arXiv:2301.02347v1 [math.OC] 6 Jan 2023

一种用于非光滑正则化最小二乘的 Levenberg–Marquardt方法

ALEKSANDR Y. ARAVKIN*、ROBERT BARALDI†、和DOMINIQUE ORBAN‡

摘要。 我们提出一种 Levenberg–Marquardt方法，用于最小化光滑非线性最小二乘项 $f(x) = \frac{1}{2}\|F(x)\|_2^2$ 与非光滑项 h之和。其中 f和 h都可以是非凸的。通过使用近端梯度法等一阶方法，最小化正则化线性最小二乘模型与 h 的模型之和，从而计算步长。在假设 F 及其雅可比矩阵利普希茨连续，且 h 适当并下半连续的条件下，我们证明了 Levenberg–Marquardt方法的信赖域变体和正则化变体都全局收敛到一阶驻点。在最坏情况下，两种方法都需要进行$O(\epsilon^{-2})$次迭代，才能使驻定性度量低于 $\epsilon \in (0, 1)$。我们报告了三个示例的数值结果：组Lasso基追踪去噪示例、非线性支持向量机，以及神经元放电中的参数估计。为了使这些示例可实现，我们详细说明了如何计算可分离 h 以及带信赖域约束的组 LASSO 的近端算子。在所有情况下，与采用自适应步长的近端梯度法和拟牛顿信赖域方法相比，Levenberg–Marquardt方法所需的外层迭代更少；这两种方法都没有利用问题的最小二乘结构。我们的结果还表明，需要采用比简单一阶方法更先进的子问题求解器。

关键词。正则化优化、非光滑优化、非凸优化、非线性最小二乘、Levenberg–Marquardt方法、近端梯度方法。

AMS学科分类。49J52,65K10, 90C53, 90C56,

1. 引言。我们考虑如下问题

$$\underset{x}{\text{minimize}} \;\; f(x) + h(x), \qquad f(x) = \tfrac{1}{2}\|F(x)\|_2^2, \tag{1.1}$$

其中， $F : \mathbb{R}^n \to \mathbb{R}^m$ 连续可微，且 $h : \mathbb{R}^n \to \mathbb{R}$ 适当且下半连续；我们允许 h 是非光滑且非凸的。在实践中， f通常是一个数据失配项，而 h 是一个正则化项，旨在促使解具备所需性质，例如稀疏性。非光滑正则化优化领域中研究的许多应用实际上具有结构（1.1），包括基追踪去噪、 [14, 28], 稀疏分解和字典学习[2], 以及稀疏总体最小二乘 [30]。然而，非光滑数值方法既未利用最小二乘结构，也无法适应一般的非光滑正则化项。

我们介绍两种用于（1.1）的方法：一种二次正则化变体和一种信赖域变体，灵感源自 Levenberg [19] 和 Marquardt [21], 的方法，分别称为

*华盛顿大学应用数学系，美国华盛顿州西雅图。电子邮箱：saravkin@uw.edu。

†桑迪亚国家实验室优化与不确定性量化部门，美国新墨西哥州阿尔伯克基，邮政信箱 5800，邮编 87125。电子邮箱： rjbaral@sandia.gov。本研究由美国能源部科学办公室先进科学计算研究办公室的约翰·冯·诺依曼奖学金资助。桑迪亚国家实验室是一所多任务实验室，由桑迪亚国家技术与工程解决方案有限责任公司管理和运营。该公司是霍尼韦尔国际公司的全资子公司，依据合同 DE-NA0003525 为美国能源部国家核安全管理局提供服务。本文介绍了客观的技术成果和分析。文中可能表达的任何主观观点或意见，并不一定代表美国能源部或美国政府的观点。

‡GERAD 与蒙特利尔理工学院数学与工业工程系，加拿大魁北克省。电子邮箱：dominique.orban@gerad.ca。本研究部分由加拿大自然科学与工程研究委员会发现基金资助。

LM 和 LMTR 。通过近似最小化更简单、随迭代变化的高斯-牛顿型非光滑模型来计算步长。我们的算法实现采用一阶方法，例如近端梯度法或 Aravkin 等人 [1], 提出的二次正则化方法，来求解子问题。信赖域方法允许采用任意信赖域范数；在实践中，范数的选择受非凸子问题可解性的影响。对于这两种算法，我们基于一种最优性度量建立了全局收敛性，该度量描述单个近端梯度步所能实现的下降量。此外，我们推导出最坏情况复杂度界：经过 $\mathcal{O}\left(1/\epsilon^2\right)$ 次迭代，可使驻点性度量低于容差 $\epsilon \in (0, 1)$，适用于 LM 和 LMTR；也就是说，目标函数中存在非光滑项时，复杂度界与光滑情形同阶。

我们提供了实现细节，并通过若干数值示例展示了所提方法的性能，包括采用组 LASSO 正则化的基追踪去噪、采用 $\ell_{1/2}^{1/2}$ - 范数正则化的非线性支持向量机，以及一个取自神经元放电 FitzHugh–Nagumo 模型的稀疏参数估计示例。在某些条件下，与以往工作相比，我们的方法表现良好Aravkin 等人 [1]。此外，我们还提供了高效的开源软件实现，将LM 和LMTR 作为 Julia 语言的一个软件包提供 [3]。我们发现，利用最小二乘结构可减少LM 和LMTR 的外层迭代次数，这也是光滑优化中的一项众所周知的优势。其代价是需要大量内层迭代，即耗费大量计算来求解子问题。因此，结果凸显了开发更复杂方法的必要性，以最小化线性最小二乘项与非光滑正则项之和。

相关研究。本研究建立在Aravkin、Baraldi 和 Orban [1]提出的框架之上。我们的信赖域 Levenberg-Marquardt 实现的收敛性和复杂度直接遵循 [1]的一般结果。据我们所知，信赖域文献并未明确涵盖这样的情形：非线性最小二乘光滑目标函数带有惩罚项以外的非光滑正则项，尽管许多应用都具有这种结构。背景介绍和详尽论述请参见[13] 。

大量文献关注凸函数和/或全局利普希茨连续的情形，例如 Cartis 等人 [11], Grapiglia 等人 [17] 及其中的参考文献。由于我们关注的假设显著更弱，因此此处不打算全面综述这部分文献。一阶方法文献中存在许多算法，例如 [12], 但很少有方法能有效利用显著的曲率信息。近端牛顿法 [18] 需要求解非平凡的近端算子，并要求海森矩阵半正定。据我们所知，同时允许 f 和 h 为非凸的参考文献为数不多，其中包括：Li 和 Lin [20],在$f + h$ 具有强制性的假设下设计了近端梯度法的加速算法；Bolte 等人 [8] 针对以下情形设计了一种交替算法：$h(x) = h_1(x_1) + h_2(x_2)$，且 (x_1, x_2) 是 x的一个划分；Stella 等人 [26] 提出了一种名为 PANOC 的线搜索有限内存 BFGS 法；Themelis 等人 [27] 基于前向-后向包络提出了一种名为 ZeroFPR 的非单调线搜索近端拟牛顿法；以及 Bot,等人 [9], 研究了一种带动量的近端法。后三种方法在 $f + h$ 满足 Kurdyka-Lojasiewicz (K-L) 假设时收敛。此外，虽然这些方法都将 (1.1) 作为特殊情形，但很少有方法利用曲率信息，而且没有一种专门针对最小二乘结构。本文提出的算法与 [1], 中的算法一样，无需上述强制性或 K-L 假设。

记号。 我们用 $\|\cdot\|$ 表示 R^n 或 R^m上的一个通用但固定的范数。由该范数定义的单位球记为 B，而 $x + \Delta\mathrm{B}$ 表示以 x 为中心、半径为$\Delta > 0$的球。对于整数 $q \geq 1$， $\|\cdot\|_q$ 是 ℓ_q-范数，B_q 是 ℓ_q-范数下的单位球。若 $A \subseteq \mathrm{R}^n$，则 $\chi(\cdot \mid A)$ 是 A的示性函数，即当 $x \in A$时其值为 0，否则为 $+\infty$ 。除非另有说明，若 A 是一个矩阵，则 $\|A\|$ 表示 A的谱范数，也就是其最大奇异值。我们用 $J(x) : \mathrm{R}^n \to \mathrm{R}^{n\times m}$ 表示 F 在 x处的雅可比矩阵。

2. 背景。

Definition 2.1（限制次微分）。考虑 $\phi : \mathbb{R}^n \to \mathbb{R}$ 和 $\bar{x} \in \mathbb{R}^n$，其中 $\phi(\bar{x}) < \infty$。我们称 $v \in \mathbb{R}^n$ 为 ϕ 在 $\bar{x}$处的正则次梯度，并记作$\hat{} v \in \partial\phi(\bar{x})$，如果

$$\liminf_{x\to\bar{x}} \frac{\phi(x) - \phi(\bar{x}) - v^T(x - \bar{x})}{\|x - \bar{x}\|_2} \geq 0.$$

正则次梯度的集合也称为Fr´echet 次微分。我们称v 为 ϕ 在 $\bar{x}$处的广义次梯度，并记作 $v \in \partial\phi(\bar{x})$，如果存在序列$\{x_k\}$ 和 $\{v_k\}$，使得

$$x_k \to \bar{x}, \quad \phi(x_k) \to \phi(\bar{x}), \quad v_k \in \hat{\partial}\phi(x^k) \text{ and } v^k \to v.$$

广义次梯度的集合称为限制次微分。

Proposition 2.2（*25*，定理 *10.1*）。如果 $\phi : R^n \to R$ 是适当的，并且在 $\bar{x}$处有一个 $\hat{}$ 局部极小值，那么 $0 \in \hat{\partial}\phi(\bar{x}) \subseteq \partial\phi(\bar{x})$。如果 ϕ 是凸的，则后一个条件也是 $\bar{x}$ 为全局极小值的充分条件。如果 $\phi = f + h$，其中 f 在 $\bar{x}$ 的某个邻域上连续可微，且 h 在 $\bar{x}$处有限，那么 $\partial\phi(\bar{x}) = \nabla f(\bar{x}) + \partial h(\bar{x})$。

$\hat{}$若 $0 \in \partial\phi(\bar{x})$，则称 $\bar{x}$ 是 ϕ的一阶驻点。在我们的假设下，

$$x \text{ is first-order stationary for (1.1)} \iff 0 \in J(x)^T F(x) + \partial h(x). \tag{2.1}$$

近端梯度法 [16] 应用于正则化目标函数 $f(x) + h(x)$，其中 f 可微，其迭代定义如下

$$x_{k+1} \in \operatorname*{prox}_{\nu h}(x_k - \nu\nabla f(x_k)) \qquad (k \geq 0), \tag{2.2}$$

其中 $\nu > 0$ 是步长，近端算子定义为

$$\operatorname*{prox}_{\nu h}(y) := \operatorname*{argmin}_u \tfrac{1}{2}\|u - y\|_2^2 + \nu h(u). \tag{2.3}$$

对于 h，在没有进一步假设的情况下，(2.3) 是一个可能为空或包含一个或多个元素的集合。迭代 (2.2) 具有以下下降性质

Lemma 2.3 (8，引理 2)。设 ∇f 是利普希茨连续的，且利普希茨常数为$L \geq 0$， h 是适当下半连续的，且 下确界 $h > -\infty$。设 $x_k \in$ 定义域h， $0 < \nu < 1/L$，并且 x_{k+1} 按照 (2.2) 定义。则，

$$(f + h)(x_{k+1}) \leq (f + h)(x_k) - \tfrac{1}{2}(\nu^{-1} - L)\|x_{k+1} - x_k\|_2^2. \tag{2.4}$$

3. 线性最小二乘。对于固定的 $\sigma \geq 0$ 和 $x \in \mathbb{R}^n$，定义

$$\varphi(s;x) := \tfrac{1}{2}\|J(x)s + F(x)\|_2^2, \tag{3.1a}$$

$$\psi(s;x) \approx h(x+s) \quad \text{with} \quad \psi(0;x) = h(x), \tag{3.1b}$$

$$m(s;x,\sigma) := \varphi(s;x) + \tfrac{1}{2}\sigma\|s\|_2^2 + \psi(s;x). \tag{3.1c}$$

考虑参数化问题及其最优解集

$$p(x,\sigma) := \min_s \; m(s;x,\sigma) \leq \varphi(0;x) + \psi(0;x) = f(x) + h(x) \tag{3.2a}$$

$$P(x,\sigma) := \operatorname{argmin}_s \; m(s;x,\sigma)。 \tag{3.2b}$$

(3.2) 的形式体现了针对 (1.1) 的 Levenberg–马夸特子问题，其中 f 和 h 分别进行建模。

特别地，$\varphi(0;x) = f(x)$ 且 $\nabla_s\varphi(0;x) = \nabla f(x)$。我们作出以下附加假设。

Model Assumption 3.1. 对于任意 $x \in \mathbb{R}^n$，$\psi(\cdot;x)$ 是适当的、下半连续的且近端有界的，即存在 $\lambda_x \in \mathbb{R}_+ \cup \{+\infty\}$，使得 $\psi(\cdot;x) + \frac{1}{2}\lambda_x^{-1}\|\cdot\|_2^2$ 有下界。此外，$\psi(0;x) = h(x)$，且 $\partial\psi(0;x) = \partial h(x)$。

在模型假设 3.1中，我们假设所选的 λ_x 是所有可能取值的上确界，并称其为近端有界性阈值，它对应于 $\psi(\cdot;x)$。特别地，$\psi(\cdot;x)$ 有下界，当且仅当 $\lambda_x = +\infty$。

根据命题 2.2，若 $\sigma \geq \lambda_x^{-1}$，

$$s \in P(x,\sigma) \quad \Longrightarrow \quad 0 \in \nabla\varphi(s;x) + \sigma s + \partial\psi(s;x).$$

我们定义

$$\xi(x,\sigma) := (f+h)(x) - p(x,\sigma). \tag{3.3}$$

由上述定义可直接得到以下驻点性判据。

Lemma 3.1. 设模型假设3.1 成立，且 $\sigma \geq \lambda_x^{-1}$。那么 $\xi(x,\sigma) = 0 \Longleftrightarrow 0 \in P(x,\sigma) \Longrightarrow x$ 是 (1.1)的一阶驻点。此外，x 是(1.1) 的一阶驻点，当且仅当 $s = 0$ 是(3.1c)的一阶驻点。

证明。首先注意到 $\xi(x,\sigma) = 0 \Longleftrightarrow p(x,\sigma) = (f+h)(x) = \varphi(0;x) + \psi(0;x)$，而这当且仅当 $0 \in P(x,\sigma)$ 成立。由命题 2.2 可得 $0 \in \partial m(0;x,\sigma) = \nabla\varphi(0;x) + \partial\psi(0;x)$，这与 (2.1) 等价。 □

下一个结果给出了（3.2）的一些性质。

Proposition 3.2. 设 模型假设 3.1 成立。定义域 p = 定义域 P =定义域 $\psi \times \{\sigma \mid \sigma \geq \lambda_x^{-1}\}$。此外，对于任意 $x \in \mathbb{R}^n$，

1. $p(x,\cdot)$是真下半连续的，且对于每个 $\sigma > \lambda_x^{-1}$，$P(x,\sigma)$非空且紧；2. 若 $\{\sigma_k\} \to \bar\sigma > \lambda_x^{-1}$以使得 $\{p(x,\sigma_k)\} \to p(x,\bar\sigma)$的方式成立，且对于每个 k，$s_k \in P(x,\sigma_k)$，则 $\{s_k\}$ 有界，且其所有极限点均属于 $P(x,\bar\sigma)$；3. $p(x,\cdot)$在任意 $\bar\sigma > \lambda_x^{-1}$处连续，且若 $\bar\sigma > 0$，则第2部分中的 $\{p(x,\sigma_k)\} \to p(x,\bar\sigma)$成立。

证明。第 1–2 部分可通过应用 [25, 定理 1.17] 得到，因为注意到 (3.1c) 关于 (x,σ) 局部一致地水平有界于 s，这是因为 $\psi(\cdot;x) + \frac{1}{2}\lambda_x^{-1}\|s\|_2^2$ 有界，且 $\varphi(s;x) + \frac{1}{2}(\sigma - \lambda_x^{-1})\|s\|_2^2$ 关于 (x,σ) 局部一致地水平有界于 s。第 3 部分也可通过应用 [25, 定理 1.17] 得到，因为注意到 (3.1c) 在任意 $\bar\sigma > \lambda_x^{-1}$ 处于 σ 连续。 □

根据 命题 3.2 第 3 部分， $\xi(x, \cdot)$ 在任意 $\bar{\sigma} > \lambda_x^{-1}$ 处均连续。

尽管 (3.1a) 是关于 f about x的自然模型，但收敛性质可以用更简单的一阶模型来表述

$$\varphi_1(s;x) := f(x) + \nabla f(x)^T s = \tfrac{1}{2}\|F(x)\|_2^2 + (J(x)^T F(x))^T s, \tag{3.4a}$$

$$m_1(s;x,\sigma) := \varphi_1(s;x) + \tfrac{1}{2}\sigma\|s\|^2 + \psi(s;x). \tag{3.4b}$$

近端梯度法的第一步（2.2）应用于同时最小化 $\varphi(s;x) + \psi(s;x)$ 和 $\varphi_1(s;x) + \psi(s;x)$ 时，步长为 $\nu > 0$ 时

$$\begin{aligned} s_1 \in & \operatorname*{prox}_{\nu\psi(\cdot;x)}(-\nu J(x)^T F(x)) \\ &= \operatorname*{argmin}_s \ \tfrac{1}{2}\|s + \nu J(x)^T F(x)\|_2^2 + \nu\psi(s;x) \\ &= \operatorname*{argmin}_s \ (J(x)^T F(x))^T s + \tfrac{1}{2}\nu^{-1}\|s\|_2^2 + \psi(s;x) \\ &= \operatorname*{argmin}_s \ m_1(s;x,\nu^{-1}). \end{aligned} \tag{3.5}$$

若 $\nu^{-1} \geq \sigma$，则 $m_1(s;x,\sigma) \leq m_1(s;x,\nu^{-1})$。因此，若 s_1 由（3.5）得到，它也会使（3.4b）下降。

与 引理 3.1 和 命题 3.2类似，我们可以定义

$$p_1(x,\sigma) := \min_s \ m_1(s;x,\sigma) \leq \varphi_1(0;s) + \psi(0;x) = f(x) + h(x) \tag{3.6a}$$

$$P_1(x,\sigma) := \operatorname*{argmin}_s \ m_1(s;x,\sigma), \tag{3.6b}$$

$$\xi_1(x,\sigma) := (f+h)(x) - p_1(x,\sigma) \geq 0, \tag{3.6c}$$

并且我们有以下结果，说明了 p_1 和 ξ_1相应的性质。证明与命题 3.2 和引理 3.3中的证明相同。

LEMMA 3.3. 设模型假设3.1 成立，且 $\sigma \geq \lambda_x^{-1}$。那么 $\xi_1(x,\sigma) = 0 \iff 0 \in P_1(x,\sigma) \implies x$ 是(1.1)的一阶驻点。此外， x 是(1.1) 的一阶驻点，当且仅当 $s = 0$ 是(3.4b)的一阶驻点。

PROPOSITION 3.4. 设模型假设3.1 成立。$\operatorname{dom} p_1 = \operatorname{dom} P_1 = \operatorname{dom}\psi \times \{\sigma \mid \sigma \geq \lambda_x^{-1}\}$。此外，对于任意 $x \in \mathbb{R}^n$，

1. $p_1(x,\cdot)$是真下半连续的，且对每个 $\sigma > \lambda_x^{-1}$， $P_1(x,\sigma)$非空且紧； *2.* 若 $\{\sigma_k\} \to \bar{\sigma} > \lambda_x^{-1}$以这样的方式使得 $\{p_1(x,\sigma_k)\} \to p_1(x,\bar{\sigma})$，并且对每个 k，$s_k \in P_1(x,\sigma_k)$，那么 $\{s_k\}$ 有界，且其所有极限点都属于 $P_1(x,\bar{\sigma})$； *3.* $p_1(x,\cdot)$在任意 $\bar{\sigma} > \lambda_x^{-1}$处连续，且若 $\bar{\sigma} > 0$，则第2部分中的 $\{p_1(x,\sigma_k)\} \to p_1(x,\bar{\sigma})$成立。

因为 $L = 0$ 对于 φ_1，引理 2.3 意味着由 s_1 所实现的下降量为 $(\varphi_1 + \psi)(s_1; x) \leq (\varphi_1 + \psi)(0; x) - \frac{1}{2}\nu^{-1}\|s_1\|^2$，可将其重新整理为

$$(f+h)(x) - (\varphi_1 + \psi)(s_1;x) \geq \tfrac{1}{2}\nu^{-1}\|s_1\|^2 \geq \tfrac{1}{2}\sigma\|s_1\|^2. \tag{3.7}$$

在 $\psi = 0$的特殊情形下， $s_1 = -\nu^{-1}\nabla f(x)$，因此 (3.7) 化简为

$$\xi_1(x,\sigma) \geq \xi_1(x,\nu^{-1}) \geq f(x) - \varphi_1(s_1;x) \geq \tfrac{1}{2}\sigma\nu^{-1}\|\nabla f(x)\|^2 \geq \tfrac{1}{2}\sigma^2\|\nabla f(x)\|^2,$$

这表明 $\sigma^{-1}(\xi_1(x, \nu^{-1}))^{1/2}$ 可用作驻点性度量。

4. 非线性最小二乘。

4.1. 正则化方法。我们首先考察 Levenberg 和 Marquardt 方法的表述，其中采用模型 (3.1c) 来计算一个步长。具体而言，考虑算法 4.1。步长 s_k 通过近似最小化 (3.1c) 在阶段 7 中计算，但在阶段 $\frac{1}{2}\sigma_k\|s_k\|^2$ 8中衡量步长的质量时，并未考虑正则化项 s_k 。子问题步长可通过继续执行近端梯度法的迭代来计算，该方法以 s_k 初始化1。这便得到 算法 4.1 的一种可能实现。

算法 4.1非光滑正则化 Levenberg-Marquardt 方法。

1：选择常数 $0 < \eta_1 \leq \eta_2 < 1$ 和 $0 < \gamma_3 \leq 1 < \gamma_1 \leq \gamma_2$。2：选择 $x_0 \in \mathbb{R}^n$，其中 h 为有限值，$\sigma_0 > 0$，计算 $F(x_0)$ 和 $h(x_0)$。3：对于 $k = 0, 1, \ldots$，执行 4：选择步长 $\nu_k < 1/(\|J(x_k)\|^2 + \sigma_k)$。5：计算在 (3.5) 中定义的 s_k,1，以及在 (3.6c) 中定义的 $\xi_1(x_k, \nu_k^{-1})$。6：按照 (3.1c) 定义 $m(s; x_k, \sigma_k)$。7：计算 (3.2b) 的一个近似解 s_k 。8：计算比值

$$\rho_k := \frac{f(x_k) + h(x_k) - (f(x_k + s_k) + h(x_k + s_k))}{\varphi(0; x_k) + \psi(0; x_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k))}.$$

9: 如果 $\rho_k \geq \eta_1$，则设置 $x_{k+1} = x_k + s_k$。否则，设置 $x_{k+1} = x_k$。10: 按照以下规则更新正则化参数

$$\sigma_{k+1} \in \begin{cases} [\gamma_3\sigma_k, \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_1\sigma_k] & \text{if } \eta_1 \leq \rho_k < \eta_2, \\ [\gamma_1\sigma_k, \gamma_2\sigma_k] & \text{if } \rho_k < \eta_1. \end{cases}$$

11: **结束循环**

可能会出现 $\sigma_k \leq \lambda_{x_k}^{-1}$的情况。在此情形下，$\psi(s_k;\ x_k) = -\infty$，因此扩展算术的规则意味着 $\rho_k = 0$，无论 $h(x_k + s_k) = +\infty$ 是否有限。因此，s_k将在阶段9被拒绝，而σ_{k+1} 将在阶段10被选为大于 σ_k 。经过有限次这样的增大后，σ_k 将超过 $\lambda_{x_k}^{-1}$，从而得到一个 $\psi(s_k;\ x_k)$有限的步。

我们的主要工作假设如下。

Problem Assumption 4.1. 残差 F 及其雅可比矩阵 J 在$\Omega := \{x \in \mathbb{R}^n \mid (f + h)(x) \leq (f + h)(x_0)\}$ 上有界且利普希茨连续，并且 h 是适当且下半连续的。

虽然问题假设 4.1 对整个$\mathbb{R}^n$ 提出了很强的要求，而且特别排除了线性最小二乘的情形，但它是 Levenberg–Marquardt 方法收敛分析中的常见假设。如果 Ω 是紧集，那么当 F 在 Ω 上 $\mathcal{C}^1$ 时，F 在 Ω 上是利普希茨连续的；并且当 F 在 Ω 上 $\mathcal{C}^2$ 时，F 在 Ω 上是利普希茨连续的。

根据问题假设 4.1，∇f 在 Ω 上是利普希茨连续的，即存在$L > 0$ 使得

$$|f(x + s) - (f(x) + \nabla f(x)^T s)| \leq \tfrac{1}{2}L\|s\|_2^2 \quad \text{for all } x,\ x + s \in \Omega. \tag{4.1}$$

我们强调，在下文中，不需要了解 L，也不需要对其进行估计。我们对模型作出的下一个假设如下。

MODEL ASSUMPTION 4.1. 存在一个常数 $\kappa_{\mathrm{m}} > 0$，使得对于所有 x 和 $s \in \mathbb{R}^n$，$|(f+h)(x+s) - (\varphi+\psi)(s;x)| \leq \kappa_{\mathrm{m}}\|s\|^2$。

模型假设 4.1 本质上是对模型的非光滑部分 ψ 所作的假设。事实上，(3.1a) 和 (4.1) 结合可得

$$\begin{aligned} |f(x+s) - \varphi(s;x)| &\leq |f(x+s) - (f(x) + \nabla f(x)^T s)| + \tfrac{1}{2}\|J(x)s\|^2| \\ &\leq \tfrac{1}{2}(L + \|J(x)\|^2)\|s\|^2. \end{aligned}$$

其中我们使用了 $f(x)$ 的定义、$\nabla f(x) = J(x)^T F(x)$ 恒等式以及 (4.1)。因此，如果 J 在 Ω 上有界，则得到

$$|f(x+s) - \varphi(s;x)| \leq \tfrac{1}{2}(L + \sup_{x\in\Omega} \|J(x)\|^2)\|s\|^2.$$

特别地，模型假设 4.1 在选取 $\kappa_{\mathrm{m}} = \frac{1}{2}(L + \sup_{x\in\Omega}\|J(x)\|^2)$ 时成立：$\psi(s;x) := h(x+s)$。

我们作出以下附加假设，并称 $\{\psi(\cdot;x_k)\}$ 为一致近端有界。

MODEL ASSUMPTION 4.2. 存在 $\lambda > 0$，使得对于所有 $k \in \mathbb{N}$，$\lambda_{x_k} \geq \lambda$。

模型假设 4.2 在以下情况下成立：h 本身是近端有界的，并且我们在$\psi(s;x_k)$处选择：$= h(x_k + s)$，并在每次迭代中如此选择。

我们的第一个结果保证 σ_k 在算法 4.1中有上界。

THEOREM 4.1. 设问题假设 4.1 以及模型假设 3.1、4.1和4.2 均成立，并设

$$\sigma_{\mathrm{succ}} := \max(2\kappa_{\mathrm{m}}/(1-\eta_2),\, \lambda^{-1}) > 0. \tag{4.2}$$

若 x_k 不是一阶驻点且 $\sigma_k \geq \sigma_{\mathrm{succ}}$，则迭代 k 非常成功，并且 $\sigma_{k+1} \leq \sigma_k$。

证明。设 s_k 为算法 k 在迭代算法 4.1中计算得到的步长。若 $\sigma_k < \lambda_{x_k}^{-1}$，$\rho_k = 0$ 如上所述，s_k 将被拒绝，而 σ_k 将增大。因此，我们假设$\sigma_k \geq \lambda^{-1} \geq \lambda_{x_k}^{-1}$。由于 x_k 不是一阶驻点，$s_k \neq 0$。由于 s_k 是 (3.2b) 的近似解，我们必有

$$\varphi(0;x_k) + \psi(0;x_k) \geq \varphi(s_k;x_k) + \tfrac{1}{2}\sigma_k\|s_k\|^2 + \psi(s_k;x_k)$$

因此，

$$\varphi(0;x_k) + \psi(0;x_k) - (\varphi(s_k;x_k) + \psi(s_k;x_k)) \geq \tfrac{1}{2}\sigma_k\|s_k\|^2. \tag{4.3}$$

模型假设 4.1 和（4.3）结合可得

$$|\rho_k - 1| = \frac{|f(x_k+s_k) + h(x_k+s_k) - (\varphi(s_k;x_k) + \psi(s_k;x_k))|}{\varphi(0;x_k) + \psi(0;x_k) - (\varphi(s_k;x_k) + \psi(s_k;x_k))} \leq \frac{2\kappa_{\mathrm{m}}\|s_k\|^2}{\sigma_k\|s_k\|^2}.$$

经 $\|s_k\|^2$ 化简后，我们得到 $\sigma_k \geq \sigma_{\mathrm{succ}} \Longrightarrow \rho_k \geq \eta_2$。 □

注意，定理 4.1 并未在其假设中明确包含问题假设 4.1，不过要使模型假设 4.1 成立，很可能需要这一假设。

有趣的是，定理 4.1 成立时，并不要求步长 s_k 满足充分下降条件。仔细考察证明可知，原因在于：任何使 $m(s;\sigma,x)$ $\varphi(\cdot;x)+\psi(\cdot;x)$ 简单下降的步长，都会使其充分下降，而无论 s_k 是如何计算得到的。

定理 4.1 保证存在一个常数 $\sigma_{\max}>0$ 使得

$$\sigma_k \le \sigma_{\max} := \min(\sigma_0, \gamma_2\sigma_{\text{succ}}) > 0 \quad \text{for all } k \in \mathbb{N}. \tag{4.4}$$

我们的下一个结果讨论有限次成功迭代发生的情形。其证明与 [13, 定理 6.4.4] 和 [1, 定理 3.5]的证明几乎完全相同，故略去。

THEOREM 4.2. 设 问题假设 4.1 以及 模型假设 3.1 和 4.1均成立。若算法 4.1 仅产生有限次成功迭代，则$x_k = x^*$ 对所有充分大的 k， x^* 是一阶临界的。

根据 Rockafellar 和 Wets [25, 定理 1.25]， $p_1(x,\ \sigma)$ 随着 σ 的增大而增大，因此，$\xi_1(x,\ \sigma)$ 随着 σ 的增大而减小。因此，由式 (4.4) 可得

$$\xi_1(x_k,\sigma_k) \ge \xi_1(x_k,\sigma_{\max}) \quad \text{for all } k \in \mathbb{N}. \tag{4.5}$$

引理 3.1、(4.5) 以及 第 3 节 末尾的说明表明，可以采用 $\xi_1(x_k,\ \sigma_{\max})^{\frac{1}{2}}$作为驻点性度量。事实上，对于给定的 $\epsilon > 0$， $\xi_1(x_k,\ \sigma_{\max}) \le \epsilon/\sigma_{\max} \Longrightarrow \sigma_k\xi_1(x_k,\ \sigma_{\max}) \le \epsilon$。

由于我们必须像 ν_k 算法 4 的 算法 4.1第 4 步那样选择 $\xi_1(x_k,\ \nu_k^{-1})$，因此我们计算 $\xi_1(x_k,\ \sigma_k)$，而不是 $0<\theta<1$。具体而言，对于给定的 $0<\theta<1$，我们令

$$\nu_k := \theta/(\|J_k\|^2 + \sigma_k). \tag{4.6}$$

在 问题假设 4.1下，存在 $\kappa_J > 0$ 使得 $\|J(x)\| \le \kappa_J$ 对所有$x \in \Omega$ 都成立。由于 算法 4.1 只生成 $x_k \in\Omega$，上述结论与式 (4.4) 可得

$$\nu_k \ge \theta/(\kappa_J^2 + \sigma_{\max}) := \nu_{\min} > 0 \quad \text{for all } k \in \mathbb{N}. \tag{4.7}$$

因此，对于所有 $k \ge 0$, $\nu_k^{-1} \le \nu^{-1}{}_{\min}$，并且

$$\xi_1(x_k,\nu_k^{-1}) \ge \xi_1(x_k,\nu_{\min}^{-1}) \quad \text{for all } k \in \mathbb{N}. \tag{4.8}$$

对于停止容差 $\epsilon \in (0,1)$，我们寻求确定 $k(\epsilon) \in \mathbb{N}$，使得

$$\xi_1(x_k,\nu_{\min}^{-1})^{\frac{1}{2}} > \epsilon \quad \text{for all } k < k(\epsilon) \quad \text{and} \quad \xi_1(x_{k(\epsilon)},\nu_{\min}^{-1})^{\frac{1}{2}} \le \epsilon. \tag{4.9}$$

定义集合

$$\mathcal{S} := \{k \in \mathbb{N} \mid \rho_k \ge \eta_1\}, \tag{4.10a}$$
$$\mathcal{S}(\epsilon) := \{k \in \mathcal{S} \mid k < k(\epsilon)\}, \tag{4.10b}$$
$$\mathcal{U}(\epsilon) := \{k \in \mathbb{N} \mid k \notin \mathcal{S} \text{ and } k < k(\epsilon)\}. \tag{4.10c}$$

为进行复杂度分析，需要假设阶段 7 的 算法 4.1 中的步计算与 $\xi_1(x_k,\ \sigma_k)$ 相关。我们作出如下假设。

STEP ASSUMPTION 4.1. 存在 $\kappa_{\mathrm{mdc}} \in (0,\ 1)$ 使得 s_k 在阶段 7 的算法 4.1 中计算得到的量满足

$$\varphi(0; x_k) + \psi(0; x_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k)) \geq \kappa_{\mathrm{mdc}} \xi_1(x_k, \nu_k^{-1}). \tag{4.11}$$

步假设 4.1 类似于信赖域方法中所使用的充分下降条件——参见 [13]。Aravkin 等人 [1] 给出了在非光滑正则化优化的信赖域方法中具体使用此类条件的实例。显然，将近端梯度法应用于（3.1c）后经过单步迭代即可满足充分下降假设。因此，在（3.1c）的极小值点处该假设也成立。于是，在 算法 4.1的步骤 7中，一种策略是继续进行近端梯度迭代，直到达到停止条件。

以下结果与 Aravkin 等人 [1], 的结果相平行，而后者又受到Cartis 等人 [11] 及其中参考文献的启发。

LEMMA 4.3. 设问题假设 *4.1* 以及模型假设 *3.1* 和 *4.1* 成立，并根据步假设 *4.1* 计算 s_k，其中根据（*4.6*）选取 ν_k。假设存在无穷多个成功迭代，并且对于所有 $x \in \mathbb{R}^n$，$f(x) + h(x) \geq (f + h)_{\mathrm{low}}$。那么，对于所有 $\epsilon \in (0, 1)$，

$$|\mathcal{S}(\epsilon)| \leq \frac{(f + h)(x_0) - (f + h)_{\mathrm{low}}}{\eta_1 \kappa_{\mathrm{mdc}} \epsilon^2} = O(\epsilon^{-2}). \tag{4.12}$$

证明。 对于 $k \in \mathcal{S}(\epsilon)$， 步假设 4.1 和（4.8）推出

$$\begin{aligned}(f + h)(x_k) - (f + h)(x_k + s_k) &\geq \eta_1(\varphi(0; x_k) + \psi(0; x_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k))) \\ &\geq \eta_1 \kappa_{\mathrm{mdc}} \xi_1(x_k, \nu_k^{-1}) \\ &\geq \eta_1 \kappa_{\mathrm{mdc}} \xi_1(x_k, \nu_{\min}^{-1}) \\ &\geq \eta_1 \kappa_{\mathrm{mdc}} \epsilon^2.\end{aligned}$$

其余证明与 [1, 引理 3.6] 的证明类似。 □

LEMMA 4.4. 在 引理 4.3 的假设下，

$$|\mathcal{U}(\epsilon)| \leq \frac{\log(\sigma_{\max}/\sigma_0)}{\log(\gamma_1)} + |\mathcal{S}(\epsilon)| \frac{|\log(\gamma_3)|}{\log(\gamma_1)} = O(\epsilon^{-2}). \tag{4.13}$$

证明。 对于每个 $k \in \mathcal{U}(\epsilon)$， $\sigma_{k+1} \geq \gamma_1 \sigma_k$；而对于每个 $k \in \mathcal{S}(\epsilon)$， $\sigma_{k+1} \geq \gamma_3 \sigma_k$。因此，若 $k(\epsilon)$ 是首次出现（4.9）的迭代，

$$\sigma_0 \gamma_1^{|\mathcal{U}(\epsilon)|} \gamma_3^{|\mathcal{S}(\epsilon)|} \leq \sigma_{k(\epsilon)-1} \leq \sigma_{\max}.$$

取对数，得

$$|\mathcal{U}(\epsilon)| \log(\gamma_1) + |\mathcal{S}(\epsilon)| \log(\gamma_3) \leq \log(\sigma_{\max}/\sigma_0).$$

重新整理，并回想 $0 < \gamma_3 < 1$，得到（4.13）。 □

结合引理 4.3 和 4.4 可得总体迭代复杂度界。

THEOREM 4.5. 在 引理 4.3的假设下，

$$|\mathcal{S}(\epsilon)| + |\mathcal{U}(\epsilon)| = O(\epsilon^{-2}). \tag{4.14}$$

换言之，定理 4.5 保证要么 $(f + h)(x_k) \to -\infty$，要么 $\liminf_{k \to \infty} \xi_1(x_k,\ \nu^{-1}{}_{\min}) = 0$。

4.2. 信赖域方法。现在，我们将 Aravkin 等人的算法 3.1[1] 应用于 (1.1)。我们假设每个 $f_i : \mathbb{R}^n \to \mathbb{R}$ 都是 $\mathcal{C}^1$，因此满足他们的问题假设 3.1。关于 x 的 f 自然模型是高斯-牛顿模型 (3.1a)，它满足 $\varphi(0; x) = f(x)$ 和 $\nabla_s \varphi(0; x) = \nabla f(x) = J(x)^T F(x)$。要求 $h(x+s)$的模型$\psi(s; x)$满足相同的模型假设 4.1；当 ∇f 利普希茨连续，或每个 f_i 均为 $\mathcal{C}^2$ 且具有有界海森矩阵时，该假设成立。在 Aravkin 等人的 [1, 算法 3.1]中，第一步近端梯度步 s_1 通过求解以下问题计算得出

$$\begin{aligned} &\underset{s}{\text{minimize}} \ \tfrac{1}{2}\|F(x)\|_2^2 + \left(J(x)^T F(x)\right)^T s + \tfrac{1}{2}\nu^{-1}\|s\|^2 + \psi(s; x) \\ &\text{subject to } \|s\| \le \Delta, \end{aligned} \tag{4.15}$$

即，

$$s_1 \in \underset{\nu\psi(\cdot;x)+\chi(\cdot|\Delta\mathbb{B})}{\text{prox}} (-\nu J(x)^T F(x)),$$

其中，对于预设常数 $\alpha > 0$，$0 < \nu < 1/(\|J(x)\|^2 + \alpha^{-1}\Delta^{-1})$。后续步骤继续进行近端梯度迭代，以计算一个近似解

$$\underset{s}{\text{minimize}} \ \tfrac{1}{2}\|J(x)s + F(x)\|_2^2 + \psi(s; x) \quad \text{subject to } \|s\| \le \min(\beta\|s_1\|, \Delta), \tag{4.16}$$

其中 $\beta \ge 1$。以上描述了针对正则化非线性最小二乘问题的一种方法的信赖域变体，该方法由 Levenberg [19]和 Marquardt [21] 提出。由于 $\psi(\cdot; x)$ 是近端有界的这一假设可以去除，因为 $\psi(\cdot; x) + \chi(\cdot \mid \Delta\mathbb{B})$ 总是有下界，因此是近端有界的，且 $\lambda_x = \infty$。(4.16) 的近似解必须满足 步假设 4.1 ，其中 $\xi_1(x, \sigma)$ 替换为

$$\hat{\xi}_1(\Delta; x, \nu) := f(x) + h(x) - \hat{p}_1(\Delta; x, \nu),$$

其中 $\hat{p}_1(\Delta; x, \nu)$ 是 (4.15) 的最优值。

在上述假设下，Aravkin 等人 证明信赖域半径Δ 不会低于阈值

$$\Delta_{\min} := \min\left(\Delta_0, \hat{\gamma}_1 \frac{\kappa_{\text{mdc}}(1-\eta_2)}{2\kappa_{\text{m}}\alpha\beta^2}\right),$$

其中 $\Delta_0 > 0$ 是初始信赖域半径，$\hat{\gamma}_1 \in (0, 1)$ 是在 Δ被拒绝的步上缩减的比例，$\eta_2 \in (0, 1)$ 是阈值，当 Δ 高于该阈值时，会在被接受的步上增大，而 κ_{mdc} 和 κ_{m} 与模型假设 4.1 和Step Assumption4.1中同名常数的作用类似。

Aravkin 等人 使用 $\xi_1(\Delta_{\min}; x, \nu)$ 作为驻点性度量。他们证明，对于任意$\epsilon \in (0, 1)$，达到

$$\hat{\xi}_1(\Delta_{\min}; x, \nu)^{\frac{1}{2}} \le \epsilon$$

所需的迭代次数为 $O(\epsilon^{-2})$，前提是 $f + h$ 有下界。完整细节请参阅 [1]。

5. 近端算子。在算法 4.1 或第 4.2 节的算法中，非光滑项 h 的典型模型为 $\psi(s; x) := h(x+s)$。如果这些算法要使用 Aravkin 等人的二次正则化方法 [1, Algorithm 6.1] 来计算一个步长，那么后者会在每次迭代中构造 $\psi(\cdot; x)$ 的模型。为简化记号，令 $\psi_k(s) := \psi(s; x_k) = h(x_k + s)$ 表示算法 4.1 或第 4.2 节算法在迭代 k 中使用的模型。

5.1. 一般近端算子。在算法 4.1中，子问题目标函数中的非光滑项为 $\psi_k(s)$。关于s_j 的典型模型可化简为$\omega_j(t) = \psi_k(s_j + t) = h(x_k + s_j + t)$，并且，所计算的步长不再采用（5.2），而是

$$t_j \in \operatorname*{argmin}_t \ \tfrac{1}{2}\nu^{-1}\|t - q\|^2 + h(x_k + s_j + t). \tag{5.1}$$

与上文相同的变量替换可得

$$v_j \in \operatorname*{argmin}_v \ \tfrac{1}{2}\nu^{-1}\|v - \bar{q}\|^2 + h(v) = \operatorname*{prox}_{\nu h}(\bar{q}),$$

无论 h 是否可分，均可得到

$$t_j \in \operatorname*{prox}_{\nu h}(\bar{q}) - (x_k + s_j).$$

第 4.2 节算法子问题目标函数中的非光滑项为 $\psi_k(s) + \chi(s;\ \Delta_k)$。对于 [1, 算法 6.1]的迭代点 s_j，用户提供模型 $\omega_j(t) := \omega(t; s_j) \approx \psi_k(s_j + t) + \chi(s_j + t \mid \Delta_k \mathbb{B})$，典型的选择为$\omega_j(t) = \psi_k(s_j + t) + \chi(s_j + t \mid \Delta_k \mathbb{B}) = h(x_k + s_j + t) + \chi(s_j + t \mid \Delta_k \mathbb{B})$。计算得到的步长为 $t_j \in \operatorname{prox}^{\nu\omega_j}(q)$，其中 $\nu > 0$ 和 $q \in \mathbb{R}^n$为某些固定值，即

$$t_j \in \operatorname*{argmin}_t \ \tfrac{1}{2}\nu^{-1}\|t - q\|^2 + h(x_k + s_j + t) + \chi(s_j + t \mid \Delta_k \mathbb{B}). \tag{5.2}$$

变量替换 $v := x_k + s_j + t$ 可将式（5.2）改写为

$$v_j \in \operatorname*{argmin}_v \ \tfrac{1}{2}\nu^{-1}\|v - \bar{q}\|^2 + h(v) + \chi(v - x_k \mid \Delta_k \mathbb{B}), \tag{5.3}$$

其中 $\bar{q} := x_k + s_j + q$，由此可恢复 $t_j = v_j - (x_k + s_j)$。

5.2. 可分离移位近端算子。如果 h 是可分离的，且信赖域由 ℓ_∞-范数定义，则该问题分解为若干子问题，而 v_j 的第 i 个分量为

$$\begin{aligned} v_{j,i} &\in \operatorname*{argmin}_{v_i} \ \tfrac{1}{2}\nu^{-1}(v_i - \bar{q}_i)^2 + h_i(v_i) + \chi(v_i - x_{k,i} \mid [-\Delta_k, \Delta_k]) \\ &= \operatorname*{argmin}_{v_i} \ \tfrac{1}{2}\nu^{-1}(v_i - \bar{q}_i)^2 + h_i(v_i) + \chi(v_i \mid [x_{k,i} - \Delta_k, x_{k,i} + \Delta_k]). \end{aligned} \tag{5.4}$$

可能出现两种情况。在第一种情况下，$x_{k,\,i} - \Delta_k < v_{j,\,i} < x_{k,\,i} + \Delta_k$，因此 $v_{j,\,i} \in \operatorname{prox}^{\nu h_i}(\bar{q}_i)$，即

$$t_{j,i} \in \operatorname*{prox}_{\nu h_i}(\bar{q}_i) - (x_{k,i} + s_{j,i}).$$

在第二种情况下，至少有一个无约束解位于 $[x_{k,i} - \Delta_k$ 之外，$x_{k,\,i} + \Delta_k]$，因此（5.4）的约束全局最小值要么是两个边界中的一个或两个，要么是位于边界之间的无约束局部最小值。

当 h 为凸函数时，约束解是距离唯一无约束全局解最近的可行点，即

$$v_{j,i} \in \operatorname*{proj}_{[x_{k,i} - \Delta_k, x_{k,i} + \Delta_k]} (\operatorname*{prox}_{\nu h_i}(\bar{q}_i)),$$

也就是说，

$$t_{j,i} \in \operatorname*{proj}_{[x_{k,i} - \Delta_k, x_{k,i} + \Delta_k]} (\operatorname*{prox}_{\nu h_i}(\bar{q}_i)) - (x_{k,i} + s_{j,i}).$$

EXAMPLE 5.1 ($\ell_{1/2}^{1/2}$ 伪范数）。考虑 $\psi(s) = \|s\|_{1/2}^{1/2} = \sum_j |s_j|^{1/2}$。当信赖域边界处于非活跃状态时，*Cao等人 [10]*将(5.4) 的解表示为

$$v_{j,i} = \begin{cases} \frac{2}{3}|\bar{q}_i|\left(1+\cos\left(\frac{2}{3}\pi - \frac{2}{3}\mu_\lambda(\bar{q}_i)\right)\right) & \bar{q}_i > p(\lambda) \\ 0 & |\bar{q}_i| \le p(\lambda) \\ -\frac{2}{3}|\bar{q}_i|\left(1+\cos\left(\frac{2}{3}\pi - \frac{2}{3}\mu_\lambda(\bar{q}_i)\right)\right) & \bar{q}_i < -p(\lambda) \end{cases}$$

其中

$$\mu_\lambda(\bar{q}_i) \coloneqq \arccos\left(\frac{\lambda}{4}\left(\frac{|\bar{q}_i|}{3}\right)^{-3/2}\right), \qquad p(\lambda) \coloneqq \frac{54^{1/3}}{4}(2\lambda)^{2/3}.$$

当信赖域约束处于活跃状态时，*Cao等人 [10]* 指出，上式给出了(5.4)的拐点。我们同时检查拐点和边界。如果拐点位于边界范围内，则选择最小值；否则，选择边界处代价函数的最小值。

5.3. 凸 h的非可分移位近端算子。本节将讨论非可分移位近端算子的示例。出发点是（5.3），其中假设 h 是闭的、适当的且凸的。我们将其重写为

$$\chi(v - x \mid \Delta\mathbb{B}) = \sup_z \langle v - x, z\rangle - \sigma_{\Delta\mathbb{B}}(z),$$

为简便起见，我们用 x 和 Δ 代替 x_k 和 Δ_k，其中支撑函数

$$\sigma_{\Delta\mathbb{B}}(z) \coloneqq \sup_d \langle d, z\rangle + \chi(d \mid \Delta\mathbb{B}).$$

代入（5.3）可得到鞍点问题

$$\min_v \sup_z \tfrac{1}{2}\nu^{-1}\|v - \bar{q}\|^2 + h(v) + \langle v - x, z\rangle - \sigma_{\Delta\mathbb{B}}(z). \tag{5.5}$$

（5.5）的目标函数关于 v 是凸的，关于 z是凹的。鞍点条件可写为

$$\begin{aligned} 0 &\in \nu^{-1}(v - \bar{q}) + \partial h(v) + z = \nu^{-1}(v - (\bar{q} - \nu z)) + \partial h(v) \\ 0 &\in v - x - \partial\sigma_{\Delta\mathbb{B}}(z). \end{aligned}$$

第一个条件意味着 $v \in \mathrm{prox}_{\nu h}(\bar{q} - \nu z)$。由 h 的凸性可知， v 是唯一的，因此我们得到

$$0 \in v - x - \partial\sigma_{\Delta\mathbb{B}}(z), \quad \text{where} \quad \operatorname*{prox}_{\nu h}(\bar{q} - \nu z) = \{v\}. \tag{5.6}$$

5.3.1. 特殊情形： ℓ_2-范数。对于 $h(\cdot) \coloneqq \lambda\|\cdot\|_2$,

$$\operatorname*{prox}_{\nu\lambda\|\cdot\|_2}(y) = \begin{cases} 0 & \text{if } \|y\| \le \nu\lambda \\ \left(1 - \frac{\nu\lambda}{\|y\|_2}\right) y & \text{if } \|y\| > \nu\lambda \end{cases}. \tag{5.7}$$

现在我们说明如何通过将5.6转化为标量求根问题来求解 z。

$$\zeta = \zeta(z) \coloneqq \|\bar{q} - \nu z\|_2.$$

有两种可能。
情形 A：若 $\zeta \le \nu\lambda$，(5.7) 得到

$$\operatorname*{prox}_{\nu\lambda\|\cdot\|_2}(\bar{q}-\nu z) = \{v\} = \{0\}.$$

(5.3) 在此情形下的最优值为 $\frac{1}{2}\nu^{-1}\|\bar{q}\|^2$。 –
情形 B：若 $\zeta > \nu\lambda$，(5.7) 得到

$$\operatorname*{prox}_{\nu\lambda\|\cdot\|_2}(\bar{q}-\nu z) = \{v\} = \left\{\left(1-\frac{\nu\lambda}{\zeta}\right)(\bar{q}-\nu z)\right\}, \tag{5.8}$$

且 (5.6) 变为

$$\begin{aligned} 0 &\in x - \left(1-\frac{\nu\lambda}{\zeta}\right)(\bar{q}-\nu z) + \partial\sigma_{\Delta\mathbb{B}}(z) \\ &= (\zeta-\nu\lambda)\frac{\nu}{\zeta}\left(z - \left(\frac{1}{\nu}\bar{q} - \frac{\zeta}{\nu(\zeta-\nu\lambda)}x\right)\right) + \partial\sigma_{\Delta\mathbb{B}}(z), \end{aligned}$$

我们将其解释为

$$z = z(\zeta) := \operatorname*{prox}_{\frac{\zeta}{\nu(\zeta-\nu\lambda)}\sigma_{\Delta\mathbb{B}}}\left(\frac{1}{\nu}\bar{q} - \frac{\zeta}{\nu(\zeta-\nu\lambda)}x\right). \tag{5.9}$$

回顾一下， [6,定理 6.46]

$$\operatorname*{prox}_{\alpha\sigma_{\Delta\mathbb{B}}}(y) = y - \alpha\operatorname*{proj}_{\Delta\mathbb{B}}(\alpha^{-1}y), \quad (\alpha > 0). \tag{5.10}$$

因此，必须能够计算到 ΔB 的投影。在我们的实现中，我们使用 B =B∞。

现在我们可以搜索满足以下条件的 ζ

$$g(\zeta) := \zeta - \|\bar{q}-\nu z(\zeta)\|_2 = 0. \tag{5.11}$$

由于投影到凸集的映射是利普希茨连续的，因此 g 在 $(\nu\lambda,+\infty$ 上也是如此).
由于 (5.3) 是强凸的，因此存在唯一解，所以 g 至多有一个根满足 $\zeta > \nu\lambda$。任何这样的 g 根都会给出 v，由 (5.8) 给出，以及 $z(\zeta$，由 (5.9) 给出，二者共同满足 (5.6)。如果 g 不存在这样的根，则必然出现情形 A。
(5.9) 与 (5.10) 联立可得

$$\bar{q}-\nu z(\zeta) = \frac{\zeta}{\zeta-\nu\lambda}\left[x + \operatorname*{proj}_{\Delta\mathbb{B}}\left(\frac{\zeta-\nu\lambda}{\zeta}\bar{q} - x\right)\right]. \tag{5.12}$$

由于 $\zeta\uparrow\infty$，$(\zeta-\nu\lambda)/\zeta\uparrow 1$，且根据连续性，(5.12) 中方括号内的项收敛到 $x+\mathrm{proj}_{\Delta\mathbb{B}}(\bar{q}-x)$。因此， $\|\bar{q}-\nu z(\zeta)\|_2 \to \|x+\mathrm{proj}_{\Delta\mathbb{B}}(\bar{q}-x)\|_2$，并且当 ζ充分大时，必有 $g(\zeta) > 0$。

为了研究 $g(\zeta)$ 在 $\zeta\downarrow\nu\lambda$ 时的情况，我们考虑几个互斥情形。
1. 若 $x \notin \Delta\mathbb{B}$，则 $\mathrm{proj}_{\Delta\mathbb{B}}(-x) \neq -x$。由于 $\zeta\downarrow\nu\lambda$，$(\zeta-\nu\lambda)/\zeta\downarrow 0$，根据连续性，方括号内的项收敛到 $x+\mathrm{proj}_{\Delta\mathbb{B}}(-x) \neq 0$。因此， $\|\bar{q}-\nu z(\zeta)\|_2 \to \infty$，并且对于充分小的 ζ，必有$g(\zeta) < 0$。

2. 接下来考虑 $x \in \operatorname{int}\Delta\mathbb{B}$ 的情况。当 ζ 足够接近 $\nu\lambda$ 时，

$$\operatorname*{proj}_{\Delta\mathbb{B}}\left(\frac{\zeta-\nu\lambda}{\zeta}\bar{q}-x\right)=\frac{\zeta-\nu\lambda}{\zeta}\bar{q}-x, \tag{5.13}$$

且 $\bar{q}-\nu z(\zeta)=\bar{q}$，即 $z(\zeta)=0$。在这种情况下，

(a) 若 $\|\bar{q}\|_2>\nu\lambda$，则对于足够接近 $\nu\lambda$ 的 ζ， $g(\zeta)<0$；

(b) 若 $\|\bar{q}\|_2\le\nu\lambda$，则对所有 $\zeta>\nu\lambda$， $g(\zeta)>0$；

3. 若 $\|x\|_\infty=\Delta$ 且投影$_{\Delta\mathbb{B}}(\bar{q}-x)=-x$，则投影$_{\Delta\mathbb{B}}(\alpha\bar{q}-x)=-x$ 对任意$\alpha>0$。在这种情况下，(5.12) 中方括号内的项始终为零，且 $\bar{q}-\nu z(\zeta)=0$。因此，对所有 $\zeta>\nu\lambda$， $g(\zeta)=\zeta>0$。

4. 如果 $\|x\|_\infty=\Delta$ 但 $\operatorname{proj}_{\Delta\mathbb{B}}(\bar{q}-x)\neq-x$，则存在两种可能的情况。射线 $\alpha\bar{q}-x$ 与 $\operatorname{int}\Delta\mathbb{B}$ 相交，或者不相交。如果相交，那么对于所有 5.13) 都成立，只要 ζ 充分接近 $\nu\lambda$，$\bar{q}-\nu z(\zeta)=\bar{q}$，且情形 2a–2b 适用。如果不相交，由利普希茨连续性可得

$$\left\|x+\operatorname*{proj}_{\Delta\mathbb{B}}\left(\frac{\zeta-\nu\lambda}{\zeta}\bar{q}-x\right)\right\|_2=\left\|\operatorname*{proj}_{\Delta\mathbb{B}}\left(\frac{\zeta-\nu\lambda}{\zeta}\bar{q}-x\right)-\operatorname*{proj}_{\Delta\mathbb{B}}(-x)\right\|_2\le\frac{\zeta-\nu\lambda}{\zeta}\|\bar{q}\|_2.$$

因此， $\|\bar{q}-\nu z(\zeta)\|_2\le\|\bar{q}\|_2$，并且

(a) 若 $\|\bar{q}\|_2>\nu\lambda$，则 $g(\zeta)\ge\zeta-\|\bar{q}\|_2>0$ 对于 $\zeta>\|\bar{q}\|_2$成立，因此在 $(\nu\lambda,\|\bar{q}\|_2]$ 中可能存在一个根。根据 (5.12)，以及 $\|y\|_2\le\sqrt{n}\|y\|_\infty$对所有 y均成立这一事实，我们还有

$$\|\bar{q}-\nu z(\zeta)\|_2\le\frac{\zeta}{\zeta-\nu\lambda}\left(\|x\|_2+\left\|\operatorname*{proj}_{\Delta\mathbb{B}}\left(\frac{\zeta-\nu\lambda}{\zeta}\bar{q}-x\right)\right\|_2\right)\le\frac{(\|x\|_2+\Delta\sqrt{n})\zeta}{\zeta-\nu\lambda},$$

因此，对于 $\zeta>\nu\lambda+2\Delta\sqrt{n}$， $g(\zeta)>0$。于是，搜索区间可能可以缩小为 $(\nu\lambda,\ \min(\nu\lambda+\|x\|_2+\Delta\sqrt{n},\ \|\bar{q}\|_2)]$。

(b) 如果 $\|\bar{q}\|\le\nu\lambda$，那么对于所有 $\zeta>\nu\lambda$， $g(\zeta)>0$。

因此，在1 和2a的情况下，保证在 $(\nu\lambda,\ +\infty)$ 中存在一个根，并且可以用二分法找到该根。通过观察 (5.12) 可得上界：

$$\|\bar{q}-\nu z(\zeta)\|\le\frac{\zeta}{\zeta-\nu}(\|x\|+\Delta),$$

因此

$$g(\zeta)=\zeta-\|\bar{q}-\nu z(\zeta)\|\ge\zeta-\frac{\zeta}{\zeta-\nu\lambda}(\|x\|+\Delta),$$

并且一旦满足 $\zeta>\|x\|+\Delta+\nu\lambda$， $g(\zeta)>0$。

在1的情况下，对 (5.12) 应用反三角不等式即可得到下界：

$$\|\bar{q}-\nu z(\zeta)\|\ge\frac{\zeta}{\zeta-\nu\lambda}(\|x\|-\Delta),$$

因此，只要 $\zeta<\nu\lambda+\|x\|-\Delta$， $g(\zeta)<0$。

在情形 2a中，下界就是 $\|\bar{q}\|$。

在情形 2b、3 和 4b中，区间 $(\nu\lambda,\ +\infty)$ 内不可能有根，因此必须出现情形 A。

只有情形 4a 需要进行根搜索，无论是否发生符号变化。如果搜索区间内不存在根，则必然出现情形 A。

5.3.2. 特殊情形：组 LASSO。 组 Lasso 惩罚项是子向量的 ℓ_2-范数之和：

$$R_g(x) = \sum_i \|x_{[i]}\|_2,$$

其中 $x_{[i]}$ 将 x 划分为互不重叠的组。 的近端算子 R_g由对每个子向量应用（5.7）构成：

$$\operatorname*{prox}_{\lambda R_g}(z)_{[i]} = \left(1 - \frac{\lambda}{\|z_{[i]}\|_2}\right)_+ z_{[i]}. \tag{5.14}$$

因此，上一节的策略可以应用于每个组。

6. 实现与数值实验。 我们对 [1] 的算法 3.1 和用于（1.1）的算法 4.1 进行了实现，采用Aravkin、Baraldi 和 Orban的二次正则化方法 R² 来计算步长。R² 可视为一种采用自适应步长的近端梯度法实现。信赖域变体使用 $\Delta_0 = 1$，并在满足以下条件时立即终止外层迭代：$\xi(\Delta_k; x_k, \nu_k)^{1/2} < \epsilon_a + \epsilon_r \xi_{1\,0}^{1/2}$，，其中 $\epsilon_a > 0$ 和 $\epsilon_r > 0$ 分别是绝对容差和相对容差，而 $\xi_{1,0}$ 是 ξ_1 在第一次迭代时的观测值。一轮内层迭代在满足以下条件时终止

$$\hat{\xi}_1(x_k + s, \hat{\sigma}_k) \leq \begin{cases} 10^{-1} & \text{if } k = 0, \\ \max(\epsilon, \min(10^{-1}, \xi_1(x_k, \sigma_k)/10)) & \text{if } k > 0, \end{cases} \tag{6.1}$$

ˆ其中，ˆσ_k 和 ξ_1 分别是 R2 内部使用的正则化参数和一阶驻点性度量。在算法 4.1 中，我们使用 $\sigma_0 = 0.01$；由于 $\sigma_{\max}$ 未知，当 $\xi_1(x_k, \sigma_k)^{1/2} < \epsilon$ 满足容差 $\epsilon > 0$ 时，我们立即终止外层迭代。内层迭代以与 (6.1) 相同的方式停止。所有算法均使用 Julia 语言 [7] 1.8 版实现，属于 RegularizedOptimization.jl 软件包[3]。移位近端算子在 ShiftedProximalOperators.jl 软件包 [5], 中实现，而测试问题则收录于 RegularizedProblems.jl 软件包 [4]中。与 Aravkin 等人 [1], 的数值结果不同，测试案例均明确实现为非线性最小二乘问题，可访问残差 $F(x)$及其雅可比矩阵，而不只是 $f(x)$的梯度：$= \frac{1}{2}\|F(x)\|_2^2$。雅可比-向量乘积和转置雅可比-向量乘积分别通过手动实现或借助前向 [24]和反向 [23] 自动微分计算。

我们将结果与R² 以及 Aravkin 等人 [1], 提出的TR拟牛顿信赖域方法进行比较，该方法没有利用（1.1）的结构。信赖域采用 ℓ_∞-范数定义，二次模型使用内存为 5 的有限内存 SR1 海森矩阵近似。在所有实验中，我们使用 $\psi(s; x) := h(x + s)$。

四种方法之间难以直接比较，因为 LM 和 LMTR 使用的梯度并不相同；它们使用的是雅可比-向量乘积和转置雅可比-向量乘积。为了进行有意义的比较，我们在下表中列出：1) 目标函数（或残差）评估次数；2) 梯度评估次数（对于 R² 和 TR）；3) 转置雅可比-向量乘积次数（对于 LM 和 LMTR，列在梯度评估项下）；4) 求解时间（秒）。理由如下：LM 和 LMTR 向 R²传入一个模型，其目标函数评估需要一次 Jv，而梯度计算需要一个 Jv 和一个 J^Tv。但请注意，后者 Jv 可以缓存并重复使用。因此，R²每次迭代需要一次 Jv，并且每次成功迭代还需要额外一次 J^Tv。

在图中，我们绘制了下降量随残差/目标函数评估次数的变化情况。

以下数值结果的总结是：利用最小二乘结构可大幅减少外层迭代次数。然而，使用 R2 等一阶方法求解子问题会消耗许多 $J^T v$。因此，我们的实验凸显了采用专门针对 (3.1c) 和 (4.16) 的更复杂子问题求解器的必要性。

6.1. 组 LASSO。 在组 LASSO 问题中，我们从线性系统 $b = Ax_T + \varepsilon$ 中观测到含噪数据，其中 $A \in \mathbf{R}^{m\times n}$ 具有正交归一行，并且 x_T 被划分为 g 组，每组中的所有元素均设为 $\{-1, 0, 1\}$ 中的某个值。组 LASSO 问题可表示为

$$\min_x \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_{1,2}, \tag{6.2}$$

其中 $h(x) = \|x\|_{1,\ 2} = \sum_{i=1}^{g} \|x_{[i]}\|_2$，即各组 ℓ_2范数之和。所有元素均为零的组标记为“非活跃”，而设为 ± 1的组则为“活跃”。令 $m = 512$，$n = 200$ 且 $\lambda = 10^{-2}$。我们指定 $g = 5$ 个可能的组（每组含 32 个元素）中的 16 个为“活跃”。噪声 $\varepsilon \sim \mathcal{N}(0,\ 0.01)$。因此，(6.2) 具有 (1.1) 的形式，其中 $F(x) = Ax - b$。我们将绝对和相对退出容差均设为 10^{-4}。每次外层迭代的子问题迭代次数上限为 100。

图 6.1 展示了各算法的解，而 表 6.1报告了相关统计数据。所有算法都得到了近似相同的解。R2 所需的函数评估次数最多，其他算法所需次数则大致相同。 表 6.1 显示，近端算子评估次数与梯度/Jacobian-向量评估次数之间存在权衡。 TR 需要进行大量近端迭代，而 LMTR 和 LM 所需次数则少得多。下文的测试案例将进一步体现这种权衡。

TABLE 6.1 *Group-LASSO (6.2) statistics for* R2, TR, LM, *and* LMTR, *and* $h(x) = \|x\|_{1,\ 2}$. *The* $\#\nabla f$ *is the number of* $J^T v$ *for* LM *and* LMTR.

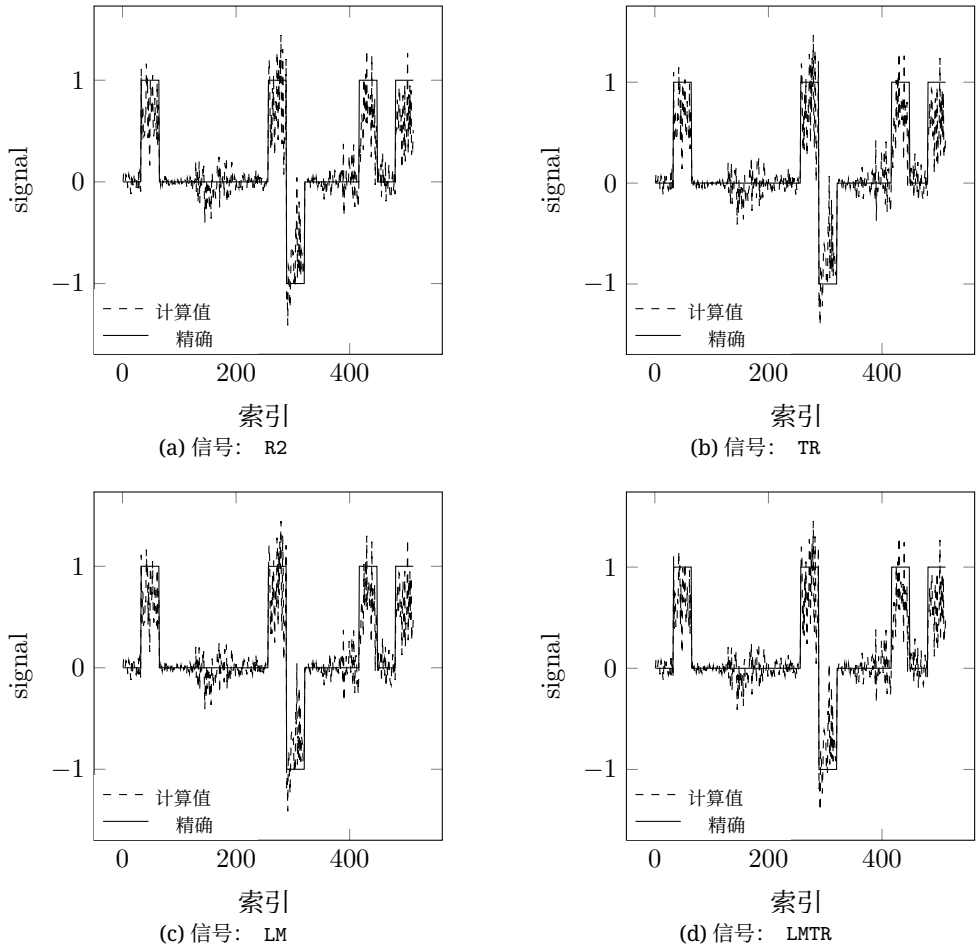
Alg	$f(x)$	$h(x)$	$(f+h)(x)$	$\|x - x_T\|_2$	$\#\ f$	$\#\ \nabla f$	# prox	t (s)
R2	0.00	0.26	0.27	0.45	113	67	113	0.02
TR	0.00	0.26	0.27	0.47	17	17	339	2.56
LM	0.00	0.26	0.27	0.46	10	647	265	0.05
LMTR	0.00	0.26	0.27	0.46	5	327	130	0.98

我们还绘制了图 6.4a中的下降历史。这些图大致相似，其中信赖域方法 TR 和 LMTR 表现最佳。

6.2. 非线性支持向量机。 现在，我们求解形式为（1.1）的图像识别问题，其中

$$F(x) = \mathbf{1} - \tanh(b \odot \langle A, x\rangle), \quad \mathbf{1} = [1, \ldots, 1]^T, \tag{6.3}$$

$A \in \mathbf{R}^{m\times n}$, $n = 784$ 是向量化后的图像尺寸，训练集中的图像数量为 $m = 13007$，测试集中的图像数量为 $m = 2163$，而 $\odot$ 表示向量之间的逐元素乘积



(a) 信号： R2 (b) 信号： TR

(c) 信号： LM (d) 信号： LMTR

FIG. 6.1. *Group-LASSO* (6.2) *solutions with* R2, TR, LM, *and* LMTR *with* $h = \lambda\|\cdot\|_1$, 2。

。我们希望使用这个非线性支持向量机，将 MNIST 数据集中的数字分类为 1 或 7，并移除所有其他数字。此外，我们还要求支持集是稀疏的，因此使用 $h(x) = \|x\|_{1/2}^{1/2}$ 作为正则化项。因此，我们的整体问题是

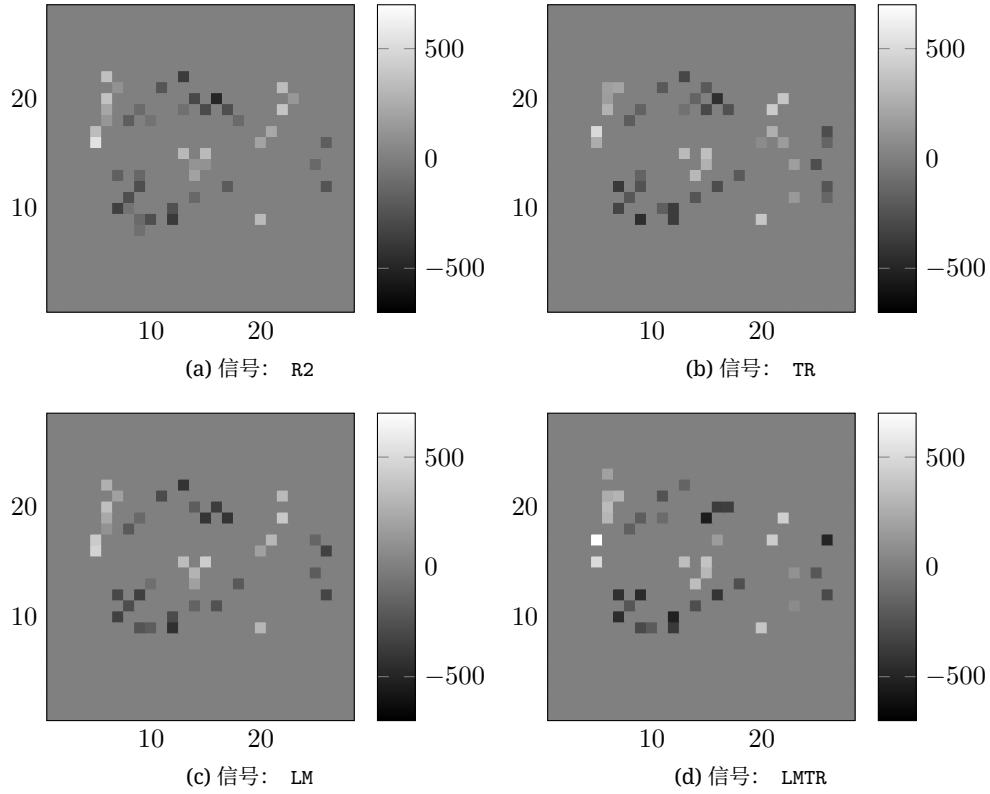
$$\min_x \ \tfrac{1}{2}\|\mathbf{1} - \tanh(b \odot \langle A, x\rangle)\|^2 + \lambda\|x\|_{1/2}^{1/2} \tag{6.4}$$

其中 $\lambda = 10^{-1}$。我们将问题初始化为 $x = \mathbf{1}^n$，使得约 50% 的数据被错误分类。我们再次将停止容差设为 10^{-4}，并将最大内层迭代次数设为 100。

图 6.2 展示了各算法的解映射，可将其理解为对判断图像究竟是 1 还是 7 最重要的像素。所有算法都产生了稀疏解；支持向量中只有约 8% 的像素非零。该问题规模很大且为非凸问题，因此，最终解中有一些像素是共有的，但总体而言彼此不同。这一点可从表 6.3所报告的统计数据中看出。 R2 仍然需要最多的函数评估。 TR 所需数量

约为 LM 和 LMTR 的 10 倍。我们再次观察到，近端算子评估次数与梯度/Jacobian-向量评估次数之间存在权衡。这里，近端算子评估的开销低于梯度或 Jv评估，因此 LM 和 LMTR 的墙钟时间更长。

我们绘制下降历史与函数/残差迭代次数之间的关系，见图 6.4b。在这里可以看到，`LM` 和 `LMTR` 在下降表现方面最佳。



(a) 信号： `R2` (b) 信号： `TR`

(c) 信号： `LM` (d) 信号： `LMTR`

FIG. 6.2。 *Nonlinear SVM* (6.4) *solutions with* `R2`, `TR`, `LM`, `LMTR`.

TABLE 6.3
Nonlinear SVM (6.4) *statistics for* `R2`, `TR`, `LM`, *and* `LMTR`. *Training/test error is with respect to the* ℓ_2*-norm.*

Alg	f	h	$f+h$	(训练，测试)	# f	# ∇f	# prox	t (秒)
R2	57.11	66.28	123.39	(99.80, 99.35)	1359	1085	1359	18.99
TR	49.80	72.37	122.17	(99.83, 99.26)	267	171	10478	6.62
LM	54.36	65.86	120.21	(99.83, 99.35)	23	3567	1276	24.98
LMTR	49.43	68.26	117.69	(99.81, 99.12)	24	3925	1420	44.32

6.3. FitzHugh-Nagumo 反问题。该问题具有形式 (1.1)，其中 $F\colon \mathrm{R}^5 \to \mathrm{R}^{2n+2}$ 定义为 $F(x) = (v(x) - \bar{v}(\bar{x})\ w(x) - \bar{w}(\bar{x}))$，其中 $v(x) = (v_1(x) \ldots, v_{n+1}(x))$ 和 $w(x) = (w_1(x) \ldots, w_{n+1}(x))$ 是离散化函数 $V(t;\ x)$ 和 $W(t;\ x)$ 的采样值，它们满足 FitzHugh 和 Nagumo 等人的神经元激活模型

$$\frac{\mathrm{d}V}{\mathrm{d}t} = (V - V^3/3 - W + x_1)x_2^{-1}, \quad \frac{\mathrm{d}W}{\mathrm{d}t} = x_2(x_3 V - x_4 W + x_5), \tag{6.5}$$

参数化为 x。采样通过对时间区间进行离散化来定义$t \in [0, 20]$ 以及初始条件（$V(0)\ W(0)$）$= (2,\ 0)$。数据（$\bar{v}(x)\ \bar{w}(x)$）通过求解（6.5）生成，其中 $\bar{x} = (0,\ 0。2,\ 1,\ 0,\ 0)$，对应于对范德波尔 [29] 振子的模拟。在我们的实验中，我们使用 $n = 100$ 并求解

$$\min_x \ \tfrac{1}{2}\|F(x)\|_2^2 + \lambda\|x\|_1, \tag{6.6}$$

其中 $h(x) = \lambda\|x\|_1$ 使用 $\lambda = 10$ 来在参数中施加稀疏性。我们的绝对停止准则为 10^{-2}，而相对停止准则设为 10^{-4}。

每个求解器得到的解见表 6.5TR 具有正确的非零参数，但数值偏差更大。相应的模拟结果见图 6.3；每种方法都能够拟合数据。

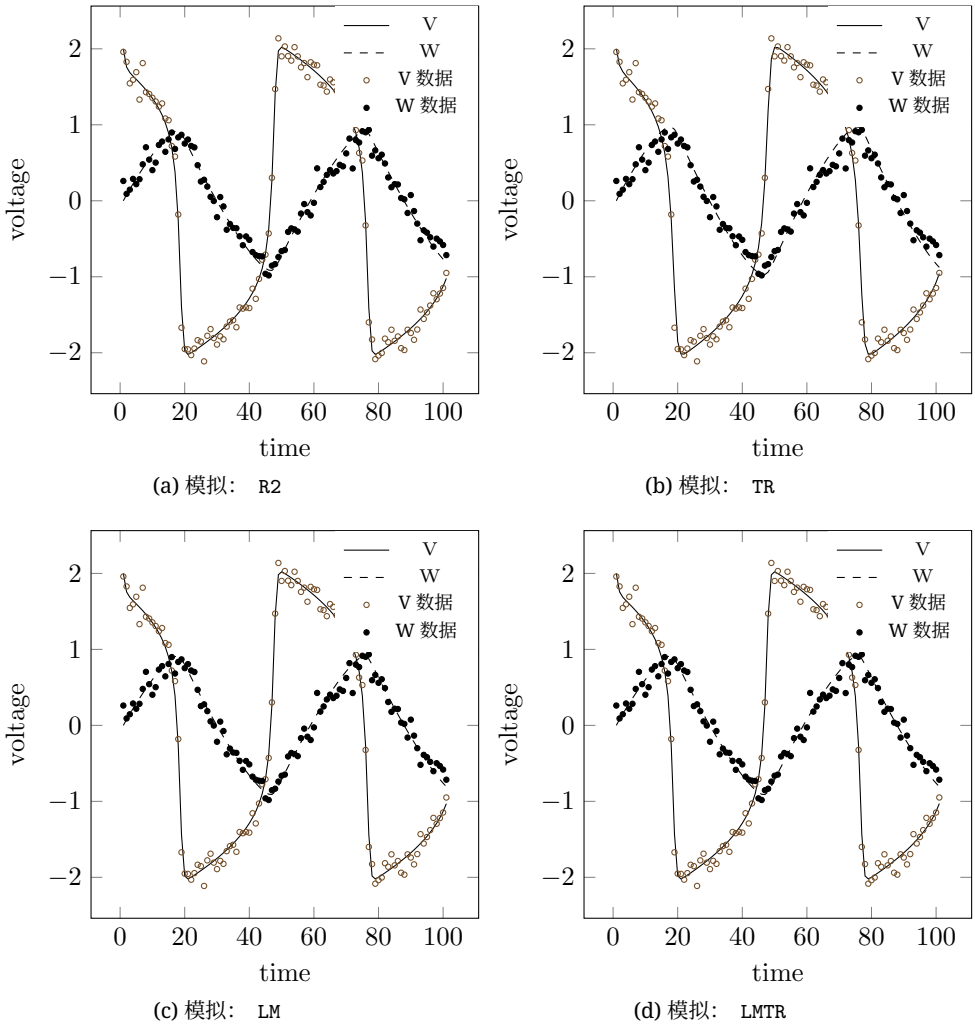
TABLE 6.5
Final parameters for the FH problem (6.6) *found by* R2, TR, LM, *and* LMTR.

True	R2	TR	LM	LMTR
0.00	0.00	0.00	0.00	0.00
0.20	0.26	0.33	0.25	0.25
1.00	0.84	0.70	0.86	0.85
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00

表 6.7 报告了各算法的统计数据，结果呈现出与此前相同的模式。最终目标函数值相当接近。LMTR 所需的目标函数评估次数最少，而 LM 求解 (6.6) 的难度较大。由于 (6.6) 中光滑项的梯度并非利普希茨连续，我们不得不为 $\sigma_{\min}$ 和 R^2 以及 LM 设置较大的，这增加了迭代次数。与支持向量机示例类似，我们可以看到 LM 和 LMTR 耗时比TR更长，这同样是因为在此示例中，计算近端算子比计算 Jv 乘积便宜得多。值得注意的是，TR 似乎对数据的拟合较差，但正则化项的值更低。

最后，图 6.4c 展示了目标函数值随目标函数迭代次数的下降情况。LMTR 再次表现最佳，而 LM 和 TR 在这一指标上表现相近。这再次说明了目标函数、梯度和近端算子计算开销之间的权衡。在 TR 和 R^2 中，近端算子计算开销高会成为限制因素；全变分正则化可视为一个测试案例，因为其近端算子本身就是一个最小化问题。

7. 讨论。与光滑优化类似，利用 f 的最小二乘结构可以显著减少外层迭代次数。挑战在于



(a) 模拟： R2　　(b) 模拟： TR

(c) 模拟： LM　　(d) 模拟： LMTR

FIG. 6.3。 *Simulation of the FH problem* (6.6) *solutions found by* R2, TR, LM, LMTR.

正如我们的数值结果所凸显的，这也是当前研究的主题：对于相关的 ψ，要么找出 (3.1c) 的闭式极小化解，要么设计出能够以更少的转置雅可比-向量乘积生成比 R^2 质量更高的步长的方法。只要子问题求解器生成的步长满足步假设 4.1，我们的收敛性质和最坏情况复杂度界就能得到保证。因此，任何步长计算机制的改进都会立即提升整体求解器的效率。在当前研究中，我们正在探索其他改进，包括对 f 和 ∇f进行非精确求值、采用非单调方法，以及对近端算子进行非精确求值。

参考文献

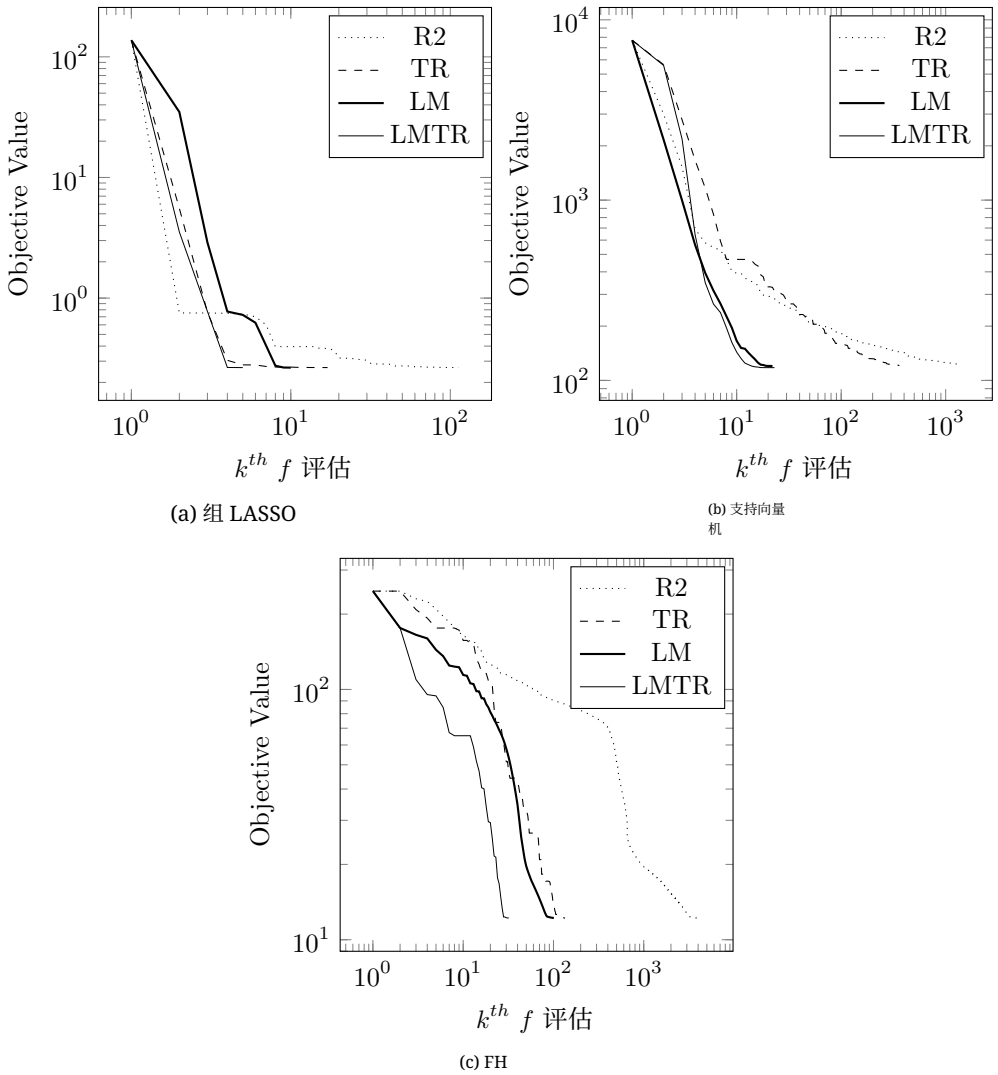
[1] A. Aravkin、R. Baraldi 和 D. Orban。一种用于非光滑正则化优化的近端拟牛顿信赖域方法。*SIAM J. Optim.*，(2):900–929，2022。

TABLE 6.7
Statistics for the FH problem (6.6) for R2, TR, LM, *and* LMTR.

Alg	f	h	$f+h$	$\|x-x_T\|_2$	# f	# ∇f	# 近端	t (s)
R2	1.24	10.91	12.15	1.58	4230	3428	4230	40.40
TR	1.87	10.31	12.17	1.93	134	77	2452	0.67
LM	1.20	11.03	12.23	1.55	101	4236	1402	20.17
LMTR	1.20	11.02	12.22	1.55	32	2006	741	10.50

[2] F. Bach、R. Jenatton、J. Mairal 和 G. Obozinski。*Optimization with Sparsity-Inducing Penalties* 4 2012 ,

*Foundations and Trends in Machine Learning*卷。now出版社，。[3] R. Baraldi 和 D. Orban。RegularizedOptimization.jl：用于正则化优化的算法。https://github.com/JuliaSmoothOptimizers/RegularizedOptimization.jl，2022年2月。[4] R. Baraldi 和 D. Orban。RegularizedProblems.jl：正则化优化的测试案例。https://github.com/JuliaSmoothOptimizers/RegularizedProblems.jl，2022年2月。[5] R. Baraldi 和 D. Orban。ShiftedProximalOperators.jl：用于正则化优-化的近端算子。https://github.com/JuliaSmoothOptimizers/ShiftedProximalOperators.jl，2022年2月。[6] A. Beck。*First Order Methods in Optimization*。工业与应用数学学会，费城，美国，2017年。[7] J. Bezanson、A. Edelman、S. Karpinski 和 V. B. Shah。Julia：一种数值计算的新方法。*SIAM Rev.*，59(1)：65–98，2017年。[8] J. Bolte、S. Sabach 和 M. Teboulle。用于非凸和非光滑问题的近端交替线性化最小化。*Math. Program.*，(146)：459—494，2014年。[9] R. I. Bot、E. R. Csetnek 和 S. L´aszl´o。用于最小化两个非凸函数之和的惯性前向–后向算法。*EURO J. Comput. Optim.*，(4)：3–25，2016年。[10] W. Cao、J. Sun 和 Z. Xu。利用 lq（q = 12，23）正则化的闭式阈值公式快速进行图像反卷积。*Journal on visual communication and image representation*，24 1 2013( )，。[11] C. Cartis、N. I. M. Gould 和 Ph. L. Toint。复合函数最小化的评估复杂度及其在非凸非线性规划中的应用。*SIAM J. Optim.*，21(4)：1721–1739，2011年。[12] P. L. Combettes 和 J.-C. Pesquet。信号处理中的近端分裂方法，载于*Fixed-point algorithms for inverse problems in science and engineering* 185 212 ，第 – 页。施普林格，2011年。[13] A. R. Conn、N. I. M. Gould 和 Ph. L. Toint。*Trust-Region Methods*。MOS-SIAM优化系列第1卷。工业与应用数学学会，费城，美国，2000年。[14] D. L. Donoho。压缩感知。*IEEE T. Inform. Theory*，52(4)：1289–1306，2006年。[15] R. FitzHugh。神经膜阈值现象的数学模型。*B. Math.* , ( )：–。*Biophys.* 17 4 257 278, 1955
[16] M. Fukushima 和 H. Mine。针对某些非凸最小化问题的广义近端点算法。*Int. J. Syst. Sci.*，12(8)：989–1000，1981年。[17] G. Grapiglia、J. Yuan 和 Y. Yuan。非线性步长控制算法：一阶和二阶最优性的复杂度界。*J. Optim. Theory and Applics.*，(171)：980—997，2016年。[18] J. D. Lee、Y. Sun 和 M. A. Saunders。用于最小化复合函数的近端牛顿型方法。*SIAM J. Optim.*，24(3)：1420–1443，2014年。[19] K. Levenberg。求解某些最小二乘问题的方法。*Q. Appl. Math.*，(2)：164–168，1944年。[20] H. Li 和 Z. Lin。用于非凸规划的加速近端梯度法，载于*Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* 15, 379 387, 2015 ，NIPS'页–，美国马萨诸塞州剑桥，。麻省理工学院出版社。[21] D. W. Marquardt。非线性参数最小二乘估计的一种算法。*Journal of the Society for Industrial and Applied Mathematics* 11 2 431 441, 1963 , ( )：–。
[22] J. Nagumo、S. Arimoto 和 S. Yoshizawa。模拟神经轴突的活跃脉冲传输线。*Proceedings of the IRE*，50(10)：2061–2070，1962年。[23] J. Revels。Julia 的反向模式自动微分。https://github.com/JuliaDiff/ReverseDiff.jl，2022年。



(a) 组 LASSO

(b) 支持向量机

(c) FH

FIG. 6.4。 *Objective decrease per objective or residual evaluation.*

[24] J. Revels、M. Lubin 和 T. Papamarkou。Julia 中的前向模式自动微分，2016年。https://arxiv.org/abs/1607.07892。[25] R. Rockafellar 和 R. Wets。 *Variational Analysis*，第317卷。施普林格出版社，1998年。[26] L. Stella、A. Themelis、P. Sopasakis 和 P. Patrinos。一种简单高效的非线性模型预测控制算法。载于 *2017 IEEE 56th Annual Conference on Decision and Control (CDC)* 1939 1944, 2017 ，第 – 页。

[27] A. Themelis、L. Stella 和 P. Patrinos。两个非凸函数之和的前向-后向包络：更多性质与非单调线搜索算法。 *SIAM J. Optim.*，28(3)：2274–2303，2018年。[28] R. Tibshirani。通过 Lasso 进行回归收缩与变量选择。 *J. Roy. Statist. Soc. Ser. B*，58(1)：267–288，1996年。[29] B. Van der Pol。Lxxxviii. 关于“弛豫振荡”。 *The London Edinburgh and Dublin , ,, Philosophical Magazine and Journal of Science* 2 11 978 992, 1926 ( )： –。[30] H. Zhu、G. Leus 和 G. B. Giannakis。针对扰动的稀疏性感知总最小二乘法

压缩采样。 *IEEE T. Signal Proces.*，59(5):2002–2016，2011。