用于认证最优 k-稀疏广义线性模型的适用于 GPU 且线性收敛的一阶方法

Jiachang Liu[*,†]、Andrea Lodi[‡]和 Soroosh Shafiee[†]

[*]企业与社会数据科学中心，康奈尔大学，伊萨卡，美国
[†]运筹学与信息工程学院，康奈尔大学，伊萨卡，美国
[‡]雅各布斯以色列理工–康奈尔研究所，康奈尔理工学院与以色列理工学院，纽约，美国，
{jiachang.liu, andrea.lodi, shafiee}@cornell.edu

arXiv:2603.01306v1 [math.OC] 1 Mar 2026

摘要

我们研究通过基数约束实现稀疏性的广义线性模型（GLM）的最优性认证问题。分支定界（BnB）框架可以利用透视松弛认证最优性，但现有求解这些松弛问题的方法计算开销很大，限制了其可扩展性。为应对这一挑战，我们将松弛问题重新表述为复合优化问题，并构建了一个兼具线性收敛性和计算效率的统一近端框架。在特定的几何正则性条件下，我们的分析揭示了原始二次增长与对偶二次衰减之间的联系，从而得出误差界，使芬切尔对偶间隙成为衡量解集逼近进展的精确指标。由此，我们提出一种基于对偶间隙的重启方案，可将广泛一类次线性收敛近端方法升级为具有理论保证的线性收敛方法，且其适用范围不限于稀疏广义线性模型。对于隐式透视正则化项，我们还推导出专用算法，可在对数线性时间内精确计算该正则化项及其近端算子，避免使用计算开销高昂的通用锥优化求解器。由此得到的迭代主要由矩阵-向量乘法构成，从而支持 GPU 加速。在合成数据集和真实世界数据集上的实验表明，对偶界计算速度提升了数个数量级，并显著增强了分支定界在大型实例上的可扩展性。

1 引言

稀疏广义线性模型（GLMs）是机器学习（ML）和统计学中的重要工具，广泛应用于医疗保健、金融、工程和科学领域。本文研究基数约束广义线性模型的全局优化；其正式表述见第 2.3 节。虽然许多可扩展方法依赖凸替代函数（例如 LASSO）或其他启发式方法，但在特征维度很高或特征高度相关的情况下，这些近似方法可能产生较差的解 [Liu 等人，2022b， 2023]。这一局限在医疗保健等高风险应用中尤为突出，因为准确率、可靠性和可解释性至关重要。因此，我们着重寻求精确ℓ_0约束形式的可认证最优解。此类问题属于 NP 难问题，标准做法是采用分支定界（BnB）求解混合整数规划，这需要在搜索树的每个节点计算有效下界。然而，标准的大 M 松弛通常较弱，导致下界不够紧且剪枝速度缓慢。为解决这一问题，透视松弛 [Ceria 和 Soares，1999，G¨unl¨uk 和 Linderoth，2010，Atamt¨urk 和 G´omez， 2023] 已成为强化节点松弛并提升 BnB 性能的常用方法。透视松弛产生的下界比标准

大 M 松弛 [Liu 等人，2025，引理 2。1] 更紧，并且通常可以表述为锥凸规划。

然而，大规模求解透视松弛颇具挑战。一种标准做法是使用内点法（IPM），但它可扩展性不佳，因为它依赖于求解线性方程组，其复杂度关于变量个数为三次，并且难以在 GPU 等现代硬件上轻松并行化。此外，IPM 无法有效热启动。由于这些局限，人们越来越有兴趣使用一阶方法 [Hazimeh 等人，2022年，Liu 等人， 2023], 来求解此类松弛，这些方法更具可扩展性，且易于热启动。然而，为了真正从 GPU 中获益，这些一阶方法必须避免无法并行化的昂贵子程序。目前尚不清楚如何设计能够为透视松弛充分利用 GPU 并行的此类一阶方法。

一阶方法的另一个缺点是收敛缓慢。从理论角度看，人们对这些方法的收敛速率缺乏了解，尤其是在获得分支定界中剪枝所需的安全下界 方面。由于迭代方法仅产生近似解，任意迭代点处的目标值是松弛最小值的上界，而不是原问题的有效下界。虽然已有技术可从不精确解推导安全下界，但这些界的收敛速率尚未得到讨论。这一点至关重要，因为获得准确下界的速度直接决定分支定界算法的效率。理想情况下，人们希望达到线性收敛速率，但尚不清楚使用一阶方法求解这些透视松弛时是否能够达到此类收敛速率。

上述计算挑战以及收敛速率方面的理论差距，共同引出了本文的核心问题，即：

核心研究问题

我们能否通过设计一种具有可证明线性收敛性的 *GPU* 友好型一阶方法，在分支定界过程中求解透视松弛，并获得有效下界？

我们对此问题的回答是肯定的。下面列出我们为肯定回答这一问题而取得的主要技术贡献。

◇ 复合重构： 我们将透视松弛重构为一个无约束凸复合问题，并引入一种由透视函数隐式定义的新型非光滑正则项。

◇ 几何分析： 我们分析了一般的复合结构，并揭示了两种有利的几何性质。具体而言，在一定正则性条件下，原始问题满足二次增长条件，而对偶问题满足二次衰减条件；这是我们提出的术语，用以表示原始二次增长的对偶对应概念。

◇ 线性收敛重启方案：利用这些几何性质，我们提出一种新颖的基于间隙的重启方案。这一通用框架广泛适用于透视松弛之外的凸复合问题，可加速一大类别的近端方法，包括固定步长、基于线搜索和无线搜索变体，从而使原始问题与对偶目标函数及序列均实现可证明的线性收敛速率。

◇ 稀疏广义线性模型的高效实现： 我们证明，隐式透视函数正则项满足所需的几何条件，并表明其函数值和近端算子均可精确且高效地在对数线性时间内计算。这使得无需求解代价高昂的锥子问题，即可使用高效、适用于 GPU 的一阶方法。

◇ 实证性能： 我们在合成数据集和真实世界数据集上验证了该方法的实际效率。在计算对偶界以及为大规模稀疏广义线性模型认证最优解时，该方法在中央处理器上实现了 1 至 2 个数量级的显著加速，并在 GPU 上进一步提升了一个数量级。

与 ***Liu*** 等人的关系 *[2025]*。Liu 等人 [2025]研究了根节点处透视松弛的近端方法。本文在该会议版本的基础上，研究每个节点处的透视松弛，加入上述几何分析和线性收敛重启方案，并报告了用于最优性认证的*GPU* 加速分支定界实验。

1.1 相关工作

机器学习中的混合整数规划。混合整数规划（MIP）已应用于医疗评分系统 [Ustun 和 Rudin，2016年、2019年，Liu 等人，2022年a], 投资组合优化 [Bienstock，1996年，Wei 等人， 2024], 非线性辨识系统 [Bertsimas 和 Gurnee，2023年，Liu 等人， 2023],决策树 [Bertsimas 和 Dunn，2017年，Hu 等人， 2019], 生存分析 [Zhang 等人，2023年，Liu 等人， 2024], 层次模型 [Bertsimas 和 Van Parys， 2020年a], 回归与分类模型 [Atamturk 和 G´omez，2020年，Bertsimas 和 Van Parys，2020年a，Bertsimas 等人，2020年，Bertsimas 和 Van Parys，2020年b，Hazimeh 和 Mazumder，2020年，Xie 和 Deng，2020年，Atamt¨urk 等人，2021年，Dedieu 等人，2021年，Hazimeh 等人，2022年，Liu 等人，2023年，Guyard 等人，2024], 图模型 [Manzour 等人，2021年，Kucukyavuz 等人， 2023], 可解释性 [Dash 等人，2018年，Lodi 和 Ram´ırez-Ayerbe， 2024],以及异常值检测 [G´omez，2021年，G´omez 和 Neto， 2025]。大多数研究旨在获得高质量的可行解，只有少数关注最优性认证 [Tillmann 等人， 2024]；我们关注稀疏广义线性模型中可扩展的最优性认证。

透视公式。类似我们的方法（第 2.3 节）这样的透视松弛，其应用可追溯至 Ceria 和 Soares [1999]。[G¨unl¨uk 和 Linderoth（2010）、Xie 和 Deng（2020年）、Wei 等人（2022年）、Bacci 等人（2024年）、Shafiee 和 Kılınc-Karzan 2024],针对可分函数提出了透视公式； [Atamt¨urk 和 G´omez（2023年）、Wei 等人（2020年、2022年）、Han 和 G´omez（2025年）、Shafiee 和 Kılınc-Karzan 2024]则针对秩一函数提出了相关公式。我们的方法采用了稀疏广义线性模型中由 Tikhonov 正则化项产生的可分透视公式。我们基于透视函数引入了一个新的隐函数，并推导出其 Fenchel 共轭、近端算子、函数求值及几何性质。

下界计算。商业混合整数规划求解器通常使用外逼近法 [Kelley, 1960] （通过割平面）迭代线性化目标函数，并求解由此得到的线性规划 [Schrijver, 1998, Wolsey, 2020]。然而，这种方法往往会产生较松的下界，尤其是在未生成高质量割时。另一种做法是使用 IPM 求解锥凸松弛 [Dikin, 1967年, Renegar, 2001, Nesterov and Nemirovskii, 1994], ，但内点法的可扩展性有限。近期方法采用一阶方法，包括次梯度下降 [Bertsimas et al., 2020], 交替方向乘子法（ADMM） [Liu et al., 2023], 以及坐标下降 [Hazimeh et al., 2022]。我们的工作沿着这一方向展开。我们的隐式表述支持使用近端梯度法，其低成本子程序可在 GPU 上高效并行。至关重要的是，我们为 BnB 中每个节点所导出的安全下界建立了可证明的全局线性收敛率，而以往针对这一问题的一阶方法并未提供这一保证。

近端梯度法 对于含光滑项的凸复合问题，近端梯度下降（PGD）的收敛速率为次线性，即 $O(1/k)$ [Chen 和 Rockafellar，1997年，Combettes 和 Wajs， 2005]。应用 Nesterov 加速可得到快速迭代收缩阈值算法（FISTA），其收敛速率得到提升，为 $O(1/k^2)$ [Nesterov，1983年，

Beck 和 Teboulle，2009年，Tseng， 2008]。近期的免参数变体，尤其是自适应梯度下降（AdProxGD） [Malitsky 和 Mishchenko， 2024] 以及自适应复合快速梯度法（AC-FGM） [Li 和 Lan， 2025], ，无需手动调节步长。尽管这些速率在一般情况下已是最优的，但在强凸性、Polyak-Lojasiewicz 条件 [Karimi 等人， 2016], 或二次增长条件 [Luo 和 Tseng，1993年，Drusvyatskiy 和 Lewis， 2018]等严格正则性条件下，一阶方法可以线性 收敛。在二次增长条件下，重启方案可以加速动量方法，方案既包括启发式方案 [O'Donoghue 和 Candes， 2015] ，也包括更系统化的方案 [Fercoq 和 Qu， 2019] ，后者基于近端梯度映射或函数值 [Roulet 和 d'Aspremont，2017]。我们提出一种基于对偶间隙的重启方案；与仅依赖原始问题信息的现有方案不同，该方案适用范围广，并能实现线性收敛。

与基于对偶的策略的关系。 我们使用对偶间隙 进行重启，这在理念上与 FDPG [Beck 和 Teboulle，2014], 一脉相承，但提供了新的视角并提高了收敛速率。FDPG 将快速迭代收缩阈值算法应用于对偶问题，使对偶目标函数达到 $\mathcal{O}(1/k^2)$，但恢复得到的原始问题仅达到 $\mathcal{O}(1/k)$。我们的做法恰好相反：优化原始问题，并构造对偶序列以监测对偶间隙。借助基于间隙的重启策略，我们确保原始序列和对偶序列均 线性收敛，优于 FDPG 的次线性速率。此外，我们不假设强凸性，这使对偶目标函数具有全局光滑性。

GPU 加速。 近期研究利用 GPU 加速连续优化问题，包括线性规划 [Applegate 等人, 2021年, Lu 等人, 2024], 二次规划 [Lu 和 Yang，2025], 以及半定规划 [Han 等人，2024]。在离散问题中利用 GPU 的一种自然方式，是在混合整数规划求解器中使用基于 GPU 的线性规划，De Rosa 等人 [2024] 在聚类问题中展示了这种方法。然而，他们的方法可能需要迭代生成大量割平面，以近似原始目标函数。与之不同，我们开发了一种定制的近端梯度法，可直接处理非线性目标函数；由于计算主要涉及矩阵向量乘法，因此适用于 GPU。其他一阶方法，例如 ADMM [Liu 等人，2023] 和坐标下降 [Hazimeh 等人，2022], 则不太适合 GPU：ADMM 需要求解线性方程组，而坐标下降本质上是串行的。

2 预备知识和问题表述

在本节中，我们回顾在分支定界框架内分析稀疏广义线性模型的透视松弛所需的数学背景。我们首先介绍全文使用的标准记号与凸分析概念。随后，我们陈述稀疏广义线性模型问题以及在分支定界树的每个节点上求解的透视松弛。接下来，我们将透视松弛改写为一个无约束复合优化问题。最后，我们给出其Fenchel对偶，它构成我们基于对偶间隙的重启方案的基础。

2.1 记号

在全文中，我们在标准欧氏空间 $\mathbb{R}^p$中进行讨论，其中两个向量 $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^p$ 的内积记为 $\boldsymbol{\beta}^\top \boldsymbol{\gamma}$，$\boldsymbol{\beta}$ 的（欧氏）范数记为$\|\boldsymbol{\beta}\|_2$。扩展实直线定义为 $\mathbb{R} := \mathbb{R} \cup \{-\infty, +\infty\}$。对任意 $p \in \mathbb{N}$，我们定义$[p] := \{1, \ldots, p\}$。我们还对任意满足 $p_2 \geq p_1$的 $p_1, p_2 \in \mathbb{N}$定义 $[p_1 : p_2] := \{p_1, \ldots, p_2\}$。有限集合 $\mathcal{J}$ 中元素的个数记为 $|\mathcal{J}|$。对于集合 $\mathcal{J} \subseteq [p]$，子向量 $\boldsymbol{\beta}_{\mathcal{J}}$由 $\boldsymbol{\beta} \in \mathbb{R}^p$ 中以 $\mathcal{J}$为下标的分量组成。例如，子向量$\boldsymbol{\beta}_{[p_1:p_2]}$ 由 $\boldsymbol{\beta}$ 中从下标 p_1 到 p_2的分量组成。我们分别用 $\mathbf{1}_p$ 和 $\mathbf{0}_p$ 表示 $\mathbb{R}^p$中的全1向量和全0向量，并在维数从

上下文中明确时省略下标。对于向量 $\boldsymbol{\beta} \in \mathbb{R}^p$，我们用记号 $|\beta_{(1)}| \geq \cdots \geq |\beta_{(p)}|$ 表示其各分量的绝对值按非递增顺序排列（并列时任意打破）。等价地，$\beta_{(j)}$ 表示 $\boldsymbol{\beta}$ 的一个分量，其绝对值在 $\{|\beta_1|, \ldots, |\beta_p|\}$ 中为第 j 大。给定集合 $\mathcal{C} \subseteq \mathbb{R}^p$，我们分别用 $\mathrm{int}(\mathcal{C})$ 和 $\mathrm{rint}(\mathcal{C})$ 表示 $\mathcal{C}$ 的内部和相对内部。$\boldsymbol{\beta} \in \mathbb{R}^p$ 与集合 $\mathcal{C}$ 之间的距离函数定义为 $\mathrm{dist}(\boldsymbol{\beta}, \mathcal{C}) := \inf_{\boldsymbol{\gamma} \in \mathcal{C}} \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2$。两个集合 $\mathcal{B}$, $\mathcal{C} \subseteq \mathbb{R}^p$ 之间的距离定义为 $\mathrm{dist}(\mathcal{B}, \mathcal{C}) = \inf_{\boldsymbol{\beta} \in \mathcal{B}} \mathrm{dist}(\boldsymbol{\beta}, \mathcal{C})$。我们用 $\delta_{\mathcal{C}}(\boldsymbol{\beta})$ 表示集合 $\mathcal{C}$ 的示性函数，其定义为：若 $\boldsymbol{\beta} \in \mathcal{C}$，则 $\delta_{\mathcal{C}}(\boldsymbol{\beta}) = 0$；否则为 $= +\infty$。

2.2 凸分析背景

对于扩展实值函数 $h: \mathbb{R}^p \to \mathbb{R}$，其（有效）定义域定义为 $\mathrm{dom}(h) := \{\boldsymbol{\beta} \in \mathbb{R}^p \mid h(\boldsymbol{\beta}) < +\infty\}$。$h$ 的上图集定义为集合 $\mathrm{epi}(h) = \{(\boldsymbol{\beta}, t) \mid h(\boldsymbol{\beta}) \leq t\}$。若函数 h 不取 $-\infty$ 且定义域非空，则称其为**真函数**。若其上图集为闭集，则称其为**闭函数**。h 的 ***Fenchel* 共轭**定义为

$$h^*(\boldsymbol{\alpha}) := \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \boldsymbol{\beta}^\top \boldsymbol{\alpha} - h(\boldsymbol{\beta}),$$

h 的**近端算子**定义为

$$\mathrm{prox}_h(\boldsymbol{\beta}) := \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^p} \frac{1}{2}\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2 + h(\boldsymbol{\gamma}).$$

若 $\mathrm{int}(\mathrm{dom}(h))$ 非空，h 在 $\mathrm{int}(\mathrm{dom}(h))$ 上可微，并且对于收敛至 $\mathrm{int}(\mathrm{dom}(h))$ 边界点的任意序列 $\{\boldsymbol{\beta}^t\} \subset \mathrm{int}(\mathrm{dom}(h))$，$\lim_{t \to +\infty} \|\nabla h(\boldsymbol{\beta}^t)\|_2 = +\infty$，则称函数 h 为**本质光滑**。若函数 h 的梯度 ∇h 存在，且在 $\mathrm{int}(\mathrm{dom}(h))$ 上局部 Lipschitz连续，则称其为**局部光滑**，也就是说，对于每个紧集 $\mathcal{B} \subset \mathrm{int}(\mathrm{dom}(h))$，都存在常数 $L_{\mathcal{B}} > 0$，使得对于所有 $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathcal{B}$，$\|\nabla h(\boldsymbol{\beta}) - \nabla h(\boldsymbol{\gamma})\|_2 \leq L_{\mathcal{B}} \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2$。

函数 h 是**分段线性二次**（PLQ）函数，当且仅当定义域 h 可以表示为有限个多面体集的并，且在每个多面体集上，h 都是二次函数；也就是说，对于并集中的每个多面体集 $\mathcal{P}$，都存在一个对称矩阵 $\boldsymbol{Q} \in \mathbb{R}^{p \times p}$、一个向量 $\boldsymbol{b} \in \mathbb{R}^p$ 和一个标量 $c \in \mathbb{R}$，使得对所有 $\boldsymbol{\beta} \in \mathcal{P}$，$h(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{Q} \boldsymbol{\beta} + \boldsymbol{b}^\top \boldsymbol{\beta} + c$。真凸函数 h 在点 $\boldsymbol{\beta} \in \mathbb{R}^p$ 处的次微分 记为 $\partial h(\boldsymbol{\beta})$，定义为满足以下条件的 $\boldsymbol{s} \in \mathbb{R}^p$ 的集合：对所有 $\boldsymbol{\gamma} \in \mathbb{R}^p$，$h(\boldsymbol{\gamma}) \geq h(\boldsymbol{\beta}) + \boldsymbol{s}^\top (\boldsymbol{\gamma} - \boldsymbol{\beta})$。

真、闭且凸的函数 h: $\mathbb{R}^p \to \mathbb{R}$ 称为相对于向量 $\boldsymbol{\nu} \in \mathbb{R}^p$ **强凸**，如果倾斜函数 $h_{\boldsymbol{\nu}}(\boldsymbol{\beta}) := h(\boldsymbol{\beta}) - \boldsymbol{\nu}^\top \boldsymbol{\beta}$ 满足**二次增长条件**。也就是说，对于任意紧集 $\mathcal{B} \subset \mathbb{R}^p$，存在常数 $\kappa > 0$ 使得

$$h_{\boldsymbol{\nu}}(\boldsymbol{\beta}) \geq (\inf h_{\boldsymbol{\nu}}) + \frac{\kappa}{2} \mathrm{dist}^2\left(\boldsymbol{\beta}, (\partial h_{\boldsymbol{\nu}})^{-1}(\mathbf{0})\right) \qquad \forall \boldsymbol{\beta} \in \mathcal{B}.$$

若对于任意使 $(\partial h_{\boldsymbol{\nu}})^{-1}(\mathbf{0})$ 非空的 $\boldsymbol{\nu} \in \mathbb{R}^p$，$h$ 都是相对于该 $\boldsymbol{\nu} \in \mathbb{R}^p$ 强凸的，则称函数 h 为 **强凸**。

2.3 稀疏广义线性模型、分支定界与透视松弛

我们的目标是求解以下稀疏广义线性模型优化问题，并达到全局最优性。 :

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ f(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{y}) + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \ : \ \|\boldsymbol{\beta}\|_\infty \leq M, \ \|\boldsymbol{\beta}\|_0 \leq k \right\}, \tag{1}$$

其中，$\boldsymbol{X} \in \mathbb{R}^{n \times p}$（含 $\boldsymbol{X} \neq \mathbf{0}$）和 $\boldsymbol{y} \in \mathbb{R}^n$ 分别表示特征矩阵和标签向量。这里，$M > 0$ 是用户指定的盒约束参数，$f: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ 是凸且可微的，$k \in \mathbb{N}$ 限制非零系数的数量，$\lambda_2 > 0$ 是正则化参数。

我们使用辅助二元变量 $\boldsymbol{z} \in \{0,1\}^p$ 和大M重构，得到以下混合整数非线性规划（MINLP）问题：

$$\min_{(\boldsymbol{\beta},\boldsymbol{z})\in\mathbb{R}^{2p}} \left\{ f(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{y}) + \lambda_2\|\boldsymbol{\beta}\|_2^2 \ : \ \boldsymbol{z}\in\{0,1\}^p,\ \mathbf{1}^\top\boldsymbol{z}\le k,\ |\beta_j|\le Mz_j,\ \forall j\in[p] \right\}. \tag{2}$$

从总体上看，混合整数 MIP 求解器使用分支定界（BnB）来解决这类问题：通过将部分z_j 变量固定为 $0/1$，递归地划分搜索空间。

在分支定界树的每个节点 $\mathcal{N}$，都可以利用透视松弛，通过求解以下凸松弛来得到更紧的下界：

$$\min_{(\boldsymbol{\beta},\boldsymbol{z})\in\mathcal{D}} \ f(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{y}) + \lambda_2 \sum_{j\in[p]} \beta_j^2/z_j, \tag{3}$$

其中，节点 $\mathcal{N}$ 处的定义域 $\mathcal{D}$ 定义为

$$\mathcal{D} := \left\{ (\boldsymbol{\beta},\boldsymbol{z})\in\mathbb{R}^{2p} \ \middle| \ \begin{array}{l} \mathbf{1}^\top\boldsymbol{z}\le k, \quad |\beta_j|\le Mz_j,\ \forall j\in[p], \\ z_j=0,\ \forall j\in\mathcal{J}_0, \quad z_j=1,\ \forall j\in\mathcal{J}_1, \quad z_j\in[0,1],\ \forall j\in\mathcal{J}_f, \end{array} \right\}. \tag{4}$$

这里，二元变量 $\boldsymbol{z}$ 被松弛为在 $[0,1]$, 中取连续值，而项 β_j^2/z_j 在 $\beta_j = z_j = 0$ 时定义为 **0**，在 $\beta_j \neq 0$ 且 $z_j = 0$时定义为 $+\infty$ 。在分支定界树中的任意节点 $\mathcal{N}$ 处，集合$\mathcal{J}_0$、 $\mathcal{J}_1$和 $\mathcal{J}_f$ 构成 $[p]$ 的一个划分，分别对应满足 $z_j = 0$、 $z_j = 1$和$z_j \in [0,1]$的指标。如果 $\mathcal{J}_0 = \mathcal{J}_1 = \emptyset$，则问题 (3) 对应于分支定界树的根节点。

2.4 复合重构与 Fenchel对偶性

我们首先将透视松弛 (3) 重构为一个无约束优化问题。为此，我们定义隐式正则化函数 $g_{\mathcal{N}}$: $\mathbb{R}^p \to \mathbb{R}$ 为

$$g_{\mathcal{N}}(\boldsymbol{\beta}) := \inf_{\boldsymbol{z}\in\mathbb{R}^p} \left\{ \frac{1}{2}\sum_{j\in[p]} \beta_j^2/z_j \ : \ (\boldsymbol{\beta},\boldsymbol{z})\in\mathcal{D} \right\}, \tag{5}$$

其中，我们采用空集上的下确界为 $+\infty$ 的约定。函数 $g_{\mathcal{N}}$涵盖了透视项 β_j^2/z_j 以及节点 $\mathcal{N}$处关于$\boldsymbol{z}$ 的基数和分支约束。因此，我们可以将 $g_{\mathcal{N}}$ 视为正则化项，并将 (3) 重写为以下复合形式：

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \Phi(\boldsymbol{\beta}) := F(\boldsymbol{X}\boldsymbol{\beta}) + G(\boldsymbol{\beta}) \right\}, \tag{6}$$

其中， $F(\boldsymbol{X}\boldsymbol{\beta}) := f(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{y})$ 是损失函数，而 $G(\boldsymbol{\beta}) := 2\lambda_2\, g_{\mathcal{N}}(\boldsymbol{\beta})$ 是正则化函数。利用标准的Fenchel对偶框架，(6) 的对偶问题为

$$\sup_{\boldsymbol{\zeta}\in\mathbb{R}^n} \left\{ \Psi(\boldsymbol{\zeta}) := -F^*(-\boldsymbol{\zeta}) - G^*\left(\boldsymbol{X}^\top\boldsymbol{\zeta}\right) \right\}, \tag{7}$$

其中， F^* 和 G^* 分别是 F 和 G的 Fenchel 共轭。

根据弱对偶性，对于任意可行对 $(\boldsymbol{\beta},\boldsymbol{\zeta})$， $\Phi(\boldsymbol{\beta}) \ge \Psi(\boldsymbol{\zeta})$，这意味着 $\Psi(\boldsymbol{\zeta})$ 为问题 (1) 提供了一个有效下界。然而，高效求解该问题并非易事。直接应用标准一阶方法通常会导致次线性收敛速率较慢。为了解决这一问题，下一节将介绍一个通用 原始-对偶算法框架。通过利用 F 和 G 的特定几何性质，我们将证明该框架能够实现线性收敛速率，从而快速计算稀疏广义线性模型问题的紧下界。

3 线性收敛的算法框架

本节针对满足特定几何正则性条件的复合优化问题，开展统一分析。我们考虑由原始与对偶目标函数定义的复合优化问题类别

$$\Phi(\boldsymbol{\beta}) = F(\boldsymbol{X}\boldsymbol{\beta}) + G(\boldsymbol{\beta}) \quad \text{and} \quad \Psi(\boldsymbol{\zeta}) = -F^*(-\boldsymbol{\zeta}) - G^*\left(\boldsymbol{X}^\top \boldsymbol{\zeta}\right).$$

首先陈述分析所需的一般正则性假设。

假设3.1。 函数 $F: \mathbb{R}^n \to \mathbb{R}$ 是凸且局部光滑的。函数$G: \mathbb{R}^p \to \mathbb{R}$ 是真、闭且凸的。此外，不存在这样的直线：沿着该直线， F 或 G 为有限且仿射的。

由于假设 F 局部光滑，其定义域必然具有非空内部。不存在这样的直线：沿着该直线， F 或 G 为有限且仿射；这一假设进一步意味着，共轭函数 F^* 和 G^* 的定义域内部非空 [Rockafellar，1970，推论13.4。2]。接下来，我们定义 $\Phi^\star := \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \Phi(\boldsymbol{\beta})$和 $\Psi^\star := \sup_{\boldsymbol{\zeta}\in\mathbb{R}^n} \Psi(\boldsymbol{\zeta})$，分别作为原始问题和对偶问题的最优目标值。为支持原始—对偶分析，我们采用以下 Slater 型假设。

假设 3.2。 集合 rint(定义域(Φ)) 和 int(定义域(Ψ)) 非空。具体而言，存在 Slater 点 $\boldsymbol{\beta}_s \in \mathbb{R}^p$ 和 $\boldsymbol{\zeta}_s \in \mathbb{R}^n$，使得 $\boldsymbol{\beta}_s \in$ rint(定义域(G))、 $\boldsymbol{X}\boldsymbol{\beta}_s \in$ int(定义域(F))、$-\boldsymbol{\zeta}_s \in$ int(定义域(F^*)) 以及 $\boldsymbol{X}^\top\boldsymbol{\zeta}_s \in$ int(定义域(G^*))。

在假设 3.2 下，广义 Fenchel 对偶定理 [Rockafellar，1970 年，推论 31.2。1] 保证原始问题和对偶问题均有解，且强对偶成立。也就是说，

$$\Phi(\boldsymbol{\beta}^\star) = \Phi^\star = \Psi^\star = \Psi(\boldsymbol{\zeta}^\star),$$

其中， $\boldsymbol{\beta}^\star$ 表示一个最优原始解， $\boldsymbol{\zeta}^\star$ 表示一个最优对偶解。我们将所有最优原始解组成的集合记为 $\mathcal{B}^\star$。此外，KKT 条件 [Rockafellar，1970 年，定理 31。3] 刻画了任意最优原始–对偶对 ($\boldsymbol{\beta}^\star$， $\boldsymbol{\zeta}^\star$)，其形式为

$$\boldsymbol{X}^\top\boldsymbol{\zeta}^\star \in \partial G(\boldsymbol{\beta}^\star), \quad \boldsymbol{X}\boldsymbol{\beta}^\star \in \partial F^*(-\boldsymbol{\zeta}^\star), \quad \boldsymbol{\beta}^\star \in \partial G^*(\boldsymbol{X}^\top\boldsymbol{\zeta}^\star), \quad -\boldsymbol{\zeta}^\star = \nabla F(\boldsymbol{X}\boldsymbol{\beta}^\star). \tag{8}$$

假设3.1的一个关键推论是， F 是本质光滑的，这意味着其共轭函数 F^* 是本质严格凸的 [Rockafellar，1970，定理26.3], 从而确保对偶解 $\boldsymbol{\zeta}^\star$ 是唯一的。我们将这些存在性与唯一性结果总结在如下引理中。

引理 3.3. *在假设 3.1 和 3.2下，原始问题和对偶问题均可解，且强对偶成立，即 $\Phi^\star = \Psi^\star$。此外，对偶问题具有唯一最优解。*

尽管引理 3.3保证了解的存在性与唯一性，但仅凭这些性质不足以获得线性收敛速率。为了实现更快的收敛，目标函数必须在最优解集附近展现特定的曲率性质。在接下来的小节中，我们通过引入更强的几何正则性条件来形式化这些要求。

3.1 结构几何

现在，我们介绍建立主要误差界所需的具体结构性假设。这些条件通过施加与透视松弛结构相容的局部曲率要求，强化了标准凸性。

假设 3.4。 以下条件成立。

(i) T 最优原始解集合 $\mathcal{B}^\star$ 是紧集，并满足 $\boldsymbol{X}\mathcal{B}^\star \subset$ 属于 dom(F) 的内部。

(ii) 函数 F 和 G 分别相对于 $-\boldsymbol{\zeta}^\star$ 和 $\boldsymbol{X}^\top\boldsymbol{\zeta}^\star$ 强凸。(iii) 次微分 $\partial F^*(-\boldsymbol{\zeta}^\star)$ 和 $\partial G^*(\boldsymbol{X}^\top\boldsymbol{\zeta}^\star)$ 是多面体集。

(iv) 最优对偶解满足 $-\boldsymbol{\zeta}^\star \in$ 属于 dom(F^*) 的内部，且 $\boldsymbol{X}^\top\boldsymbol{\zeta}^\star \in$ 属于 dom(G^*) 的内部。

假设 3.4 (i)–(iii) 构成一阶方法原始分析中常用的二次增长条件 [Zhou 和 So，2017年，Drusvyatskiy 和 Lewis， 2018]。相比之下，假设 3.4 (iv) 针对的是对偶目标函数的结构性质，这一视角在现有文献中受到的关注较少。在这些假设共同成立时，我们建立以下基本几何界。

定理 3.5。 在假设 *3.1–3.4* 下，令 $\boldsymbol{\zeta}^\star$ 表示唯一的对偶最优解。对于满足 $\boldsymbol{\zeta}^\star \in \mathcal{Z}$, $-\mathcal{Z} \subset \mathrm{int}(\mathrm{dom}(F^*))$ *and* $\boldsymbol{X}^\top\mathcal{Z} \subset \mathrm{int}(\mathrm{dom}(G^*))$ 的任意紧集 $\mathcal{B} \subset \mathbb{R}^p$ 和 $\mathcal{Z} \subset \mathbb{R}^n$，存在常数 $\alpha_{\mathcal{B}}$, $\kappa_{\mathcal{Z}} > 0$，使得

$$\Phi(\boldsymbol{\beta}) \geq \Phi^\star + \frac{\alpha_{\mathcal{B}}}{2}\,\mathrm{dist}^2(\boldsymbol{\beta}, \mathcal{B}^\star),\ \forall \boldsymbol{\beta} \in \mathcal{B}, \quad \textit{and} \quad -\Psi(\boldsymbol{\zeta}) \leq -\Psi^\star + \frac{\kappa_{\mathcal{Z}}}{2}\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^\star\|^2,\ \forall \boldsymbol{\zeta} \in \mathcal{Z}.$$

图 1 展示了定理 3.5 所确立的原始二次增长和对偶二次衰减条件。定理 3.5 揭示了一种显著的对称性：原始目标函数 $\Phi(\boldsymbol{\beta})$ 至少以 二次速度远离最优解集 $\mathcal{B}^\star$ 增长，而对偶目标函数 $\Psi(\boldsymbol{\zeta})$ 至多以 二次速度远离唯一的对偶解 $\boldsymbol{\zeta}^\star$ 衰减。因此，我们将假设 3.4 (iv) 所保证的性质称为二次衰减条件，并将其视为假设 3.4 (i)–(iii) 中提出的原始二次增长条件的对偶对应性质。

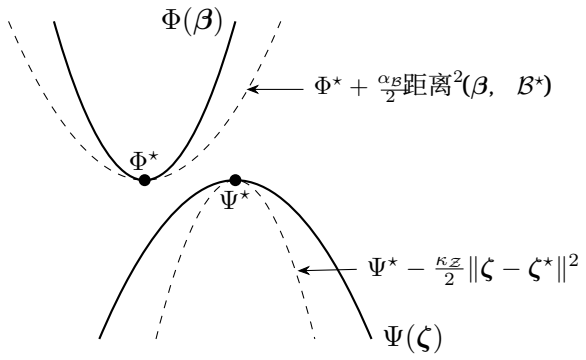

图 1：原始目标函数 $\Phi(\boldsymbol{\beta})$ 的二次增长和对偶目标函数 $\Psi(\boldsymbol{\zeta})$ 的二次衰减示意图。

定理 3.5 的证明依赖于关于强凸函数的两个技术性引理。第一个引理建立了次微分误差界。

引理 3.6. 设 h 是一个真闭凸函数，且不存在使h 有限且为仿射函数的直线。若 h 相对于某个 $\boldsymbol{\nu} \in \mathrm{int}(\mathrm{dom}(h^*))$强凸，则对于任意满足 $\boldsymbol{\nu} \in \mathcal{U}$的紧集 $\mathcal{U} \subset \mathrm{int}(\mathrm{dom}(h^*))$，存在常数 $\alpha_{\mathcal{U}} > 0$，使得

$$\mathrm{dist}\left(\partial h^*\left(\boldsymbol{\mu}\right), \partial h^*\left(\boldsymbol{\nu}\right)\right) \leq \frac{2}{\alpha_{\mathcal{U}}}\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2, \quad \forall \boldsymbol{\mu} \in \mathcal{U}.$$

引理 3.6 的证明见附录 B.1。引理 3.6 可视为 Drusvyatskiy 和 Lewis，2018年，定理 3.3] 中次微分误差界的对偶对应结果。Drusvyatskiy 和 Lewis [2018] 分析原始次微分 ∂h，而我们的结果刻画了共轭次微分 ∂h^*的行为。值得注意的是，不等式方向相反，这反映了 Fenchel 对偶形式中固有的符号翻转。基于这一误差界，我们推导出共轭函数值的二次上界。

引理 3.7. 设 h 是一个真闭凸函数，且不存在使 h 有限且仿射的直线。若 h 相对于某个 $\boldsymbol{\nu} \in$ 内部(定义域(h^*))是强凸的，则对任意满足 $\boldsymbol{\nu} \in \mathcal{U}$ 的紧集 $\mathcal{U} \subset$ 内部(定义域(h^*))，存在常数 $\alpha_{\mathcal{U}} > 0$ 和向量$\boldsymbol{s} \in \partial h^*(\boldsymbol{\nu})$ 使得

$$h^*(\boldsymbol{\mu}) \le h^*(\boldsymbol{\nu}) + \boldsymbol{s}^\top(\boldsymbol{\mu} - \boldsymbol{\nu}) + \frac{\alpha_{\mathcal{U}}}{2}\|\boldsymbol{\mu} - \boldsymbol{\nu}\|^2, \quad \forall \boldsymbol{\mu} \in \mathcal{U}.$$

引理 3.7 的证明见附录 B.2。现在我们可以借助引理 3.7 证明定理 3.5。定理 3.5 的证明见附录 B.3。上文所建立的原始二次增长与对偶二次衰减构成了我们问题结构的基本几何性质。

3.2 原始–对偶误差界

虽然定理 3.5独立地刻画了原始与对偶景观，我们现在建立它们之间的严格联系。为便于这一分析，我们首先引入关于原始损失函数的如下正则性条件。

假设 3.8。 损失函数 F 是严格凸的。

在该假设下，我们可以将原始最优性间隙 $\Phi(\boldsymbol{\beta}) - \Phi^\star$ 与两个关键的对偶量联系起来：对偶最优距离 $\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^\star\|_2^2$ 和对偶最优性间隙$\Psi^\star - \Psi(\boldsymbol{\zeta})$。这些量之间的联系依赖于对偶变量的一种特定构造。受式 (8) 中 KKT 条件 $-\boldsymbol{\zeta}^\star = \nabla F(\boldsymbol{X}\boldsymbol{\beta}^\star)$ 的启发，我们通过以下映射定义与任意原始可行解 $\boldsymbol{\beta}$ 相关联的诱导对偶向量 $\boldsymbol{\zeta}$

$$\boldsymbol{\zeta} := -\nabla F(\boldsymbol{X}\boldsymbol{\beta}).$$

在下面的定理中，我们证明，在该映射下，原始最优性间隙严格控制对偶误差，从而使原始问题的收敛能够有效推动对偶收敛。

定理 3.9。 在假设 *3.1–3.8* 下，令 $\mathcal{B}^\star$ 表示原始最优解集合，令 $\boldsymbol{\zeta}^\star$ 表示唯一的对偶最优解。令 $\mathcal{B} \subset$定义域(G) 为满足$\mathcal{B}^\star \subseteq \mathcal{B}$的紧集。定义相应的对偶像集 $\mathcal{Z} := \{-\nabla F(\boldsymbol{X}\boldsymbol{\beta}) \mid \boldsymbol{\beta} \in \mathcal{B}\}$，并假设内部条件 $-\mathcal{Z} \subset$ 内部(定义域(F^*)) 和 $\boldsymbol{X}^\top\mathcal{Z} \subset$ 内部(定义域(G^*)) 成立。那么，存在常数 $\sigma_{\mathcal{B}}$、$\kappa_{\mathcal{B}} > 0$，使得对于任意原始向量 $\boldsymbol{\beta} \in \mathcal{B}$ 及其诱导对偶向量$\boldsymbol{\zeta} := -\nabla F(\boldsymbol{X}\boldsymbol{\beta})$，我们有

$$\frac{\sigma_{\mathcal{B}}}{2}\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^\star\|_2^2 \le \Phi(\boldsymbol{\beta}) - \Phi^\star \quad and \quad \Psi(\boldsymbol{\zeta}^\star) - \Psi(\boldsymbol{\zeta}) \le \frac{\kappa_{\mathcal{B}}}{\sigma_{\mathcal{B}}}\big(\Phi(\boldsymbol{\beta}) - \Phi^\star\big).$$

定理 3.9 的证明见附录 B.4。为了对定理 3.9 中的第一个不等式建立几何直觉，我们将其与光滑凸优化中的标准结果作类比。回顾一下，如果函数 h 在全局上是 L-光滑的，那么其 Fenchel 共轭 h^* 是 $(1/L)$-强凸的 [Beck，2017，定理 5.10]。此外，对于任意 $\boldsymbol{z} \in \mathbb{R}^p$ 以及极小化子$\boldsymbol{z}^\star = \arg\min h(\boldsymbol{z})$，$h$的光滑性意味着最优性间隙 [存在一个下界8]Beck，2017，定理 5。

$$\frac{1}{2L}\|\nabla h(\boldsymbol{z})\|_2^2 \le h(\boldsymbol{z}) - h(\boldsymbol{z}^\star).$$

如果我们将 $\nabla h(z)$ 解释为对偶变量，那么这个不等式类似于定理 3.9 中的第一个界限。不过，我们的结果将这种关系扩展到了更复杂的情形：原始目标函数 $F+G$ 具有非光滑复合结构，而损失函数 F 仅局部 光滑。

此外，定理 3.9 具有一项关键的算法意义：原始目标函数 $\Phi(\beta^t)$ 的收敛严格控制着对偶序列 ζ^t 和对偶目标函数 $\Psi(\zeta^t)$ 的收敛。这种关系不受生成原始序列的具体算法影响。因此，如果某个算法保证 $\Phi(\beta^t)$ 以特定速率收敛到 $\Phi^\star$，那么对偶序列和对偶目标值也会继承相同的收敛速率。

3.3 基于间隙重启的线性收敛

为了推导出适用于多种一阶方法的统一分析，我们首先对通用求解器的性质进行形式化描述。我们不针对单一算法的特定更新规则定制分析，而是假设可以使用一种求解器，它能在某个紧区域内生成迭代点，并提供标准的收敛保证。

假设 3.10。令 $\mathcal{B}\subseteq \operatorname{dom} G$ 为一个满足 $\mathcal{B}^\star\subseteq\mathcal{B}$ 的紧集。存在一种近端一阶算法，可生成可行原始序列 $\{\beta^t\}_{t\geq 0}\subset\mathcal{B}$ 及其对应的对偶序列 $\zeta^t:=-\nabla F(\boldsymbol{X}\beta^t)$。我们假设该算法依赖于 G 的近端算子。此外，该算法保证相对于初始化 β^0具有由速率函数 r 表征的收敛速率：$\mathbb{N}\to\mathbb{R}_+$，即

$$\Phi(\beta^t)-\Phi^\star\leq r(t)\cdot \operatorname{dist}^2(\beta^0,\mathcal{B}^\star)\quad \forall t\geq 1. \tag{9}$$

广泛的一阶方法类别均满足假设 3.10，包括标准投影梯度下降（PGD）[Beck，2017年，第 10 节；2], 快速迭代收缩阈值算法（FISTA）等加速变体 [Beck，2017年，第 10 节；7], 以及 AdProxGD [Malitsky 和 Mishchenko， 2024] 和 AC-FGM[Li 和 Lan， 2025]等无参数方法。我们首先指出，任何满足该假设的算法都会自动为对偶间隙提供收敛保证。结合式 (9) 中的算法收敛速率与定理 3.9 中得出的结构性界，我们推导出以下对偶间隙。

命题 3.11。在假设*3.1–3.10*下，令 $\alpha_{\mathcal{B}},\sigma_{\mathcal{B}},\kappa_{\mathcal{B}}>0$ 为定理 *3.9* 中得出的几何常数。则，原对偶迭代 $\{(\beta^t,\zeta^t)\}_{t\geq 1}$满足以下对偶间隙

$$\Phi(\beta^t)-\Psi(\zeta^t)\leq\left(\frac{2r(t)(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}{\alpha_{\mathcal{B}}}\right)\big(\Phi(\beta^0)-\Psi(\zeta^0)\big). \tag{10}$$

命题 3.11 的证明见附录 B.5。命题 3.11 证明，对偶间隙的收敛速率与 $r(t)$ 相同；对于标准一阶方法，该速率为次线性。一个显著的例外是采用固定步长的投影梯度下降，它在二次增长条件下会自动实现原始问题的线性收敛速率 [Zhou和So，2017年，Drusvyatskiy 和 Lewis， 2018]。这与命题 3.11 一致，因为它直接意味着投影梯度下降的对偶间隙具有线性收敛速率。

相比之下，快速迭代收缩阈值算法等加速方法通常需要重启方案，以防止振荡并使原始序列实现线性收敛 [O'Donoghue 和 Candes，2015年，Fercoq 和 Qu， 2019]。此外，近期的自适应方法 [Malitsky 和 Mishchenko，2024年，Li 和 Lan， 2025] 通常只能达到次线性收敛速率，而在二次增长条件下，它们对于对偶目标函数的理论保证仍基本未得到探索。为确保这一更广泛类别的算法实现线性收敛（按照标准的 Q-线性意义），我们采用一种由对偶间隙明确控制的重启策略。核心思路是，仅运行算法直至对偶间隙缩小至固定因子 $\eta>1$，然后重启算法，并将当前迭代点作为新的初始化。由于对偶间隙是误差的严格上界，因此可以确保每个重启周期内误差都按几何级数收缩。

定理 3.12。假设命题 *3.11* 的条件成立，且 $r(t)$ 单调递减。令 $(\boldsymbol{\beta}^{\text{init}}, \boldsymbol{\zeta}^{\text{init}})$ 为初始原始–对偶对。对于任意目标收缩因子$\eta > 1$，将重启间隔 $t_{\max}$ 定义为

$$t_{\max} := \left\lceil r^{-1}\left(\frac{\alpha_{\mathcal{B}}}{2\eta(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}\right)\right\rceil, \tag{11}$$

其中，r^{-1} 是速率函数的逆函数。如果算法每 $t_{\max}$ 次迭代重启一次（重置 $t=0$ 并将 $\boldsymbol{\beta}^0$ 设置为当前迭代点），则重启点处的对偶间隙序列以线性速率收敛，且

$$\Phi(\boldsymbol{\beta}^{s\cdot t_{\max}}) - \Psi(\boldsymbol{\zeta}^{s\cdot t_{\max}}) \le \eta^{-s}\left(\Phi(\boldsymbol{\beta}^{\text{init}}) - \Psi(\boldsymbol{\zeta}^{\text{init}})\right).$$

定理 3.12 的证明见附录 B.6。定理 3.12 的一个重要推论是，对偶间隙的线性收敛必然意味着原始目标值和对偶目标值以及相应迭代点的线性收敛。为说明这一点，注意到根据弱对偶性（$\Psi(\boldsymbol{\zeta}) \le \Psi^\star = \Phi^\star$），对偶间隙是原始最优性间隙的严格上界

$$\Phi\left(\boldsymbol{\beta}^{s\cdot t_{\max}}\right) - \Phi^\star \le \Phi\left(\boldsymbol{\beta}^{s\cdot t_{\max}}\right) - \Psi\left(\boldsymbol{\zeta}^{s\cdot t_{\max}}\right) \le \eta^{-s}\left(\Phi(\boldsymbol{\beta}^0) - \Psi(\boldsymbol{\zeta}^0)\right).$$

这就证明了原始目标值的线性收敛。根据定理 3.5 和定理 3.9，原始迭代点 $\{\boldsymbol{\beta}^t\}$、对偶迭代点 $\{\boldsymbol{\zeta}^t\}$ 以及对偶目标值 $\{\Psi(\boldsymbol{\zeta}^t)\}$ 也都以线性速率收敛。

最后，我们通过量化 PGD、快速迭代收缩阈值算法和 AC-FGM 的具体迭代复杂度来结束本节。对于 AC-FGM，标准次线性速率取决于用户指定的参数$c \in (1, \infty)$ 和 $\rho \in (0, 1-\sqrt{6}/3]$，其中 c用于初始化期间的线搜索过程。此外，L_0 表示算法使用的光滑常数初始低估值。为简化下文比较中的记号，我们省略几何常数 κ（对偶衰减）、L（原始光滑性，相对于 $F(\boldsymbol{X}\cdot)$）、α（原始增长性）和 σ（对偶强凸性）中的下标 $\mathcal{B}$。表 1 总结了速率函数 $r(t)$，以及应用基于对偶间隙的重启方案后得到的复杂度。至关重要的是，该重启方案将加速方法（快速迭代收缩阈值算法和 AC-FGM）最坏情况下的次线性收敛转化为线性速率，其尺度与条件数 $(\sqrt{\kappa L/\alpha\sigma})$ 的平方根 成正比。与标准 PGD 线性速率（其尺度与条件数成正比）相比，这在理论上有显著提升。

表 1：标准速率函数 $r(t)$（仅假设 F 的光滑性时），以及应用基于对偶间隙的重启方案后达到 ϵ 准确率所需的相应复杂度。

算法	速率函数 $r(t)$	重启后的复杂度
PGD	$\dfrac{L}{2t}$	$\mathcal{O}\left(\dfrac{\kappa L}{\alpha\sigma}\log\left(\dfrac{1}{\epsilon}\right)\right)$
快速迭代收缩阈值算法	$\dfrac{2L}{(t+1)^2}$	$\mathcal{O}\left(\sqrt{\dfrac{\kappa L}{\alpha\sigma}}\log\left(\dfrac{1}{\epsilon}\right)\right)$
AC-FGM	$\dfrac{15cL}{2t\,(t+1)\;\rho(1-\rho)}$	$\mathcal{O}\left(\sqrt{\dfrac{\kappa L}{\alpha\sigma}}\log\left(\dfrac{1}{\epsilon}\right)+\log_c\left(\dfrac{L}{L_0}\right)\right)$

3.4 实践中的重启

定理 3.12 基于几何常数 $\sigma_{\mathcal{B}}$、$\alpha_{\mathcal{B}}$ 和 $\kappa_{\mathcal{B}}$，以及特定的速率函数 $r(t)$，提供了理论保证。在实践中，这些参数通常是未知的，或难以估计。然而，应用重启

方案并不要求事先知道这些常数。由于对偶间隙可以计算，我们可以在运行时间内对其进行监测，并动态触发重启。

具体而言，实际实现中需要在每次迭代时检查条件 $\Phi(\boldsymbol{\beta}^t) - \Psi(\boldsymbol{\zeta}^t) \le \eta^{-1}(\Phi(\boldsymbol{\beta}^0) - \Psi(\boldsymbol{\zeta}^0))$。理论分析保证该条件至多在 $t_{\max}$ 次迭代内满足，其中 $t_{\max}$ 的定义见 (11)。为了使原始问题–对偶间隙达到 ϵ准确率，所需重启总次数 $s_{\max}$ 的上界为

$$s_{\max} \le \left\lceil \log_\eta \left(\frac{\Phi(\boldsymbol{\beta}^{\text{init}}) - \Psi(\boldsymbol{\zeta}^{\text{init}})}{\epsilon} \right) \right\rceil,$$

该上界关于 $1/\epsilon$呈对数增长。因此，迭代总次数的上界为 b y

$$T_{\text{total}} \le s_{\max} \cdot t_{\max} \le \left\lceil \log_\eta \left(\frac{\Phi(\boldsymbol{\beta}^{\text{init}}) - \Psi(\boldsymbol{\zeta}^{\text{init}})}{\epsilon} \right) \right\rceil \cdot \left\lceil r^{-1} \left(\frac{\alpha_{\mathcal{B}}}{2\eta(1 + \kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})} \right) \right\rceil.$$

下面的定理给出了多项式速率模型下的最优重启参数 η 。

定理 3.13. 假设收敛速率表现为 $r(t) \approx C/t^q$ ，其中 $q > 0$ 和$C > 0$为某些常数。那么，常数意义下最优的重启参数选择为 $\eta = e^q$。

定理 3.13 的证明见附录 B.7。特别地， $q = 2$ （加速方法中很常见，此时 $r(t) = \Theta(1/t^2)$）可得 $\eta = e^2$。不过请注意，这一最优选择是根据最坏情况推导得出的，未必与 η的最佳经验选择一致。例如，在实践中可能会观察到 $\eta = e^3$ 的表现优于 $\eta = e^2$ ，这至少有三个原因。首先，对于加速方法，界 $r(t) = \Theta(1/t^2)$ 通常并不紧：如果在我们关注的问题实例上，一个周期内的有效下降表现为 $r(t) \approx C/t^q$ ，其中 $q > 2$ 为某个常数，那么上述同样的计算表明，选择更大的 η （即 $\eta = e^q$）更为合适。其次，我们的推导将每次重启视为在迭代预算 $t_{\max}$之外“免费”。实际上，重启加速方法会舍弃其动量，并引入一个短暂的“预热”阶段，之后加速阶段才会开始发挥作用。如果将这一影响建模为每次重启额外产生 t_0 次迭代的开销（或者更一般地，在上界中加入依赖于重启的常数），那么较大的 η 会减少重启次数，从而减少支付这项开销的次数。第三，我们的实际方案依据观测到的对偶间隙下降来触发重启，而不是依据最坏情况界：如果间隙通常远早于 $t_{\max}$便下降至原来的 η^{-1} 倍，那么增大 η 可以大幅减少重启次数，同时只会略微增加周期的典型长度。综合来看，这些因素说明， $\eta = e^2$ 最适合被视为理想化多项式速率模型下的最优选择，而在经验上， $\eta = e^3$ 等更大的取值可能更合适。

4 稀疏广义线性模型的高效实现

第 3 节中的算法框架保证了通用复合问题的线性收敛，前提是这些问题满足假设 3.1–3.8。在本节中，我们将该框架应用于分支定界树的每个节点 $\mathcal{N}$ 处的稀疏广义线性模型透视松弛。

回顾第2.4节，该问题采用复合形式，其中 $F(\boldsymbol{X}\boldsymbol{\beta}) := f(\boldsymbol{X}\boldsymbol{\beta},\ \boldsymbol{y})$，且 $G(\boldsymbol{\beta}) := 2\lambda_2 g_{\mathcal{N}}(\boldsymbol{\beta})$，其中 $g_{\mathcal{N}}$是定义如下的隐式正则化项

$$g_{\mathcal{N}}(\boldsymbol{\beta}) = \inf_{\boldsymbol{z} \in \mathbb{R}^p} \left\{ \frac{1}{2} \sum_{j \in [p]} \beta_j^2 / z_j \ :\ (\boldsymbol{\beta}, \boldsymbol{z}) \in \mathcal{D} \right\}.$$

这里，式（4）中定义的定义域 $\mathcal{D}$，通过指定节点 $\mathcal{N}$ 处哪些二元变量固定为 0 或 1、哪些变量保持自由，来编码分支约束。

为充分发挥我们框架的理论保证与实际性能，我们将在以下小节中探讨三个关键方面。首先，我们证明 $g_{\mathcal{N}}$ 具备必要的几何性质。其次，我们说明可以高效计算 $g_{\mathcal{N}}$ 的近端算子，从而确保每次迭代的计算成本较低。最后，我们通过证明标准广义线性模型损失函数天然满足所需的光滑性和严格凸性假设，验证原始-对偶框架可以应用。

4.1 隐式正则化项的几何性质

我们首先建立隐式正则化项 $g_{\mathcal{N}}$ 的基本几何性质。首先，我们验证 $g_{\mathcal{N}}$ 是一个具有有界定义域的有效凸正则项，这是最优解集紧致性的前提。

引理 4.1. (5) 中的隐函数 $g_{\mathcal{N}}$ 是真、闭且凸的，并具有紧定义域。

引理 4.1 的证明见附录 B.8。下一个引理给出了 g^* 的闭式表达式，其中函数 $\mathrm{TopSum}_k(\cdot)$ 表示最大的前 k 个元素之和，而 $\boldsymbol{H}_M(\boldsymbol{\alpha}) := (H_M(\alpha_1) \ldots, H_M(\alpha_p))$ 是对每个坐标应用的 Huber 损失函数，其中

$$H_M(\alpha_j) := \begin{cases} \frac{1}{2}\alpha_j^2 & \text{if } |\alpha_j| \le M \\ M|\alpha_j| - \frac{1}{2}M^2 & \text{if } |\alpha_j| > M. \end{cases} \tag{12}$$

引理 4.2. $g_{\mathcal{N}}$ 的 *Fenchel* 共轭由下式给出

$$g_{\mathcal{N}}^*(\boldsymbol{\alpha}) = \sum_{j \in \mathcal{J}_1} H_M(\alpha_j) \; + \mathrm{TopSum}_{k-|\mathcal{J}_1|}\left(\boldsymbol{H}_M\left(\boldsymbol{\alpha}_{\mathcal{J}_f}\right)\right). \tag{13}$$

引理 4.2 的证明见附录 B.9。(13) 中的第二项表明，当我们从 原始问题 空间转移到 对偶空间时， $g_{\mathcal{N}}$ 的变分形式被转化为显式表达式。引理 4.2 中的闭式表达式表明，$g_{\mathcal{N}}^*$ 完全由保持分段线性二次结构的运算构成，即有限求和、排序和二次样条。这一观察在下面的引理中得到形式化。

引理 4.3。 函数 $g_{\mathcal{N}}$ 和 $g_{\mathcal{N}}^*$ 都是分段线性二次函数且是强凸的。

引理 4.3 的证明见附录 B.10。该结果证实，透视正则项满足我们算法框架严格的几何要求。有关 $g_{\mathcal{N}}$ 及其共轭函数 $g_{\mathcal{N}}^*$ 的分段线性二次结构的可视化示意，请参见图 2。我们将在第 4.3 节的命题 4.6 中利用这一性质。

4.2 隐式正则化项及其近端算子的高效计算

在确立隐式正则化项的几何正则性之后，我们现在讨论其计算可处理性。我们的算法框架需要执行两项关键操作：(i) 计算函数值 $g_{\mathcal{N}}(\boldsymbol{\beta})$，以便监控重启时的对偶间隙；(ii) 在近端梯度法中计算其近端算子。

我们首先讨论函数求值。需要注意的是，通过通用的二阶锥规划（SOCP）求解器直接求解（5）中的最小化问题，计算开销很大；而且，由于内点法具有迭代性质，只能得到近似解。相比之下，利用定义域 $\mathcal{D}$ 的特定结构，我们推导出一种专用算法，能够在有限步内精确地 计算 $g_{\mathcal{N}}(\boldsymbol{\beta})$，其最坏情况下的计算复杂度关于 p 呈对数线性。整体算法见算法1。

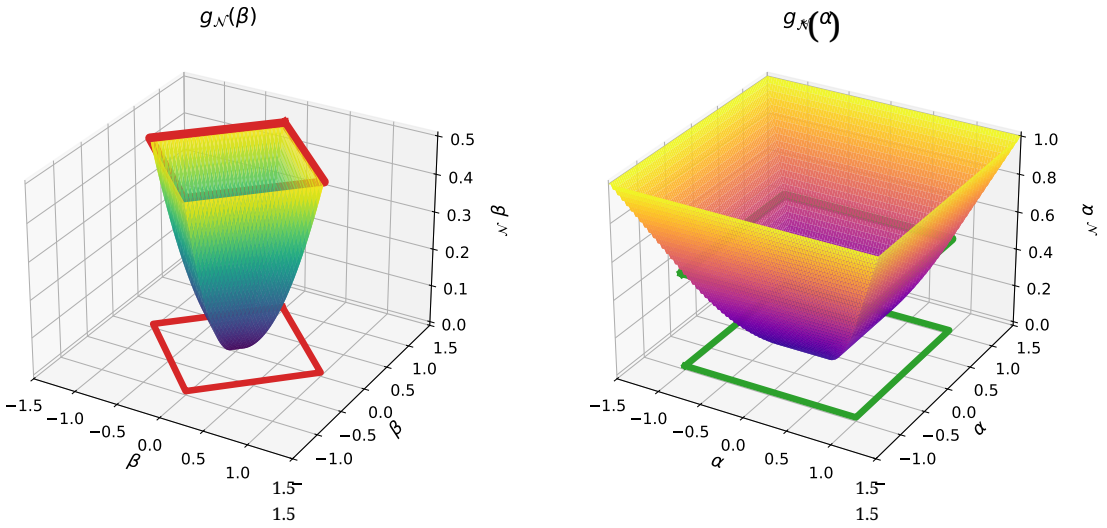

图 2：根节点（$p = 2,\ \ k = 1,\ \ M = 1$）处 $g_{\mathcal{N}}$（左）和 $g^*_{\mathcal{N}}$（右）的三维图。左：红色菱形标记定义域($g_{\mathcal{N}}$)的边界 $|\beta_1| + |\beta_2| = M$（顶部位于 $g_{\mathcal{N}}(\boldsymbol{\beta}) = \frac{1}{2}M^2$，底部为投影）。右：绿色方块标记转变集 $\max(|\alpha_1|, |\alpha_2|) = M$（顶部位于$g^*_{\mathcal{N}}(\boldsymbol{\alpha}) = \frac{1}{2}M^2$，底部为投影）；$g^*_{\mathcal{N}}$在内部二次增长，在外部线性增长。

算法1： 在节点 $\mathcal{N}$处计算 $g_{\mathcal{N}}(\boldsymbol{\beta})$，其关联指标集为 $\mathcal{J}_0, \mathcal{J}_1, \mathcal{J}_f$

数据： 向量 $\boldsymbol{\beta} \in$ 定义域($g_{\mathcal{N}}$)，基数参数 $k \in [p]$

结果： $g_{\mathcal{N}}\ (\boldsymbol{\beta})$

集合 $\bar{k} \leftarrow k - |\mathcal{J}_1|$，$\bar{p} \leftarrow |\mathcal{J}_f|$

集合 $\bar{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}_{\mathcal{J}_f}$，$\boldsymbol{\omega} \leftarrow \mathbf{0}_{\bar{k}}$，$\theta \leftarrow \sum_{j\in[\bar{p}]} |\bar{\beta}_j|$

排序 $\bar{\boldsymbol{\beta}}$ 使得 $|\bar{\beta}_{(1)}| \geq \ldots \geq |\bar{\beta}_{(\bar{k})}| \geq \max_{j\in[\bar{k}+1:\ \bar{p}]} \{|\bar{\beta}_j|\}$

针对 $j \leftarrow 1$ **至** $\bar{k}$ **do**

  $\overline{\theta} \leftarrow \theta / (\ \bar{k} - j + 1)$

  如果 $\theta \geq |\bar{\beta}_{(j)}|$ **那么** $\omega_{[j:\bar{k}]} \leftarrow \overline{\theta}$ **中断**

  否则 $\omega_j \leftarrow |\bar{\beta}_{(j)}|$ 以及 $\theta \leftarrow \theta - |\bar{\beta}_{(j)}|$

返回 $\frac{1}{2}\sum_{j\in\mathcal{J}_1} \beta_j^2 + \frac{1}{2}\sum_{j\in[\bar{k}]} \omega_j^2$

The **定理 4.4。** 对于任意 $\boldsymbol{\beta} \in$ 定义域($g_{\mathcal{N}}$)，令 $\bar{\boldsymbol{\beta}} := \boldsymbol{\beta}_{\mathcal{J}_f}$，$\bar{p} := |\mathcal{J}_f|$，以及 $\bar{k} := k - |\mathcal{J}_1|$。该 n,

$$g_{\mathcal{N}}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{j\in\mathcal{J}_1} \beta_j^2 + \frac{1}{2}\min_{\boldsymbol{\omega}} \left\{ \sum_{j=1}^{\bar{k}} \omega_j^2 \;\middle|\; \begin{array}{l} \boldsymbol{\omega} \geq \mathbf{0}, \quad M \geq \omega_1 \geq \cdots \geq \omega_{\bar{k}} \geq 0, \\ \omega_{\bar{k}+1} = \cdots = \omega_{\bar{p}} = 0, \quad \boldsymbol{\omega} \succeq_m |\bar{\boldsymbol{\beta}}| \end{array} \right\}.$$

这里，$\boldsymbol{\omega} \succeq_m |\boldsymbol{\beta}|$ 表示 $\boldsymbol{\omega}$ 主序于 $|\boldsymbol{\beta}|$，即，

$$\sum_{j=1}^{\ell} \omega_j \geq \sum_{j=1}^{\ell} |\bar{\beta}_{(j)}| \quad \forall \ell = 1, \ldots, \bar{p} - 1, \qquad \textit{and} \qquad \sum_{j=1}^{\bar{p}} \omega_j = \sum_{j=1}^{\bar{p}} |\bar{\beta}_{(j)}|.$$

算法*1*计算上述问题的一个极小化子 $\boldsymbol{\omega}$，从而以 $\mathcal{O}(p + \bar{p}\log\ k + k)$ 的计算复杂度，精确地计算 $g_{\mathcal{N}}$ $(\boldsymbol{\beta})$ 。

定理 4.4 的证明见附录 B.11。算法1的高效性源于利用了主序表示与凸包表示之间的等价性。

算法2： 在节点 $\mathcal{N}$ 处，使用相关指标集 $\mathcal{J}_0$，$\mathcal{J}_1$，$\mathcal{J}_f$计算 $\mathrm{prox}_{\rho g_{\mathcal{N}}^*}(\boldsymbol{\beta})$

数据： 向量 $\boldsymbol{\beta}\in\mathbb{R}^p$、标量 $\rho>0$、基数参数 $k\in[p]$、盒参数 $M>0$

结果： 近端算子 $\mathrm{prox}_{\rho g_{\mathcal{N}}^*}(\boldsymbol{\beta})$

集合 $\bar{k}\leftarrow k-|\mathcal{J}_1|$，$\bar{p}\leftarrow|\mathcal{J}_f|$

集合 $\bar{\boldsymbol{\beta}}\leftarrow\boldsymbol{\beta}_{\mathcal{J}_f}$，$\boldsymbol{\rho}\in\mathbb{R}_+^{\bar{p}}$ 若 $j\in\{1,2,\ldots,$ 则为 $\rho_j\leftarrow\rho_{\bar{k}}\}$，否则为 $\rho_j\leftarrow 0$。

排序 $\bar{\boldsymbol{\beta}}$，其中置换 $\boldsymbol{\pi}$ 属于 $[\bar{p}]$，且满足 $|\bar{\beta}_{\pi(1)}|\geq\ldots\geq|\bar{\beta}_{\pi(\bar{p})}|\geq 0$

对于 $j\leftarrow 1$ **到** $\bar{p}$，**执行** $\hat{\alpha}_j\leftarrow$ 近端算子$_{\rho_j H_M}(\bar{\beta}_{\pi(j)})$

$\mathcal{W}\leftarrow\{[1,1],[2,2]\ldots,[\bar{p},\bar{p}]\}$。

当 $\exists[j_1,j_2],\ [j_2+1,j_3]\in\mathcal{W}$ s。t。 $\hat{\alpha}_{j_1}<\hat{\alpha}_{j_3}$ **执行**

  $\mathcal{W}\leftarrow\mathcal{W}\setminus\{[j_1,j_2]\}\setminus\{[j_2+1,j_3]\}$

  $\varrho\leftarrow\left(\sum_{j=j_1}^{j_3}\rho_j\right)/(j_3-j_1+1)$

  $\xi\leftarrow\left(\sum_{j=j_1}^{j_3}\bar{\beta}_{\pi(j)}\right)/(j_3-j_1+1)$

  $\hat{\alpha}_{[j_1:\ j_3]}\leftarrow$近端算子$_{\rho H_M}(\xi)$

  $\mathcal{W}\leftarrow\mathcal{W}\cup\{[j_1,j_3]\}$

集合 $\boldsymbol{\alpha}\in\mathbb{R}^p$ 与 $\boldsymbol{\alpha}_{\mathcal{J}_0}=\boldsymbol{\beta}_{\mathcal{J}_0}$，$\boldsymbol{\alpha}_{\mathcal{J}_1}=[$近端算子$_{\rho H_M}(\beta_j)]_{j\in\mathcal{J}_1}$，$\boldsymbol{\alpha}_{\mathcal{J}_f}=\mathrm{sgn}(\bar{\boldsymbol{\beta}})\odot\boldsymbol{\pi}^{-1}(\hat{\boldsymbol{\alpha}})$

返回 $\boldsymbol{\alpha}$

这一联系的直观示意请见图3。算法1并非求解一般的凸优化问题，而是高效地找到一个主序于 $|\boldsymbol{\beta}|$的稀疏向量 $\boldsymbol{\omega}$，将可能复杂的 SOCP 求解转化为简单的索引搜索。

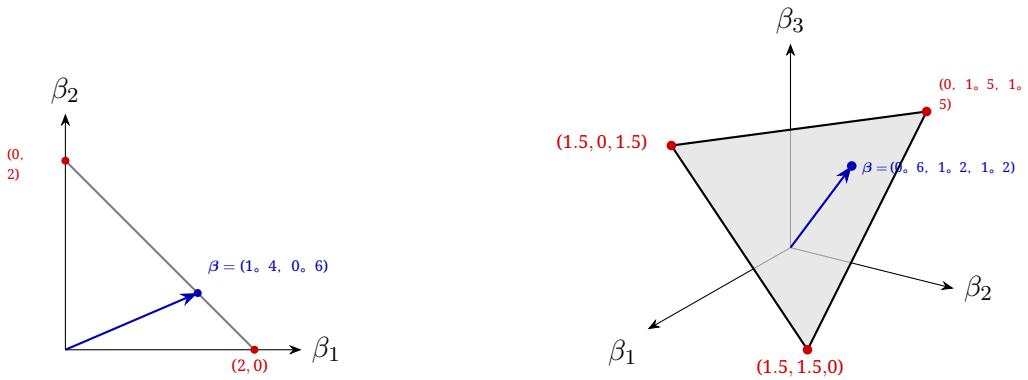

图3：向量 $\boldsymbol{\omega}$ 主序 $\boldsymbol{\beta}$ 当且仅当 $\boldsymbol{\beta}\in\mathrm{conv}(\mathrm{perm}(\boldsymbol{\omega}))$. 左图（$p=2$，$k=1$）：凸包是连接 (2，0) 和 (0，2) 的线段，其中包含 $\boldsymbol{\beta}=(1.4,\ 0.6)$。右图（$p=3$，$k=2$）：凸包是顶点为 (1.5，1.5，0)、(1.5，0，1.5) 和 (0，1.5，1.5) 的三角形，其中包含$\boldsymbol{\beta}=(0.6,\ 1.2,\ 1.2)$。给定 $\boldsymbol{\beta}$，算法1 构造出这样的 k-稀疏向量 $\boldsymbol{\omega}$。

我们现在转向近端算子的计算。由于 $g_{\mathcal{N}}$的隐式定义，直接计算近端算子$\rho g_{\mathcal{N}}$ 颇具挑战。然而，借助引理 4.2 所建立的共轭函数 $g_{\mathcal{N}}^*$ 的显式结构，我们可以高效地计算该共轭函数的近端算子，并通过扩展莫罗分解恢复所需的解。

定理 4.5. 对于任意 $\boldsymbol{\beta}\in\mathbb{R}^p$ 和 $\rho>0$，算法2以计算复杂度 $\mathcal{O}(p+\bar{p}\log\bar{p})$精确计算近端算子 $\rho g_{\mathcal{N}}^*(\boldsymbol{\beta})$，其中 $\bar{p}=|\mathcal{J}_f|$和 $k=k-|\mathcal{J}_1|$。此外，算法2可用作子程序，通过扩展莫罗分解精确计算近端算子$\rho g_{\mathcal{N}}(\boldsymbol{\beta})$

$$\mathrm{prox}_{\rho g_{\mathcal{N}}}(\boldsymbol{\beta})=\boldsymbol{\beta}-\rho\,\mathrm{prox}_{\rho^{-1}g_{\mathcal{N}}^*}\left(\rho^{-1}\boldsymbol{\beta}\right). \tag{14}$$

定理 4.5 的证明见附录 B.12。核心思想是将近端算子$\rho g_{\mathcal{N}}^{\star}$ 重新表述为一个广义保序回归问题，即在由 top-k 和算子诱导的保序约束下，最小化 Huber 损失¯函数之和。这一广义保序回归问题在统计学中已有充分研究，并可利用相邻违背者合并算法（PAVA） [Busing, 2022]高效求解。在算法2中，PAVA 在第6–11行执行。具体而言，在 while 循环中通过一种高效的分块上下调整过程识别相邻违背者，其详细实现见附录中的算法 3。

定理 4.4 和定理 4.5 中确立的计算结果，直接满足假设 3.10 所提出的算法要求。具体而言，高效计算近端算子，使得能够以较低的单次迭代成本执行标准近端一阶方法。同时，精确计算 $g_{\mathcal{N}}(\boldsymbol{\beta})$ 可以精确监控对偶间隙，而这正是所提出的重启方案实现线性收敛的关键。

4.3 标准广义线性模型的收敛保证

在确立算法在计算上的可行性和高效性之后，我们现在转而探讨问题结构在理论上的有效性。为了应用定理 3.12 的线性收敛保证，我们必须验证透视松弛问题 (6) 满足假设 3.1–3.8 中列出的正则性条件。下面的命题证实，在损失函数满足标准且易于验证的条件时，这些假设均成立。

命题 4.6. *考虑原始目标函数 $\Phi(\boldsymbol{\beta}) = F(\boldsymbol{X}\boldsymbol{\beta}) + G(\boldsymbol{\beta})$，其中 $F(\boldsymbol{z}) := f(\boldsymbol{z}, \boldsymbol{y})$满足 $f\colon\ \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ 在 $\boldsymbol{z}$ 上严格凸、局部光滑且强凸，并且 $G(\boldsymbol{\beta}) := 2\lambda_2\, g_{\mathcal{N}}(\boldsymbol{\beta})$ 中，$\lambda_2 > 0$ 和 $g_{\mathcal{N}}$ 按 (5) 定义。则假设 3.1–3.8 均成立。*

命题 4.6 的证明见附录 B.13。最后，我们就这些结果的适用性作几点重要说明。首先，命题 4.6 中确立的正则性条件适用于标准广义线性模型损失函数，包括最小二乘、逻辑回归及其他常见损失函数。我们在附录 B.14 中给出了详细验证。其次，关于定理 3.12 中的重启方案，分析依赖于一个包含迭代点的紧集 $\mathcal{B}$ 。对于透视松弛问题，由于引理 4.1，正则化项 $g_{\mathcal{N}}$ 的定义域 $\mathcal{B}$ 本身就是紧集。因此，我们只需定义 $= \mathrm{dom}\ g_{\mathcal{N}}$，即可确保几何常数在全局范围内均有定义。最后，根据定理 4.4 和定理 4.5， G 及其近端算子可以精确计算，这确保了所提出的线性收敛框架不仅在理论上可靠，而且在大规模分支定界应用中也具有很高的计算效率。

5 实验

在前面的章节中，我们提出了一种有理论依据的重启方案和一系列低复杂度子程序，以突破一阶方法在计算安全下界方面的计算能力限制。本节中，我们在合成数据集和真实世界数据集上开展了大量实验，以评估这些方法的实际效果。我们将所提方法与现有方法及最先进的求解器进行基准比较，以回答以下实证问题。

◇函数与近端评估效率：与通用 SOCP 求解器相比，我们用于计算 $g_{\mathcal{N}}$ 和近端算子$\rho g_{\mathcal{N}}$ 的专用算法运行时间如何？◇ 下界计算效率：与最先进的锥优化求解器相比，我们提出的一阶方法求解透视松弛 (3) 的速度有多快？我们还研究了算法框架的两项具体性质。

- 收敛速率：采用基于间隙的重启方案后，我们的方法是否表现出理论所预测的线性收敛速率?

- 硬件加速：我们的方法是否适用于 GPU? 利用 GPU 并行计算可以获得多大的加速比?

◇ 最优性认证：将我们的方法集成到分支定界框架后，与现有混合整数规划求解器相比，它能在多大程度上加速全局最优性的认证过程?

我们使用 Python 实现了所有算法。为确保最高效率，我们使用 Numba [Lam 等人， 2015] 对用于计算 $g_{\mathcal{N}}$ 和近端算子$\rho g_{\mathcal{N}}$ 的专用子程序进行即时编译。对于凸基线，我们将其与以下最先进的商业和开源二阶锥规划求解器进行比较：Gurobi [Gurobi Optimization, LLC， 2025], MOSEK [MOSEK ApS， 2025], SCS [奥多诺霍等人， 2023], 以及 Clarabel [Goulart 和 Chen, 2024], 并使用 cvxpy 包 [Diamond 和 Boyd， 2016] 作为接口。在使用分支定界框架进行最优性认证的实验中，我们直接将其与混合整数规划求解器 Gurobi 和 MOSEK 进行比较。

5.1 我们评估 $g_{\mathcal{N}}(\cdot)$ 和 $\mathbf{prox}_{\rho g_{\mathcal{N}}}(\cdot)$有多快?

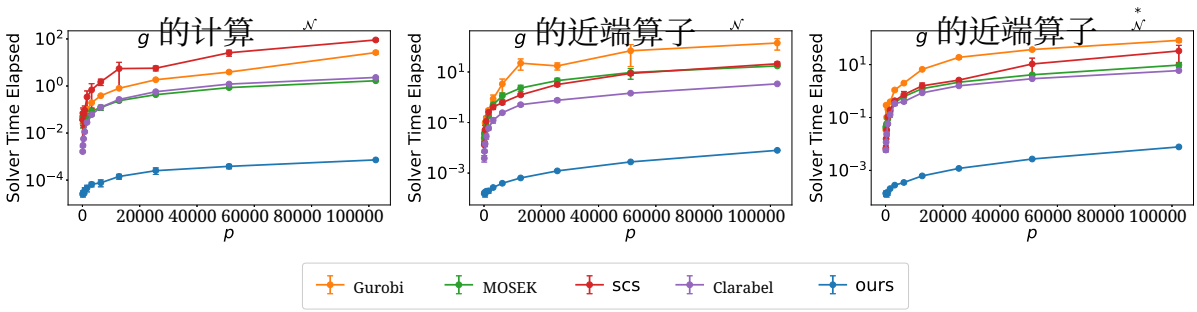

图4：评估 $g_{\mathcal{N}}$ 以及 $g_{\mathcal{N}}$ 和 $g^*_{\mathcal{N}}$的近端算子的运行时间比较，其中 $\mathcal{N}$ 是根节点。基线方法直接求解相应的 SOCP。

我们将用于评估 $g_{\mathcal{N}}$ 及其近端算子的子程序，与直接求解相应 SOCP 的二阶锥规划求解器进行比较，其中 $\mathcal{N}$ 是根节点。对于近端算子，我们同时评估 $\mathbf{prox}_{\rho g_{\mathcal{N}}}(\gamma)$ 和 $\mathbf{prox}_{\rho^{-1}g^*_{\mathcal{N}}}(\rho^{-1}\gamma)$。我们报告各方法的运行时间。每种设置重复 5 次（使用不同的随机种子），并报告均值和标准差（以误差棒表示）。我们在按如下方式生成的合成数据集上测试不同的方法：在 $\mathbb{R}^p$ 中采样输入向量 $\gamma \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$，其中 $\boldsymbol{I}_p$ 表示维度为 p 的单位矩阵。我们改变 $p \in \{2^0, \ldots, 2^{10}\} \times 10^2$，并设置 $k = 10$、 $M = 1.0$ 和$\rho = 1.0$。评估 $g_{\mathcal{N}}$ 时，我们先调用算法2来设置 $\beta = \mathbf{prox}_{\rho g_{\mathcal{N}}}(\gamma)$，以确保$\beta \in \operatorname{dom} g_{\mathcal{N}}$，满足算法1的要求。

结果如图4所示，凸显了我们方法的优越性。与传统的 SOCP 求解器相比，我们的算法1在计算 $g_{\mathcal{N}}$时实现了三个数量级的加速；算法2结合公式（14）在计算近端算子时实现了两个数量级的加速。例如，当 $p = 102400$时，基线方法计算近端算子需要数秒到数分钟，而我们的方法分别仅需 $\approx 10^{-3}$ 秒计算 $g_{\mathcal{N}}$和 $\approx 10^{-2}$ 秒计算近端算子。这种速度优势结合我们方法的精确性，在求解透视松弛时将带来显著的整体加速。请注意，我们的方法返回的是精确的值。这确保我们能够精确地计算原始目标函数 $\Phi(\beta)$，并应用基于对偶间隙的重启方案，而无需任何

近似误差。现在，我们能够高效地计算 $g_{\mathcal{N}}$ 的函数值及其近端算子，接下来将评测计算下界的速度。

5.2 我们计算下界的速度有多快?

当 $\mathcal{N}$ 为根节点时，我们将所提方法与当前最先进的 SOCP 求解器进行比较，以求解原始混合整数非线性规划问题的透视松弛。对于我们的方法，我们采用配备基于对偶间隙的重启方案（缩减因子为 $\eta = e^3$）的快速迭代收缩阈值算法（带线搜索），来求解问题 (3)，并使用我们提出的算法计算 $g_{\mathcal{N}}$ 的函数值及其近端算子。当原始-对偶间隙（分支定界的间隙）达到容差 $\epsilon = 10^{-6}$，或运行时间超过 1800 秒时，所有方法均停止。我们在合成数据集上进行线性回归和逻辑回归任务的评估。有关数据生成过程的详细信息，请参阅附录 C.1。在本实验中，我们改变特征维度 $p \in \{1000, 2000, 4000, 8000, 16000\}$。我们通过一个称为 n-to-p ratio（即样本数与特征数之比）的参数来控制样本量。我们通过一个称为 ρ的参数来控制特征相关性。我们将 n-to-p ratio 设为 1。0，特征相关性 ρ 设为 0。5，盒约束 M 设为 2，非零系数的数量（以及基数约束）设为 10，并将 ℓ_2 正则化系数 λ_2 设为 1。0。同样，我们报告基于使用不同随机种子进行的 5 次重复模拟所得的运行时间均值和标准差。

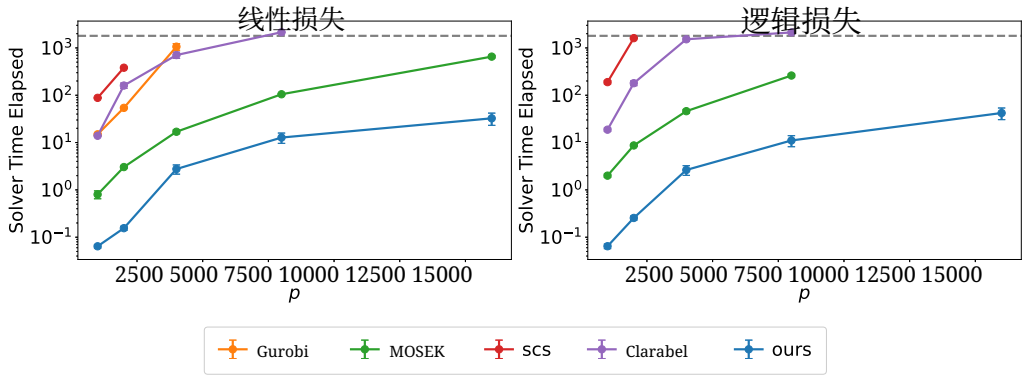

图 5：在根节点求解问题 (3) 的透视松弛的运行时间比较。我们将 $M = 2.0$、$\lambda_2 = 1.0$ 和 n-to-p ratio 设为 1。Gurobi 无法求解带基数约束的逻辑回归问题的松弛。

图 5 所示的结果表明，我们的方法比最快的锥规划求解器（MOSEK）快一个数量级以上。在规模最大的测试实例（$n = 16000$和 $p = 16000$）上，对于回归和分类任务，我们的方法都能在 100 秒内达到目标容差（10^{-6}）；而大多数基线方法在 1800 秒的时间限制内无法收敛。我们还开展了其他扰动研究，例如考察样本数与特征数之比（n-to-p）、盒约束 M 和 ℓ_2 正则化系数 λ_2，相关结果见附录 D.1。结果一致表明，在各种问题设置下，相较于现有的 SOCP 求解器，我们的方法具有显著的计算优势。加速主要源于两个因素。首先，函数值计算和近端算子求值的单次迭代计算成本都很低，这一点得到上一节理论分析和实证结果的支持。其次，我们的重启方案显著提高了近端梯度法的收敛速率，接下来我们将通过计算实验对此进行量化。

5.2.1 我们的重启方案能在多大程度上加速收敛?

我们在快速迭代收缩阈值算法（FISTA）之上实现了重启方案（分别采用固定步长和线搜索），并在 [Beck、 2017] 以及 ACFGM（自适应步长） [Li 和 Lan， 2025]的基础上进行了实现。对于 ACFGM，我们按照 Li 和 Lan 的推论 2 中所述实现了算法 [2025], ，其中 $\alpha = 0.1$。我们在回归和逻辑回归任务上运行了不同的方法，所用参数和数据生成过程与上一小节相同。每当对偶间隙缩小至原来的 η^{-1} 倍时，我们就重启加速方法，其中 $\eta = e$。作为比较，我们还运行了这些方法未采用重启方案的原始版本，以及常用的近端梯度法和采用基于函数值的启发式重启的 FISTA [O'Donoghue 和 Candes， 2015]。图 6 和图 7 展示了我们的基于对偶间隙的重启方案所带来的优势。当损失函数 F 光滑时，众所周知，加速近端梯度法在原始目标函数值上具有 $O(1/T^2)$ 的次线性收敛速率，其中 T 为迭代次数；两幅图左侧图中的结果证实了这一点。采用我们的重启方案后，我们观察到，所有加速方法在原始目标函数值、对偶目标函数值（中间图）和对偶间隙（右图）上均实现了线性收敛，这与我们的理论分析一致。由于对偶目标函数值可用作安全下界，据我们所知，这是首次通过实证展示，对于这一类混合整数非线性规划，使用一阶方法计算安全下界能够实现可证明的线性收敛。

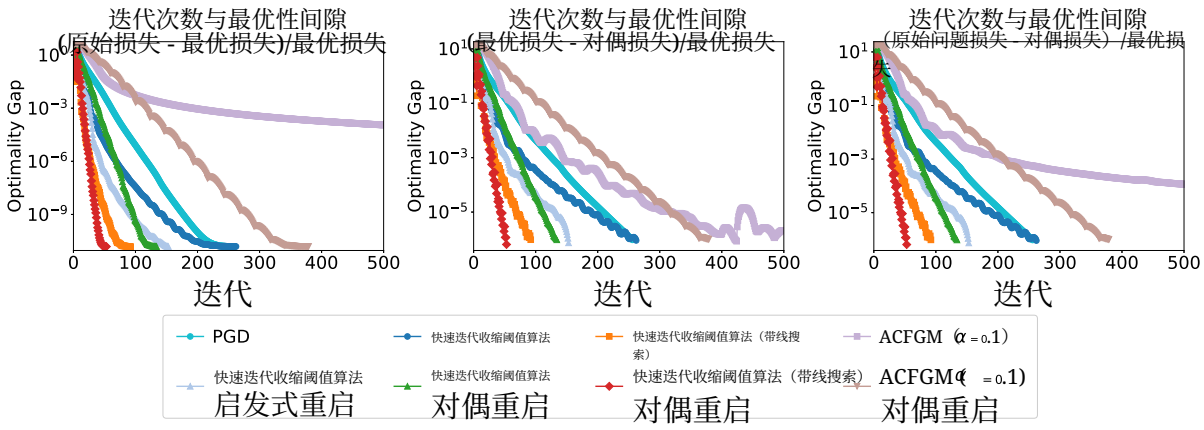

图 6：在使用线性回归损失求解问题 3 中的透视松弛时，加速方法（有重启和无重启）的收敛速度对比， $n = 16000$、$p = 16000$、 $k = 10$、 $\rho = 0.5$、 $\lambda_2 = 1.0$ 和 $M = 2.0$。

正如我们在理论部分所讨论的，只要满足 $\eta > 1$，基于对偶间隙的重启方案就能实现线性收敛。在 D.2 中，我们提供了更多数值结果，表明通过将 η 从 e 提高到 e^2或 e^3，可以改善重启方案的实际表现；在这种情况下，基于对偶间隙的重启可以媲美甚至优于启发式重启方法。为进一步说明我们的重启方案并不限于透视松弛，我们还将其用于加速近端梯度法，以求解 LASSO 问题（包括受 ℓ_1-球约束和 ℓ_1-正则化的形式）；详情见附录 D.3。

5.2.2 GPU能在多大程度上加速下界计算?

为公平比较，所有先前结果均通过在中央处理器上运行实验得出。这里，我们展示 GPU 进一步加速下界计算的潜力。由于函数值计算和近端求值的单次迭代复杂度较低，矩阵向量乘法在我们的近端算法中占主导地位，因此可以在 GPU 上加速。我们使用

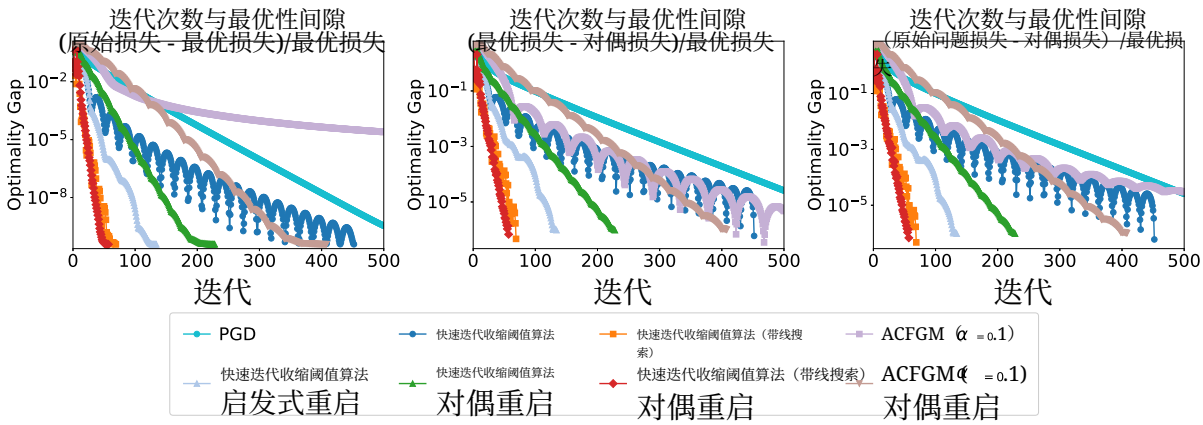

图 7：在使用逻辑回归损失求解问题 3 中的透视松弛时，加速方法（重启与不重启）的收敛速度比较，$n = 16000$、 $p = 16000$、 $k = 10$、 $\rho = 0.5$、 $\lambda_2 = 1.0$ 和 $M = 2.0$。

CuPy [Okuta等人， 2017] 进行 GPU 矩阵向量乘法。每次迭代中，我们将γ 和 β 从 GPU 传输到中央处理器，以使用我们的 Numba 实现计算近端算子$\rho g_{\mathcal{N}}(\gamma)$ 和 $g_{\mathcal{N}}(\beta)$，然后将结果传回 GPU 进行矩阵向量乘法。我们重新运行上一小节中的实验，比较仅使用中央处理器的运行与采用 GPU 加速矩阵向量乘法的运行。总运行时间汇总于表2和表3。对于线性回归和逻辑回归，GPU 实现均使高维实例（$p \geq 8000$）的运行时间额外降低了一个数量级。

表2：我们的方法在线性回归任务上的 GPU 加速。上、下两行分别对应运行时间（秒）的均值和标准差。

p	1000	2000	4000	8000	16000
我们的方法 中央处理器	0.070 (0.001)	0.127 (0.002)	0.447 (0.075)	5.239 (0.632)	12.491 (1.914)
我们的方法 GPU	0.296 (0.003)	0.282 (0.002)	0.266 (0.001)	0.444 (0.002)	0.858 (0.003)

表 3：我们的方法在逻辑回归任务上的 GPU 加速。上、下两行分别对应运行时间（秒）的均值和标准差。

p	1000	2000	4000	8000	16000
我们的方法 中央处理器	0.069 (0.006)	0.124 (0.002)	0.629 (0.028)	4.566 (0.129)	12.075 (0.824)
我们的方法 GPU	0.275 (0.013)	0.256 (0.015)	0.249 (0.013)	0.434 (0.017)	0.997 (0.029)

5.3 我们的下界计算能以多快的速度证明最优性?

最后，我们展示了本方法高效计算安全下界的能力如何实现大规模数据集的高效最优性认证。请注意，这只是首次尝试

将我们已有的方法嵌入分支定界框架。目标并非打造一个功能完备的混合整数规划求解器，而是展示使用本方法计算下界以加速分支定界过程的潜力。我们将下界计算集成到一个基础分支定界框架中，优先通过下界计算进行节点剪枝，同时有意省略高级混合整数规划启发式（例如割平面、预求解程序），以评估本方法的影响。我们将该方法与 Gurobi 和 MOSEK 进行基准比较，并报告运行时间、最优性间隙以及探索的节点数。对于逻辑回归任务上的 Gurobi，我们使用外逼近法来逼近逻辑损失，并将其记为 gurobiOA。这与图 5 并不矛盾：对于大规模逻辑回归松弛，在 Gurobi 中仅通过割平面将根节点松弛求解至最优在计算上并不可行。对于 Gurobi 和 MOSEK，我们都向其提供从本分支定界框架在根节点获得的同一当前最优解，以确保公平比较。我们在合成数据集和真实世界数据集上进行测试，包括 Santander 客户交易预测数据集 [Piedra 等人， 2019]用于线性回归，以及 DOROTHEA 药物发现数据集 [Asuncion 和 Newman， 2007] 用于逻辑回归。对于所有方法，我们将最优性间隙容差设为 10^{-6}，最大运行时间设为 7200 秒，最大内存设为 100 GB。详细的实验配置，包括数据集描述、预处理、超参数选择以及分支定界实现细节，见附录 C.3。对于真实世界数据集，我们使用束搜索启发式，通过五折交叉验证的网格搜索选择 (λ_2, k, M)；参见附录 D.4.2。

下面，我们详细说明在分支定界框架中使用我们的方法计算安全下界的两个关键方面，包括动态终止策略和热启动技术。具体而言，我们采用动态终止策略，通过提前分支或剪枝节点，避免不必要的近端梯度迭代。这样便无需等待完全收敛，也不必进行高精度下界计算。具体规则如下：

1. 如果透视松弛当前的原始损失 低于当前最优解的损失（$\hat{}\boldsymbol{\beta}$），即， $\Phi\left(\boldsymbol{\beta}^t\right) \leq f\left(\boldsymbol{X}\hat{}\boldsymbol{\beta}, \boldsymbol{y}\right) + \lambda_2\|\hat{}\boldsymbol{\beta}\|_2^2$，我们就提前终止并进入下一分支步骤，因为该下界永远不会优于当前最优解的损失，无法据此剪枝当前节点。

2. 如果透视松弛当前的对偶损失 高于当前最优解的损失（$\hat{}\boldsymbol{\beta}$），即， $\Psi\left(\boldsymbol{\zeta}^t\right) \geq f\left(\boldsymbol{X}\hat{}\boldsymbol{\beta}, \boldsymbol{y}\right) + \lambda_2\|\hat{}\boldsymbol{\beta}\|_2^2$，我们就停止下界计算并立即剪枝该节点，因为其子树中不存在更优解。

此外，我们还使用热启动来加速下界计算。具体而言，我们将父节点处透视松弛问题的解，作为当前节点处透视松弛问题的初始解。

结果见表 4–7。就运行时间和最优性间隙而言，我们的方法优于 Gurobi 和 MOSEK，且往往优势显著。特别是在较大规模的实例中，Gurobi 和 MOSEK 难以在时间限制内消除最优性间隙，而我们的方法始终能达到 0% 的最优性间隙。在所有方法都能达到最优性的实例中，我们的方法通常快一到两个数量级。

我们将这一优异表现归因于本方法能够高效计算安全下界，而这对在分支定界框架中有效剪枝至关重要。为支持这一观点，我们报告了另外两个与分支定界过程相关的指标：探索的节点数，以及计算下界所花费的时间。在本方法和 MOSEK 均达到 0% 最优性间隙的实例上，本方法探索的节点数与之相当，但计算下界所花费的时间显著更少，这有助于提升整体加速比。我们在 D.4 中提供了关于具有更大盒约束的合成数据集（$M = 3$）的更多结果。结果与 $M = 2$ 时的情况一致——本方法在运行时间和最优性间隙方面均优于 Gurobi 和 MOSEK，且优势往往十分显著。

表4：在合成（线性）数据上运行分支定界的结果，使用 $k = 10$, $\rho = 0.5$, $\lambda_2 = 1.0$, $M = 2.0$, $n/p = 1.0$。括号中为下界时间。OOM = 内存不足，TL = 时间 > 7200秒。

	ours			Gurobi			MOSEK		
p	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
1000	9 (1)	0.00	105	146	0.00	237	112	0.00	113
2000	6 (1)	0.00	59	562	0.00	100	470	0.00	57
4000	12 (6)	0.00	31	3146	0.00	34	2909	0.00	31
8000	25 (18)	0.00	21	TL	100.00	1	TL	21.00	4
16000	72 (53)	0.00	21	TL	100.00	1	OOM	OOM	OOM

表 5：使用 $k = 10$、 $\rho = 0.5$、 $\lambda_2 = 1.0$、 $M = 2.0$、 $n/p = 1.0$ 的合成数据（逻辑回归）上的分支定界结果。括号中显示下界时间。OOM = 表示内存不足，TL = 时间为 > 7200秒。

	ours			gurobiOA			MOSEK		
p	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
1000	140（7）	0.00	1121	TL	54.17	597933	386	0.00	1083
2000	110 (13)	0.00	595	TL	66.26	582581	1188	0.00	583
4000	126 (60)	0.00	219	TL	54.04	255320	2917	0.00	235
8000	180 (106)	0.00	131	TL	46.95	81405	OOM	OOM	OOM
16000	281 (187)	0.00	67	TL	35.04	21479	OOM	OOM	OOM

表6：Santander（线性）上的分支定界结果， $n = 4459$, $p = 4735$, $\lambda_2 = 1.0$, $M = 10$。括号中为下界时间。OOM = 表示内存不足，TL = 表示时间限制，为 > 7200秒。

	ours			Gurobi			MOSEK		
k	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
6	345 (259)	0.00	1359	TL	100.00	1	TL	0.33	437
7	536 (395)	0.00	2125	TL	100.00	1	TL	0.43	497
8	1206 (908)	0.00	4393	TL	100.00	1	TL	0.36	393
9	3060 (2281)	0.00	10085	TL	100.00	1	TL	0.49	303
10	时间限制 (5213)	0.01	18775	TL	100.00	1	TL	0.52	369

表 7：Dorothea（逻辑回归）上的分支定界结果， $n = 2300$, $p = 89989$, $\lambda_2 = 1.0$, $M = 10$。括号中为下界时间。内存溢出 = 内存不足，时间限制 = 时间 > 7200秒。

	ours			gurobiOA			MOSEK		
k	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
5	19 (15)	0.00	1	683	0.00	938	1075	0.00	0
10	66 (48)	0.00	23	2304	0.00	3118	OOM	OOM	OOM
15	116 (76)	0.00	33	3005	0.00	3209	OOM	OOM	OOM
20	223（142）	0.00	61	TL	0.07	3680	OOM	OOM	OOM
25	629（487）	0.00	167	TL	0.12	3635	OOM	OOM	OOM

最后，我们展示了分支定界框架的 GPU 加速。我们保持分支定界的实现不变，仅将下界计算程序替换为上一小节介绍的 GPU 加速版本。总运行时间汇总于表 8–11。为确保比较公平，中央处理器和 GPU 运行使用同一台机器（NVIDIA A100）。与根节点松弛实验的情况一样，对于线性和逻辑回归任务，GPU 加速使高维实例（表10）的下界计算运行时间最多缩短一个数量级，这

进一步加快了分支定界认证速度。这并不总能转化为总分支定界时间数量级的提升，因为仍有部分工作受限于中央处理器（例如每个节点上的可行解搜索）；完全采用原生 GPU 分支定界实现超出了本文的研究范围。对于规模较小的实例，由于 CPU–GPU 传输开销，GPU 带来的收益较小。在 D.4 中给出了盒约束更大的合成数据集（$M = 3$）上的其他结果，这些结果与 $M = 2$ 的结果一致。

表 8：线性回归中分支定界的中央处理器与 GPU 结果，$k = 10$、$\rho = 0.5$、$\lambda_2 = 1.0$、$M = 2.0$、$n/p = 1.0$。括号内为下界时间。

	我们的方法 中央处理器			我们的方法 GPU		
p	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
1000	8 (1)	0.00	105	10 (2)	0.00	105
2000	6 (1)	0.00	59	6 (1)	0.00	59
4000	4 (1)	0.00	31	4 (1)	0.00	31
8000	12 (7)	0.00	21	6 (1)	0.00	21
16000	31 (20)	0.00	21	15 (2)	0.00	21

表 9：逻辑回归中分支定界的中央处理器与 GPU 结果（$k = 10$、$\rho = 0.5$、$\lambda_2 = 1.0$、$M = 2.0$、$n/p = 1.0$）。括号中为下界时间。

	我们的方法 中央处理器			我们的方法: GPU		
p	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
1000	132 (6)	0.00	1121	141 (20)	0.00	1123
2000	94 (7)	0.00	595	99 (11)	0.00	595
4000	64 (14)	0.00	219	55 (5)	0.00	219
8000	111 (45)	0.00	131	68 (5)	0.00	131
16000	160 (72)	0.00	67	88 (8)	0.00	67

表10：分支定界中央处理器与GPU在Santander（线性）上的结果，$n = 4459$, $p = 4735$, $\lambda_2 = 1.0$, $M = 10$。括号显示下界时间。

	我们的方法中央处理器			我们的方法 GPU		
k	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
6	169 (107)	0.00	1357	83 (23)	0.00	1363
7	267 (166)	0.00	2149	145 (46)	0.00	2143
8	600 (368)	0.00	4385	337 (111)	0.00	4387
9	1655 (1058)	0.00	10087	838 (267)	0.00	10121
10	5808 (3536)	0.00	34397	3135 (909)	0.00	34379

6 结论

我们提出了一种一阶近端框架，用于求解基数约束 GLM 问题的透视松弛。通过利用原始与对偶目标函数的几何结构，我们提出了一种新的基于对偶间隙的重启方案。该方案不仅适用于特定的透视松弛设定，还可广泛应用于其他场景，并能加速算法，使其在原始问题和对偶问题上均可实现有理论保证的线性收敛速率。这首次表明，可以通过线性收敛的一阶方法获得安全下界。此外，我们还利用问题独特的数学结构，设计了

表 11：在 Dorothea（逻辑回归）上的分支定界中央处理器与 GPU 结果，$n = 2300$, $p = 89989$, $\lambda_2 = 1.0$,$M = 10$。括号中显示下界时间。

	我们的中央处理器			我们的方法（GPU）		
k	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
5	12 (11)	0.00	1	3 (1)	0.00	1
10	52 (46)	0.00	23	9 (4)	0.00	23
15	82 (69)	0.00	33	21 (6)	0.00	33
20	167 (135)	0.00	61	37 (11)	0.00	63
25	466 (403)	0.00	167	84 (30)	0.00	165

定制子程序，以高效计算隐式定义的正则化项 g_N 的函数值和近端算子；这也为 GPU 加速铺平了道路，因为计算量最大的运算变成了矩阵向量乘法。大量实证结果表明，我们的方法比最先进的求解器快 1–2 个数量级，使其成为可集成到下一代混合整数规划求解器中的实用高性能组件。

致谢

本研究通过分配编号 CIS250029 使用了国家超级计算应用中心的 Delta 系统。该分配由“高级网络基础设施协调生态系统：服务与支持”（ACCESS）计划提供支持，该计划由美国国家科学基金会的 #2138259、#2138286、#2138307、#2137603 和 #2138296 号拨款资助。

参考文献

David Applegate、Mateo D´ıaz、Oliver Hinder、Haihao Lu、Miles Lubin、Brendan O’Donoghue 和 Warren Schudy。使用原始-对偶混合梯度进行实用的大规模线性规划。载于神经信息处理系统进展，第 20243–20257 页，2021年。

Andreas Argyriou、Rina Foygel 和 Nathan Srebro。使用 k-支持范数进行稀疏预测。载于神经信息处理系统进展，第 1457–1465 页，2012年。

Arthur Asuncion 与 David Newman。UCI 机器学习资料库，2007 年 .

Alper Atamturk 与 Andr´es G´omez。基于透视松弛的 ℓ_0-回归安全筛选规则。载于第*37*届国际机器学习会议论文集，第421–430页，2020年。

Alper Atamt¨urk 与 Andr´es G´omez。带指示变量的二次优化的超模性与有效不等式。数学规划，201(1–2):295–338，2023年。

Alper Atamt¨urk、Andr´es G´omez 与 Shaoning Han。稀疏平滑信号估计：ℓ_0-模型的凸化。机器学习研究期刊，22(52):1–43，2021年。

Tiziano Bacci、Antonio Frangioni、Claudio Gentile 与 Kostas Tavlaridis-Gyparakis。带爬坡约束的机组组合问题的新混合整数非线性规划模型。运筹学，72(5):2153–2167，2024年。

Amir Beck。优化中的一阶方法。SIAM，2017年。

Amir Beck 与 Marc Teboulle。线性逆问题的快速迭代收缩阈值算法。*SIAM* 成像科学期刊，2(1):183–202，2009年。Amir Beck 与 Marc Teboulle。凸极小化及其应用的快速对偶近端梯度算法。《运筹学快报》，42(1):1–6，2014年。Dimitris Bertsimas 与 Jack Dunn。最优分类树。《机器学习》，106(7): 1039–1082，2017年。Dimitris Bertsimas 与 Wes Gurnee。通过混合整数优化学习稀疏非线性动力学。《非线性动力学》，111(7):6585–6604，2023年。Dimitris Bertsimas 与 Bart Van Parys。带多项式的稀疏分层回归。《机器学习》，109(5):973–997，2020年a。Dimitris Bertsimas 与 Bart Van Parys。稀疏高维回归：精确可扩展算法与相变。《统计年鉴》，48(1):300–323，2020年b。Dimitris Bertsimas、Jean Pauphilet 与 Bart Van Parys。稀疏回归：可扩展算法与实证性能。《统计科学》，35(4):555–578，2020年。Daniel Bienstock。一类混合整数二次规划问题的计算研究。数学规划，74(2):121–140，1996年。Frank MTA Busing。单调回归：一种简单而快速 $O(n)$的 PAVA 实现。《统计软件杂志》，102（代码片段 1）:1–25，2022年。Sebasti´an Ceria 与 Jo˜ao Soares。析取凸优化的凸规划。数学规划，86(3):595–614，1999年。George HG Chen 与 R Tyrrell Rockafellar。前向–后向分裂中的收敛速率。*SIAM* 优化期刊，7(2):421–444，1997年。Patrick L Combettes 与 Val´erie R Wajs。基于近端前向–后向分裂的信号恢复。《多尺度建模与仿真》，4(4):1168–1200，2005年。Cassio F Dantas、Emmanuel Soubies 与 C´edric F´evotte。扩展间隙安全筛选的边界。机器学习研究期刊，22(236):1–57，2021年。Sanjeeb Dash、Oktay G¨unl¨uk 与 Dennis Wei。通过列生成得到布尔决策规则。载于神经信息处理系统进展，第4655–4665页，2018年。Antonio De Rosa、Aida Khajavirad 与 Yakun Wang。论线性规划对于 K均值聚类的效力。*arXiv:2402.01061*，2024年。Antoine Dedieu、Hussein Hazimeh 与 Rahul Mazumder。学习稀疏分类器：连续优化与混合整数优化视角。机器学习研究期刊，22(135): 1–47，2021年。Steven Diamond 与 Stephen Boyd。CVXPY：一种面向凸优化的 Python 嵌入式建模语言。机器学习研究期刊，17(83):1–5，2016年。II Dikin。线性规划与二次规划问题的迭代求解。《科学院学报》，174(4):747–748，1967年。Dmitriy Drusvyatskiy 与 Adrian S Lewis。近端方法的误差界、二次增长与线性收敛。《运筹学数学》，43(3):919–948，2018年。

Olivier Fercoq 和 Zheng Qu. 局部二次增长条件下加速梯度方法的自适应重启。*IMA* 数值分析杂志，39(4):2069–2095，2019年。Olivier Fercoq、Alexandre Gramfort 和 Joseph Salmon. 注意对偶间隙：更安全的 Lasso 规则。载于第 *32* 届国际机器学习会议论文集，第 333–342 页，2015年。Rafal Goebel 和 Ralph Tyrell Rockafellar. 凸函数的局部强凸性与梯度的局部 Lipschitz 连续性。凸分析杂志，15(2):263–270，2008年。Andr´es G´omez. 通过混合整数锥二次优化检测时间序列中的异常值。*SIAM* 优化期刊，31(3):1897–1925，2021年。Andr´es G´omez 和 Jos´e Neto. 回归中的异常值检测：锥二次形式。*INFORMS* 计算杂志（即将出版），2025年。Paul J. Goulart 和 Yuwen Chen. Clarabel：用于带二次目标函数的锥规划的内点求解器。*arxiv:2405.12762*，2024年。Oktay G¨unl¨uk 和 Jeff Linderoth. 带指示变量的混合整数非线性规划的透视重构。数学规划，124(1–2):183–205，2010年。Gurobi Optimization, LLC. Gurobi 优化器参考手册，2025年。Th´éo Guyard、C´édric Herzet、Cl´ément Elvira 和 Ayse-Nur Arslan. 一种用于 ℓ_0正则化问题的新型分支定界剪枝框架。载于第 *41* 届国际机器学习会议论文集，第 48077–48096 页，2024年。Qiushi Han、Zhenwei Lin、Hanwen Liu、Caihua Chen、Qi Deng、Dongdong Ge 和 Yinyu Ye. 在 GPU 上加速基于低秩分解的半定规划算法。*arxiv:2407.15049*，2024年。Shaoning Han 和 Andr´es G´omez. 带指示变量的低秩函数的紧凑扩展形式。《运筹学数学》，50(3):1992–2016，2025年。Hussein Hazimeh 和 Rahul Mazumder. 快速最佳子集选择：坐标下降与局部组合优化算法。运筹学，68(5):1517–1537，2020年。Hussein Hazimeh、Rahul Mazumder 和 Ali Saab. 大规模稀疏回归：植根于一阶优化的分支定界。数学规划，196(1):347–388，2022年。Jean-Baptiste Hiriart-Urruty 和 Claude Lemar´echal. 凸分析基础。Springer，2004。Xiyang Hu、Cynthia Rudin 和 Margo Seltzer. 最优稀疏决策树。载于神经信息处理系统进展，第 7267–7275 页，2019年。Hamed Karimi、Julie Nutini 和 Mark Schmidt. Polyak-Lojasiewicz 条件下梯度法和近端梯度法的线性收敛。载于欧洲机器学习与数据库知识发现联合会议，第 795–811 页，2016年。James E Kelley, Jr. 求解凸规划的切平面法。工业与应用数学学会杂志，8(4):703–712，1960年。Jinhak Kim、Mohit Tawarmalani 和 Jean-Philippe P Richard. 置换不变集合的凸化及其在稀疏主成分分析中的应用。《运筹学数学》，47(4):2547–2584，2022年。

Simge Kucukyavuz、Ali Shojaie、Hasan Manzour、魏林川 和 Hao-Hsiang Wu. 用于学习贝叶斯网络的一致二阶锥整数规划。机器学习研究期刊，24(322):1–38，2023年。Siu K wan Lam、Antoine Pitrou 和 Stanley Seibert. Numba：基于 LLVM 的 Python 即时编译器。载于 Hal Finkel（编者），高性能计算中的 *LLVM* 编译器基础设施研讨会，第 7:1–7:6 页，2015年。Tianjiao Li 和 Guanghui Lan. 一种无需线搜索的简单一致最优凸优化方法。数学规划（即将出版），2025年。Wu Li. 分段凸二次规划的误差界及其应用。*SIAM* 控制与优化杂志，33(5):1510–1529，1995年。Jiachang Liu、Chudi Zhong、Boxuan Li、Margo Seltzer 和 Cynthia Rudin. Fasterrisk：快速、准确且可解释的风险评分。载于神经信息处理系统进展，第 17760–17773 页，2022a。Jiachang Liu、Chudi Zhong、Margo Seltzer 和 Cynthia Rudin. 广义线性模型与加性模型的快速稀疏分类。载于国际人工智能与统计学会议，第 9304–9333 页，2022b。Jiachang Liu、Sam Rosen、Chudi Zhong 和 Cynthia Rudin. OKRidge：可扩展的最优 k 稀疏岭回归。载于神经信息处理系统进展，第 41076–41258 页，2023年。Jiachang Liu、Rui Zhang 和 Cynthia Rudin. FastSurvival：训练 Cox 比例风险模型时隐藏的计算优势。载于神经信息处理系统进展，第 87712–87765 页，2024年。Ji achang Liu、Soroosh Shafiee 和 Andrea Lodi. 用于认证最优 k 稀疏广义线性模型的可扩展一阶方法。载于第 *42* 届国际机器学习会议论文集，第 39455–39481 页，2025年。Andrea Lodi 和 Jasone Ram´ırez-Ayerbe. 通过列生成实现一对多反事实解释。*arXiv:2402.09473*，2024年。Haihao Lu 和 Jinwen Yang. 一种适用于大规模凸二次规划的实用且最优的一阶方法。*arxiv:2311.07710*，2025年。Haihao Lu、Jinwen Yang、Haodong Hu、Qi Huangfu、Jinsong Liu、Tianhao Liu、Yinyu Ye、Chuwen Zhang 和 Dongdong Ge. cuPDLP-C：使用 C 语言实现的强化版 cuPDLP 线性规划算法。*arxiv:2312.14832*，2024年。Zhi-Quan Luo 和 Paul Tseng. 可行下降法的误差界与收敛性分析：一种通用方法。运筹学年鉴，46–47(1):157–178，1993年。Yura Malitsky 和 Konstantin Mishchenko. 凸优化的自适应近端梯度法。载于神经信息处理系统进展，第 100670–100697 页，2024年。Hasan Manzour、Simge K¨u˙c¨ukyavuz、Hao-Hsiang Wu 和 Ali Shojaie. 用整数规划从连续数据中学习有向无环图。*INFORMS* 优化杂志，3 (1):46–73，2021年。MOSEK ApS. *MATLAB* 的 *MOSEK* 优化工具箱手册。版本 *11.0.4*，2025年。

尤里·涅斯捷罗夫和阿尔卡季·涅米罗夫斯基。《凸规划中的内点多项式算法》。SIAM，1994。尤里·E·涅斯捷罗夫。求解凸规划问题的一种方法，其收敛速率为$O(1/k^2)$。《苏联科学院通报》，269(3):543–547，1983。Brendan O'Donoghue 和埃马纽埃尔·坎德斯。用于加速梯度方案的自适应重启。《计算数学基础》，15(3):715–732，2015。Brendan O'Donoghue、埃里克·朱、尼尔·帕里克和Stephen Boyd。SCS：分裂锥规划求解器，3.2.4版，2023。奥田亮介、海野雄也、西野大辅、肥田翔平和克里斯曼·卢米斯。CuPy：用于英伟达 GPU 计算的 NumPy 兼容库。发表于神经信息处理系统进展中的机器学习系统研讨会，2017。梅赛德斯·皮耶德拉、索希尔·戴恩和索拉娅·希门尼斯。桑坦德客户交易预测。
https://kaggle.com/competitions/santander-customer-transaction-prediction，2019。Kaggle竞赛。詹姆斯·雷内加。《凸优化中内点法的数学视角》。SIAM，2001。R·泰瑞尔·洛克菲勒。《凸分析》。普林斯顿大学出版社，1970。R Tyrrell Rockafellar 和罗杰·J-B·韦茨。《变分分析》。Springer，2009。文森特·鲁莱和亚历山大·达斯普雷蒙。尖锐性、重启与加速。发表于神经信息处理系统进展，第1119–1129页，2017。A·施赖弗。《线性规划与整数规划理论》。John Wiley & Sons，1998。Soroosh Shafiee 和 Fatma Kılınc-Karzan。带指示变量的秩一函数约束优化。《数学规划》，208(1–2):533–579，2024。安德烈亚斯·M·蒂尔曼、Daniel Bienstock、Andrea Lodi 和亚历山德拉·施瓦茨。基数最小化、约束与正则化：综述。《*SIAM*评论》，66(3):403–477，2024。Paul Tseng。关于凸凹优化的加速近端梯度法。技术报告，华盛顿大学数学系，2008。贝尔克·乌斯通和Cynthia Rudin。用于优化医疗评分系统的超稀疏线性整数模型。《机器学习》，102(3):349–391，2016。贝尔克·乌斯通和Cynthia Rudin。学习优化风险评分。《机器学习研究期刊》，20(150):1–75，2019。魏林川、Andr´es G´omez 和 Simge K¨uc¨ukyavuz。关于带指示变量的约束二次优化问题的凸化。发表于整数规划与组合优化，第433–447页，2020。魏林川、Andr´es G´omez 和 Simge K¨uc¨ukyavuz。带指示变量的约束凸优化问题的理想表述。《数学规划》，192(1):57–88，2022。魏林川、Alper Atamt¨urk、Andr´es G´omez 和 Simge K¨uc¨ukyavuz。关于带指示变量的凸二次优化问题的凸包。《数学规划》，204 (1–2):703–737，2024。

Laurence A Wolsey。《整数规划》。John Wiley & Sons，2020年。

Weijun Xie 和 Xinwei Deng. 稀疏岭回归的可扩展算法. *SIAM* 优化期刊, 30(4):3359–3386, 2020年.

Rui Zhang、Rui Xin、Margo Seltzer 和 Cynthia Rudin. 最优稀疏回归树. 收录于*AAAI*人工智能会议，第 11270–11279 页，2023年.

Zirui Zhou 和 Anthony Man-Cho So。结构化凸优化问题误差界的统一方法。数学规划，165(2):689–728，2017年。

附录

目录

A 补充算法 **31**

B 证明 **31**

B.1 Fenchel 共轭函数次梯度映射的误差界（引理 3.6）. . . . . 31 B.2 Fenchel 共轭函数的局部二次上界（引理 3.7）. . . . . 32 B.3 原始二次增长与对偶二次衰减条件（定理 3.5）33 B.4 原始-对偶关系（定理 3.9）.34 B.5 对偶间隙的几何衰减（命题 3.11）.35 B.6 基于间隙的重启下的线性收敛（定理 3.12）.36 B.7 （理论上的）重启参数的最优选择（定理 3.13）. . . . . 36 B.8 分支定界节点处的凸分析性质（引理 4.1）.37 B.9 分支定界节点处的 Fenchel 共轭表示（引理 4.2）. . . . . 38 B.10 分支定界节点处的分段线性二次与坚实凸结构（引理 4.3）. . . . . 38 B.11 Compute-g-Value 算法的正确性（定理 4.4）.39 B.12 Evaluate-Prox 算法的正确性（定理 4.5）.41 B.13 框架的广义线性模型正则性条件（命题 4.6）. 42 B.14 常见广义线性模型损失函数的验证.43

C 实验设置详情 **45**

C.1 求解透视松弛的设置.45 C.2 展示重启方案在 LASSO 问题上通用性的设置 46 C.3 最优性认证的设置.46 C.4 计算平台.46

D 附加实验结果 **47**

D.1 求解透视松弛.47 D.2 基于对偶性的重启方案中的收缩因子.50 D.3 除透视松弛问题外，我们的重启方案还能惠及其他问题吗？ 51 D.4 最优性认证.53

A 补充算法

PAVA算法的一个关键组成部分是能够合并相邻块并快速更新解。为完整起见，算法 3 给出了算法2中用于执行此合并操作的高效上下块算法。

算法 3: 用于算法2中第5–10行的高效上下块算法

数据: 向量 $\boldsymbol{\mu} \in \mathbb{R}^p$、非负权重 $\boldsymbol{\rho} \in \mathbb{R}^p_+$、向量 $\hat{\boldsymbol{\alpha}}$ 及其 $\hat{\alpha}_j = \text{近端算子}_{\rho_j H_M}(|\mu_j|)$ 对于 $j \in [p]$、基数 $k \in \mathbb{N}$ 和盒约束 $M > 0$。

结果:向量 $\hat{\boldsymbol{\alpha}}$。

初始化 $b = 1$, $P_1 = \rho_1$, $S_1 = |\mu_1|$, $N_b = 1$, α_1, $r_1 = 1$

$\alpha_{\text{prev}} = \hat{\alpha}_1$, $j = 2$

当 $j \le p$ **执行**

  $b = b + 1$

  $P_b = \rho_j$, $S_b = |\mu_j|$, $N_b = 1$, $\alpha = \hat{\alpha}_j$

  如果 $\alpha > \alpha_{prev}$ **then**

    $b = b - 1$

    $P_b = P_b + \rho_j, \quad S_b = S_b + |\mu_j|, \quad N_b = N_b + 1, \quad \alpha = \text{近端算子}_{\frac{P_b}{N_b} H_M}(\frac{S_b}{N_b})$

    当 $j < n$ 且 $\alpha \le \hat{\alpha}_j$ **执行**

      $j = j + 1$

      $P_b = P_b + \rho_j, \quad S_b = S_b + |\mu_j|, \quad N_b = N_b + 1, \quad \alpha = \text{近端算子}_{\frac{P_b}{N_b} H_M}(\frac{S_b}{N_b})$

    当 $b > 1$ 且 $\alpha_{b-1} < \alpha$ **执行时**

      $b = b - 1$

      $P_b = P_b + P_{b+1}, \quad S_b = S_b + S_{b+1}, \quad N_b = N_b + N_{b+1}, \quad \alpha = \text{近端算子}_{\frac{P_b}{N_b} H_M}(\frac{S_b}{N_b})$

  $\alpha_b = \alpha, \quad r_b = j$

  $\alpha_{\text{prev}} = \alpha$, $j = j + 1$

对于 $l = 1, \ldots, b$ **do**

  $\hat{\alpha}_{[r_{l-1}+1:r_l]} = \alpha_l$

返回$\hat{\boldsymbol{\nu}}$

B 证明

本节给出论文中主要理论结果的完整证明。为方便起见，我们在给出每个结果的证明之前先重述该结果。

B.1 Fenchel 共轭函数次梯度映射的误差界（引理 3.6）

引理 *3.6.* 设 h 是一个真闭凸函数，且不存在使 h 有限且为仿射函数的直线。若 h 相对于某个 $\boldsymbol{\nu} \in \text{int}(\text{dom}(h^*))$ 强凸，则对于任意满足 $\boldsymbol{\nu} \in \mathcal{U}$ 的紧集 $\mathcal{U} \subset \text{int}(\text{dom}(h^*))$，存在一个常数 $\alpha_{\mathcal{U}} > 0$，使得

$$\text{dist}\left(\partial h^*(\boldsymbol{\mu}), \partial h^*(\boldsymbol{\nu})\right) \le \frac{2}{\alpha_{\mathcal{U}}} \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2, \qquad \forall \boldsymbol{\mu} \in \mathcal{U}.$$

证明。 证明：由于不存在使 h 有限且为仿射函数的直线，h^* 的定义域具有非空内部 [（Rockafellar, 1970，推论 13.4）。定义 h 相对于 $\boldsymbol{\nu}$ 的倾斜函数为 $h_{\boldsymbol{\nu}}(\boldsymbol{\gamma}) := h(\boldsymbol{\gamma}) - \boldsymbol{\nu}^\top \boldsymbol{\gamma}$。由于 h 是真、闭且凸的，其共轭函数 h^* 也是真、闭且凸的 [（Rockafellar，1970，定理 12）。此外，由于 $\mathcal{U} \subset$ 属于 int(dom(h^*)),

[（Rockafellar，1970，定理 24）7] 蕴含 $\partial h^*(\mathcal{U})$ 非空且紧。由于 h 相对于 $\boldsymbol{\nu}$ 是强凸的，因此对于紧集 $\partial h^*(\mathcal{U})$，存在常数 $\alpha_{\mathcal{U}} > 0$，使得对于任意 $\boldsymbol{\gamma} \in \partial h^*(\mathcal{U})$，函数 $h_{\boldsymbol{\nu}}$ 满足二次增长条件

$$h_{\boldsymbol{\nu}}(\boldsymbol{\gamma}) \geq \inf(h_{\boldsymbol{\nu}}) + \frac{\alpha_{\mathcal{U}}}{2} \operatorname{dist}^2\left(\boldsymbol{\gamma}, (\partial h_{\boldsymbol{\nu}})^{-1}(\mathbf{0})\right).$$

根据 [Drusvyatskiy 和 Lewis，2018年，定理 3.3], 上述二次增长条件蕴含次微分误差界不等式

$$\operatorname{dist}\left(\boldsymbol{\gamma}, (\partial h_{\boldsymbol{\nu}})^{-1}(\mathbf{0})\right) \leq \frac{2}{\alpha_{\mathcal{U}}} \operatorname{dist}(\mathbf{0}, \partial h_{\boldsymbol{\nu}}(\boldsymbol{\gamma})).$$

为简化该不等式，首先注意到倾斜函数的次微分满足$\partial h_{\boldsymbol{\nu}}(\boldsymbol{\gamma}) = \partial h(\boldsymbol{\gamma}) - \boldsymbol{\nu}$，这意味着 $\operatorname{dist}(\mathbf{0}, \partial h_{\boldsymbol{\nu}}(\boldsymbol{\gamma})) = \operatorname{dist}(\boldsymbol{\nu}, \partial h(\boldsymbol{\gamma}))$。其次，根据 [Rockafellar，1970年，推论 23.5.1], 恒等式 $(\partial h_{\boldsymbol{\nu}})^{-1} = \partial h^*_{\boldsymbol{\nu}}$ 成立。此外，由于倾斜函数的共轭函数满足 $h^*_{\boldsymbol{\nu}}(\cdot) = h^*(\cdot + \boldsymbol{\nu})$，我们有 $(\partial h_{\boldsymbol{\nu}})^{-1}(\mathbf{0}) = \partial h^*_{\boldsymbol{\nu}}(\mathbf{0}) = \partial h^*(\boldsymbol{\nu})$。将这两个关系代入误差界可得

$$\operatorname{dist}(\boldsymbol{\gamma}, \partial h^*(\boldsymbol{\nu})) \leq \frac{2}{\alpha_{\mathcal{U}}} \operatorname{dist}(\boldsymbol{\nu}, \partial h(\boldsymbol{\gamma})), \quad \forall \boldsymbol{\gamma} \in \partial h^*(\mathcal{U}). \tag{B.1}$$

给定某个 $\boldsymbol{\mu} \in \mathcal{U}$，取某个 $\boldsymbol{\beta} \in \partial h^*(\boldsymbol{\mu})$。于是有

$$\operatorname{dist}(\partial h^*(\boldsymbol{\mu}), \partial h^*(\boldsymbol{\nu})) \leq \operatorname{dist}(\boldsymbol{\beta}, \partial h^*(\boldsymbol{\nu})) \leq \frac{2}{\alpha_{\mathcal{U}}} \operatorname{dist}(\boldsymbol{\nu}, \partial h(\boldsymbol{\beta})) \leq \frac{2}{\alpha_{\mathcal{U}}} \operatorname{dist}(\boldsymbol{\nu}, \boldsymbol{\mu}),$$

其中，第一个不等式由 $\boldsymbol{\beta} \in \partial h^*(\boldsymbol{\mu})$ 得出，第二个不等式由 $\boldsymbol{\beta} \in \partial h^*(\mathcal{U})$ 和不等式 B.1 得出。最后一个不等式成立，是因为 $\boldsymbol{\beta} \in \partial h^*(\boldsymbol{\mu})$ 给出 $\boldsymbol{\mu} \in \partial h(\boldsymbol{\beta})$ [Rockafellar，1970，推论 23.5.1]。利用 $\operatorname{dist}(\boldsymbol{\nu}, \boldsymbol{\mu}) = \|\boldsymbol{\nu} - \boldsymbol{\mu}\|_2$，证明得证。

□

B.2 Fenchel 共轭的局部二次上界（引理 3.7）

引理 *3.7*。设 h 是一个真闭凸函数，且不存在使 h 有限且为仿射函数的直线。若 h 相对于某个 $\boldsymbol{\nu} \in \operatorname{int}(\operatorname{dom}(h^*))$ 强凸，则对于任意满足 $\boldsymbol{\nu} \in \mathcal{U}$ 的紧集 $\mathcal{U} \subset \operatorname{int}(\operatorname{dom}(h^*))$，存在常数 $\alpha_{\mathcal{U}} > 0$ 和一个向量 $\boldsymbol{s} \in \partial h^*(\boldsymbol{\nu})$，使得

$$h^*(\boldsymbol{\mu}) \leq h^*(\boldsymbol{\nu}) + \boldsymbol{s}^\top(\boldsymbol{\mu} - \boldsymbol{\nu}) + \frac{\alpha_{\mathcal{U}}}{2}\|\boldsymbol{\mu} - \boldsymbol{\nu}\|^2, \quad \forall \boldsymbol{\mu} \in \mathcal{U}.$$

证明。证明 回顾引理 3.6 的证明可知，h^* 是真、闭且凸的，并且 $\operatorname{dom}(h^*)$ 是凸集，且内部非空。由于凸集的内部本身也是凸集 [Rockafellar, 1970, 定理 6.2], $\operatorname{int}(\operatorname{dom}(h^*))$ 是凸的。令 $\mathcal{K}$ 表示 $\mathcal{U}$的凸包。由于 $\mathcal{U}$ 是紧集，且包含于凸集 $\operatorname{int}(\operatorname{dom}(h^*))$ 中，因此 $\mathcal{K}$ 是紧集 [Rockafellar, 1970, 定理 17.2] 并满足 $\mathcal{K} \subset \operatorname{int}(\operatorname{dom}(h^*))$。

由于 $\boldsymbol{\nu} \in \mathcal{U}$，对于任意固定的 $\boldsymbol{\mu} \in \mathcal{U}$，连接 $\boldsymbol{\nu}$与 $\boldsymbol{\mu}$的整条线段都属于$\mathcal{K}$。根据凸函数积分形式的中值定理 [Hiriart-Urruty 和 Lemar´echal, 2004, 定理 2.3.4], ，我们有

$$h^*(\boldsymbol{\mu}) - h^*(\boldsymbol{\nu}) = \int_0^1 \boldsymbol{s}_t^\top(\boldsymbol{\mu} - \boldsymbol{\nu})\,\mathrm{d}t$$

对于任意次梯度的选取 $\boldsymbol{s}_t \in \partial h^*(\boldsymbol{\nu} + t(\boldsymbol{\mu} - \boldsymbol{\nu}))$。注意，由于 h^* 是真凸函数，它在其定义域 [的内部几乎处处可微（Rockafellar，1970，定理 25.5]）。因此，对于几乎所有 $t \in [0, 1]$, $\boldsymbol{s}_t$ 都是唯一的，从而确保该路径可测且积分有定义。

现在，我们构造一条伴随路径 $\boldsymbol{r}_t$ ，以界定该积分。对于每个 $t \in [0,1]$，令$\boldsymbol{r}_t := \text{proj}_{\partial h^*(\boldsymbol{\nu})}(\boldsymbol{s}_t)$ 为 $\boldsymbol{s}_t$ 在紧凸集 $\partial h^*(\boldsymbol{\nu})$ 上的投影。由于次梯度映射在紧集 $\mathcal{K}$ [上有界（Rockafellar, 1970，定理 24.7]），$\boldsymbol{s}_t$ 有界。此外，由于投影映射连续且 $\boldsymbol{s}_t$ 可测，$\boldsymbol{r}_t$ 既有界又可测。因此，积分 $\boldsymbol{s}_0 := \int_0^1 \boldsymbol{r}_t \mathrm{d}t$ 有定义。

利用中值定理等式，我们可以写为：

$$\begin{aligned} h^*(\boldsymbol{\mu}) - h^*(\boldsymbol{\nu}) &= \int_0^1 (\boldsymbol{s}_t - \boldsymbol{r}_t + \boldsymbol{r}_t)^\top (\boldsymbol{\mu} - \boldsymbol{\nu}) \, \mathrm{d}t \\ &\leq \left(\int_0^1 \|\boldsymbol{s}_t - \boldsymbol{r}_t\|_2 \cdot \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2 \, \mathrm{d}t \right) + (\boldsymbol{\mu} - \boldsymbol{\nu})^\top \boldsymbol{s}_0, \end{aligned} \tag{B.2}$$

其中，不等式由柯西-施瓦茨不等式得出。

根据投影的定义，$\|\boldsymbol{s}_t - \boldsymbol{r}_t\|_2 = \text{dist}(\boldsymbol{s}_t, \partial h^*(\boldsymbol{\nu}))$。由于 $\mathcal{K}$ 是凸且紧的，并且 $\boldsymbol{s}_t \in \partial h^*(\boldsymbol{\nu} + t(\boldsymbol{\mu} - \boldsymbol{\nu}))$ 满足 $\boldsymbol{\nu} + t(\boldsymbol{\mu} - \boldsymbol{\nu}) \in \mathcal{K}$，引理 3.6 表明存在一个常数 $\alpha_{\mathcal{K}} > 0$ ，使得

$$\|\boldsymbol{s}_t - \boldsymbol{r}_t\|_2 = \text{dist}\left(\boldsymbol{s}_t, \partial h^*(\boldsymbol{\nu})\right) \leq \frac{2}{\alpha_{\mathcal{K}}} \|\boldsymbol{\nu} + t(\boldsymbol{\mu} - \boldsymbol{\nu}) - \boldsymbol{\nu}\|_2 = \frac{2}{\alpha_{\mathcal{K}}} t \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2$$

对任意 $t \in [0,1]$ 均成立。因此，我们可以将第一个积分项界定如下

$$\int_0^1 \|\boldsymbol{s}_t - \boldsymbol{r}_t\|_2 \cdot \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2 \, \mathrm{d}t \leq \int_0^1 \frac{2t}{\alpha_{\mathcal{K}}} \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2^2 \, \mathrm{d}t = \frac{1}{\alpha_{\mathcal{K}}} \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2^2.$$

接下来只需证明 $\boldsymbol{s}_0 \in \partial h^*(\boldsymbol{\nu})$。为反证起见，假设 $\boldsymbol{s}_0 \notin \partial h^*(\boldsymbol{\nu})$。由于 $\partial h^*(\boldsymbol{\nu})$ 非空、紧且凸，根据分离超平面定理（Rockafellar，1970，定理 11.4], ），存在一个向量 $\boldsymbol{a}$，将 $\boldsymbol{s}_0$ 与 $\partial h^*(\boldsymbol{\nu})$ 严格分离，即

$$\boldsymbol{a}^\top \boldsymbol{s}_0 > \sup_{\boldsymbol{s} \in \partial h^*(\boldsymbol{\nu})} \boldsymbol{a}^\top \boldsymbol{s}.$$

然而，由于对所有 $t \in [0,1]$ 都有 $\boldsymbol{r}_t \in \partial h^*(\boldsymbol{\nu})$，我们还得到 $\boldsymbol{a}^\top \boldsymbol{r}_t \leq \sup_{\boldsymbol{s} \in \partial h^*(\boldsymbol{\nu})} \boldsymbol{a}^\top \boldsymbol{s}$。对该不等式在 $t \in [0,1]$ 上积分，得到

$$\boldsymbol{a}^\top \boldsymbol{s}_0 = \int_0^1 \boldsymbol{a}^\top \boldsymbol{r}_t \mathrm{d}t \leq \sup_{\boldsymbol{s} \in \partial h^*(\boldsymbol{\nu})} \boldsymbol{a}^\top \boldsymbol{s},$$

这就导致了矛盾。因此，必然有 $\boldsymbol{s}_0 \in \partial h^*(\boldsymbol{\nu})$。将这两个积分项的上述结果代回 (B.2)，得到

$$h^*(\boldsymbol{\mu}) - h^*(\boldsymbol{\nu}) \leq \frac{1}{\alpha_{\mathcal{K}}} \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2^2 + \boldsymbol{s}_0^\top (\boldsymbol{\mu} - \boldsymbol{\nu}).$$

我们通过定义 $\alpha_{\mathcal{U}} := \frac{2}{\alpha_{\mathcal{K}}}$ 和 $\boldsymbol{s} := \boldsymbol{s}_0$ 来完成证明。 □

B.3 原始二次增长与对偶二次衰减条件（定理 3.5）

定理 *3.5.* 在假设 3.1–3.4 下，令 $\boldsymbol{\zeta}^\star$ 表示唯一的对偶最优解。对于任意满足 $\boldsymbol{\zeta}^\star \in \mathcal{Z}$、$-\mathcal{Z} \subset \text{int}(\text{dom}(F^*))$ 且$\boldsymbol{X}^\top \mathcal{Z} \subset \text{int}(\text{dom}(G^*))$的紧集 $\mathcal{B} \subset \mathbb{R}^p$ 和 $\mathcal{Z} \subset \mathbb{R}^n$ ，存在常数 $\alpha_{\mathcal{B}}$，$\kappa_{\mathcal{Z}} > 0$ ，使得

$$\Phi(\boldsymbol{\beta}) \geq \Phi^\star + \frac{\alpha_{\mathcal{B}}}{2} \text{dist}^2(\boldsymbol{\beta}, \mathcal{B}^\star), \forall \boldsymbol{\beta} \in \mathcal{B}, \quad \text{and} \quad -\Psi(\boldsymbol{\zeta}) \leq -\Psi^\star + \frac{\kappa_{\mathcal{Z}}}{2} \|\boldsymbol{\zeta} - \boldsymbol{\zeta}^\star\|^2, \forall \boldsymbol{\zeta} \in \mathcal{Z}.$$

证明. 定理 3.5 的证明 在假设 3.1 和 3.2 下，引理 3.3 保证对偶解 $\zeta^\star$ 存在且唯一。我们将证明分为两部分。

第一个不等式直接由 [Drusvyatskiy 和 Lewis, 2018] 中的误差界结果得出。具体而言，[Drusvyatskiy 和 Lewis, 2018, Corollary 4.3] 的结构性要求由假设 3.4 (ii) 和假设 3.4 (iii) 满足。此外，假设 3.4 (i) 保证最优解集 $\mathcal{B}^\star$ 是紧的。这种紧致性确保增长常数 $\alpha_{\mathcal{B}}$ 严格为正，并且在紧集 $\mathcal{B}$ 上一致。至此，第一个结论的证明完成。

接下来，我们通过将引理 3.7 应用于对偶目标函数分量来推导第二个不等式。首先，考虑 F^*。令 $h = F$ 和 $\boldsymbol{\nu} = -\boldsymbol{\zeta}^\star$。假设 3.4 (iv) 确保 $-\boldsymbol{\zeta}^\star \in$ 属于定义域内部(dom(F^*))，这满足引理 3.7 的条件。此外，假设 3.4 (ii) 提供了坚实凸性。因此，存在 $\boldsymbol{s}_1 \in \partial F^*(-\boldsymbol{\zeta}^\star)$ 和 $\alpha_1 > 0$，使得对所有 $\boldsymbol{\zeta} \in \mathcal{Z}$:

$$F^*(-\boldsymbol{\zeta}) \leq F^*(-\boldsymbol{\zeta}^\star) + \boldsymbol{s}_1^\top (-\boldsymbol{\zeta} + \boldsymbol{\zeta}^\star) + \frac{\alpha_1}{2}\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^\star\|_2^2. \tag{B.3}$$

接下来，考虑 G^*。令 $h = G$ 和 $\boldsymbol{\nu} = \boldsymbol{X}^\top\boldsymbol{\zeta}^\star$。根据假设 3.4 (iv)，有 $\boldsymbol{X}^\top\boldsymbol{\zeta}^\star \in$int(dom($G^*$))。再次应用引理 3.7，可得存在 $\boldsymbol{s}_2 \in \partial G^*(\boldsymbol{X}^\top\boldsymbol{\zeta}^\star)$ 和 $\alpha_2 > 0$，使得对所有 $\boldsymbol{\zeta} \in \mathcal{Z}$:

$$\begin{aligned} G^*(\boldsymbol{X}^\top\boldsymbol{\zeta}) &\leq G^*(\boldsymbol{X}^\top\boldsymbol{\zeta}^\star) + \boldsymbol{s}_2^\top \left(\boldsymbol{X}^\top\boldsymbol{\zeta} - \boldsymbol{X}^\top\boldsymbol{\zeta}^\star\right) + \frac{\alpha_2}{2}\|\boldsymbol{X}^\top\boldsymbol{\zeta} - \boldsymbol{X}^\top\boldsymbol{\zeta}^\star\|_2^2 \\ &\leq G^*(\boldsymbol{X}^\top\boldsymbol{\zeta}^\star) + (\boldsymbol{X}\boldsymbol{s}_2)^\top (\boldsymbol{\zeta} - \boldsymbol{\zeta}^\star) + \frac{\alpha_2\|\boldsymbol{X}\|_2^2}{2}\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^\star\|_2^2. \end{aligned} \tag{B.4}$$

Co将 (B.3) 和 (B.4) 与定义 $-\Psi(\boldsymbol{\zeta}) = F^*(-\boldsymbol{\zeta}) + G^*(\boldsymbol{X}^\top\boldsymbol{\zeta})$ 结合，可得

$$-\Psi(\boldsymbol{\zeta}) \leq -\Psi^\star + (-\boldsymbol{s}_1 + \boldsymbol{X}\boldsymbol{s}_2)^\top (\boldsymbol{\zeta} - \boldsymbol{\zeta}^\star) + \frac{\alpha_1 + \alpha_2\|\boldsymbol{X}\|_2^2}{2}\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^\star\|_2^2.$$

我们声称 $-\boldsymbol{s}_1 + \boldsymbol{X}\boldsymbol{s}_2 = \boldsymbol{0}$。令 $\boldsymbol{d} := -\boldsymbol{s}_1 + \boldsymbol{X}\boldsymbol{s}_2$。由于假设 3.2 保证 int(dom($\Psi$)) 非空，且 $\boldsymbol{\zeta}^\star$ 是一个最优解，因此当步长 $t > 0$充分小时，点 $\boldsymbol{\zeta}_t := \boldsymbol{\zeta}^\star - t\boldsymbol{d}$ 仍然可行。将 $\boldsymbol{\zeta}_t$ 代入不等式，得到

$$-\Psi(\boldsymbol{\zeta}_t) \leq -\Psi^\star - t\|\boldsymbol{d}\|_2^2 + \mathcal{O}(t^2).$$

如果 $\boldsymbol{d} \neq \boldsymbol{0}$，那么对于足够小的 t，我们会有 $-\Psi(\boldsymbol{\zeta}_t) < -\Psi^\star$。然而，$\Psi^\star$ 是最优值，因此 $\Psi(\boldsymbol{\zeta}_t) \leq \Psi^\star$，这就产生了矛盾。因此，必有 $-\boldsymbol{s}_1 + \boldsymbol{X}\boldsymbol{s}_2 = \boldsymbol{0}$。于是，令 $\kappa_{\mathcal{Z}} := \alpha_1 + \alpha_2\|\boldsymbol{X}\|_2^2$，即可得出第二个结论。 □

B.4 原始-对偶关系（定理 3.9）

定理 *3.9*。 在假设 3.1–3.8 下，令 $\mathcal{B}^\star$ 表示原始最优解的集合，$\boldsymbol{\zeta}^\star$ 表示唯一的对偶最优解。令 $\mathcal{B} \subseteq$dom(G) 为一个紧集，使得$\mathcal{B}^\star \subseteq \mathcal{B}$。定义相应的对偶像集 $\mathcal{Z} := \{-\nabla F(\boldsymbol{X}\boldsymbol{\beta}) \mid \boldsymbol{\beta} \in \mathcal{B}\}$，并假设内部条件 $-\mathcal{Z} \subset$ int(dom(F^*)) 和 $\boldsymbol{X}^\top\mathcal{Z} \subset$ int(dom(G^*)) 成立。则存在常数 $\sigma_{\mathcal{B}}$，$\kappa_{\mathcal{B}} > 0$，使得对于任意原始向量 $\boldsymbol{\beta} \in \mathcal{B}$ 及其诱导对偶向量$\boldsymbol{\zeta} := -\nabla F(\boldsymbol{X}\boldsymbol{\beta})$，有

$$\frac{\sigma_{\mathcal{B}}}{2}\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^\star\|_2^2 \leq \Phi(\boldsymbol{\beta}) - \Phi^\star \quad \text{and} \quad \Psi(\boldsymbol{\zeta}^\star) - \Psi(\boldsymbol{\zeta}) \leq \frac{\kappa_{\mathcal{B}}}{\sigma_{\mathcal{B}}}\big(\Phi(\boldsymbol{\beta}) - \Phi^\star\big).$$

证明。 定理 3.9 的证明 我们首先证明对偶距离与原始间隙之间的第一个不等式。根据定义，$\boldsymbol{\zeta} = -\nabla F(\boldsymbol{X}\boldsymbol{\beta})$。根据假设 3.8，函数 F^* 是本质光滑的 [Rockafellar, 1970, 定理 26.3]。因此，利用假设 $-\mathcal{Z} \subset$ int(dom(F^*))，

函数 F^* 在 $-\zeta$处可微。应用芬切尔恒等式 [Rockafellar，1970，定理 23.5], ，我们有

$$-\zeta = \nabla F(\boldsymbol{X}\boldsymbol{\beta}) \quad \Longleftrightarrow \quad \nabla F^*(-\zeta) = \boldsymbol{X}\boldsymbol{\beta}.$$

由于 F 是局部光滑的（假设3.1），其共轭函数 F^* 是局部强凸的 [（Goebel 和 Rockafellar，2008，定理4.1]）。具体而言，在紧集 $-\mathcal{Z}$ 上对点 $-\zeta^\star$ 和 $-\zeta$应用强凸性的定义，存在一个常数 $\sigma_{\mathcal{B}} > 0$ ，使得

$$F^*(-\zeta^\star) \geq F^*(-\zeta) + \nabla F^*(-\zeta)^\top (\zeta - \zeta^\star) + \frac{\sigma_{\mathcal{B}}}{2}\|\zeta - \zeta^\star\|_2^2.$$

代入 $\nabla F^*(-\zeta) = \boldsymbol{X}\boldsymbol{\beta}$ 并整理各项，我们得到

$$\frac{\sigma_{\mathcal{B}}}{2}\|\zeta - \zeta^\star\|_2^2 \leq F^*(-\zeta^\star) - F^*(-\zeta) - (\zeta^\star - \zeta)^\top \boldsymbol{X}\boldsymbol{\beta}. \tag{B.5}$$

现在，考虑原始最优性间隙 $\Phi(\boldsymbol{\beta}) - \Phi^\star$。利用恒等式 $\Phi(\boldsymbol{\beta}) = F(\boldsymbol{X}\boldsymbol{\beta}) + G(\boldsymbol{\beta})$和 $\Phi^\star = -F^*(-\zeta^\star) - G^*(\boldsymbol{X}^\top\zeta^\star)$，我们有

$$\begin{aligned}\Phi(\boldsymbol{\beta}) - \Phi^\star &= F(\boldsymbol{X}\boldsymbol{\beta}) + F^*(-\zeta^\star) + G(\boldsymbol{\beta}) + G^*(\boldsymbol{X}^\top\zeta^\star) \\ &= \left[F^*(-\zeta^\star) - F^*(-\zeta) - \zeta^\top \boldsymbol{X}\boldsymbol{\beta}\right] + G(\boldsymbol{\beta}) + G^*(\boldsymbol{X}^\top\zeta^\star),\end{aligned}$$

其中，由于 $\zeta = -\nabla F(\boldsymbol{X}\boldsymbol{\beta})$，等号成立，因此芬切尔不等式取等，且我们有$F(\boldsymbol{X}\boldsymbol{\beta}) = -F^*(-\zeta) - \zeta^\top \boldsymbol{X}\boldsymbol{\beta}$。为匹配式（B.5）的形式，加上并减去$(\zeta^\star)^\top \boldsymbol{X}\boldsymbol{\beta}$，得到

$$\begin{aligned}\Phi(\boldsymbol{\beta}) - \Phi^\star &= \left[F^*(-\zeta^\star) - F^*(-\zeta) - (\zeta^\star - \zeta)^\top \boldsymbol{X}\boldsymbol{\beta}\right] + \left[G(\boldsymbol{\beta}) + G^*(\boldsymbol{X}^\top\zeta^\star) - (\zeta^\star)^\top \boldsymbol{X}\boldsymbol{\beta}\right] \\ &\geq \frac{\sigma_{\mathcal{B}}}{2}\|\zeta - \zeta^\star\|_2^2,\end{aligned}$$

其中，该不等式成立是因为式 (B.5)，以及对 G 和配对 $(\boldsymbol{\beta},\ \boldsymbol{X}^\top\zeta^\star)$ 应用 Fenchel 不等式。这证明了第一个结论。

对于第二个不等式，我们应用定理 3.5。由于 $\zeta \in \mathcal{Z}$ 和定理 3.5 的假设均成立，存在一个常数 $\kappa_{\mathcal{B}} > 0$ ，使得

$$\Psi(\zeta^\star) - \Psi(\zeta) \leq \frac{\kappa_{\mathcal{B}}}{2}\|\zeta - \zeta^\star\|_2^2.$$

将其与第一个不等式结合，即可得到所需结果。 □

定理 3.9 建立的结构关系表明，对偶间隙 $\Phi(\boldsymbol{\beta}^t) - \Psi(\zeta^t)$ 可作为到最优性距离的稳健替代指标。我们现在利用这一性质，设计一种通用的重启方案，以实现线性收敛。

B.5 对偶间隙的几何衰减（命题 3.11）

命题 *3.11*。 在假设3.1–3.10下，设 $\alpha_{\mathcal{B}},\ \sigma_{\mathcal{B}},\ \kappa_{\mathcal{B}} > 0$ 为定理 3.9 中导出的几何常数。则原对偶迭代 $\{(\boldsymbol{\beta}^t,\ \zeta^t)\}_{t\geq 1}$ 满足对偶间隙

$$\Phi(\boldsymbol{\beta}^t) - \Psi(\zeta^t) \leq \left(\frac{2r(t)(1 + \kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}{\alpha_{\mathcal{B}}}\right)\left(\Phi(\boldsymbol{\beta}^0) - \Psi(\zeta^0)\right).$$

证明。 命题 3.11 的证明：由定理 3.9 可得不等式 $\Psi(\zeta^\star) - \Psi(\zeta^t) \leq \frac{\kappa_{\mathcal{B}}}{\sigma_{\mathcal{B}}}(\Phi(\boldsymbol{\beta}^t) - \Phi^\star)$。在两边加上 $\Phi(\boldsymbol{\beta}^t) - \Phi^\star$，由于强对偶成立，我们可以界定总对偶间隙

$$\Phi(\boldsymbol{\beta}^t) - \Psi(\zeta^t) \leq \left(1 + \frac{\kappa_{\mathcal{B}}}{\sigma_{\mathcal{B}}}\right)\left(\Phi(\boldsymbol{\beta}^t) - \Phi^\star\right).$$

将 (9) 中的算法速率假设代入该不等式，可得

$$\Phi(\boldsymbol{\beta}^t) - \Psi(\boldsymbol{\zeta}^t) \le r(t)\left(1 + \frac{\kappa_{\mathcal{B}}}{\sigma_{\mathcal{B}}}\right)\operatorname{dist}^2(\boldsymbol{\beta}^0, \mathcal{B}^\star) \le \left(\frac{2r(t)(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}{\alpha_{\mathcal{B}}}\right)(\Phi(\boldsymbol{\beta}^0) - \Phi^\star),$$

其中，第二个不等式由定理 3.5 得出。证明最后指出，根据弱对偶性，我们有 $\Phi(\boldsymbol{\beta}^0) - \Phi^\star \le \Phi(\boldsymbol{\beta}^0) - \Psi(\boldsymbol{\zeta}^0)$。 □

B.6 基于间隙的重启下的线性收敛（定理 3.12）

定理 *3.12*。 假设命题 3.11 的条件成立，且 $r(t)$ 单调递减。令 $(\boldsymbol{\beta}^{\text{init}}, \boldsymbol{\zeta}^{\text{init}})$ 为初始原始–对偶对。对于任意目标收缩因子$\eta > 1$，定义重启间隔 $t_{\max}$ 为

$$t_{\max} := \left\lceil r^{-1}\left(\frac{\alpha_{\mathcal{B}}}{2\eta(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}\right)\right\rceil,$$

其中， r^{-1} 是速率函数的逆函数。如果算法每经过 $t_{\max}$ 次迭代便重启一次（重置 $t=0$ 并将 $\boldsymbol{\beta}^0$ 设为当前迭代点），则重启点处的对偶间隙序列以如下速率线性收敛

$$\Phi(\boldsymbol{\beta}^{s\cdot t_{\max}}) - \Psi(\boldsymbol{\zeta}^{s\cdot t_{\max}}) \le \eta^{-s}\left(\Phi(\boldsymbol{\beta}^{\text{init}}) - \Psi(\boldsymbol{\zeta}^{\text{init}})\right).$$

证明。 定理 3.12 的证明 从第一个重启周期开始，该周期起始于 $t=0$，结束于$t = t_{\max}$。根据命题 3.11，迭代 $t_{\max}$ 时的对偶间隙满足

$$\Phi(\boldsymbol{\beta}^{t_{\max}}) - \Psi(\boldsymbol{\zeta}^{t_{\max}}) \le \left[\frac{2r(t_{\max})(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}{\alpha_{\mathcal{B}}}\right]\left(\Phi(\boldsymbol{\beta}^{\text{init}}) - \Psi(\boldsymbol{\zeta}^{\text{init}})\right).$$

现在我们 分析收缩因子。根据重启间隔 $t_{\max}$ 的定义，我们有

$$t_{\max} \ge r^{-1}\left(\frac{\alpha_{\mathcal{B}}}{2\eta(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}\right).$$

由于 $r(\cdot)$ 是单调递减函数，对两边应用 r 会使不等号反向 y:

$$r(t_{\max}) \le \frac{\alpha_{\mathcal{B}}}{2\eta(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}.$$

将此上界代入对偶间隙不等式中，几何常数相互抵消，我们得到

$$\begin{aligned}\Phi(\boldsymbol{\beta}^{t_{\max}}) - \Psi(\boldsymbol{\zeta}^{t_{\max}}) &\le \left[\frac{2(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}{\alpha_{\mathcal{B}}}\cdot\frac{\alpha_{\mathcal{B}}}{2\eta(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}\right]\left(\Phi(\boldsymbol{\beta}^{\text{init}}) - \Psi(\boldsymbol{\zeta}^{\text{init}})\right)\\ &= \eta^{-1}\left(\Phi(\boldsymbol{\beta}^{\text{init}}) - \Psi(\boldsymbol{\zeta}^{\text{init}})\right).\end{aligned}$$

这证明经过一个完整的重启周期后，对偶间隙至少按因子 η^{-1} 收缩。由于算法在每个周期结束时都会重新初始化（将新的 $\boldsymbol{\beta}^0$ 设为当前迭代点），这种收缩可以递归应用。因此，经过 s 次重启（总计跨越 $s\cdot t_{\max}$ 次迭代）后，我们得到定理陈述中的界。证明完毕。 □

B.7（理论）重启参数的最优选择（定理 3.13）

定理 *3.13*。 假设收敛速率的形式为 $r(t) \approx C/t^q$，其中 $q>0$ 和 $C>0$。那么，常数最优的重启参数取值为 $\eta = e^q$。

证明。 证明：给定原始-对偶间隙的容差水平 $\epsilon > 0$ 以及某个重启衰减因子 $\eta > 1$，若忽略向上取整操作，迭代次数近似为

$$T_{\max} \approx \log_{\eta}\left(\frac{\Phi(\boldsymbol{\beta}^{\text{init}}) - \Psi(\boldsymbol{\zeta}^{\text{init}})}{\epsilon}\right) \cdot r^{-1}\left(\frac{\alpha_{\mathcal{B}}}{2\eta(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}\right)$$

在假设 $r(t) \approx C/t^q$ 的情况下，我们有

$$T_{\max} \approx \log_{\eta}\left(\frac{\Phi(\boldsymbol{\beta}^{\text{init}}) - \Psi(\boldsymbol{\zeta}^{\text{init}})}{\epsilon}\right) \cdot \left(\frac{2C\eta(1+\kappa_{\mathcal{B}}/\sigma_{\mathcal{B}})}{\alpha_{\mathcal{B}}}\right)^{1/q}.$$

忽略不依赖于 η的项后，迭代复杂度为

$$T_{\max} \propto \log_{\eta}\left(\frac{1}{\epsilon}\right)\eta^{1/q} = \log\left(\frac{1}{\epsilon}\right)\frac{\eta^{1/q}}{\log\eta}.$$

对 $\eta^{1/q}/\log\,\eta$ 关于 η求导并令其等于零，我们有

$$\frac{d}{d\eta}\left(\frac{\eta^{1/q}}{\log\eta}\right) = \frac{\eta^{1/q-1}\left(\frac{1}{q}\log\eta - 1\right)}{(\log\eta)^2} = 0 \quad\Longleftrightarrow\quad \log\eta = q \quad\Longleftrightarrow\quad \eta = e^q.$$

$\eta^{1/q}/\log\,\eta$的二阶导数为

$$\frac{d^2}{d\eta^2}\left(\frac{\eta^{1/q}}{\log\eta}\right) = \frac{\eta^{1/q-2}}{(\log\eta)^3}\left(\frac{1-q}{q^2}(\log\eta)^2 + \frac{q-2}{q}\log\eta + 2\right), \qquad (\eta > 1,\ q > 0).$$

在 $\eta = e^q$ 处求值（使得 $\log\eta = q$），得到

$$\left.\frac{d^2}{d\eta^2}\left(\frac{\eta^{1/q}}{\log\eta}\right)\right|_{\eta=e^q} = \frac{(e^q)^{1/q-2}}{q^3}\left(\frac{1-q}{q^2}q^2 + \frac{q-2}{q}q + 2\right) = \frac{e^{1-2q}}{q^3} > 0.$$

因此， $\eta = e^q$使 $\eta^{1/q}/\log\,\eta$ 取得局部最小值，且这是唯一的局部最小值，所以它是 $\eta > 1$ 的全局最小值，也是重启衰减因子的最优选择。 □

B.8 分支定界节点处的凸分析性质（引理 4.1）

引理 *4.1.* 式 (5) 中的隐函数 $g_{\mathcal{N}}$ 是真、闭且凸的，且定义域紧。

证明. 证明 定义函数 $H\colon \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ 如下

$$H(\boldsymbol{\beta}, \boldsymbol{z}) = \frac{1}{2}\sum_{j\in[p]} \beta_j^2/z_j + \delta_{\mathcal{D}}(\boldsymbol{\beta}, \boldsymbol{z}).$$

首先，注意到每一项 β_j^2/z_j 都是凸二次函数的透视函数，因此是真、闭且凸的 [Shafiee 和 Kılınc-Karzan，2024年，引理 1 & 2]。其次，由于定义域 $\mathcal{D}$ 是非空的有界多面体，示性函数 $\delta_{\mathcal{D}}$ 也是真、闭且凸的。因此，由于 H 是若干真、闭且凸函数之和，它也保有这些性质 [Rockafellar，1970，定理 9.3]。

函数 $g_{\mathcal{N}}$ 定义为关于 $\boldsymbol{z}$对 H 进行部分极小化。由于定义域 $\mathcal{D}$ 有界， H 的衰退函数的定义域仅包含零向量。因此，Rockafellar [1970,定理 9.2]的衰退条件得以满足，这意味着 $g_{\mathcal{N}}$ 是真、闭且凸的。最后， $g_{\mathcal{N}}$ 的定义域对应于将紧集 $\mathcal{D}$ 投影到 $\boldsymbol{\beta}$坐标上。由于投影是线性映射，且 $\mathcal{D}$ 是紧多面体，dom($g_{\mathcal{N}}$) 是紧多面体 [Rockafellar，1970，定理 19.3]。

□

B.9 分支定界节点处的 Fenchel 共轭表示（引理 4.2 ）

引理 *4.2*。 $g_{\mathcal{N}}$ 的 Fenchel 共轭为

$$g_{\mathcal{N}}^{*}(\boldsymbol{\alpha})=\sum_{j\in\mathcal{J}_1}H_M(\alpha_j)\ +\mathrm{TopSum}_{k-|\mathcal{J}_1|}\left(\boldsymbol{H}_M\left(\boldsymbol{\alpha}_{\mathcal{J}_f}\right)\right).$$

证明。引理 4.2 的证明 回顾一下，Huber 损失函数具有如下变分形式

$$H_M(\alpha)=\sup_{\beta\in\mathbb{R}}\left\{\alpha\beta-\beta^2/2\ :\ |\beta|\le M\right\}.$$

根据 Fenchel 共轭的定义，我们有

$$g_{\mathcal{N}}^{*}(\boldsymbol{\alpha})=\sup_{(\boldsymbol{\beta},\boldsymbol{z})\in\mathcal{D}}\ \boldsymbol{\alpha}^{\top}\boldsymbol{\beta}-\frac{1}{2}\sum_{j\in[p]}\frac{\beta_j^2}{z_j}.$$

利用定义域 $\mathcal{D}$ 的定义，我们可以将上确界按划分$\mathcal{J}_0$、 $\mathcal{J}_1$、 $\mathcal{J}_f$分解。

$$\begin{aligned}g_{\mathcal{N}}^{*}(\boldsymbol{\alpha})=&\sup\left\{\sum_{j\in\mathcal{J}_0}\alpha_j\beta_j-\frac{\beta_j^2}{2z_j}\ :\ z_j=\beta_j=0,\ \forall j\in\mathcal{J}_0\right\}\\&+\sup\left\{\sum_{j\in\mathcal{J}_1}\alpha_j\beta_j-\frac{\beta_j^2}{2z_j}\ :\ z_j=1,|\beta_j|\le M,\ \forall j\in\mathcal{J}_1\right\}\\&+\sup\left\{\sum_{j\in\mathcal{J}_f}\alpha_j\beta_j-\frac{\beta_j^2}{2z_j}\ :\ \sum_{j\in\mathcal{J}_f}z_j\le k-|\mathcal{J}_1|,\ z_j\in[0,1],\ |\beta_j|\le Mz_j,\ \forall j\in\mathcal{J}_f\right\}.\end{aligned}$$

第一项显然为零。根据 Huber 损失的变分形式，第二项的和为 $\sum_{j\in\mathcal{J}_1}H_M(\alpha_j)$。对于第三项，在固定 z_j 的情况下对 β_j 进行优化可得

$$\sup_{|\beta_j|\le Mz_j}\left\{\alpha_j\beta_j-\frac{\beta_j^2}{2z_j}\right\}=z_j\sup_{|\beta_j/z_j|\le M}\left\{\alpha_j\frac{\beta_j}{z_j}-\frac{1}{2}\left(\frac{\beta_j}{z_j}\right)^2\right\}=z_jH_M(\alpha_j).$$

代入后，第三个上确界变为关于$\boldsymbol{z}_{\mathcal{J}_f}$的以下线性规划

$$\sup\left\{\sum_{j\in\mathcal{J}_f}z_jH_M(\alpha_j):\sum_{j\in\mathcal{J}_f}z_j\le k-|\mathcal{J}_1|,\ z_j\in[0,1],\forall j\in\mathcal{J}_f\right\}=\mathrm{TopSum}_{k-|\mathcal{J}_1|}\left(\boldsymbol{H}_M\left(\boldsymbol{\alpha}_{\mathcal{J}_f}\right)\right),\tag{B.6}$$

其中，最后一个等式源于这样一个事实：最优策略是将 $z_j=1$ 分配给非负权重 $H_M(\alpha_j)$中数值最大的那些索引。对这三部分求和，即完成证明。

□

B.10 分支定界节点处的分段线性二次与坚实凸结构（引理 4.3 ）

引理 *4.3.* 函数 $g_{\mathcal{N}}$ 和 $g_{\mathcal{N}}^{*}$ 均为分段线性二次函数且强凸。

证明。 证明 我们首先证明 $g_{\mathcal{N}}^{*}$是分段线性二次函数。回顾一下， $g_{\mathcal{N}}^{*}(\boldsymbol{\alpha})=h_1(\boldsymbol{\alpha})+h_2(\boldsymbol{\alpha})$，其中 $h_1(\boldsymbol{\alpha}):=\sum_{j\in\mathcal{J}_1}H_M(\alpha_j)$，以及 $h_2(\boldsymbol{\alpha}):=\mathrm{TopSum}_{\bar{k}}\left(\boldsymbol{H}_M(\boldsymbol{\alpha}_{\mathcal{J}_f})\right)$。根据定义，Huber函数 H_M是分段线性二次函数。由于凸分段线性二次函数的类别在加法下是闭的 [Rockafellar 和

Wets，2009年，练习10。22]， h_1 是分段线性二次函数。对于第二项，令 $\bar{p} = |\mathcal{J}_f|$ 和 $k = k - |\mathcal{J}_1|$。我们将与 $\boldsymbol{\alpha}_{\mathcal{J}_f}$ 相关的子空间 $\mathbb{R}^{\bar{p}}$ 划分为由 l索引的 $\binom{\bar{p}}{\bar{k}}$ 个排序区域

$$\mathcal{D}_l := \{\boldsymbol{\alpha} \in \mathbb{R}^p \mid |\alpha_i| \geq |\alpha_j|,\ \forall i \in \mathcal{I}_l,\ \forall j \in \mathcal{J}_f \setminus \mathcal{I}_l\}.$$

在每个区域 $\mathcal{D}_l$ 内，TopSum算子会将其选取固定为集合 $\mathcal{I}_l$。我们进一步根据 Huber函数的阈值对每个 $\mathcal{D}_l$ 进行细分。定义区间如下:

$$I_1 := (M, +\infty), \quad I_2 := [0, M], \quad I_3 := [-M, 0), \quad I_4 := (-\infty, -M].$$

我们将 $\mathcal{D}_l$ 划分为 $4^{\bar{p}}$ 个多面体胞腔 $\mathcal{D}_{l,\ \boldsymbol{m}}$，并以向量 $\boldsymbol{m} \in \{1,2,3,4\}^{\bar{p}}$ 为索引，使得对每个 $j \in \mathcal{J}_f$， $\alpha_j \in I_{m_j}$。这些集合的并满足 $\bigcup_l \bigcup_{\boldsymbol{m}} \mathcal{D}_{l,\ \boldsymbol{m}} = \mathbb{R}^p = \mathrm{dom}(g_{\mathcal{N}}^*)$。

在任意特定的胞腔 $\mathcal{D}_{l,\ \boldsymbol{m}}$ 上，函数 h_2 可化简为集合 $\mathcal{I}_l$ 上一组固定的 Huber项之和。具体而言，取决于区间索引 m_j 对应于二次区域（$|\alpha_j| \leq M$）还是线性区域（$|\alpha_j| > M$），该函数的形式为:

$$h_2(\boldsymbol{\alpha}) = \sum_{j \in \mathcal{I}_l \cap \mathcal{Q}_{\boldsymbol{m}}} \frac{1}{2}\alpha_j^2 + \sum_{j \in \mathcal{I}_l \cap \mathcal{L}_{\boldsymbol{m}}} \left(M|\alpha_j| - \frac{1}{2}M^2 \right),$$

其中， $\mathcal{Q}_{\boldsymbol{m}}$ 和 $\mathcal{L}_{\boldsymbol{m}}$ 分别是映射到二次区间和线性区间的索引集合。由于该表达式仅由线性项和二次项组成，h_2 是分段线性二次函数。分段线性二次函数之和仍为分段线性二次函数（Rockafellar 和 Wets，2009，习题 10.22]），因此 $g_{\mathcal{N}}^*$ 是分段线性二次函数。

接下来证明 $g_{\mathcal{N}}$ 是分段线性二次的。根据引理 4.1， $g_{\mathcal{N}}$ 是真、闭且凸的，因此$g_{\mathcal{N}}^*$ 也是真、闭且凸的，且 $g_{\mathcal{N}} = (g_{\mathcal{N}}^*)^*$ [Rockafellar，1970，定理 12.2]。由于真、闭且凸的分段线性二次函数的 Fenchel 共轭也为分段线性二次函数 [Rockafellar 和 Wets，2009年，定理11.14(b)]，可知 $g_{\mathcal{N}}$ 是分段线性二次的。

最后，我们证明 $g_{\mathcal{N}}$和 $g_{\mathcal{N}}^*$具有坚实凸性。由于这两个函数都是凸的分段线性二次函数，其倾斜形式定义为 $h_{\boldsymbol{\nu}}(\cdot) := h(\cdot) - \boldsymbol{\nu}^\top(\cdot)$，对于任意向量 $\boldsymbol{\nu} \in \mathbb{R}^p$仍显然是凸的分段线性二次函数。众所周知，真、闭且凸的分段线性二次函数在任意紧集上都满足二次增长条件 [Li，1995，定理 2.7]。因此， $g_{\mathcal{N}}$ 和 $g_{\mathcal{N}}^*$ 都满足相对于任意向量 $\boldsymbol{\nu}$的坚实凸性定义。证明完毕。

□

B.11 计算 g 值算法的正确性（定理 4.4)

定理 *4.4.* 对于任意 $\boldsymbol{\beta} \in \mathrm{dom}(g_{\mathcal{N}})$，令 $\boldsymbol{\beta} := \boldsymbol{\beta}_{\mathcal{J}_f}$, $\bar{p} := |\mathcal{J}_f|$，且 $k := k - |\mathcal{J}_1|$。则

$$g_{\mathcal{N}}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{j \in \mathcal{J}_1} \beta_j^2 + \frac{1}{2}\min_{\boldsymbol{\omega}} \left\{ \sum_{j=1}^{\bar{k}} \omega_j^2 \;\middle|\; \begin{array}{l} \boldsymbol{\omega} \geq \mathbf{0}, \quad M \geq \omega_1 \geq \cdots \geq \omega_{\bar{k}} \geq 0, \\ \omega_{\bar{k}+1} = \cdots = \omega_{\bar{p}} = 0, \quad \boldsymbol{\omega} \succeq_m |\bar{\boldsymbol{\beta}}| \end{array} \right\}.$$

这里， $\boldsymbol{\omega} \succeq_m |\boldsymbol{\beta}|$ 表示 $\boldsymbol{\omega}$ 主序于 $|\boldsymbol{\beta}|$，即，

$$\sum_{j=1}^{\ell} \omega_j \geq \sum_{j=1}^{\ell} |\bar{\beta}_{(j)}| \quad \forall \ell = 1, \ldots, \bar{p} - 1, \qquad \text{and} \qquad \sum_{j=1}^{\bar{p}} \omega_j = \sum_{j=1}^{\bar{p}} |\bar{\beta}_{(j)}|.$$

算法1计算出上述问题的一个极小化子 $\boldsymbol{\omega}$，从而精确计算 $g_{\mathcal{N}}(\boldsymbol{\beta})$，其计算复杂度为 $\mathcal{O}(p + \bar{p}\ \log\ k + k)$。

证明。 定理 4.4 的证明 我们首先通过根据节点 $\mathcal{N}$ 定义的划分 $\mathcal{J}_0, \mathcal{J}_1, \mathcal{J}_f$ 分解问题，简化对 $g_{\mathcal{N}}(\boldsymbol{\beta})$ 的求值。根据定义域 $\mathcal{D}$ 和函数 $g_{\mathcal{N}}$ 的定义，式 (5) 中的优化问题可分解如下

$$g_{\mathcal{N}}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{j \in \mathcal{J}_0} \frac{\beta_j^2}{0} + \frac{1}{2}\sum_{j \in \mathcal{J}_1} \beta_j^2 + \frac{1}{2}\min_{\boldsymbol{z}_{\mathcal{J}_f} \in \mathbb{R}^{\bar{p}}} \left\{ \sum_{j \in \mathcal{J}_f} \frac{\beta_j^2}{z_j} \;:\; \begin{array}{l} z_j \in [0,1],\ \forall j \in \mathcal{J}_f, \\ |\beta_j| \leq M z_j,\ \forall j \in \mathcal{J}_f, \\ \sum_{j \in \mathcal{J}_f} z_j \leq \bar{k}, \end{array} \right\},$$

其中 $k := k - |\mathcal{J}_1|$ 且 $\bar{p} := |\mathcal{J}_f|$。注意，当且仅当对所有 $j \in \mathcal{J}_0$ 都有 $\beta_j = 0$ 时，第一项才是有限的（此时其值为零）。给定 $\boldsymbol{\beta}$ 后，第二项为常数。因此，核心难点在于第三项。为简化记法，我们将第三项（忽略1 2）重新表述为

$$\min_{\bar{\boldsymbol{z}} \in \mathbb{R}^{\bar{p}}} \left\{ \sum_{j \in [\bar{p}]} \frac{\bar{\beta}_j^2}{\bar{z}_j} \;:\; \begin{array}{l} \bar{\boldsymbol{z}} \in [0,1]^{\bar{p}}, \quad \mathbf{1}^\top \bar{\boldsymbol{z}} \le \bar{k}, \\ |\bar{\beta}_j| \le M \bar{z}_j, \ \forall j \in [\bar{p}], \end{array} \right\}, \tag{B.7}$$

其中 $\bar{\boldsymbol{\beta}} := \boldsymbol{\beta}_{\mathcal{J}_f}$ 且 $\bar{\boldsymbol{z}} := \boldsymbol{z}_{\mathcal{J}_f}$。由于目标函数和约束仅通过自由变量的绝对值 $\bar{\boldsymbol{\beta}}$ 来依赖于 $\bar{\ }$，且具有置换不变性，因此我们将自由变量的绝对值按如下方式排序 $|\beta_{(1)}| \ge |\beta_{(2)}| \ge \cdots \ge |\beta_{(\bar{p})}|$。该表达式等价于半个平方 k-支撑范数的变分刻画（Argyriou 等人，2012年；Kim 等人， 2022]）。

现在，我们不再直接处理 $\bar{\boldsymbol{z}}$ ，而是通过基于主序的凸包表示，仅基于 $\boldsymbol{\beta}$ 推导一种计算方法。考虑以下三个集合：

- $\mathcal{S}_1 := \left\{ (t, \bar{\boldsymbol{\beta}}) \,\middle|\, \|\bar{\boldsymbol{\beta}}\|_2^2 \le t, \, \|\bar{\boldsymbol{\beta}}\|_\infty \le M, \, \|\bar{\boldsymbol{\beta}}\|_0 \le \bar{k} \right\}$,
- $\mathcal{S}_2 := \left\{ (t, \bar{\boldsymbol{\beta}}) \,\middle|\, \exists \bar{\boldsymbol{z}} \in [0,1]^{\bar{p}} \text{ s.t. } \sum_{j \in [\bar{p}]} \bar{\beta}_j^2 / \bar{z}_j \le t, \ \mathbf{1}^\top \bar{\boldsymbol{z}} \le \bar{k}, \ |\bar{\beta}_j| \le M \bar{z}_j \ \forall j \in [\bar{p}] \right\}$,
- $\mathcal{S}_3 := \left\{ (t, \bar{\boldsymbol{\beta}}) \,\middle|\, \exists \boldsymbol{\omega} \in \mathbb{R}^{\bar{p}} \text{ s.t. } \begin{array}{l} \boldsymbol{\omega} \ge \mathbf{0}, \quad \sum_{j \in [\bar{p}]} \omega_j^2 \le t, \quad M \ge \omega_1 \ge \cdots \ge \omega_{\bar{k}} \ge 0, \\ \omega_{\bar{k}+1} = \cdots = \omega_{\bar{p}} = 0, \quad \boldsymbol{\omega} \succeq_m |\bar{\boldsymbol{\beta}}| \end{array} \right\}$,

其中 $|\boldsymbol{\beta}|$ 表示逐元素绝对值， $\boldsymbol{\omega} \succeq_m |\boldsymbol{\beta}|$ 表示 $\boldsymbol{\omega}$ 主序于 $|\boldsymbol{\beta}|$，即，

$$\sum_{j=1}^{\ell} \omega_j \ge \sum_{j=1}^{\ell} |\bar{\beta}_{(j)}| \quad \forall \ell = 1, \ldots, \bar{p} - 1, \qquad \text{and} \qquad \sum_{j=1}^{\bar{p}} \omega_j = \sum_{j=1}^{\bar{p}} |\bar{\beta}_{(j)}|.$$

如果 $\mathcal{S}_2$ 具有二元约束 $\bar{\boldsymbol{z}} \in \{0,1\}^{\bar{p}}$，那么它将与 $\mathcal{S}_1$ 重合。将 $\bar{z}_j \in \{0,1\}$ 放宽为 $\bar{z}_j \in [0,1]$ 可得到（闭）凸包，因此 $\mathcal{S}_2$ 是 $\mathcal{S}_1$ [的闭凸包（Shafiee and Kılınc-Karzan， 2024]）。此外，由于 $\mathcal{S}_1$ 在符号变换和置换下保持不变，Kim 等人 [2022,的定理 4] 表明， $\mathcal{S}_3$ 也是 $\mathcal{S}_1$ 的闭凸包。因此， $\mathcal{S}_2 = \mathcal{S}_3$，且 (B.7) 的最优值可写为

$$\begin{aligned} \min \left\{ t \,\middle|\, (t, \bar{\boldsymbol{\beta}}) \in \mathcal{S}_2 \right\} &= \min \left\{ t \,\middle|\, (t, \bar{\boldsymbol{\beta}}) \in \mathcal{S}_3 \right\} \\ &= \min_{\boldsymbol{\omega}} \left\{ \sum_{j=1}^{\bar{k}} \omega_j^2 \;\middle|\; \begin{array}{l} \boldsymbol{\omega} \ge \mathbf{0}, \quad M \ge \omega_1 \ge \cdots \ge \omega_{\bar{k}} \ge 0, \\ \omega_{\bar{k}+1} = \cdots = \omega_{\bar{p}} = 0, \quad \boldsymbol{\omega} \succeq_m |\bar{\boldsymbol{\beta}}| \end{array} \right\}. \end{aligned} \tag{B.8}$$

接下来，我们证明算法1构造了 (B.8) 中的极小化子。令 $\theta := \sum_{j=1}^{\bar{p}} |\bar{\beta}_{(j)}|$ 表示总和。由于 $\boldsymbol{\omega} \succeq_m |\boldsymbol{\beta}|$，我们必有 $\omega_1 \ge |\beta_{(1)}|$。此外，总和相等以及排序关系 $\omega_1 \ge \cdots \ge \omega_{\bar{k}}$ 意味着 $\bar{k}\,\omega_1 \ge \sum_{j=1}^{\bar{k}} \omega_j = \theta$，因此 $\omega_1 \ge \theta / \bar{k}$。因此， $\omega_1 \ge \max\left(|\bar{\beta}_{(1)}|, \theta/\bar{k}\right)$。如果 $\theta/\bar{k} \ge |\bar{\beta}_{(1)}|$，那么选择 $\omega_1 = \cdots = \omega_{\bar{k}} = \theta/\bar{k}$ 可在所有总和固定为 θ 的向量中使 $\sum_{j=1}^{\bar{k}} \omega_j^2$ 最小（根据 $x \mapsto x^2$ 的凸性）。否则，极小化子必须满足 $\omega_1 = |\beta_{(1)}|$，我们便在剩余坐标上重复相同的论证，并将 $\bar{\ }$ 更新为剩余总和 $\theta \leftarrow \theta - |\beta_{(1)}|$。这就得到了算法1中实现的迭代“剥离”规则：在第 j 次迭代中，计算候选平均值 $\theta := \theta/(k - j + 1)$；如果 $\theta \ge |\beta_{(j)}|$，则令 $\omega_j = \cdots = \omega_{\bar{k}} = \theta$ 并停止；否则令 $\omega_j = |\beta_{(j)}|$ 并更新 $\theta \leftarrow \theta - |\beta_{(j)}|$。返回的目标函数值等于 (B.8) 的最优值，因此也等于 (B.7) 的最优值。

最后，上述构造同样给出了与 Kim 等人 [2022, 命题 6] 中相同的熟知闭式表达式。令 $\kappa \in \{0, \ldots, k-1\}$ 表示“剥离”坐标的个数，即，

算法将其设为 $\omega_j = |\beta_{(j)}|$ 的指标个数。于是 $\omega_1 = |\beta_{(1)}|$, ……, $\omega_\kappa = |\beta_{(\kappa)}|$ 与 $\omega_{\kappa+1} = \cdots = \omega_{\bar{k}} = \frac{1}{\bar{k}-\kappa}\sum_{j=\kappa+1}^{\bar{p}} |\bar{\beta}_{(j)}|$。代入 (B.8) 可得

$$\sum_{j=1}^{\kappa} \bar{\beta}_{(j)}^2 + \frac{1}{\bar{k}-\kappa}\left(\sum_{j=\kappa+1}^{\bar{p}} |\bar{\beta}_{(j)}|\right)^2,$$

其中 κ 是满足该条件的唯一整数

$$|\bar{\beta}_{(\kappa)}| > \frac{1}{\bar{k}-\kappa}\sum_{j=\kappa+1}^{\bar{p}} |\bar{\beta}_{(j)}| \geq |\bar{\beta}_{(\kappa+1)}|.$$

在原始的 $\bar{z}$表述（B.7）中，同一分裂指标 κ 对应于约束 $\bar{z}_j \leq 1$ 处于有效状态（$\bar{z}^\star_{(j)} = 1$）的坐标个数，而对于 $j > \kappa$，剩余的预算则按比例分配为 $\bar{z}^\star_{(j)} \propto |\beta_{(j)}|$。

该算法的计算成本主要由对自由变量 $\mathcal{J}_f$的运算所主导。提取 $\boldsymbol{\beta}$ 需要 $\mathcal{O}(\bar{p})$时间。排序步骤通常需要 $\mathcal{O}(\bar{p}\log\ \bar{p})$；然而，由于最优解实际上仅依赖于前 k 个分量以及残差和，因此可通过部分排序优化为 $\mathcal{O}(\bar{p}\log\ k)$。随后用于确定 κ 的线性扫描以及最终求和最多执行 $\mathcal{O}(k)$次运算。计入处理 固定指标所需的 $\mathcal{O}(p)$时间后，总复杂度为 $\mathcal{O}(p + \bar{p}\log\ k + k)$。至此证明完成。

□

B.12 Evaluate-Prox 算法的正确性（定理 4.5）

定理 *4.5.* 对于任意 $\boldsymbol{\beta} \in \mathbb{R}^p$ 和 $\rho > 0$，算法2可精确计算 $\mathrm{prox}_{\rho g_{\mathcal{N}}^*}(\boldsymbol{\beta})$，其计算复杂度为 $\mathcal{O}(p + \bar{p}\log\ \bar{p})$，其中 $\bar{p} = |\mathcal{J}_f|$ 和 $k = k - |\mathcal{J}_1|$。此外，通过扩展莫罗分解，算法2可用作子程序来精确计算 $\mathrm{prox}_{\rho g_{\mathcal{N}}}(\boldsymbol{\beta})$

$$\mathrm{prox}_{\rho g_{\mathcal{N}}}(\boldsymbol{\beta}) = \boldsymbol{\beta} - \rho\,\mathrm{prox}_{\rho^{-1} g_{\mathcal{N}}^*}\left(\rho^{-1}\boldsymbol{\beta}\right).$$

证明. 定理 4.5 的证明 原始问题更新式 (14) 直接由扩展莫罗分解得出 [Beck，2017，定理 6.45]。因此，我们重点讨论如何计算 $\mathrm{prox}_{\rho g_{\mathcal{N}}^*}$。

回顾引理 4.2，$g_{\mathcal{N}}^*$可在指标集 $\mathcal{J}_0$、$\mathcal{J}_1$、和 $\mathcal{J}_f$ 上可分解。因此，近端优化问题可按这些指标集分解。令

$$\boldsymbol{\alpha}^\star := \mathrm{prox}_{\rho g_{\mathcal{N}}^*}(\boldsymbol{\beta}) = \operatorname*{arg\,min}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\ \frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2^2 + \rho g_{\mathcal{N}}^*(\boldsymbol{\alpha}).$$

利用 $g_{\mathcal{N}}^*$ 的可分离形式，目标函数可写为

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2^2 + \rho g_{\mathcal{N}}^*(\boldsymbol{\alpha}) =& \frac{1}{2}\|\boldsymbol{\alpha}_{\mathcal{J}_0} - \boldsymbol{\beta}_{\mathcal{J}_0}\|_2^2 \quad + \quad \frac{1}{2}\|\boldsymbol{\alpha}_{\mathcal{J}_1} - \boldsymbol{\beta}_{\mathcal{J}_1}\|_2^2 + \rho\sum_{j\in\mathcal{J}_1} H_M(\alpha_j) \\ &+ \frac{1}{2}\|\boldsymbol{\alpha}_{\mathcal{J}_f} - \boldsymbol{\beta}_{\mathcal{J}_f}\|_2^2 + \rho\,\mathrm{TopSum}_{\bar{k}}\left(\boldsymbol{H}_M(\boldsymbol{\alpha}_{\mathcal{J}_f})\right),\end{aligned}$$

其中 $k = k - |\mathcal{J}_1|$。前两个子问题可平凡地由 $\boldsymbol{\alpha}^\star_{\mathcal{J}_0} = \boldsymbol{\beta}_{\mathcal{J}_0}$ 和 $\boldsymbol{\alpha}^\star_{\mathcal{J}_1} =$[近端算子$_{\rho H_M}(\beta_j)]_{j\in\mathcal{J}_1}$ 求解，即对每个 $j \in \mathcal{J}_1$ 的 Huber 损失的标准近端算子。

核心难点在于求解最后一个优化问题。为简化记号，令 $\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}_{\mathcal{J}_f}$ 和 $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_{\mathcal{J}_f}$。我们的目标是求解重新表述后的问题

$$\min_{\bar{\boldsymbol{\alpha}}\in\mathbb{R}^{\bar{p}}}\ \frac{1}{2}\|\bar{\boldsymbol{\alpha}} - \bar{\boldsymbol{\beta}}\|_2^2 + \rho\,\mathrm{TopSum}_{\bar{k}}\left(\boldsymbol{H}_M(\bar{\boldsymbol{\alpha}})\right).$$

我们利用最优解 $\bar{\boldsymbol{\alpha}}^\star$的两个结构性质来简化问题。首先，由于Huber 损失是偶函数，因此有 $\operatorname{sgn}(\bar{\boldsymbol{\alpha}}^\star) = \operatorname{sgn}(\boldsymbol{\beta})$。其次，如果 $|\beta_i| \geq |\beta_j|$，那么$|\bar{\alpha}_i^\star| \geq |\bar{\alpha}_j^\star|$。为进行反证，假设相反的情况成立 $|\bar{\alpha}_i^\star| < |\bar{\alpha}_j^\star|$。那么我们可以交换 $\alpha_i^\star$ 和 $\alpha_j^\star$，以进一步降低损失函数，这与解 $\bar{\boldsymbol{\alpha}}^\star$ 已经是最优解这一事实相矛盾。

利用这些性质和 $\operatorname{TopSum}_{\bar{k}}$ 算子的定义，我们可以将上述最小化问题改写为关于绝对值 $|\bar{\boldsymbol{\alpha}}|$ 的广义保序回归问题，如下所示

$$\min_{\boldsymbol{\nu}\in\mathbb{R}_+^{\bar{p}}} \quad \sum_{j=1}^{\bar{p}} \frac{1}{2}(\nu_j - |\bar{\beta}_{\pi(j)}|)^2 + \rho_j H_M(\nu_j) \quad \text{s.t.} \quad \nu_1 \geq \nu_2 \geq \cdots \geq \nu_{\bar{p}} \geq 0, \tag{B.9}$$

其中，$\beta_{\pi(j)}$ 表示 $|\boldsymbol{\beta}|$ 中按绝对值大小排列的第 j 大元素，权重定义为 $\rho_j = \rho$（对于 $j \leq k$）以及 $\rho_j = 0$（对于 $j > k$）。如果 $\boldsymbol{\nu}^\star$ 是 (B.9) 的解，我们可以通过 $\bar{\boldsymbol{\alpha}}^\star = \operatorname{sign}(\boldsymbol{\beta}) \odot \boldsymbol{\nu}^\star$恢复 $\bar{\boldsymbol{\alpha}}^\star$。

优化问题 (B.9) 在总序约束下最小化一组凸函数之和。排序后，可以使用 [Busing, 2022] 中介绍的 PAVA算法在线性时间内精确求解。算法2实现了这一过程。第 6–11 行执行 PAVA 步骤，迭代合并违反单调性约束$\nu_j \geq \nu_{j+1}$的相邻块。由于排序需要 $\mathcal{O}(\bar{p}\log\ \bar{p})$ 时间，而 PAVA 扫描需要 $\mathcal{O}(\bar{p})$ 时间，总复杂度由排序主导。证明至此完成。

□

B.13 广义线性模型正则性条件（框架的命题 4.6）

命题 *4.6.* 考虑原始目标函数 $\Phi(\boldsymbol{\beta}) = F(\boldsymbol{X}\boldsymbol{\beta}) + G(\boldsymbol{\beta})$，其中 $F(\boldsymbol{z}) := f(\boldsymbol{z}, \boldsymbol{y})$，且 $f\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ 在 $\boldsymbol{z}$中严格凸、局部光滑且强凸，并且$G(\boldsymbol{\beta}) := 2\lambda_2\, g_{\mathcal{N}}(\boldsymbol{\beta})$，其中 $\lambda_2 > 0$ 和 $g_{\mathcal{N}}$定义见 (5)。那么，假设 3.1–3.8 均成立。

证明. 命题 4.6 的证明 由于 $f(\cdot, \boldsymbol{y})$ 关于其第一个自变量 $\boldsymbol{z}$严格凸，因此不存在使 F 有限且为仿射的直线。此外，由于 f 是实值、凸且局部光滑的，F 显然是真、闭且凸的。此外，引理 4.1 证明了$g_{\mathcal{N}}$（或等价地，G）是真、闭且凸的，且其定义域是紧的。由于 $\operatorname{dom}(G)$ 是紧的，不存在使 G 有限且为仿射的直线。因此，假设 3.1 和 3.8 均成立。

由于 F 局部光滑且为实值函数，因此处处可微，并且 $\operatorname{int}(\operatorname{dom}(F)) = \mathbb{R}^n$。$G$ 的定义域由盒约束和基数约束定义，这些约束构成一个非空多面体，且其相对内部非空。因此，$\operatorname{rint}(\operatorname{dom}(\Phi))$ 非空，保证 Slater 点 $\boldsymbol{\beta}_s$的存在。对于对偶问题，由于 $\operatorname{dom}(G)$ 紧，其共轭函数 G^* 处处为实值且有限，这意味着 $\operatorname{int}(\operatorname{dom}(G^*)) = \mathbb{R}^p$。由于不存在使 F 有限且为仿射函数的直线，$\operatorname{int}(\operatorname{dom}(F^*))$ 非空 [Rockafellar, 1970, Corollary 13.4.2]。因此，内部的交集非空，保证 Slater 点$\boldsymbol{\zeta}_s$的存在。由此，假设 3.2 得以满足。

最优原始解集 $\mathcal{B}^\star$ 是 $\operatorname{dom}(G)$ 的闭子集。由引理 4.1 可知，$\operatorname{dom}(G)$ 紧，$\mathcal{B}^\star$因此也是紧集。由于 $\operatorname{dom}(F) = \mathbb{R}^n$，包含关系 $\boldsymbol{X}\mathcal{B}^\star \subset \operatorname{int}(\operatorname{dom}(F))$ 显然成立。这验证了假设 3.4 (i)。此外，根据引理的假设，F 是强凸的。借助引理 4.3，G 也为强凸函数。这验证了假设 3.4 (ii)。

由于根据引理 4.3，G^* 是分段线性二次函数，其次微分是一个多面体集合。根据假设 3.8，F 是严格凸的。由于 F 局部光滑且为实值函数，因此处处可微，且 $\operatorname{dom}(\partial F) = \mathbb{R}^n$，所以 F本质严格凸。因此，根据 Rockafellar（1970，定理 26.3]），F^*本质光滑。由于 F^* 本质光滑，根据 Rockafellar（1970，定理 26.1]），[意味着 $\operatorname{dom}(\partial F^*) = \operatorname{int}(\operatorname{dom}(F^*))$。于是，KKT 条件 $\boldsymbol{X}\boldsymbol{\beta}^\star \in \partial F^*(-\boldsymbol{\zeta}^\star)$ 得出 $-\boldsymbol{\zeta}^\star \in \operatorname{int}(\operatorname{dom}(F^*))$。由于 $\operatorname{dom}(G^*) = \mathbb{R}^p$，$\boldsymbol{X}^\top\boldsymbol{\zeta}^\star \in \operatorname{int}(\operatorname{dom}(G^*))$显然成立。这验证了假设 3.4 (iv)。此外，$-\boldsymbol{\zeta}^\star \in \operatorname{int}(\operatorname{dom}(F^*))$意味着 $\partial F^*(-\boldsymbol{\zeta}^\star) = \{\nabla F^*(-\boldsymbol{\zeta}^\star)\}$

根据 [Rockafellar（1970，定理 26.1], ），这是一个多面体集合。这验证了假设 3.4 (iii)，并完成了证明。 □

B.14 常见广义线性模型损失函数的验证

本附录验证若干标准广义线性模型损失函数满足命题 4.6 中假设的正则性条件，即 F 严格凸、局部光滑且强凸。全文中，我们用 $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} \in \mathbb{R}^n$ 表示预测向量，并将损失写作$F(\boldsymbol{z}) := f(\boldsymbol{z}, \boldsymbol{y})$。对于多项逻辑回归，我们将 $\boldsymbol{z}$视为固定参考类别后堆叠而成的类别得分向量，详见引理 B.5。

引理 B.1 (线性回归（最小二乘）)。 令 $\boldsymbol{y} \in \mathbb{R}^n$ ，并定义最小二乘损失

$$F(\boldsymbol{z}) := \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{y}\|_2^2, \qquad \boldsymbol{z} \in \mathbb{R}^n.$$

于是， F 严格凸、全局 光滑且强凸。

证明。 证明：函数 F 二阶连续可微，且

$$\nabla F(\boldsymbol{z}) = \boldsymbol{z} - \boldsymbol{y}, \qquad \nabla^2 F(\boldsymbol{z}) = \boldsymbol{I} \quad \text{for all } \boldsymbol{z} \in \mathbb{R}^n.$$

因此， F 在 $\mathbb{R}^n$上是 1-强凸且 1-光滑的，这意味着它严格凸且局部光滑。此外，对于任意 $\boldsymbol{\nu} \in \mathbb{R}^n$，倾斜函数 $F_{\boldsymbol{\nu}}(\boldsymbol{z}) := F(\boldsymbol{z}) - \boldsymbol{\nu}^\top \boldsymbol{z}$具有相同的海森矩阵 $\nabla^2 F_{\boldsymbol{\nu}} \equiv \boldsymbol{I}$，因此是 1-强凸的。强凸性意味着满足二次增长条件，因此根据第 3 节的定义， F 是强凸的。 □

引理 B.2 (逻辑回归)。 令 $\boldsymbol{y} \in \{-1, 1\}^n$ ，并定义逻辑损失

$$F(\boldsymbol{z}) := \sum_{i=1}^{n} \log(1 + \exp(-y_i z_i)), \qquad \boldsymbol{z} \in \mathbb{R}^n.$$

那么， F 是严格凸、全局光滑且强凸的。

证明。 证明 函数 F 二阶连续可微，其梯度和海森矩阵的各分量如下

$$\big(\nabla F(\boldsymbol{z})\big)_i = -\frac{y_i}{1 + \exp(y_i z_i)}, \qquad \nabla^2 F(\boldsymbol{z}) = \operatorname{diag}\left(\left\{\frac{\exp(y_i z_i)}{(1 + \exp(y_i z_i))^2}\right\}_{i=1}^{n}\right).$$

对于所有有限的 $\boldsymbol{z}$，都有 $\nabla^2 F(\boldsymbol{z}) \succ \boldsymbol{0}$，因此 F 在 $\mathbb{R}^n$上严格凸。

为证明全局光滑性，注意到由于 $\|\nabla^2 F(\boldsymbol{z})\|_2$ 等于最大的对角元素，并且

$$0 < \frac{\exp(y_i z_i)}{(1 + \exp(y_i z_i))^2} \le \frac{1}{4} \quad \text{for all } \boldsymbol{z} \in \mathbb{R}^n.$$

因此，我们有 $\sup_{\boldsymbol{z} \in \mathbb{R}^n} \|\nabla^2 F(\boldsymbol{z})\|_2 \le 1/4$，这意味着 F 是全局光滑的。

对于坚实凸性，请注意，海森矩阵处处连续且正定。因此，根据紧致性， $\mathcal{B}$ 上的最小特征值满足

$$m_{\mathcal{B}} := \inf_{\boldsymbol{z} \in \mathcal{B}} \lambda_{\min}\big(\nabla^2 F(\boldsymbol{z})\big) > 0.$$

因此， F在 $\mathcal{B}$上是 $m_{\mathcal{B}}$-强凸的。对于任意 $\boldsymbol{\nu} \in \mathbb{R}^n$，倾斜函数 $F_{\boldsymbol{\nu}}(\boldsymbol{z}) = F(\boldsymbol{z}) - \boldsymbol{\nu}^\top \boldsymbol{z}$具有相同的海森矩阵，并且在 $\mathcal{B}$上也是 $m_{\mathcal{B}}$-强凸的，这意味着 $\mathcal{B}$上的二次增长条件。因此， F 是强凸的。 □

引理 B.3（泊松回归）。 设 $\boldsymbol{y} \in \mathbb{R}^n$，对所有 $i \in [n]$均有 $y_i \in \{0, 1, 2, \ldots\}$，并定义泊松损失

$$F(\boldsymbol{z}) := \sum_{i=1}^{n} \left(\exp(z_i) - y_i z_i\right), \qquad \boldsymbol{z} \in \mathbb{R}^n.$$

那么， F 是严格凸、局部光滑且强凸的。

证明。 证明：函数 F 二阶连续可微，且

$$\left(\nabla F(\boldsymbol{z})\right)_i = \exp(z_i) - y_i, \qquad \nabla^2 F(\boldsymbol{z}) = \operatorname{diag}(\{\exp(z_i)\}_{i=1}^n).$$

由于对所有 $z_i \in \mathbb{R}$，$\exp(z_i) > 0$，因此对所有 $\boldsymbol{z}$， $\nabla^2 F(\boldsymbol{z}) \succ \mathbf{0}$，并且 F 在 $\mathbb{R}^n$上严格凸。令 $\mathcal{B} \subset \mathbb{R}^n$ 为紧集，并定义

$$\underline{z} := \min_{\boldsymbol{z} \in \mathcal{B}} \min_{i \in [n]} z_i, \qquad \overline{z} := \max_{\boldsymbol{z} \in \mathcal{B}} \max_{i \in [n]} z_i,$$

由紧致性可知，它们都是有限的。于是，对所有 $\boldsymbol{z} \in \mathcal{B}$,

$$e^{\underline{z}} \boldsymbol{I} \preceq \nabla^2 F(\boldsymbol{z}) \preceq e^{\overline{z}} \boldsymbol{I}.$$

上界意味着 ∇F 在 $\mathcal{B}$ 上是利普希茨的（局部光滑性），而下界意味着 F 在 $\mathcal{B}$上是 e^z-强凸的。与引理 B.2 中一样，倾斜函数在任意紧集上的强凸性可推出二次增长条件，因此 F 是强凸的。

□

引理 B.4（伽马（指数）回归）。 设 $\boldsymbol{y} \in \mathbb{R}^n$，且对所有 $i \in [n]$， $y_i > 0$，并定义伽马（指数）回归损失

$$F(\boldsymbol{z}) := \sum_{i=1}^{n} \left(y_i \exp(-z_i) + z_i\right), \qquad \boldsymbol{z} \in \mathbb{R}^n.$$

那么， F 是严格凸、局部光滑且强凸的。

证明。 证明 函数 F 二次连续可微，且满足

$$\left(\nabla F(\boldsymbol{z})\right)_i = -y_i \exp(-z_i) + 1, \qquad \nabla^2 F(\boldsymbol{z}) = \operatorname{diag}(\{y_i \exp(-z_i)\}_{i=1}^n).$$

由于对于所有 $z_i \in \mathbb{R}$， $y_i > 0$ 和 $\exp(-z_i) > 0$，因此对于所有 $\boldsymbol{z}$， $\nabla^2 F(\boldsymbol{z}) \succ \mathbf{0}$，从而 F 在 $\mathbb{R}^n$ 上严格凸。

设 $\mathcal{B} \subset \mathbb{R}^n$ 为紧集，并定义

$$\underline{z} := \min_{\boldsymbol{z} \in \mathcal{B}} \min_{i \in [n]} z_i, \qquad \overline{z} := \max_{\boldsymbol{z} \in \mathcal{B}} \max_{i \in [n]} z_i,$$

由紧致性可知它们是有限的。于是，对于所有 $\boldsymbol{z} \in \mathcal{B}$,

$$\operatorname{diag}\left(\left\{y_i e^{-\overline{z}}\right\}_{i=1}^n\right) \preceq \nabla^2 F(\boldsymbol{z}) \preceq \operatorname{diag}\left(\left\{y_i e^{-\underline{z}}\right\}_{i=1}^n\right).$$

上界意味着 ∇F 在 $\mathcal{B}$ 上是利普希茨的（局部光滑性），下界意味着 F 在 $\mathcal{B}$上是强凸的。如引理 B.2 所示，倾斜函数在任意紧集上的强凸性会导出二次增长条件，因此 F 是强凸的。

□

引理 B.5（多项逻辑回归）。 固定一个整数 $K \geq 2$，假设类别 K 的系数向量为零。对于每个 $i \in [n]$，令 $\boldsymbol{y}_i \in \{0,1\}^{K-1}$ 满足 $\mathbf{1}^\top \boldsymbol{y}_i \leq 1$。定义$\boldsymbol{z}_i \in \mathbb{R}^{K-1}$，并考虑多项逻辑回归损失

$$F(\boldsymbol{z}) := \sum_{i=1}^{n} \Big(\log\Big(1 + \mathbf{1}^\top \exp(\boldsymbol{z}_i)\Big) - \boldsymbol{y}_i^\top \boldsymbol{z}_i \Big),$$

其中 $\boldsymbol{z} := (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \in \mathbb{R}^{n(K-1)}$ 将行向量 $\boldsymbol{z}_i$ 堆叠起来。于是 F 关于 $\boldsymbol{z}$ 是严格凸的、局部光滑的，并且是强凸的。

证明。 证明 对每个 $i \in [n]$，定义

$$\boldsymbol{p}_i := \frac{\exp(\boldsymbol{z}_i)}{1 + \mathbf{1}^\top \exp(\boldsymbol{z}_i)} \in \mathbb{R}^{K-1}, \qquad p_{i0} := \frac{1}{1 + \mathbf{1}^\top \exp(\boldsymbol{z}_i)} \in (0,1),$$

从而 $p_{i0} + \mathbf{1}^\top \boldsymbol{p}_i = 1$ 且所有元素均为严格正。于是 F 二次连续可微，其海森矩阵为分块对角，其中第 i 个分块由下式给出

$$\nabla^2_{\boldsymbol{z}_i} F(\boldsymbol{z}) = \operatorname{diag}\big(\boldsymbol{p}_i\big) - \boldsymbol{p}_i \boldsymbol{p}_i^\top,$$

对任意非零 $\boldsymbol{v} \in \mathbb{R}^{K-1}$，我们有

$$\boldsymbol{v}^\top \nabla^2_{\boldsymbol{z}_i} F(\boldsymbol{z})\, \boldsymbol{v} = \sum_{l=1}^{K-1} p_{il} v_l^2 - \Big(\sum_{l=1}^{K-1} p_{il} v_l \Big)^2,$$

这是一个随机变量的方差，该随机变量以概率 p_{il} 取值 v_l，以概率 $p_{i0} > 0$取值 0。只要 $\boldsymbol{v} \neq \mathbf{0}$，该方差就严格为正，因此每个分块都是正定的，从而对所有 $\boldsymbol{z}$ 有 $\nabla^2 F(\boldsymbol{z}) \succ \mathbf{0}$。故 F 是严格凸的。

设 $\mathcal{B}$ 为 $\boldsymbol{z}$ 空间的任意紧子集。由于 $\nabla^2 F$ 连续，$\|\nabla^2 F(\boldsymbol{z})\|_2$ 在 $\mathcal{B}$ 上有上界，这意味着 ∇F 在 $\mathcal{B}$ 上是利普希茨的（局部光滑性）。此外，由于对所有 $\boldsymbol{z}$ 有 $\nabla^2 F(\boldsymbol{z}) \succ \mathbf{0}$，且 $\mathcal{B}$ 是紧的，故存在 $m_{\mathcal{B}} > 0$ 使得对所有 $\boldsymbol{z} \in \mathcal{B}$ 有 $\nabla^2 F(\boldsymbol{z}) \succeq m_{\mathcal{B}} \boldsymbol{I}$，因此 F 在 $\mathcal{B}$上是强凸的。如引理 B.2 所述，倾斜函数在任意紧集上的强凸性给出二次增长条件，从而 F 是强凸的。 □

C 实验设置细节

C.1 求解透视松弛的设置

我们按照以下步骤生成合成数据集。首先，从高斯分布 $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ 中采样每个特征向量 $\boldsymbol{x}_i \in \mathbb{R}^p$，其中协方差矩阵的元素为 $\Sigma_{jl} = \sigma^{|j-l|}$。变量 $\sigma \in (0,1)$ 控制特征相关性：增大 σ 会使设计矩阵 $\boldsymbol{X}$ 中的特征列相关性更强。在实验部分中，我们始终设定 $\sigma = 0.5$。接下来，我们构造稀疏系数向量 $\boldsymbol{\beta}^*$，其中有 k 个等间隔的非零元素；当 $j \bmod (p/k) = 0$ 时，$\beta_j^* = 1$，否则 $\beta_j^* = 0$。完成这两个步骤后，我们构造预测向量 $\boldsymbol{y}$。如果损失函数是平方误差损失（回归任务），则设定 $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}^* + \epsilon_i$，其中 ϵ_i 是高斯随机噪声，其 $\epsilon_i \sim \mathcal{N}$ 为 $(0, \frac{\|\boldsymbol{X}\boldsymbol{\beta}^*\|}{\mathrm{SNR}})$；SNR 表示信噪比。在所有实验中，我们均取 $\mathrm{SNR} = 5$。如果损失函数是逻辑损失（分类任务），则设定 $y_i \sim Bern(\boldsymbol{x}_i^T \boldsymbol{\beta}^* + \epsilon_i)$，其中 $Bern(P)$ 是伯努利随机变量，其 $\mathbb{P}(y_i = 1)$ 为 $= P$，且 $\mathbb{P}(y_i = -1)$ 为 $= 1 - P$。

C.2 Se于 LASSO 问题上展示重启方案的通用性 s

为说明我们的重启方案不仅适用于透视松弛，我们考虑两种 LASSO 型复合问题：

$$\begin{aligned}&\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad f(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{y})+\lambda_1\|\boldsymbol{\beta}\|_1 \quad (\ell_1\text{-regularized}),\\&\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad f(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{y})+\delta_{\|\boldsymbol{\beta}\|_1\le C}(\boldsymbol{\beta}) \quad (\ell_1\text{-ball constrained}),\end{aligned}$$

其中， $\delta_{\|\boldsymbol{\beta}\|_1\le C}(\boldsymbol{\beta})$ 是半径为 C 的 ℓ_1-范数球的示性函数。我们采用与附录 C.1 中相同的合成数据生成过程，并固定 $n=p=16000$、 $k=10$ 和 $\sigma=0.5$，系数幅值为 $\beta_j^*\in\{0,1\}$，随机种子为 0。在 ℓ_1-正则化实验中，我们设置 $\lambda_1=1.0$；在 ℓ_1-球约束实验中，我们设置 $C=10.0$。

C.3 最优性认证的设置

数据集与预处理 我们在合成数据集和真实世界数据集上进行实验。对于合成数据集，我们采用与前一小节求解透视松弛时相同的数据生成过程，其中 n 与 p 的比值为 1.0，非零系数的数量为 $k=10$，特征相关参数为 $\sigma=0.5$。我们改变特征维度 $p\in\{1000,2000,4000,8000,16000\}$。对于真实世界数据集，我们使用 Santander 客户交易预测数据集 [Piedra 等人， 2019] 用于线性回归；并使用 DOROTHEA 数据集 Asuncion 和 Newman [2007]用于逻辑回归。

Santander 数据集有 4459 个样本，原本有 4992 个特征。剪除冗余特征后，我们有 4735 个特征。DOROTHEA数据集有 1950 个样本和 100000 个特征。剪除冗余特征后，我们有 91598 个特征。对于 Santander 和 DOROTHEA 数据集，我们将每个特征中心化，使其均值为 0，范数为 1。

超参数的选择 对于合成数据集，我们设置盒约束 $M=2$、 ℓ_2 正则化系数 $\lambda_2=1.0$以及基数约束 $k=10$。对于真实数据集，我们进行 5 折交叉验证（不包含盒约束），以选择 λ_2 的最优超参数以及 k 值的范围。对于两个数据集，交叉验证得到的最佳 λ_2 均为 1。0。交叉验证得到的最佳 k 对于 Santander 约为 10，对于 DOROTHEA 约为 5。使用盒约束运行分支定界时，我们将盒约束 M 设置为交叉验证得到的 $\boldsymbol{\beta}$ 最大幅值的约两倍。根据这些交叉验证结果，我们设置 $M=10$。对于 Santander，我们测试 $k\in\{5,6,7,8,9,10\}$；对于 DOROTHEA，我们测试 $k\in\{5,10,15,20,25\}$。

分支定界对于我们的方法，我们基于 Liu 等人 [2023]的代码库实现了一个定制的分支定界（BnB）框架。我们采用与 Liu 等人 [2023]相同的启发式方法来寻找可行解并进行分支。为了在每个节点获得高质量的可行解，我们运行束搜索，束宽度设为 5。进行分支时，我们使用每个节点上束搜索得到的最佳解，并在其非零系数中选择一个变量进行分支：若将该变量设为 0，会导致损失增加最多。这样可以在 BnB 早期优先处理重要变量。关键区别在于，我们使用近端梯度法计算每个节点的安全下界，以便进行剪枝。对于近端方法，我们采用带线搜索的快速迭代收缩阈值算法（FISTA），重启因子为 $\eta=e^3$，其实证性能最佳。

C.4 计算平台

为了研究 GPU 能在多大程度上加速我们的计算，我们在 Nvidia A100 上分别运行了中央处理器和 GPU 实现的实验。对于其他所有实验，我们使用

AMD Milan 上的中央处理器实现，中央处理器速度为 2.45 GHz，配备 8 个核心。

D 附加实验结果

D.1 求解透视函数松弛

我们通过改变盒约束 M、 ℓ_2 正则化系数 λ_2 以及样本数与特征数之比（n 对 p），研究求解透视函数松弛时的扰动情况。除非另有说明，所有设置均与图 5（$M = 2.0$、 $\lambda_2 = 1.0$、 n 对 p 比例 $= 1$）一致。

图 D.1 和图 D.2 分别将盒约束收紧和放宽至 $M = 1.2$ 和 $M = 5.0$。图 D.3 和图 D.4 将 ℓ_2 惩罚项分别改为 $\lambda_2 = 0.1$ 和 $\lambda_2 = 10$。图 D.5 和图 D.6 将 n 对 p 的比例调整为 10.0 和 0.1，同时保持 $M = 2.0$ 和 $\lambda_2 = 1.0$。

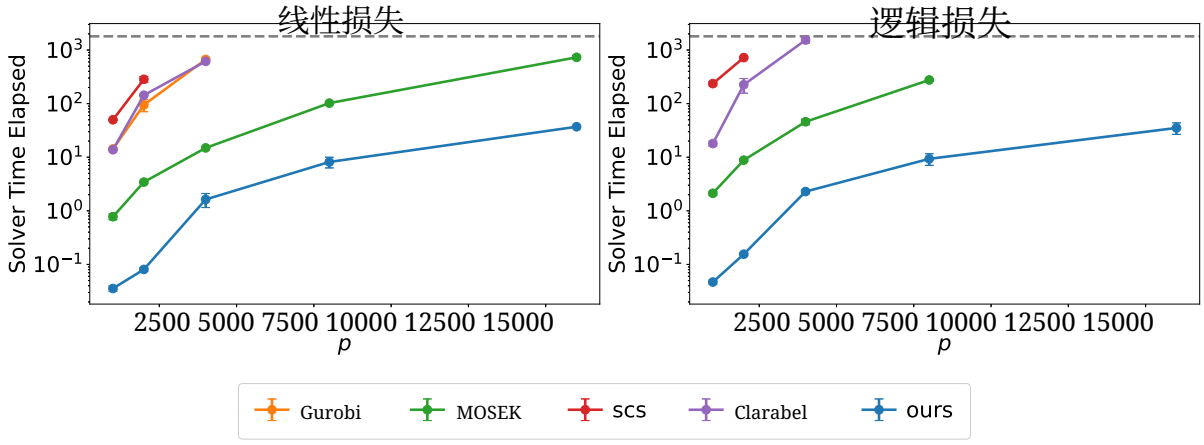

图 D.1：将盒约束收紧至 $M = 1.2$ 时求解问题 3 的运行时间比较。

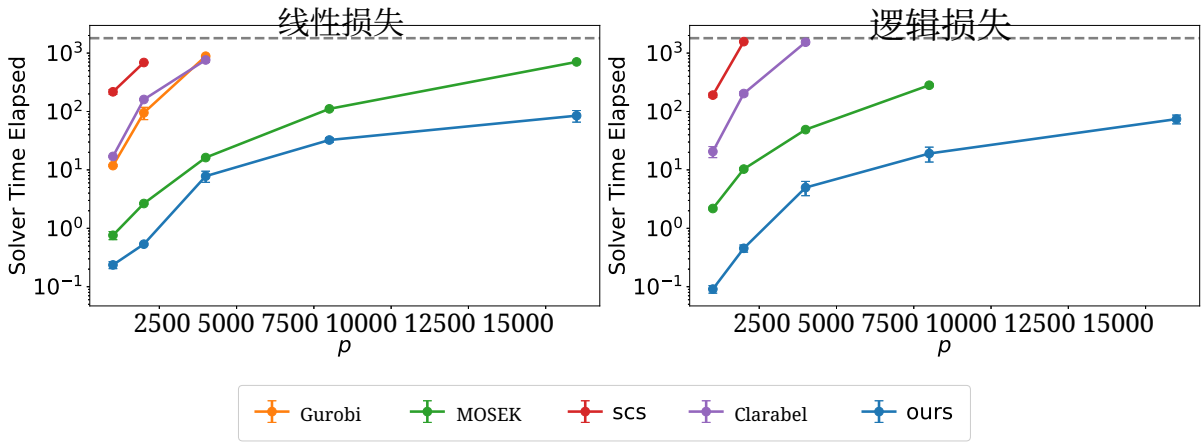

图 D.2：将盒约束放宽至 $M = 5.0$ 时求解问题 3 的运行时间比较。

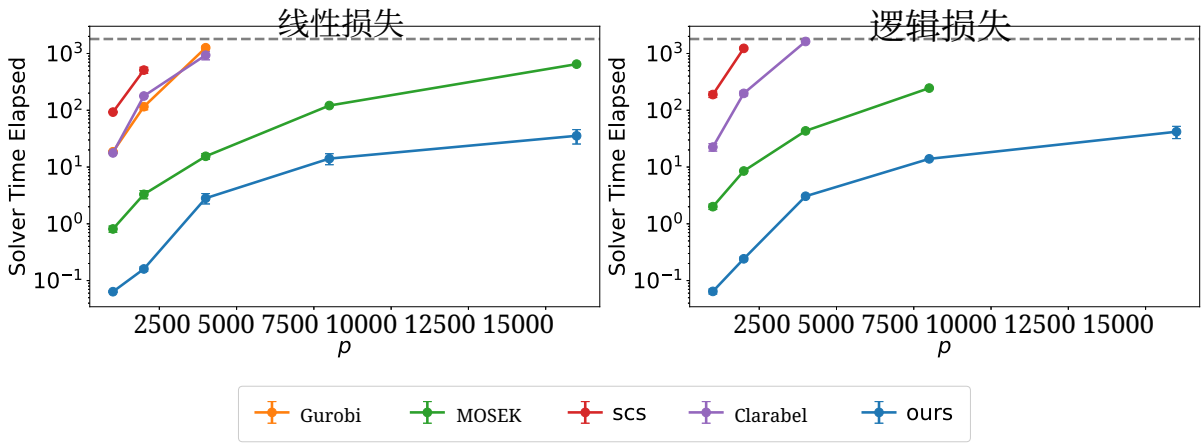

图 D.3：将 ℓ_2 正则化降低至 $\lambda_2 = 0.1$ 时求解问题 3 的运行时间比较。

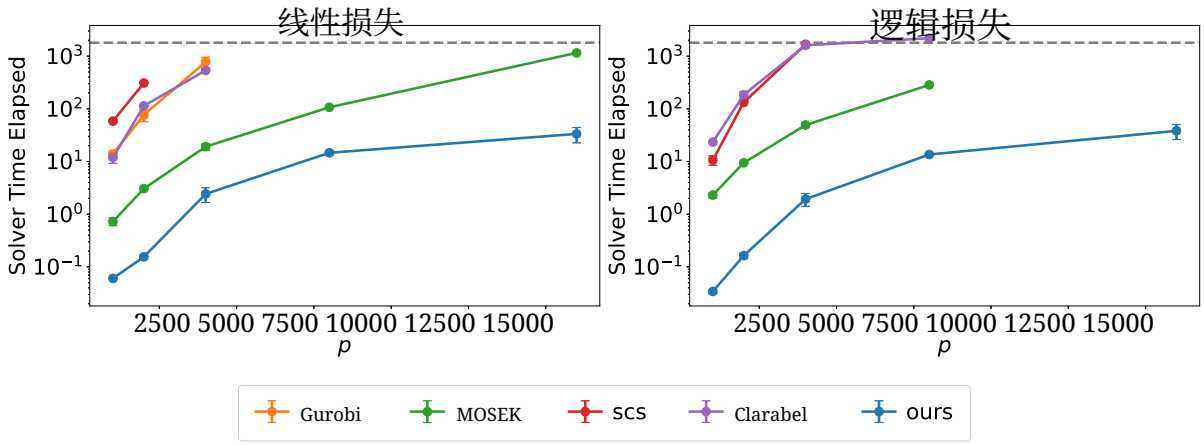

图 D.4：将 ℓ_2 正则化提高至 $\lambda_2 = 10$ 时求解问题 3 的运行时间比较。

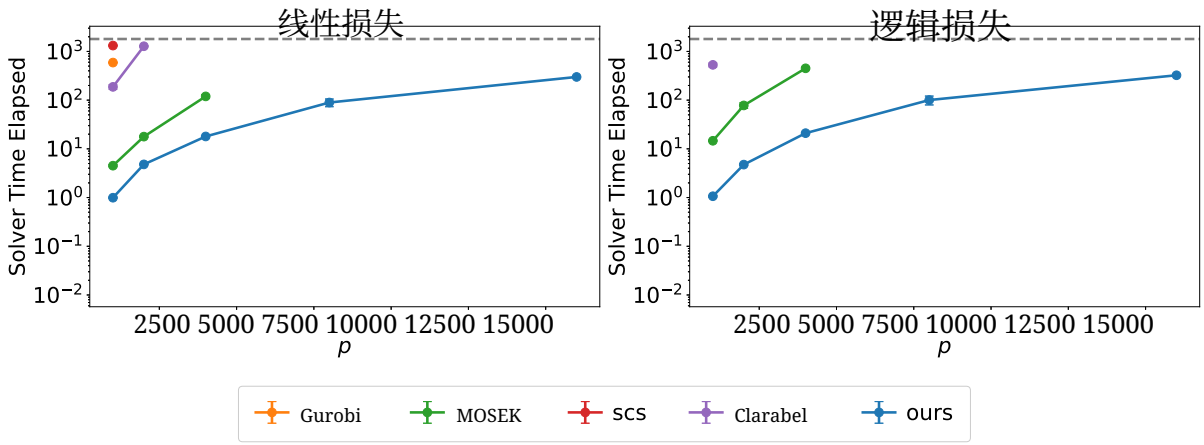

图 D.5：当 n 与 p 之比增大至 10.0 时，求解问题 3 的运行时间比较。

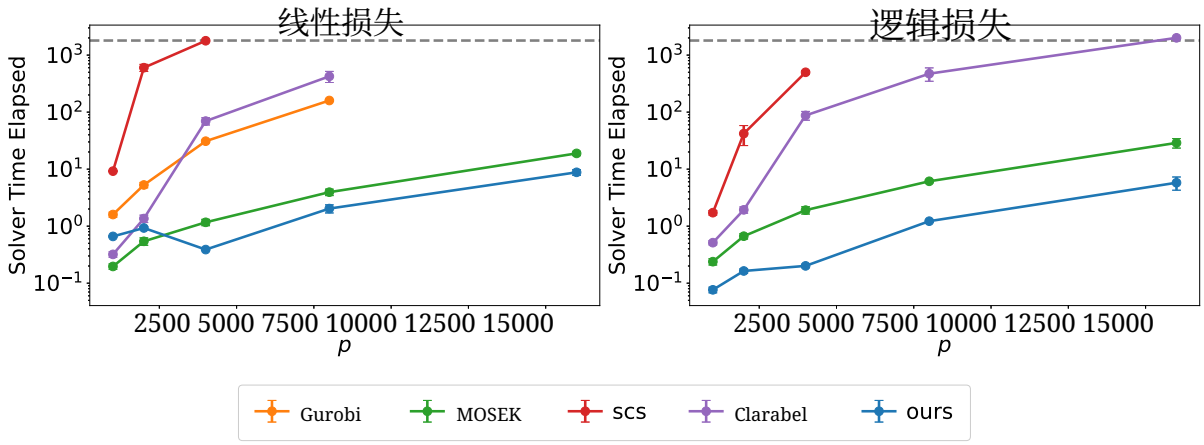

图 D.6：当 n 与 p 之比减小至 0.1 时，求解问题 3 的运行时间比较。

D.2 基于收缩因子的基于对偶性的重启方案

在正文中，我们指出，基于对偶性的重启方案适用于任意下降因子 $\eta > 1$，但 η 的选择会影响实际运行速度。我们比较了 $\eta \in \{e, e^2, e^3\}$，以展示这一影响。结果见图 D.7 和图 D.8。通过将 η 从 e 改为 e^2 和 e^3，我们能够提升所有方法的实际运行速度，这与定理 3.13 的理论预测以及第 3.3 节末尾的讨论一致。

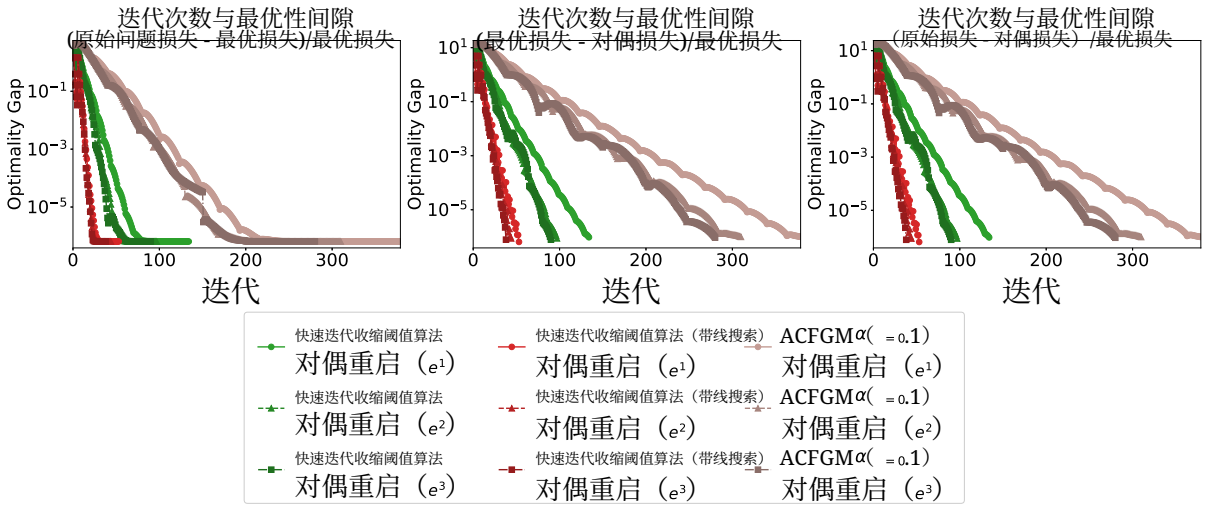

图 D.7：在使用线性回归损失求解问题 3 中的透视松弛时，比较快速迭代收缩阈值算法（采用和不采用线搜索）与 ACFGM（采用基于对偶间隙的重启）的收敛速度，$\eta = e$、$\eta = e^2$ 和 $\eta = e^3$，$n = 16000$、$p = 16000$、$k = 10$、$\rho = 0.5$、$\lambda_2 = 1.0$ 和 $M = 2.0$。

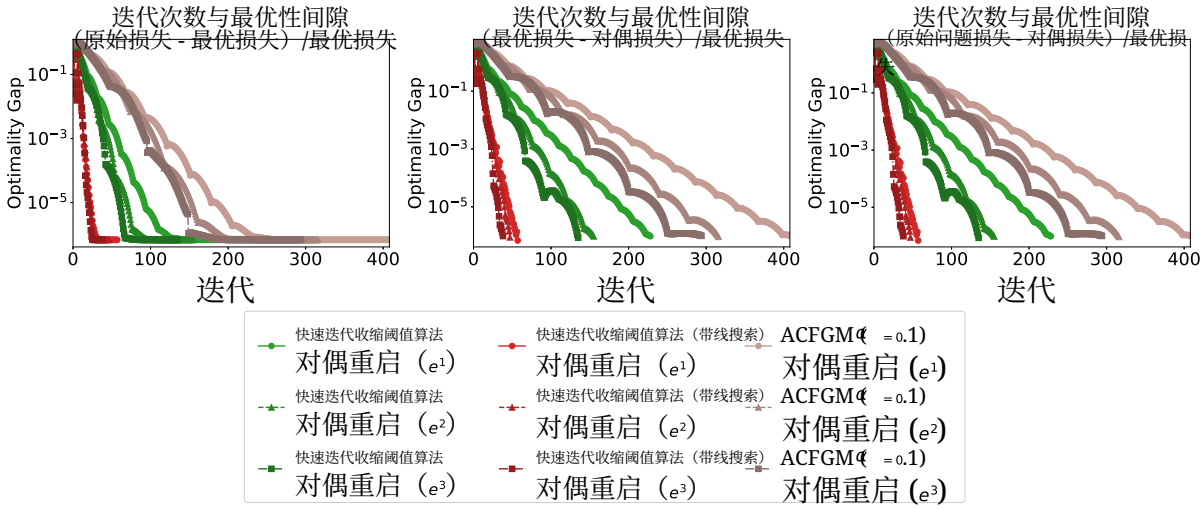

图 D.8：在使用逻辑回归损失求解问题 3 中的透视松弛时，比较快速迭代收缩阈值算法（采用和不采用线搜索）与 ACFGM（采用基于对偶间隙的重启）的收敛速度，$\eta = e$、$\eta = e^2$ 和 $\eta = e^3$，$n = 16000$、$p = 16000$、$k = 10$、$\rho = 0.5$、$\lambda_2 = 1.0$ 和 $M = 2.0$。

D.3 我们的重启方案能否使其他问题受益，而不仅限于透视松弛？

为了展示基于对偶间隙的重启方案在透视松弛之外更广泛的适用性，我们将其用于加速近端梯度法，以求解 LASSO 问题。我们报告了两种形式的结果：一种是受 ℓ_1-球约束的形式，其中 $G(\beta) := \delta_{\|\beta\|_1 \le C}(\beta)$；另一种是带 ℓ_1 正则化的形式，其中 $G(\beta) := \lambda_1 \|\beta\|_1$。实验设置和参数选择见附录 C.2。结果见图D.9–图D.12。与正文中求解透视松弛的情况类似，我们观察到，我们的重启方案能够加速快速迭代收缩阈值算法和 ACFGM，使其达到线性收敛速率。

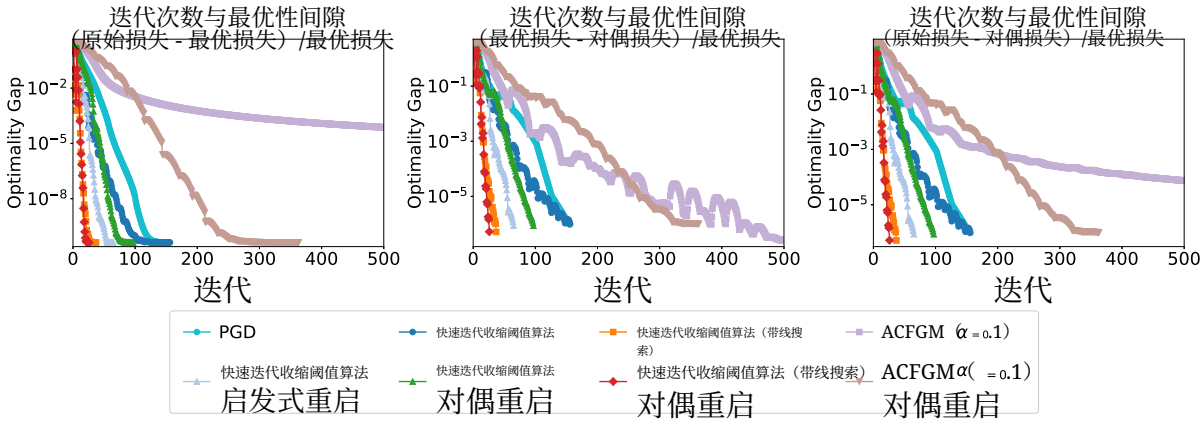

图 D.9：使用加速方法（有无重启）求解带最小二乘损失的 ℓ_1-球约束问题时的收敛速度比较。

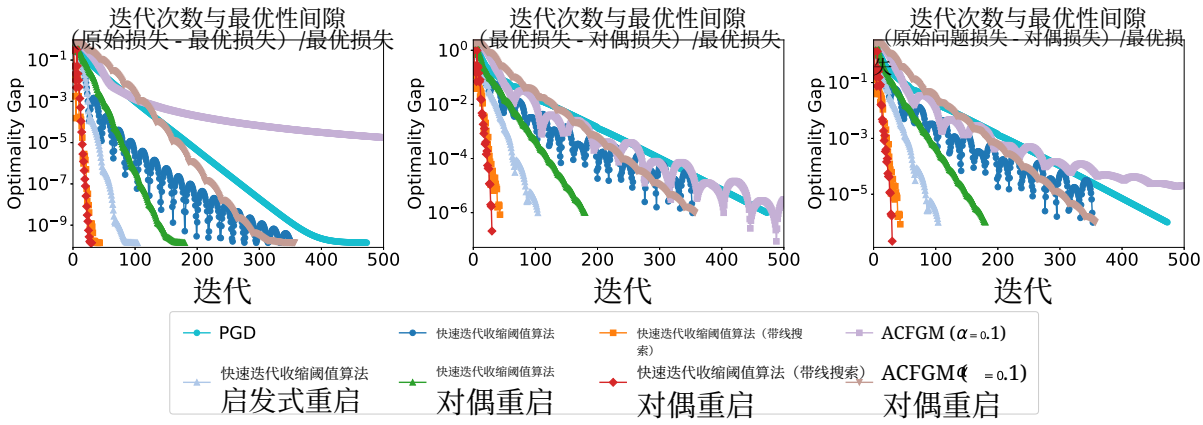

图 D.10：使用加速方法（有重启和无重启）求解带有逻辑回归损失的 ℓ_1-球约束问题时的收敛速度比较。

对偶可行性：针对 ℓ_1正则化问题。 对于 ℓ_1正则化表述，正则化项 $G(\beta) = \lambda_1 \|\beta\|_1$ 的 Fenchel 共轭为 $G^*(u) = \delta_{\|u\|_\infty \le \lambda_1}(u)$，因此仅当 $\|X^\top \zeta\|_\infty \le \lambda_1$ 时对偶目标函数才是有限的。因此，对偶序列 $\zeta^t := -\nabla F\left(X\beta^t\right)$可能不可行，此时计算对偶目标函数没有意义。遵循 LASSO 文献中的标准做法，我们通过一个简单的缩放（径向投影）步骤强制对偶可行性，$\widehat{\zeta} := \zeta \cdot \min\left\{1, \frac{\lambda_1}{\|X^\top \zeta\|_\infty}\right\}$，这保证了 $\|X^\top \widehat{\zeta}\|_\infty \le \lambda_1$，并给出一个有效（尽管可能较松）的对偶下界，用于计算对偶间隙 [Fercoq 等，2015，

Dantas 等， 2021]。因此，当将我们的基于间隙的重启方案应用于 ℓ_1 正则化问题时，用于重启的间隙应被视为非精确证书；尽管如此，它在实践中仍能提供有效的重启信号，从而产生下文所观察到的线性收敛行为。

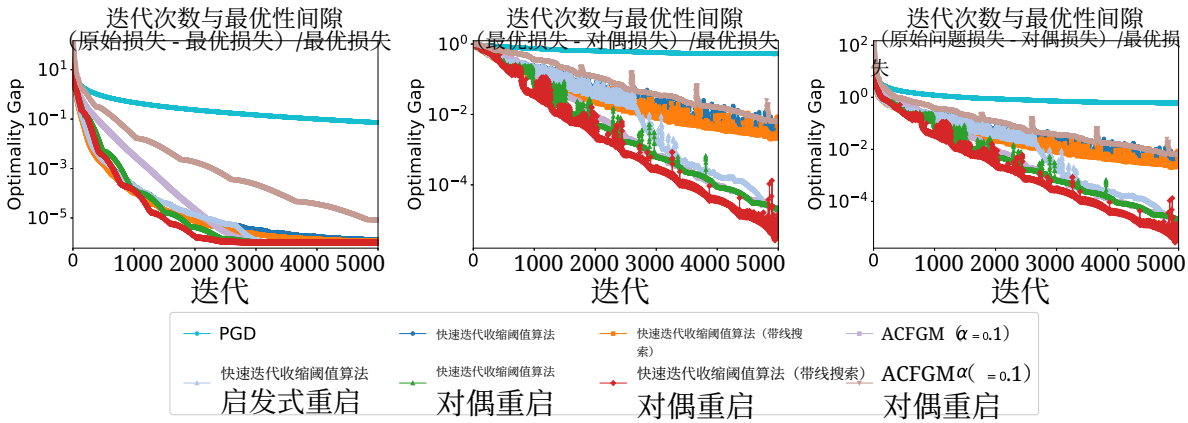

图 D.11：使用最小二乘损失求解 ℓ_1-正则化问题时，加速方法（有重启和无重启）的收敛速度比较。

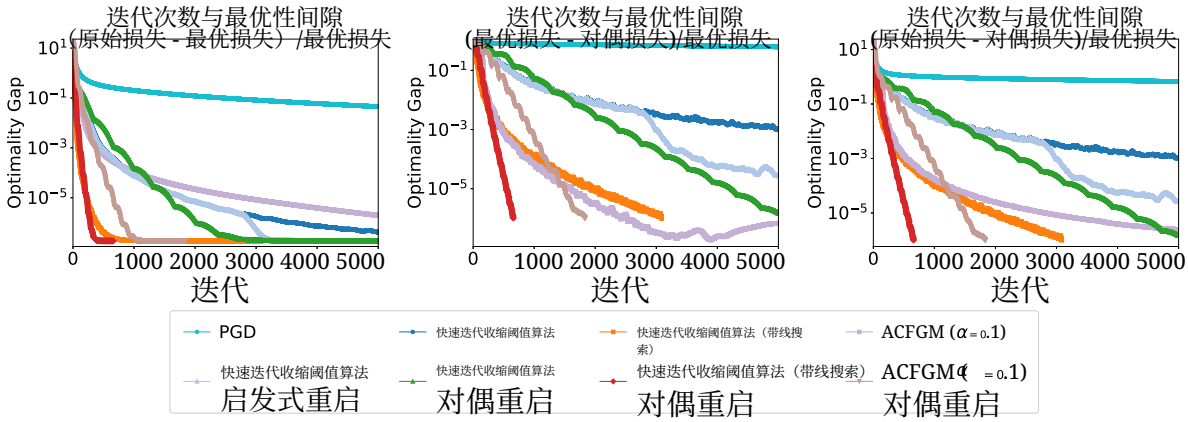

图 D.12：使用逻辑回归损失求解 ℓ_1-正则化问题时，加速方法（有重启和无重启）的收敛速度比较。

D.4 最优性认证

D.4.1 $M = 3$合成结果

我们报告了在采用更大盒约束 $M = 3.0$的合成数据集上的补充分支定界结果：表 D.1–D.2 展示中央处理器基线方法，表 D.3–D.4 展示中央处理器与 GPU 的对比。

表 D.1：采用 $k = 10$, $\rho = 0.5$, $\lambda_2 = 1.0$, $M = 3.0$, $n/p = 1.0$ 的合成（线性）数据上的分支定界结果。括号内为下界时间。OOM = 内存溢出，TL = 时间 > 7200s。

	ours			Gurobi			MOSEK		
p	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
1000	31 (3)	0.00	355	227	0.00	653	208	0.00	361
2000	16 (2)	0.00	161	2091	0.00	1544	758	0.00	171
4000	27 (15)	0.00	77	4574	0.00	97	3612	0.00	73
8000	63 (45)	0.00	49	TL	100.00	1	TL	33.57	4
16000	151 (115)	0.00	33	TL	100.00	1	OOM	OOM	OOM

表 D.2：使用 $k = 10$, $\rho = 0.5$, $\lambda_2 = 1.0$, $M = 3.0$, $n/p = 1.0$的合成数据（逻辑回归）上的分支定界结果。括号内为下界时间。OOM = 内存不足，TL = 时间 > 7200秒。

	ours			gurobiOA			MOSEK		
p	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
1000	771（37）	0.00	6049	TL	67.29	417432	1871	0.00	5847
2000	630（81）	0.00	3451	TL	87.29	443958	7031	0.00	3427
4000	384（168）	0.00	855	TL	82.12	269137	TL	13.97	569
8000	666（404）	0.00	459	TL	69.86	93806	OOM	OOM	OOM
16000	1031（677）	0.00	207	TL	56.55	35414	OOM	OOM	OOM

表 D.3：$k = 10$、$\rho = 0.5$、$\lambda_2 = 1.0$、$M = 3.0$、$n/p = 1.0$ 的线性回归分支定界中央处理器与 GPU 结果。括号中显示下界时间。

	我们的方法 中央处理器			我们的方法 GPU		
p	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
1000	29 (2)	0.00	355	36 (9)	0.00	355
2000	15 (2)	0.00	161	18 (4)	0.00	161
4000	12 (4)	0.00	77	9 (2)	0.00	77
8000	44 (30)	0.00	49	15 (2)	0.00	49
16000	78 (53)	0.00	33	26 (4)	0.00	33

D.4.2 为真实世界数据集选择交叉验证参数 λ_2、k 和 M

为真实世界数据集选择超参数（λ_2, k, M），我们进行了 5 折交叉验证研究。对于 SantanderBank 和 DOROTHEA，我们评估 $\lambda_2 \in \{0.01, 0.1, 1.0, 2.0, 3.0, 10.0\}$和 $k \in \{2, 3, \ldots, 30\}$；在此网格搜索步骤中，我们固定 $M = 100$，使盒约束实际上处于非激活状态，不会影响模型选择。对于每个（λ_2, k）组合，我们使用 Liu 等人 [2023] 的束搜索方法，在每个训练折上计算一个启发式解，并报告训练折和验证折上的性能（线性回归使用平方误差；逻辑回归使用对数损失、准确率和曲线下面积），将其表示为 λ_2 和 k的函数；见图 D.13

表 D.4：使用 $k = 10$、 $\rho = 0.5$、 $\lambda_2 = 1.0$、$M = 3.0$、 $n/p = 1.0$ 的逻辑回归的分支定界中央处理器与 GPU 结果。括号内为下界时间。

	我们的方法 中央处理器			我们的方法 GPU		
p	时间（秒）	间隙（%）	节点数	时间（秒）	间隙（%）	节点数
1000	689 (38)	0.00	6049	758 (108)	0.00	6049
2000	536 (39)	0.00	3451	556 (65)	0.00	3451
4000	251 (62)	0.00	855	203 (21)	0.00	855
8000	412 (176)	0.00	459	246 (18)	0.00	459
16000	533 (276)	0.00	207	282（23）	0.00	207

（SantanderBank）和图 D.14–D.16（DOROTHEA）。根据这些结果，我们为 SantanderBank 选择（λ_2，k）=（1.0，10），为 DOROTHEA 选择（λ_2，k）=（1.0，6）；在这些设置下，束搜索解中的最大系数绝对值分别为 4.96 和 4.95，因此我们在实验中设置 $M = 10.0$，以确保盒约束不会影响解的质量。

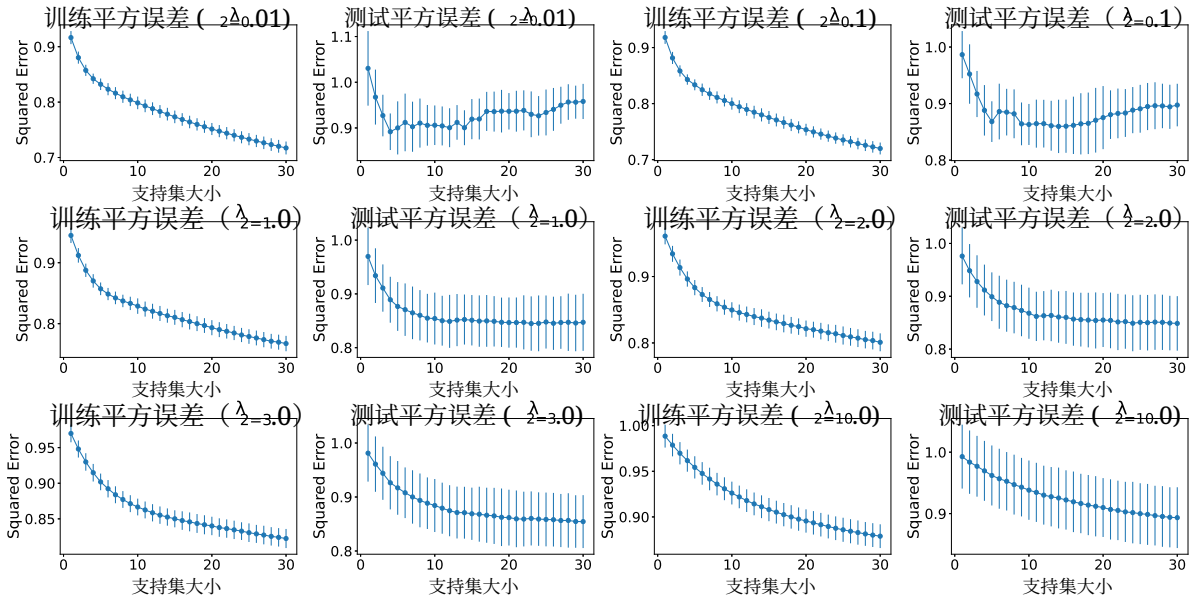

图 D.13：在 SantanderBank数据集上使用 $\lambda_2 \in \{0.01, 0.1, 1.0, 2.0, 3.0, 10.0\}$进行束搜索交叉验证的结果。

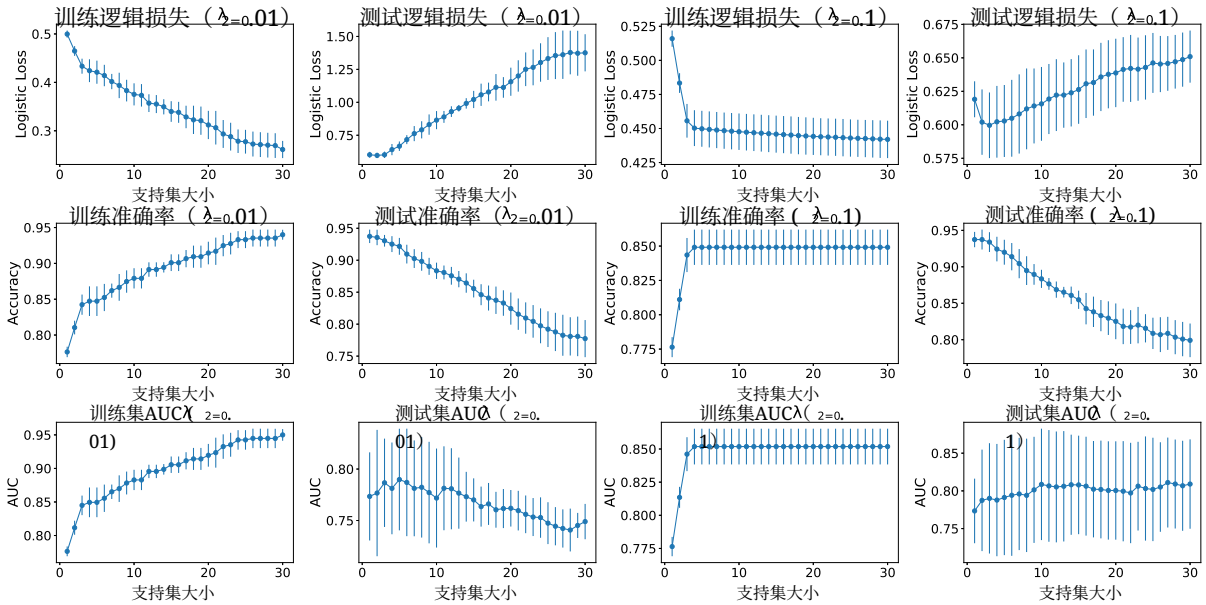

图 D.14：在 DOROTHEA数据集上使用 $\lambda_2 \in \{0.01, 0.1\}$进行束搜索交叉验证的结果。

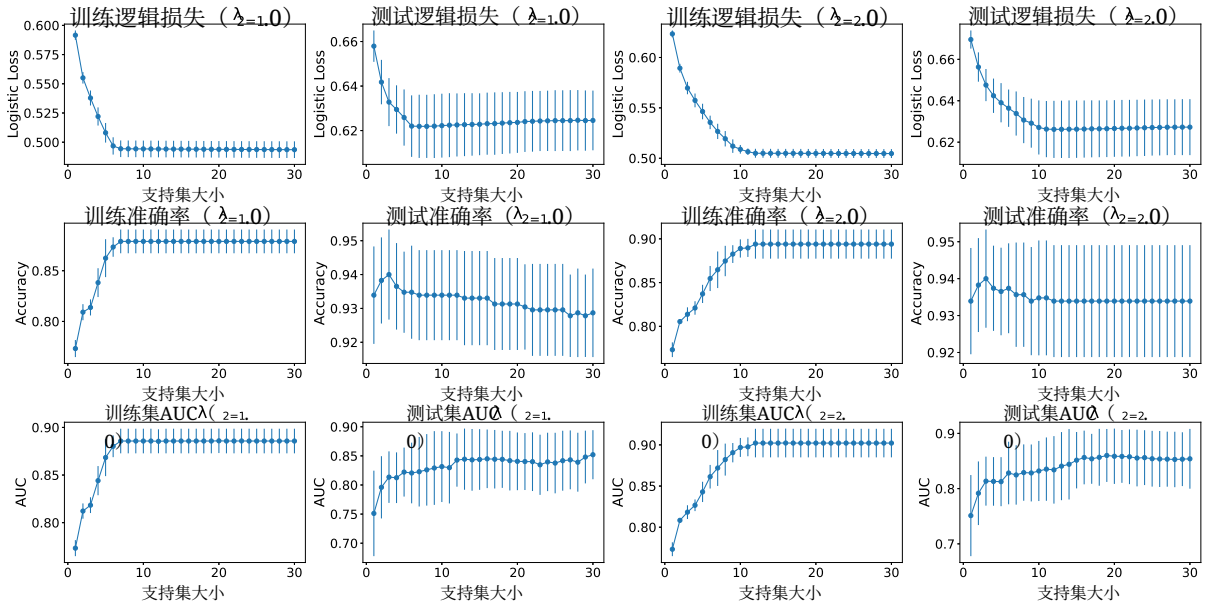

图 D.15：在 DOROTHEA 数据集上使用 $\lambda_2 \in \{1.0, 2.0\}$ 进行束搜索交叉验证的结果。

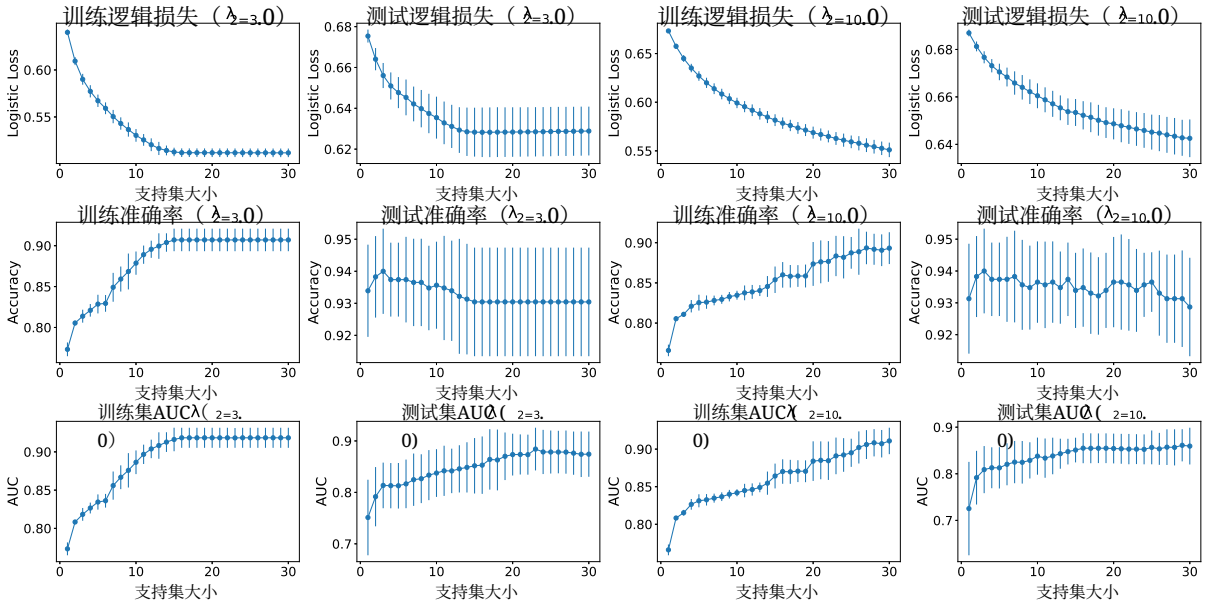

图 D.16：在 DOROTHEA 数据集上使用 $\lambda_2 \in \{3.0, 10.0\}$ 进行束搜索交叉验证的结果。