 Gotoh J, Takeda A, Tono K. DC formulations and algorithms for sparse optimization problems. *Math Program*, 2018, 169: 141–176
- 65 Gramfort A, Kowalski M. Improving M/EEG source localization with an inter-condition sparse prior. In: *Proceedings of the IEEE International Symposium on Biomedicine Imaging*. Piscataway: IEEE, 2009, 141–144
- 66 Hale E T, Yin W T, Zhang Y. Fixed-point continuation for ℓ_1 -minimization: Methodology and convergence. *SIAM J Optim*, 2008, 19: 1107–1130
- 67 Hazimeh H, Mazumder R, Radchenko P. Grouped variable selection with discrete optimization: Computational and statistical perspectives. *Ann Statist*, 2023, 51: 1–32
- 68 Hoyer P O. Non-negative sparse coding. In: *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*. Piscataway: IEEE, 2002, 557–565
- 69 Hu X L, Hu Y H, Wu F J, et al. Integration of single-cell multi-omics for gene regulatory network inference. *Comput Struct Biotechnol J*, 2020, 18: 1925–1938
- 70 Hu Y H, Hu X L, Yang X Q. On convergence of iterative thresholding algorithms to approximate sparse solution for composite nonconvex optimization. *Math Program*, 2024, 2024: 1–26
- 71 Hu Y H, Li C, Meng K W, et al. Group sparse optimization via $\ell_{p,q}$ regularization. *J Mach Learn Res*, 2017, 18: 1–52
- 72 Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. *Statist Sci*, 2012, 27: 481–499
- 73 Huang J, Ma S, Xie H, et al. A group bridge approach for variable selection. *Biometrika*, 2009, 96: 339–355
- 74 Huang J Z, Zhang T. The benefit of group sparsity. *Ann Statist*, 2010, 38: 1978–2004
- 75 Huang X L, Shi L, Yan M. Nonconvex sorted ℓ_1 minimization for sparse approximation. *J Oper Res Soc China*, 2015, 3: 207–229
- 76 Huang X X, Yang X Q. A unified augmented Lagrangian approach to duality and exact penalization. *Math Oper Res*, 2003, 28: 533–552
- 77 Hurley N, Rickard S. Comparing measures of sparsity. *IEEE Trans Inform Theory*, 2009, 55: 4723–4741
- 78 Jia X D, Zhao M, Di Y, et al. Sparse filtering with the generalized ℓ_p/ℓ_q norm and its applications to the condition monitoring of rotating machinery. *Mech Syst Signal Processing*, 2018, 102: 198–213
- 79 Jia Z H, Gao X, Cai X J, et al. Local linear convergence of the alternating direction method of multipliers for nonconvex separable optimization problems. *J Optim Theory Appl*, 2021, 188: 1–25
- 80 Jiang W H, Nie F P, Huang H. Robust dictionary learning with capped ℓ_1 -norm. In: *Proceedings of the 24th International Joint Conference on Artificial Intelligence*. Washington, D. C.: AAAI, 2015, 3590–3596
- 81 Jiao Y L, Jin B T, Lu X L. Group sparse recovery via the ℓ^0 (ℓ^2) penalty: Theory and algorithm. *IEEE Trans Signal Process*, 2016, 65: 998–1012
- 82 Jiao Y L, Jin B T, Lu X L. Iterative soft/hard thresholding with homotopy continuation for sparse recovery. *IEEE Signal Process Lett*, 2017, 24: 784–788
- 83 Kanzow C, Lechner T. Globalized inexact proximal Newton-type methods for nonconvex composite functions. *Comput Optim Appl*, 2021, 78: 377–410
- 84 Kim S J, Koh K, Lustig M, et al. An interior-point method for large-scale ℓ_1 -regularized least squares. *IEEE J Sel Top Signal Process*, 2007, 1: 606–617
- 85 Koep N, Behboodi A, Mathar R. The restricted isometry property of block diagonal matrices for group-sparse signal recovery. *Appl Comput Harmon Anal*, 2022, 60: 333–367
- 86 Krishnan D, Tay T, Fergus R. Blind deconvolution using a normalized sparsity measure. In: *Proceedings of the IEEE*

- Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2011, 233–240
- 87 Lai M J, Xu Y Y, Yin W T. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM J Numer Anal*, 2013, 51: 927–957
 - 88 Laria J C, Carmen Aguilera-Morillo M, Lillo R E. An iterative sparse-group lasso. *J Comput Graph Statist*, 2019, 28: 722–731
 - 89 Le Thi H A, Pham Dinh T, Le H M, et al. DC approximation approaches for sparse optimization. *European J Oper Res*, 2015, 244: 26–46
 - 90 Le Thi H A, Tao P D. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Ann Oper Res*, 2005, 133: 23–46
 - 91 Lee J D, Sun Y K, Saunders M A. Proximal Newton-type methods for minimizing composite functions. *SIAM J Optim*, 2014, 24: 1420–1443
 - 92 Lee S, Oh M, Kim Y. Sparse optimization for nonconvex group penalized estimation. *J Stat Comput Simul*, 2016, 86: 597–610
 - 93 Li W J, Bian W, Toh K C. Difference-of-convex algorithms for a class of sparse group ℓ_0 regularized optimization problems. *SIAM J Optim*, 2022, 32: 1614–1641
 - 94 Li X X, Hu Y H, Li C, et al. Sparse estimation via lower-order penalty optimization methods in high-dimensional linear regression. *J Global Optim*, 2023, 85: 315–349
 - 95 Li X X, Mo L L, Yuan X M, et al. Linearized alternating direction method of multipliers for sparse group and fused LASSO models. *Comput Statist Data Anal*, 2014, 79: 203–221
 - 96 Liang D, Liu B, Wang J J, et al. Accelerating SENSE using compressed sensing. *Magn Reson Med*, 2009, 62: 1574–1584
 - 97 Lin J H, Li S. Block sparse recovery via mixed ℓ_2/ℓ_1 minimization. *Acta Math Sin (Engl Ser)*, 2013, 29: 1401–1412
 - 98 Ling Q, Wen Z W, Yin W T. Decentralized jointly sparse optimization by reweighted ℓ_q minimization. *IEEE Trans Signal Process*, 2012, 61: 1165–1170
 - 99 Liu L L, Chin S P, Tran T D. JOBS: Joint-sparse optimization from bootstrap samples. In: *Proceedings of the IEEE International Symposium on Information Theory*. Piscataway: IEEE, 2019, 2689–2693
 - 100 Lou Y F, Yin P H, He Q, et al. Computing sparse representation in a highly coherent dictionary based on difference of L_1 and L_2 . *J Sci Comput*, 2015, 64: 178–196
 - 101 Lu Z B, Li H Q, Li W P. Image deblocking via group sparsity optimization. In: *Proceedings of the IEEE International Symposium on Circuits and Systems*. Piscataway: IEEE, 2015, 1582–1585
 - 102 Lu Z S, Li X R. Sparse recovery via partial regularization: Models, theory, and algorithms. *Math Oper Res*, 2018, 43: 1290–1316
 - 103 Luo Z Q, Pang J S, Ralph D. *Mathematical Programs with Equilibrium Constraints*. Cambridge: Cambridge Univ Press, 1996
 - 104 Lustig M, Donoho D, Pauly J M. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn Reson Med*, 2007, 58: 1182–1195
 - 105 Lv J, Fan Y Y. A unified approach to model selection and sparse recovery using regularized least squares. *Ann Statist*, 2009, 37: 3498–3528
 - 106 Maly J, Palzer L. Analysis of hard-thresholding for distributed compressed sensing with one-bit measurements. *Inf Inference*, 2020, 9: 455–471
 - 107 Meier L, van de Geer S, Bühlmann P. The group lasso for logistic regression. *J R Stat Soc Ser B Stat Methodol*, 2008, 70: 53–71
 - 108 Meinshausen N, Yu B. Lasso-type recovery of sparse representations for high-dimensional data. *Ann Statist*, 2009, 37: 246–270
 - 109 Meng N, Zhao Y B. Newton-step-based hard thresholding algorithms for sparse signal recovery. *IEEE Trans Signal Process*, 2020, 68: 6594–6606
 - 110 Mishali M, Eldar Y C. Reduce and boost: Recovering arbitrary sets of jointly sparse vectors. *IEEE Trans Signal Process*, 2008, 56: 4692–4702
 - 111 Natarajan B K. Sparse approximate solutions to linear systems. *SIAM J Comput*, 1995, 24: 227–234
 - 112 Needell D, Tropp J A. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl Comput Harmon Anal*, 2009, 26: 301–321
 - 113 Nikolova M. Local strong homogeneity of a regularized estimator. *SIAM J Appl Math*, 2000, 61: 633–658
 - 114 Nikolova M, Ng M K, Tam C P. Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. *IEEE Trans Image Process*, 2010, 19: 3073–3088
 - 115 Obozinski G, Jacob L, Vert J P. Group Lasso with overlaps: The latent group Lasso approach. *arXiv:1110.0413*, 2011
 - 116 Obozinski G, Taskar B, Jordan M I. Joint covariate selection and joint subspace selection for multiple classification problems. *Stat Comput*, 2010, 20: 231–252
 - 117 Pan H, Jing Z L, Li M Z. Robust image restoration via random projection and partial sorted ℓ_p norm. *Neurocomputing*, 2017, 222: 72–80

- 118 Pan L L, Chen X J. Group sparse optimization for images recovery using capped folded concave functions. *SIAM J Imag Sci*, 2021, 14: 1–25
- 119 Pham Dinh T, Le Thi H A. Recent advances in DC programming and DCA. In: *Transactions on Computational Intelligence XIII. Lecture Notes in Computer Science*, vol. 8342. Berlin: Springer, 2014, 1–37
- 120 Pham M Q, Oudompheng B, Mars J I, et al. A noise-robust method with smoothed ℓ_1/ℓ_2 regularization for sparse moving-source mapping. *Signal Process*, 2017, 135: 96–106
- 121 Qin J, Hu Y H, Xu F, et al. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods*, 2014, 67: 294–303
- 122 Qin J, Hu Y H, Yao J C, et al. Cell fate conversion prediction by group sparse optimization method utilizing single-cell and bulk OMICs data. *Briefings Bioinf*, 2021, 22: bbab311
- 123 Qin Z W, Scheinberg K, Goldfarb D. Efficient block-coordinate descent algorithms for the Group Lasso. *Math Program Comput*, 2013, 5: 143–169
- 124 Rahimi Y, Wang C, Dong H B, et al. A scale-invariant approach for sparse signal recovery. *SIAM J Sci Comput*, 2019, 41: A3649–A3672
- 125 Rakotomamonjy A, Flamary R, Gasso G. DC proximal Newton for nonconvex optimization problems. *IEEE Trans Neural Netw Learn Syst*, 2015, 27: 636–647
- 126 Rao X B, Lau V K N. Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems. *IEEE Trans Signal Process*, 2014, 62: 3261–3271
- 127 Raskutti G, Wainwright M J, Yu B. Restricted eigenvalue properties for correlated Gaussian designs. *J Mach Learn Res*, 2010, 11: 2241–2259
- 128 Repetti A, Pham M Q, Duval L, et al. Euclid in a taxicab: Sparse blind deconvolution with smoothed ℓ_1/ℓ_2 regularization. *IEEE Signal Process Lett*, 2014, 22: 539–543
- 129 Rudelson M, Zhou S H. Reconstruction from anisotropic random measurements. *IEEE Trans Inform Theory*, 2013, 59: 3434–3447
- 130 Sahoo S K, Makur A. Signal recovery from random measurements via extended orthogonal matching pursuit. *IEEE Trans Signal Process*, 2015, 63: 2572–2581
- 131 Scardapane S, Comminiello D, Hussain A, et al. Group sparse regularization for deep neural networks. *Neurocomputing*, 2017, 241: 81–89
- 132 Shannon C E. Communication in the presence of noise. *Proc IRE*, 1949, 37: 10–21
- 133 Shevade S K, Keerthi S S. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 2003, 19: 2246–2253
- 134 Simon N, Friedman J, Hastie T, et al. A sparse-group Lasso. *J Comput Graph Statist*, 2013, 22: 231–245
- 135 Song C B, Xia S T. Sparse signal recovery by ℓ_q minimization under restricted isometry property. *IEEE Signal Process Lett*, 2014, 21: 1154–1158
- 136 Sprechmann P, Ramirez I, Sapiro G, et al. C-HiLasso: A collaborative hierarchical sparse modeling framework. *IEEE Trans Signal Process*, 2011, 59: 4183–4198
- 137 Stojnic M, Parvaresh F, Hassibi B. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Trans Signal Process*, 2009, 57: 3075–3085
- 138 Su W J, Candès E. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann Statist*, 2016, 44: 1038–1068
- 139 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Methodol*, 1996, 58: 267–288
- 140 Tropp J A. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans Inform Theory*, 2004, 50: 2231–2242
- 141 Tropp J A. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Process*, 2006, 86: 589–602
- 142 Tropp J A, Gilbert A C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans Inform Theory*, 2007, 53: 4655–4666
- 143 Tropp J A, Gilbert A C, Strauss M J. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Process*, 2006, 86: 572–588
- 144 van de Geer S A, Bühlmann P. On the conditions used to prove oracle results for the Lasso. *Electron J Stat*, 2009, 3: 1360–1392
- 145 van den Berg E, Friedlander M P. Theoretical and empirical results for recovery from multiple measurements. *IEEE Trans Inform Theory*, 2010, 56: 2516–2527
- 146 van den Berg E, Schmidt M, Friedlander M P, et al. Group sparsity via linear-time projection. Technical Report TR-2008-09. Vancouver: University of British Columbia, 2008
- 147 Vincent M, Hansen N R. Sparse group lasso and high dimensional multinomial classification. *Comput Statist Data Anal*, 2014, 71: 771–786
- 148 Wang C, Yan M, Rahimi Y, et al. Accelerated schemes for the L_1/L_2 minimization. *IEEE Trans Signal Process*, 2020, 68: 2660–2669

- 149 Wang H F, Gao Y, Shi Y H, et al. Group-based alternating direction method of multipliers for distributed linear classification. *IEEE Trans Cybern*, 2017, 47: 3568–3582
- 150 Wang H F, Kong L C, Tao J Y. The linearized alternating direction method of multipliers for sparse group LAD model. *Optim Lett*, 2019, 13: 505–525
- 151 Wang H F, Zhang F, Shi Y M, et al. Nonconvex and nonsmooth sparse optimization via adaptively iterative reweighted methods. *J Global Optim*, 2021, 81: 717–748
- 152 Wang J H, Hu Y H, Li C, et al. Linear convergence of CQ algorithms and applications in gene regulatory network inference. *Inverse Problems*, 2017, 33: 055017
- 153 Wang L F, Chen G, Li H Z. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 2007, 23: 1486–1494
- 154 Wang Q, Zhang X G, Wu Y, et al. Nonconvex weighted ℓ_p minimization based group sparse representation framework for image denoising. *IEEE Signal Process Lett*, 2017, 24: 1686–1690
- 155 Wang R, Xiu N H, Zhang C. Greedy projected gradient-Newton method for sparse logistic regression. *IEEE Trans Neural Netw Learn Syst*, 2019, 31: 527–538
- 156 Wang Y, Wang J J, Xu Z B. On recovery of block-sparse signals via mixed ℓ_2/ℓ_q ($0 < q \leq 1$) norm minimization. *EURASIP J Adv Signal Process*, 2013, 2013: 76
- 157 Wang Y, Yin W T, Zeng J S. Global convergence of ADMM in nonconvex nonsmooth optimization. *J Sci Comput*, 2019, 78: 29–63
- 158 Wei F R, Zhu H X. Group coordinate descent algorithms for nonconvex penalized regression. *Comput Statist Data Anal*, 2012, 56: 316–326
- 159 Wen J M, Li D F, Zhu F M. Stable recovery of sparse signals via l_p -minimization. *Appl Comput Harmon Anal*, 2015, 38: 161–176
- 160 Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation. *IEEE Trans Pattern Anal Mach Intell*, 2008, 31: 210–227
- 161 Xie Z X, Xu Y. Sparse group LASSO based uncertain feature selection. *Int J Mach Learn Cyber*, 2014, 5: 201–210
- 162 Xu Y Y, Yin W T. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *J Sci Comput*, 2017, 72: 700–734
- 163 Xu Z B, Chang X Y, Xu F M, et al. $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Trans Neural Netw Learn Syst*, 2012, 23: 1013–1027
- 164 Yang J F, Zhang Y. Alternating direction algorithms for ℓ_1 -problems in compressive sensing. *SIAM J Sci Comput*, 2011, 33: 250–278
- 165 Yap P T, Zhang Y, Shen D G. Multi-tissue decomposition of diffusion MRI signals via ℓ_0 sparse-group estimation. *IEEE Trans Image Process*, 2016, 25: 4340–4353
- 166 Ye J P, Liu J. Sparse methods for biomedical data. *SIGKDD Explor Newsl*, 2012, 14: 4–15
- 167 Yin P H, Lou Y F, He Q, et al. Minimization of ℓ_{1-2} for compressed sensing. *SIAM J Sci Comput*, 2015, 37: A536–A563
- 168 Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Ser B Stat Methodol*, 2006, 68: 49–67
- 169 Yuan X T, Liu X B, Yan S C. Visual classification with multitask joint sparse representation. *IEEE Trans Image Process*, 2012, 21: 4349–4360
- 170 Yun J H, Zheng P, Yang E, et al. Trimming the ℓ_1 regularizer: Statistical analysis, optimization, and applications to deep learning. In: *Proceedings of the 36th International Conference on Machine Learning*. Cambridge: PMLR, 2019, 7242–7251
- 171 Zhang C H. Nearly unbiased variable selection under minimax concave penalty. *Ann Statist*, 2010, 38: 894–942
- 172 Zhang C H, Zhang T. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist Sci*, 2012, 27: 576–593
- 173 Zhang H M, Qian F, Shang F H, et al. Global convergence guarantees of (A)GIST for a family of nonconvex sparse learning problems. *IEEE Trans Cybern*, 2022, 52: 3276–3288
- 174 Zhang P H, Wang R, Xiu N H. Multinomial logistic regression classifier via $l_{q,0}$ -proximal Newton algorithm. *Neuro-computing*, 2022, 468: 148–164
- 175 Zhang T. Analysis of multi-stage convex relaxation for sparse regularization. *J Mach Learn Res*, 2010, 11: 1081–1107
- 176 Zhang T. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Trans Inform Theory*, 2011, 57: 4689–4708
- 177 Zhang T. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Trans Inform Theory*, 2011, 57: 6215–6221
- 178 Zhang X, Peng D T. Solving constrained nonsmooth group sparse optimization via group capped- ℓ_1 relaxation and group smoothing proximal gradient algorithm. *Comput Optim Appl*, 2022, 83: 801–844
- 179 Zhang Y. Theory of compressive sensing via ℓ_1 -minimization: A non-RIP analysis and extensions. *J Oper Res Soc China*, 2013, 1: 79–105

- 180 Zhang Y, Zhang N, Sun D F, et al. An efficient Hessian based algorithm for solving large-scale sparse group Lasso problems. *Math Program*, 2020, 179: 223–263
- 181 Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann Statist*, 2009, 37: 3468–3497
- 182 Zhao Y B. RSP-based analysis for sparsest and least ℓ_1 -norm solutions to underdetermined linear systems. *IEEE Trans Signal Process*, 2013, 61: 5777–5788
- 183 Zhao Y B. Equivalence and strong equivalence between the sparsest and least ℓ_1 -norm nonnegative solutions of linear systems and their applications. *J Oper Res Soc China*, 2014, 2: 171–193
- 184 Zhao Y B. Optimal k -thresholding algorithms for sparse optimization problems. *SIAM J Optim*, 2020, 30: 31–55
- 185 Zhao Y B, Li D. A theoretical analysis of sparse recovery stability of Dantzig selector and LASSO. *arXiv:1711.03783*, 2017
- 186 Zhao Y B, Luo Z Q. Analysis of optimal thresholding algorithms for compressed sensing. *Signal Process*, 2021, 187: 108148
- 187 Zhao Y B, Luo Z Q. Natural thresholding algorithms for signal recovery with sparsity. *IEEE Open J Signal Process*, 2022, 3: 417–431
- 188 Zhao Y B, Xu C L. 1-Bit compressive sensing: Reformulation and RRSP-based sign recovery theory. *Sci China Math*, 2016, 59: 2049–2074
- 189 Zhou H, Sehl M E, Sinsheimer J S, et al. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 2010, 26: 2375–2382
- 190 Zhou S L, Xiu N H, Qi H D. Global and quadratic convergence of Newton hard-thresholding pursuit. *J Mach Learn Res*, 2021, 22: 599–643
- 191 Zhou Y, Han J H, Yuan X H, et al. Inverse sparse group lasso model for robust object tracking. *IEEE Trans Multimedia*, 2017, 19: 1798–1810
- 192 Zuo W M, Meng D Y, Zhang L, et al. A generalized iterated shrinkage algorithm for non-convex sparse coding. In: *Proceedings of the IEEE International Conference on Computer Vision*. Piscataway: IEEE, 2013, 217–224

Theory and algorithms of structured sparse optimization problems

Yaohua Hu, Yufan Li, Yanyan Liu & Jing Qin

Abstract Structured sparse optimization is one of the most popular research topics in optimization, which has been widely used in a wide range of fields, such as compressed sensing, signal and image processing, machine learning, and biological sciences. In this paper, we first introduce the development of sparse optimization, group sparse optimization, mixed sparse optimization, and joint sparse optimization in recent years. We summarize the theoretical studies and optimization algorithms for structured sparse optimization, including the consistency theory and convergence theory. Finally, we also propose some interesting and important problems in structured sparse optimization that are worthy of further investigation. Generally speaking, a great deal of effort has been achieved in the development of structured sparse optimization; however, there are still several important problems to be resolved.

Keywords sparse optimization, group sparse optimization, consistency theory, first-order algorithm, second-order algorithm, convergence theory

MSC(2020) 65K05, 49M37, 90C26

doi: 10.1360/SSM-2024-0032