无名稿件 No.（由编辑稍后插入）

乘积域上的无投影方法

Immanuel Bomze · Francesco Rinaldi · Damiano Zeffiro

arXiv:2302.04839v2 [math.OC] 6 Dec 2023

收稿日期和接收日期将在稍后插入

摘要 无投影块坐标方法避免了每次迭代的高计算成本，同时利用了乘积域的特定问题结构。Frank-Wolfe 类方法是此类方法中最受欢迎的方法之一。然而，正如文献中所观察到的，经典 Frank-Wolfe 理论与块坐标情形之间存在理论空白；即使对于强凸目标，后者也无法保证线性收敛率。此外，先前的大多数研究都集中于凸目标函数。本文同时研究非凸情形，通过将一种全新且完整发展的收敛理论与新颖的活跃集识别结果相结合，缩小了上述理论空白；这些结果确保能够以高效的方式利用解所固有的稀疏性。初步数值实验似乎验证了我们的方法，并且还显示出在非凸情形下获得全局解的良好前景。

关键词 无投影优化 · 一阶优化 · 块坐标下降 t

数学主题分类（2010） MSC90C06 · MSC90C26 · MSC90C30

1 引言

我们考虑如下问题

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})\,, \tag{1}$$

目标函数 f 具有 L-Lipschitz 连续的梯度，且可行集 $\mathcal{C}\subseteq\mathbb{R}^n$ 闭且凸。此外，我们假设 $\mathcal{C}$ 具有块可分离性，即

$$\mathcal{C}=\mathcal{C}_{(1)}\times\ldots\times\mathcal{C}_{(m)} \tag{2}$$

Immanuel Bomze 维也纳运筹学中心与数据科学研究网络，维也纳大学 电子邮箱：immanuel.bomze@univie.ac.at Francesco Rinaldi 数学系，帕多瓦大学 电子邮箱：rinaldi@math.unipd.it Damiano Zeffiro 数学系，帕多瓦大学 电子邮箱：damiano.zeffiro@studenti.unipd.it

$\mathcal{C}_{(i)} \subset \mathbb{R}^{n_i}$对 $i \in [1{:}\ m]$闭且凸，当然还有 $\sum_{i=1}^m n_i = n$。

注意，问题 (1) 属于复合优化问题这一类

$$\min_{\mathbf{x}\in\mathcal{C}} [f(\mathbf{x}) + g(\mathbf{x})] \tag{3}$$

其中 f 光滑，且 $g(\mathbf{x}) = \sum_{i=1}^m \chi_{\mathcal{C}_{(i)}}(\mathbf{x}^{(i)})$ 为凸函数并具有分块可分性（例如，参见 [37]，其中概述了此类问题的求解方法）；这里 $\chi_D : \mathbb{R}^d \to [0, +\infty]$ 表示凸集 $D \subseteq \mathbb{R}^d$ 的示性函数，对于分块向量 $\mathbf{x} \in \mathbb{R}^n = \mathbb{R}^{n_1} \times ... \times \mathbb{R}^{n_m}$，我们用 $\mathbf{x}^{(i)} \in \mathbb{R}^{n_i}$ 表示对应于第 i 个块的分量，因此 $\mathbf{x} = (\mathbf{x}^{(1)},\ ...,\ \mathbf{x}^{(m)})$。

这类问题广泛应用于现实场景，例如交通分配、 [31],结构化支持向量机 [28], 基于迹范数的张量补全 [32], 降秩非参数回归 [22],半松弛最优传输 [23], 结构化子模最小化 [26],组融合 Lasso [1], 以及字典学习 [18]。

块坐标梯度下降（BCGD）策略（例如，参见 [4]）是凸情形下求解问题（1）的一种标准方法。处理非凸目标函数时，这些方法仍可作为有效工具，在概率全局优化框架中进行局部搜索（更多详情例如参见 [33] ）。BCGD 方法的原理很容易理解：这些方法在每次迭代时，针对一组变量构建原函数的适当模型，然后对与该组变量相关的可行集进行投影。

基于投影的策略（更多详情例如参见 $[6, 19, 21, 40]$ ）虽然在实践中也常以块坐标方式广泛使用（例如，参见 [35]），但即便是在流多面体、核范数球、比尔霍夫多面体或置换多面体等结构化集合上进行投影，也可能代价高昂（例如，参见 [20]）。因此，近年来，在处理此类结构化约束时，无投影方法（例如，参见 $[14, 25, 29]$）得到了广泛应用。

这些方法只需依赖一个合适的预言机，在每次迭代时对函数在原始可行集上的线性近似进行最小化，并返回其中一个点

$$\mathrm{argmin}_{\mathbf{x}\in\mathcal{C}} \langle \mathbf{g}, \mathbf{x} \rangle.$$

当 $\mathcal{C}$ 按式 (2) 定义时，得益于可行集的块可分结构，该问题可分解为 m 个相互独立的问题。进而，各块上的问题可以并行求解；文献中对此进行了广泛研究，尤其是在交通分配领域（例如，参见 [31]）。在大数据背景下，对所有变量进行完整更新仍可能构成计算瓶颈，需要在实践中妥善处理。因此，近期提出了经典 Frank-Wolfe（FW）方法的块坐标变体（例如，参见 $[28, 36, 41]$）。[28] 针对结构化支持向量机训练提出了这种方法：它在每次迭代时随机选择一个块，并对该块执行一次 Frank–Wolfe 更新。 [36], 介绍了该算法的若干改进，例如自适应分块采样、使用成对方向和远离步方向，或采用预言机调用缓存；显然，这些改进都是按顺序进行的。

然而，如果希望利用现代多核架构或分布式集群的优势，也有并行和分布式版本的块坐标 Frank-Wolfe 算法可供使用[41]。值得强调的是，上述所有论文仅考虑凸规划问题，并将随机采样变体作为主要的块选择策略。

此外，正如 [36], 中所指出的，Frank-Wolfe 变体（例如成对 Frank-Wolfe 和远离步 Frank-Wolfe 算法）的标准收敛性分析无法轻易扩展到块坐标情形。特别是，在这一设置下，针对 Frank-Wolfe 应用于强凸目标时广为人知的线性收敛速率保证尚未得到扩展

到 [14] 及其中所引用文献中的变体。这主要是由于在块坐标框架下处理不良/短步（即那些无法带来良好进展、但为保证迭代点的可行性而采取的步长）存在困难。在 [36], 中，作者因此在不存在不良步这一强假设下，将 Frank-Wolfe 变体的收敛性分析扩展到了块坐标设置，并声称要在一般情形下完成分析、弥合 Frank-Wolfe 与 BCFW 之间的差距，需要采用新的证明技术。

本文聚焦于非凸情形，并定义了一种新的一般性块坐标算法框架，使块选择策略和类Frank-Wolfe方向的使用都具有灵活性。这种灵活性主要得益于我们在各块中执行近似最小化的方式。每次迭代时，至少选定一个块后，我们会使用短步链（SSC）过程（见 [39],），该过程会跳过连续短步中的梯度计算，直至满足适当条件，从而完成对选定块的近似最小化。

关于块选择策略，我们探讨了三种不同的方案。我们考虑的第一种是并行或类 Jacobi 策略（例如，见 [5]），即对所有块执行 SSC 过程。与在整个变量空间中使用 SSC（例如，见 [39]）相比，这显然降低了计算负担，并最终能够利用多核架构并行执行这些任务。第二种方案是随机采样（例如，见 [28]），即每次迭代时，在随机选取的块子集上执行 SSC 过程。最后，我们给出 Gauss-Southwell 规则的一种变体（例如，见 [34]）：我们在所有块上执行 SSC，然后选择最违反最优性条件的块。这种贪心规则可能会使目标函数取得更大进展，因为它利用一阶信息来选择合适的块；但原则上，它比前面提到的其他方案计算开销更大（注意，每次迭代时都会对所有块执行 SSC）。

此外，我们还考虑以下无投影策略：离去步Frank-Wolfe（AFW）、成对Frank-Wolfe（PFW）以及带面内方向的Frank-Wolfe方法（FDFW）；更多细节请参见例如[39] 及其中的参考文献。AFW 和 PFW 策略依赖于一组“基本原子” A，使得 $\mathcal{C} = \text{conv}(A)$。给定 A，对于基点$\mathbf{x} \in \mathcal{C}$，我们可以定义

$$S_{\mathbf{x}} = \{S \subset A : \mathbf{x} \text{ is a proper convex combination of all the elements in } S\},$$

给定点$\mathbf{x}$的可能活跃集族。对于$\mathbf{x} \in \mathcal{C}$ 和 $S \in S_{\mathbf{x}}$，$\mathbf{d}^{\text{PFW}}$ 是相对于活跃集 S 和梯度 $-\mathbf{g}$ 的 PFW方向，当且仅当

$$\mathbf{d}^{\text{PFW}} = \mathbf{s} - \mathbf{q} \text{ with } \mathbf{s} \in \text{argmax}_{\mathbf{s}\in\mathcal{C}} \langle \mathbf{s}, \mathbf{g}\rangle \text{ and } \mathbf{q} \in \text{argmin}_{\mathbf{q}\in S} \langle \mathbf{q}, \mathbf{g}\rangle . \tag{4}$$

类似地，给定$\mathbf{x} \in \mathcal{C}$ 和 $S \in S_{\mathbf{x}}$，$\mathbf{d}^{\text{AFW}}$ 是相对于活跃集 S 和梯度 $-\mathbf{g}$ 的 AFW 方向，当且仅当

$$\mathbf{d}^{\text{AFW}} \in \text{argmax}\{\langle \mathbf{g}, \mathbf{d}\rangle : \mathbf{d} \in \{\mathbf{d}^{\text{FW}}, \mathbf{d}^{\text{AS}}\}\}, \tag{5}$$

其中，$\mathbf{d}^{\text{FW}}$ 是经典的 Frank-Wolfe 方向

$$\mathbf{d}^{\text{FW}} = \mathbf{s} - \mathbf{x} \text{ with } \mathbf{s} \in \text{argmax}_{\mathbf{s}\in\mathcal{C}} \langle \mathbf{s}, \mathbf{g}\rangle , \tag{6}$$

且$\mathbf{d}^{\text{AS}}$ 是远离方向

$$\mathbf{d}^{\text{AS}} = \mathbf{x} - \mathbf{q} \text{ with } \mathbf{q} \in \text{argmin}_{\mathbf{q}\in S} \langle \mathbf{q}, \mathbf{g}\rangle . \tag{7}$$

FDFW 只需要当前点 $\mathbf{x}$ 和梯度 $-\mathbf{g}$ 来选择下降方向（即不需要跟踪活跃集），其定义为

$$\mathbf{d}^{F} = \mathbf{x} - \mathbf{x}_F \text{ with } \mathbf{x}_F \in \text{argmin}\{\langle \mathbf{g}, \mathbf{y}\rangle : \mathbf{y} \in \mathcal{F}(\mathbf{x})\}$$

对于 $\mathcal{F}(\mathbf{x})$，包含 $\mathbf{x}$ 的最小面。然后，该选择准则类似于 AFW 所采用的准则：

$$\mathbf{d}^{\mathrm{FD}} \in \operatorname{argmax}\{\langle \mathbf{g}, \mathbf{d}\rangle : \mathbf{d} \in \{\mathbf{d}^{F}, \mathbf{d}^{\mathrm{FW}}\}\}. \tag{8}$$

从理论角度来看，这一新的算法框架使我们能够给出：

– 对于任意的块选择策略和类 FW 方向，都能获得局部线性收敛速率。该结果基于 Kurdyka-Lojasiewicz（KL）性质（例如，参见 [3], [7] 和 [8]）以及量身定制的角度条件（例如，参见 [39]）。得益于框架中处理短步的方式，我们得以将针对 FW 变体的分析推广到块坐标情形，从而弥补 [36] 中指出的相关理论空缺。

– 对于定义可行集 $\mathcal{C}$ 的笛卡尔积的特定结构，在采用合适的无投影策略（即使用 AFW 方向）且目标函数为一般光滑非凸目标函数时，可获得局部活跃集识别结果（例如，参见 [12,13,15,24]）。特别地，我们证明了该框架能够在有限时间内识别出解的支撑集。这一理论性质有助于降低当前问题的维数，从而减少优化过程的总体计算成本。

据我们所知，这是首次针对块坐标 FW 变体同时给出（无坏步的）线性收敛速率和活跃集识别结果。具体而言，我们解决了 [36] 上文讨论的开放问题，证明 FW 变体的线性收敛速率确实可以扩展到块坐标设置。此外，我们的结果首次保证了在无投影优化方法的相关文献中，无需定制的活跃集策略，即可在单次迭代 中识别局部活跃集。

我们还报告了关于一类具有块可分可行集的结构化问题的初步数值结果。结果表明，所提出的框架优于经典的块坐标 FW；同时，得益于其灵活性，该框架还可以有效地嵌入概率全局优化框架，从而显著提升其性能。

本文的组织结构如下。第 2 节介绍我们所提出的新算法框架的详细内容。第 3 节对其收敛性质进行了深入分析。第 4 节给出了活跃集识别结果。第 5 节报告了初步数值结果，重点分析我们框架的局部识别能力和收敛性质。最后，第 6 节给出总结性评述。

1.1 符号说明

对于闭凸集 $C \subset \mathbb{R}^h$，我们用 $\pi(C, \mathbf{x})$ 表示将 $\mathbf{x} \in \mathbb{R}^h$ 投影到 C上的结果，并用 $T_C(\mathbf{x})$ 表示 C 在 $\mathbf{x} \in C$ 处的切空间。对于 $\mathbf{g} \in \mathbb{R}^h$，我们也用 $\pi_{\mathbf{x}}(\mathbf{g})$ 简写为$\|\pi(T_C(\mathbf{x}),\mathbf{g})\|$。对于 $\mathbf{y} \neq \mathbf{o}$，我们将向量 $\mathbf{y}/\|\mathbf{y}\|$ 记作 $\hat{}\mathbf{y}$；否则，$\hat{}\mathbf{y} = \mathbf{o}$。最后，我们分别用$B_r(\mathbf{x})$ 和 $B_r(\mathbf{x})$ 表示以 $\mathbf{x}$ 为中心、半径为 r 的闭球和开球。

2 一种新的块坐标无投影方法

本文考虑的块坐标框架在每次迭代中，对若干块应用下文描述的短步链（SSC）过程 [39],（即算法 2）。具体方案见算法 1；回顾记号 $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots\ldots, \mathbf{x}^{(m)})$，其中 $\mathbf{x}^{(i)} \in \mathcal{C}_{(i)}$，所有 $i \in [1{:}\ m]$。

在算法 1 中，我们在每次迭代执行两个主要操作。首先，在第 3 步中，我们根据给定的块选择策略，选择合适的块子集 $\mathcal{M}_k$。然后，我们更新（步骤

算法 1 采用短步链（SSC）过程的块坐标法

1：$\mathbf{x}_0 \in \mathcal{C}$、 $k = 0$。2：如果 $\mathbf{x}_k$ 是驻点，则停止。3：选择 $\mathcal{M}_k \subset [1\colon m]$。4：对于所有 $i \notin \mathcal{M}_k$，设置 $\mathbf{x}_{k+1}^{(i)} = \mathbf{x}_k^{(i)}$ 5：对于所有 $i \in \mathcal{M}_k$，设置 $\mathbf{x}_{k+1}^{(i)} = \mathrm{SSC}(\mathbf{x}_k^{(i)}, -\nabla f(\mathbf{x}_k)^{(i)})$ 6：$k = k + 1$。转到第 2 步。

4 和 5）通过 SSC 过程更新与选定块相关的变量，同时保持其他块中的所有变量不变。

现在，我们简要回顾 [39], 中的 SSC 过程。该过程旨在连续出现不良步时重复利用梯度，直到满足适当的停止条件，如算法 2 所示。

算法 2 短步链过程 – SSC($\bar{\mathbf{x}}$, $\mathbf{g}$)

1: **初始化**。$\mathbf{y}_0 = \bar{\mathbf{x}}$, $j = 0$2: 选择 $\mathbf{d}_j \in \mathcal{A}(\mathbf{y}_j, \mathbf{g})$, $\alpha_{\max}^{(j)} \in \alpha_{\max}(\mathbf{y}_j, \mathbf{d}_j)$ 3: **如果** $\mathbf{d}_j = 0$ **则返回** $\mathbf{y}_j$4: **结束 if**5: 计算一个辅助步长 β_j6: 令 $\alpha_j = \min(\alpha_{\max}^{(j)}, \beta_j)$ 7: $\mathbf{y}_{j+1} = \mathbf{y}_j + \alpha_j \mathbf{d}_j$8: **如果** $\alpha_j = \beta_j$ **则返回** $\mathbf{y}_{j+1}$9: **结束 if**10: $j = j + 1$, 转到第 2 步

通过 $\mathcal{A}$，我们表示一种无投影策略，用于在应用 SSC 的块上为光滑函数生成一阶可行下降方向（例如 FW、PFW、AFW 方向）。由于在 SSC 过程中梯度 $-\mathbf{g}$ 保持不变，因此不难看出，该过程相当于应用 $\mathcal{A}$，以合适的步长和停止条件最小化线性化目标函数 $f_{\mathbf{g}}(\mathbf{z}) = \langle -\mathbf{g}, \mathbf{z} - \bar{\mathbf{x}} \rangle + f(\bar{\mathbf{x}})$。更具体地说，在进行驻点检查（见步骤 2–4）后，步长 α_j 取辅助步长 $\beta_j > 0$ 和最大步长 $\alpha_{\max}^{(j)}$ 中的较小值（我们始终假设最大步长严格为正）。步骤 7 生成的点 $\mathbf{y}_{j+1}$ 始终是可行的，因为 $\alpha_j \leq \alpha_{\max}^{(j)}$。请注意，如果 SSC 中使用的方法 $\mathcal{A}$ 执行 FW 步（FW 步的定义见公式 (6)），那么 SSC 将终止，此时 $\alpha_j = \beta_j$，或者 $\mathbf{y}_{j+1}$ 是 $f_{\mathbf{g}}$ 的全局极小点。

因此，辅助步长 β_j （参见 SSC 过程的步骤 5）定义为信赖域在 $\mathbf{y}_j$ 处的最大可行步长

$$\Omega_j = B_{\|\mathbf{g}\|/2L}(\bar{\mathbf{x}} + \frac{\mathbf{g}}{2L}) \cap B_{\langle \mathbf{g}, \hat{\mathbf{d}}_j \rangle / L}(\bar{\mathbf{x}}) \,. \tag{9}$$

这保证了充分下降条件

$$f(\mathbf{y}_j) \leq f(\mathbf{x}_k) - \frac{L}{2} \|\mathbf{x}_k - \mathbf{y}_j\|^2 \tag{10}$$

因此，SSC 中的 f 会单调下降。更多详情请参见 [39]。

2.1 块选择策略

正如引言中简要提到的，我们将在分析中考虑三种不同的块选择策略。第一种是并行或类 Jacobi 策略（例如，参见 [5]）。在这种情况下，我们在每次迭代中选择所有块。正如我们已经观察到的，这比一次处理整个变量空间的计算成本更低。此外，还可以考虑利用多核架构并行执行这些任务。下面给出该策略的定义：

定义 1（并行选择）令 $\mathcal{M}_k = [1:m]$。

第二种策略是 GS 规则的一种变体（参见例如 [34]）：我们首先在所有块中执行 SSC，然后选择违反最优性条件程度最大的块。其正式定义如下。

定义 2（Gauss-Southwell（GS）选择）令 $\mathcal{M}_k = \{i(k)\}$，其中

$$i(k) \in \operatorname{argmax}_{i\in[1:m]} \langle \mathbf{g}^{(i)}, \mathrm{SSC}(\mathbf{x}_k^{(i)}, -\nabla f(\mathbf{x}_k)^{(i)}) - \mathbf{x}_k^{(i)} \rangle .$$

最后，我们采用随机采样（参见例如 [28]）。在每个时刻，我们随机生成一个索引采用均匀概率分布进行迭代。本例中的定义如下：

定义 3（随机采样）集合 $\mathcal{M}_k = \{i(k)\}$，其中 $i(k)$ 的索引在 $[1$ 中均匀随机选取：$m]$。

3 收敛性分析

本节分析算法框架的收敛性质。具体而言，我们证明，在对块适当定义角度条件，并且目标函数满足局部 KL 条件的情况下，无论采用何种块选择策略，都能获得线性收敛速率。本节给出的收敛性分析将 [39] 中的结果扩展到了块坐标设置，这是一项颇具挑战的任务，绝非易事。因此，这一扩展需要一些新的论证，下面将介绍这些论证，并在附录中详细阐述。

我们的收敛框架采用了 [38,39] 中引入的角度条件。该条件确保该方法选取的下降方向的斜率在一个常数因子范围内达到最优。现在我们回顾这一角度条件。对于 $\mathbf{x} \in \mathcal{C}$ 和 $\mathbf{g} \in \mathbb{R}^n$，我们首先将方向斜率下界定义为

$$\mathrm{DSB}_{\mathcal{A}}(\mathcal{C}, \mathbf{x}, \mathbf{g}) = \inf_{\mathbf{d}\in\mathcal{A}(\mathbf{x},\mathbf{g})} \frac{\langle \mathbf{g}, \mathbf{d}\rangle}{\pi_{\mathbf{x}}(\mathbf{g})\|\mathbf{d}\|}, \tag{11}$$

如果$\mathbf{x}$ 不是 $-\mathbf{g}$ 的驻点，否则我们令 $\mathrm{DSB}_{\mathcal{A}}(\mathcal{C}, \mathbf{x}, \mathbf{g}) = 1$。随后，我们将斜率下界定义为

$$\mathrm{SB}_{\mathcal{A}}(\mathcal{C}, P) = \inf_{\substack{\mathbf{g}\in\mathbb{R}^n \\ \mathbf{x}\in P}} \mathrm{DSB}_{\mathcal{A}}(\mathcal{C}, \mathbf{x}, \mathbf{g}) = \inf_{\substack{\mathbf{g}:\pi_{\mathbf{x}}(\mathbf{g})\neq 0 \\ \mathbf{x}\in P}} \mathrm{DSB}_{\mathcal{A}}(\mathcal{C}, \mathbf{x}, \mathbf{g}) \,. \tag{12}$$

我们将 $\mathrm{SB}_{\mathcal{A}}(\mathcal{C})$ 作为 $\mathrm{SB}_{\mathcal{A}}(\mathcal{C}, \mathcal{C})$ 的简写，并称该方法满足角度条件$\mathcal{A}$，如果

$$\mathrm{SB}_{\mathcal{A}}(\mathcal{C}) = \tau > 0 \,. \tag{13}$$

备注 *1* 当 $\mathcal{C}$ 为多胞体时，AFW、PFW 和 FDFW 均满足角度条件。[39], 中给出了这一结果的详细证明，并介绍了其他一些满足角度条件的方法，这些方法适用于具有光滑边界的凸集，详见 [38]。

现在我们给出用于分析该算法收敛性的局部 KL 条件。此前也曾使用过相同的条件，见 [39, 假设 2.1]。

假设 1 给定一个驻点 $\mathbf{x}_* \in \mathcal{C}$，存在 $\eta, \delta > 0$，使得对于每个 $\mathbf{x} \in B_\delta(\mathbf{x}_*)$，若满足 $f(\mathbf{x}_*) < f(\mathbf{x}) < f(\mathbf{x}_*) + \eta$，则有

$$\pi_{\mathbf{x}}(-\nabla f(\mathbf{x})) \geq \sqrt{2\mu}[f(\mathbf{x}) - f(\mathbf{x}_*)]^{\frac{1}{2}}\,. \tag{14}$$

在处理凸规划问题时，解集上的指数为 2 的 H¨older 误差界意味着条件 (14) 成立，见 [9, 推论 6]。因此，在处理 μ-强凸函数时（例如，见 [27]），我们的假设成立；特别是，对于引言中讨论的 [36] 开放问题所涉及的情形，假设也成立。不过，值得强调的是，误差界 (14) 在多种凸和非凸情形下都成立（关于此问题的详细讨论，见 [39] ）。一个与我们的分析相关的有趣例子是 f为（非凸）二次函数的情形，即 $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$，且 $\mathcal{C}$ 为多胞体。

现在报告我们的主要收敛结果。

定理 1 设假设 *1* 在 $\mathbf{x}_*$ 处成立。考虑由算法 *1* 生成的序列 $\{\mathbf{x}_k\}$。假设：

– 对于相同的 $\tau > 0$，每个块都满足角度条件 (13)；– *SSC* 过程总能在有限步内终止。– $f(\mathbf{x}_*)$ 是包含 $\{\mathbf{x} \in \mathcal{C}\colon f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ 的连通分支中的一个极小值，该连通分支包含 $\mathbf{x}_0$。

˜那么，存在 $\delta > 0$，使得若 $\mathbf{x}_0 \in B_{\tilde{\delta}}(\mathbf{x}_*)$：

– 对于并行块选择策略，有

$$f(\mathbf{x}_k) - f(\mathbf{x}_*) \leq (q_P)^k [f(\mathbf{x}_0) - f(\mathbf{x}_*)]\,, \tag{15}$$

且 $\mathbf{x}_k \to \tilde{\mathbf{x}}_*$ 满足

$$\|\mathbf{x}_k - \tilde{\mathbf{x}}_*\| \leq \frac{\sqrt{2 - 2q_P}}{\sqrt{L}(1 - \sqrt{q_P})}\,(q_P)^{\frac{k}{2}}[f(\mathbf{x}_0) - f(\tilde{\mathbf{x}}_*)]\,, \tag{16}$$

for

$$q_P = 1 - \frac{\mu\tau^2}{4L(1+\tau)^2}.$$

– 对于 *GS* 块选择策略，我们有

$$f(\mathbf{x}_k) - f(\mathbf{x}_*) \leq (q_{GS})^k [f(\mathbf{x}_0) - f(\mathbf{x}_*)]\,, \tag{17}$$

以及$\mathbf{x}_k \to \tilde{\mathbf{x}}_*$ 与

$$\|\mathbf{x}_k - \tilde{\mathbf{x}}_*\| \leq \frac{\sqrt{2 - 2q_{GS}}}{\sqrt{L}(1 - \sqrt{q_{GS}})}\,(q_{GS})^{\frac{k}{2}}[f(\mathbf{x}_0) - f(\tilde{\mathbf{x}}_*)]\,, \tag{18}$$

for

$$q_{GS} = 1 - \frac{\mu\tau^2}{4mL(1+\tau)^2}\,,$$

– 对于随机块选择策略，在附加条件

$$\min\{f(\mathbf{x}) : \|\mathbf{x}-\mathbf{x}_*\| = \delta\} > f(\mathbf{x}_*) \tag{19}$$

对于某个 $\delta > 0$，有

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}_*)] \le (q_R)^k [f(\mathbf{x}_0) - f(\mathbf{x}_*)]\,, \tag{20}$$

并且 $\mathbf{x}_k \to \tilde{}\mathbf{x}_*$ 几乎必然与

$$\mathbb{E}[\|\mathbf{x}_k - \tilde{\mathbf{x}}_*\|] \le \frac{\sqrt{2-2q_R}}{\sqrt{L}(1-\sqrt{q_R})}\,(q_R)^{\frac{k}{2}}[f(\mathbf{x}_0) - f(\tilde{\mathbf{x}}_*)] \tag{21}$$

对于 $q_R = q_{GS}$。

该收敛结果将 [39, 定理 4.2] 扩展到了我们的块坐标设置。然而，由于 SSC 在这里独立地应用于不同的块，我们无法直接应用 [39]中的结果。相反，我们利用乘积域的切锥结构，结合应用于单个块的 SSC 的性质：

$$T_{\mathcal{C}}(\mathbf{x}) = T_{\mathcal{C}_{(1)}}(\mathbf{x}^{(1)}) \times \cdots \times T_{\mathcal{C}_{(m)}}(\mathbf{x}^{(m)})\,. \tag{22}$$

这要求证明 SSC 生成的序列具有比 [39]中所述更强的性质。相关细节及对 [39] 中相关结果的引用可见附录。另一方面，SSC 过程的有限终止性可由 [38,39], 中证明的结果直接保证，尤其是对于应用于多胞体上的 AFW、PFW 和 FDFW。

注 *2* 如果可行集 $\mathcal{C}$ 是一个多胞体，并且假设目标函数 f 在算法生成的每个点上都满足条件 **(14)**，其中固定 $f(\mathbf{x}_*)$，那么在 SSC 中使用 AFW（PFW 或 FDFW）的算法 **1** 将以如上所述的速率收敛。对于 μ-强凸函数，条件 **(14)** 成立，因此在这些情况下，我们的算法会以定理 **1** 给出的速率全局收敛。

注 *3* 并行策略和 GS策略具有相同的收敛速率，但常数不同。具体而言，决定 GS 情况下收敛速率的常数取决于所使用的块数（块数越多，收敛速率越差），且大于并行情况下的对应常数。

注 *4* 随机块选择策略与 GS策略具有相同的收敛速率，但这是以期望形式给出的。具体而言，决定收敛速率的常数与 GS 情况下的相同，因此取决于所使用的块数。请注意，在这种情况下，还需要对 $\mathbf{x}_*$ 作出额外的技术假设 **(19)**。

4 活跃集识别

现在，我们给出适用于我们框架的活跃集识别结果。我们只关注在 SSC 中采用 AFW 的算法 **1**，并假设严格互补性成立，且笛卡尔积中的集合具有特定结构：

$$\mathcal{C} = \Delta^{n_1} \times \ldots \times \Delta^{n_m}, \tag{23}$$

因此，集合 $\mathcal{C}_{(i)}$ 是 $(n_i - 1)$ 维标准单纯形

$$\Delta^{n_i} = \{\mathbf{x} \in \mathbb{R}^{n_i}_+ : \mathbf{x}^\top \mathbf{e}^{(i)} = 1\}\,, \quad i \in [1\!:\!m]\,,$$

对于 $\mathbf{e} \in \mathbb{R}^n$，即各分量均等于 **1** 的向量。现在我们给出主要识别结果。详细证明见附录。

定理 2 在关于 $\mathcal{C}$ 的上述假设下，设 $\mathcal{A}^{(i)}$ 为 $i \in [1\colon\ m]$ 的 *AFW*，并假设在 $\mathbf{x}_* \in \mathcal{C}$ 处满足严格互补条件。

– 如果 $\{\mathbf{x}_k\}$ 由采用并行选择的算法 *1* 生成，那么存在$\mathbf{x}_*$ 的一个邻域 U，使得如果 $\mathbf{x}_k \in U$，那么 $\operatorname{supp}(\mathbf{x}^{k+1}) = \operatorname{supp}(\mathbf{x}_*)$。– 如果 $\{\mathbf{x}_k\}$ 由采用随机选择或 *GS* 选择的算法 *1* 生成，那么存在$\mathbf{x}_*$ 的一个邻域 U，使得如果 $\mathbf{x}_k \in U$，那么 $\operatorname{supp}(\mathbf{x}_{k+1}^{i(k)}) = \operatorname{supp}(\mathbf{x}_*^{i(k)})$。

当我们的算法生成的序列收敛到点 $\mathbf{x}_*$时，很容易看出，对于足够大的 k，迭代点的支撑集与$\mathbf{x}_*$ 的最终支撑集相同。

推论 1 在上述关于 $\mathcal{C}$的假设下，令 $\mathcal{A}^{(i)}$ 为 $i \in [1\colon m]$的 *AFW*，并假设$\mathbf{x}_* \in \mathcal{C}$处满足严格互补条件。若$\mathbf{x}_k \to \mathbf{x}_*$（几乎必然），则对于并行选择和 *GS* 选择（随机采样），当 k 足够大时，有$\operatorname{supp}(\mathbf{x}_k) = \operatorname{supp}(\mathbf{x}_*)$。

这一结果具有重要的实际意义，尤其是在处理稀疏优化问题时。当 k 较大时，算法的迭代点具有固定的支撑集，因此此时我们只需关注少数几个支撑集分量，而忽略其他分量。由此，我们可以在算法中嵌入复杂的工具（例如缓存策略、二阶方法），从而显著提升最终的运算速度。

5 数值结果

这里，我们报告非凸二次优化问题 Multi-StQP [16] 的一些初步数值结果。该问题定义在若干个（此处相同的）单纯形的乘积上，即

$$\min\left\{\mathbf{x}^\top \bar{\mathrm{Q}}\mathbf{x} : \mathbf{x} \in (\Delta^l)^m\right\}. \tag{24}$$

矩阵 Q 的生成方式确保问题 (24) 的解具有稀疏但非顶点的分量。事实上，这正是 FW 变体已证明更为有效的情形 [15,39]。为了获得所需性质，我们考虑对随机 StQP 的一种扰动 [11]。给定 $\{\bar{\mathrm{Q}}_i\}_{i\in[1:m]}$，表示 m 个可能的 StQP，其中 $\bar{\mathrm{Q}}_i \in \mathbb{R}^{l\times l}$ 对于 $i \in [1\colon\ m]$，相应的样本空间为 $[1\colon m]$ 的随机 StQP 定义为

$$\max\left\{\sum_{i=1}^{m} p_i \mathbf{y}_i^\top \bar{\mathrm{Q}}_i \mathbf{y}_i : \mathbf{y}_i \in \Delta^l \quad \text{for all } i \in [1\colon m]\right\}. \tag{25}$$

其中 p_{-i} StQP 的概率 i。等价地，(25) 是问题 (24) 的一个实例，其中 Q =Q，f or

$$\bar{\mathrm{Q}} = \begin{bmatrix} -p_1\bar{\mathrm{Q}}_1 & 0 & \cdots & 0 \\ 0 & -p_2\bar{\mathrm{Q}}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -p_m\bar{\mathrm{Q}}_m \end{bmatrix}. \tag{26}$$

在我们的测试中，我们向随机 StQP 中加入了一个将各块耦合起来的扰动。更准确地说，矩阵 Q 被设为 Q $+\varepsilon$Q，其中 Q 是一个各元素相互独立且服从标准高斯分布的随机矩阵。系数 ε被设为 $\frac{1}{2m^2}$。我们令 $\bar{\mathrm{Q}}_i = \bar{\mathrm{A}}_i + \alpha \mathrm{I}_l$，其中 $\alpha = \overline{0.5}$且 $\bar{\mathrm{A}}_i$是 Erd˝os-R´enyi 随机图的邻接矩阵，其中每一对顶点彼此独立地以概率 p 通过一条边连接。因此，对于 $i \in [1\colon\ m]$ 该问题

$$\min\left\{-\mathbf{y}^\top \bar{\mathrm{Q}}_i \mathbf{y} : \mathbf{y} \in \Delta^l\right\} \tag{27}$$

是一个正则化最大团 formulation，其中每个极大团对应于一个唯一的严格局部最大化点，其支撑集等于该团的顶点；反之亦然（参见 [10] 及其中的参考文献）。概率 p 设置如下

$$p = \binom{l}{s}^{\frac{2}{s(s-1)}}, \tag{28}$$

对于 s，取最接近 $0.4l$ 的整数，使得大小为 $\approx 0.4l$ 的团的期望数量为 1（参见例如 [2]）。请注意，扰动项 Q 确保问题 (24) 无法通过分别优化每个块来求解。

这里我们指出，关于如何从较小的实例出发构造大型 StQP，同时保留其解的结构，文献 [17] 中已有讨论。然而，虽然由此得到的问题在较大问题的可行集上可以解耦，但它们仍然在较小实例的可行集的乘积上解耦；就我们的目的而言，这等同于块对角结构。

我们总共测试了四种方法：采用并行、GS 和随机更新的 AFW + SSC（分别为 PAFW+ SSC、GSAFW + SSC、BCAFW + SSC），以及采用随机更新的 FW（BCFW，与 [28] 中介绍的块坐标 FW 相同）。我们的测试重点关注方法的局部识别能力和收敛性质。

代码使用 Python 和 numpy 软件包编写，测试在配备 32GB 内存的 Intel Core i9-12900KS CPU 3.40GHz 上进行。数值测试相关代码可通过以下链接获取：

https://github.com/DamianoZeffiro/Projection-free-product-domain.

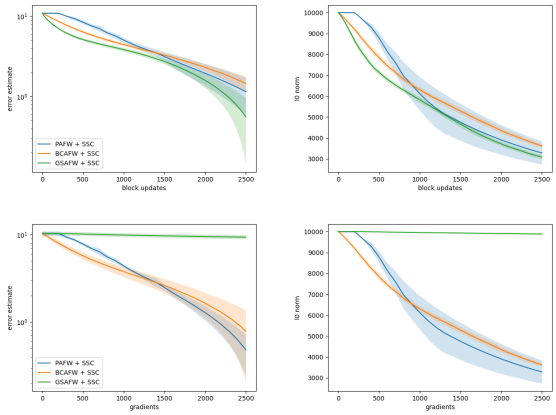

图 1：使用多起点比较 GSAFW +SSC、PAFW + SSC 和 BCAFW + SSC.$l = m = 100$。

5.1 多起点

我们首先采用多起点方法，将 20 次运行的结果取平均值；对于目标函数的 5 次随机初始化，每次选择 4 个起始点。

我们同时衡量迭代点的最优性间隙（误差估计）和稀疏性（非零分量的数量，ℓ_0 范数），并在图中报告平均值和标准差。计算最优性间隙时使用的估计全局最优值，是通过将 10^{-5} 从算法找到的最佳局部解中减去得到的。我们主要考察块梯度计算方面的性能；与以往研究（例如，参见 [28]）一样，每当在某个块上执行一次 SSC，就计为一次梯度计算。在一些涉及 GSAFW + SSC 的测试中，我们改为考察块更新；每当算法在某个块上修改当前迭代点，就计为一次块更新。需要特别指出的是，由于每次块更新时梯度保持不变，且只需在 SSC 开始时进行一次线性最小化，因此，对于我们考察的 FW 变体而言，我们的算法所需的梯度计算次数也等于各块上的线性最小化次数。

我们首先比较 PAFW + SSC、BCAFW + SSC 和 GSAFW +SSC（图 1）。如预期，GSAFW + SSC 在块更新方面表现良好，但在块梯度计算方面表现很差，因为每次迭代都必须计算 m 个梯度才能更新单个块。随后，我们比较了 PAFW + SSC、BCAFW + SSC 和 BCFW。结果（图 2）清楚地表明，PAFW + SSC 和 BCAFW +SSC 的表现优于 BCFW。所有这些发现都与第 7.2 节中描述的理论结果一致。

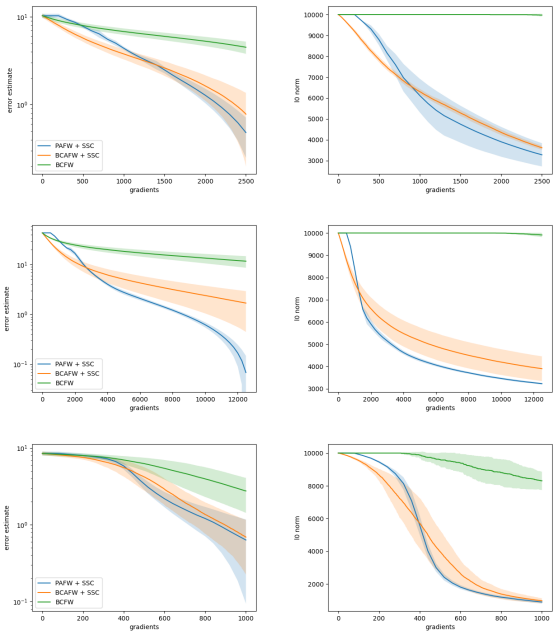

图 2：使用多起点对 BCFW、PAFW + SSC 和 BCAFW + SSC 进行比较。第一行中为 $l = m =$100，第二行中为 $l = 40$ 和 $m = 250$，第三行中为 $l = 250$ 和 $m = 40$。

5.2 单调盆地跳跃

接下来，我们考虑算法 3 中描述的单调盆地跳跃方法（例如，参见 [30,33]）。该方法计算一个局部优化器 $\mathbf{x}_{*,i}$，使其接近当前迭代点 $\bar{}\mathbf{x}_i$（步骤 2）。其中存在一个局部优化算法，给定输入 $\mathcal{M}$ 和 $\bar{}\mathbf{x}_i$ 后，子程序 LO 会返回从 $\bar{}\mathbf{x}_i$ 开始应用 $\mathcal{M}$ 所得到的结果，并采用合适的停止准则；在我们的情形中，该准则由梯度计算次数上限给定，设为 $10m$。在前 i 次迭代中找到的最佳点序列 $\{\bar{}\mathbf{x}_{*,i}\}$ 会在步骤 3 更新；在步骤 5 中，选择 $\bar{}\mathbf{x}_{i+1}$，使其位于 $\bar{}\mathbf{x}_{*,i}$ 的某个邻域内。邻域 $B(\mathbf{x}\ \gamma)$ 对于 $\mathbf{x} \in \mathcal{C}$ 和 $\gamma \in (0,1]$ 定义为

$$B(\mathbf{x},\gamma) = \{\mathbf{x} + \gamma(\mathbf{y} - \mathbf{x}) : \mathbf{y} \in \mathcal{C}\}\,. \tag{29}$$

在测试中，我们在 $\mathcal{C}$ 中均匀随机地选取 $\mathbf{y}$，并令 $\bar{}\mathbf{x}_{i+1} = \bar{}\mathbf{x}_i + \gamma(\mathbf{y} - \bar{}\mathbf{x}_i)$，其中 $\gamma = 0.25$。

算法 3 单调盆地跳跃策略

1: $\bar{}\mathbf{x}_0 \in \mathcal{C}$, $i_{\max} \in \mathbb{N}$, $\gamma \in [0,1]$, $i = 0$。令 $\bar{}\mathbf{x}$ $_{0,-1} = \bar{}\mathbf{x}$: 计算局部 2: 优化器 $\mathbf{x}_{*,i} = \mathrm{LO}(\mathcal{M}, \mathbf{x}_i)$ 3: 若 $f(\mathbf{x}_{*,i}) < f(\bar{}\mathbf{x}_{*,i-1})$，则令 $\bar{}\mathbf{x}_{*,i} = \mathbf{x}_{*,i}$；否则令 $\bar{}\mathbf{x}_{*,i} = \bar{}\mathbf{x}_{*,i-1}$。4: 若 $i = i_{\max}$，则停止。5: 在 $B(\bar{}\mathbf{x}_{*,i}, \gamma)$ 中随机选取 $\bar{}\mathbf{x}_{i+1}$。6: 令 $i = i + 1$。转到第 2 步。

我们在步骤 2 中作为子程序考虑的方法包括 PAFW + SSC、BCAFW + SSC 和 BCFW。我们设定 $i_{\max} = 9$，并运行 10 次算法 3，每次都随机初始化起点。我们再次绘制 $\tilde{}$ $\tilde{}$ 平均值和标准差，其中 $\{f(\bar{}\mathbf{x}_{*,i}) - f^*\}$，并通过 f^* 估计全局最优值（即从所找到的最佳解中减去 10^{-1} 所得）。

结果再次表明，PAFW + SSC 和 BCAFW + SSC 找到的解优于 BCFW；在大多数实例中，如果 $l \le m$，BCAFW + SSC 的表现优于 PAFW + SSC。

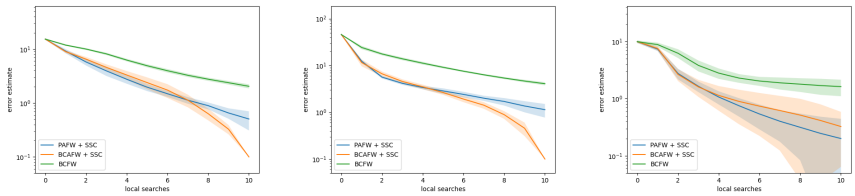

图 3：使用单调盆地跳跃比较 BCFW、PAFW + SSC 和 BCAFW + SSC。从左到右：$l = m = 100$；$l = 40$ 和 $m = 250$；$l = 250$ 和 $m = 40$。

6 结论

对于乘积域上的一个相当一般的优化问题，我们为块坐标 FW 变体提出了一套看似新颖的收敛理论。在广为接受的条件下，该理论既能保证目标值收敛，也能保证迭代点（局部）线性收敛。对于

μ-强凸目标而言，收敛是全局的，但我们主要关注非凸情形。若采用块的随机化选择，所有结果均针对期望，并且需要额外的技术假设。和通常一样，常数和收敛速率由梯度映射的利普希茨常数 L 、局部 Kurdyka-Lojasiewicz 条件中使用的常数 μ ，以及所谓角度条件中的参数 τ 来表示。

对于乘积域的一种特定结构以及合适选择的无投影策略（采用远离步来确定搜索方向的 FW 方法），这些结果还辅以一项活跃集识别结果：我们证明，该框架能在有限次迭代内识别解的支撑集。

据我们所知，这是首次针对（无需坏步的）块坐标 FW 变体同时给出线性收敛速率和活跃集识别结果，旨在缩小 [36] 中所观察到的研究空白。

在初步实验中，数值证据清楚地表明，利用结构信息的策略具有优势。在精心排除了简单实例的随机生成非凸 Multi-StQP 上，我们的方法（采用并行更新或随机更新的 AFW，二者均结合短步链策略 SSC）优于采用随机更新的块坐标 FW 方法。

我们采用两种实验设置——纯多起点和单调盆地跳跃——来检验所报告观察结果的稳健性。相同的效果似乎依然存在。

实例构造的灵感来自 StQP 的一种随机变体，其中域维数 l 和可能情景的数量 m 均有所变化。在 $l \leq m$ 的情况下，与并行变体相比，AFW 与随机更新和 SSC 的组合似乎略占优势。与 m 相比，当 l 较大时，这种效果似乎并不会出现，但这并不影响其相对于传统块坐标 FW 方法的优越性。

7 附录

7.1 证明

在 本节余下部分中，我们始终假设 SSC 会在有限步数内终止 ps
且 t 满足角度条件。
第 一个引理与单块设置有关，并加强了已证明的某些性质 d
对于 命题 4.1] 中的 SSC [39,。

引理 1 对于固定的 $i \in [1: \ m]$，令 $\{\mathbf{w}_k\} = \{\mathbf{x}_k^{(i)}\}$，并令 $\mathbf{w}_{k+1} = \mathrm{SSC}(\mathbf{w}_k, \ \mathbf{g})$。那么存在˜$\mathbf{w}_k \in \{\mathbf{y}_j\}_{j=0}^T$ 使得

$$\|\mathbf{w}_{k+1} - \mathbf{w}_k\| \geq \frac{\tau}{L} \|\pi(T_{\mathcal{C}_{(i)}}(\tilde{\mathbf{w}}_k), \mathbf{g})\| \tag{30}$$

and

$$\|\tilde{\mathbf{w}}_k - \mathbf{w}_k\| \leq \|\mathbf{w}_{k+1} - \mathbf{w}_k\|, \tag{31}$$

$$\langle \mathbf{g}, \tilde{\mathbf{w}}_k - \mathbf{w}_k \rangle \leq \langle \mathbf{g}, \mathbf{w}_{k+1} - \mathbf{w}_k \rangle . \tag{32}$$

此外，我们有

$$L\|\mathbf{y} - \mathbf{w}_k\|^2 \leq \langle \mathbf{g}, \mathbf{y} - \mathbf{w}_k \rangle \tag{33}$$

对于 $y \in \{\mathbf{w}^{k+1}, \ \tilde{}\mathbf{w}_k\}$。

证明 令 $B = B_{\|\mathbf{g}\|/2L}(\mathbf{x} + \mathbf{g}2L)$，并令 T满足$\mathbf{w}_{k+1} = \mathbf{y}_T$。根据 [39, (4.4)]，我们有 (33) 对每个$\mathbf{z} \in B$ （代替$\mathbf{y}$）都成立，因此如所需，对每个

$$\mathbf{y} \in \{\mathbf{w}_{k+1}, \tilde{\mathbf{w}}_k\} \subset \{\mathbf{y}_j : j \in [0{:}T]\} \subset \bar{B}\,.$$

现在令 $\tilde{p}_j = \|\pi(T_{\mathcal{C}_{(i)}}(\mathbf{y}_j),\ \mathbf{g})\|$。注意，若 $\tilde{\mathbf{w}}_k = \mathbf{y}_l$，则

$$\frac{\tau}{L}\,\|\pi(T_{\mathcal{C}_{(i)}}(\tilde{\mathbf{w}}_k), \mathbf{g})\| = \frac{\tau}{L}\,\tilde{p}_l \le \frac{1}{L}\langle \mathbf{g}, \hat{\mathbf{d}}_l\rangle\,, \tag{34}$$

其中，该不等式由 $\langle \mathbf{g},\ _l\rangle \ge \hat{d}\mathrm{DSB}_{\mathcal{A}}(\mathcal{C}_{(i)},\ \mathbf{y}_l,\ \mathbf{g}) \ge \mathrm{SB}_{\mathcal{A}}(\mathcal{C}_{(i)}) = \tau$ 得出。因此，为了证明 (30)，在证明的其余部分，只需证明

$$\langle \mathbf{g}, \hat{\mathbf{d}}_l\rangle \le L\|\mathbf{w}_{k+1} - \mathbf{w}_k\|\,. \tag{35}$$

此外，根据 SSC 的定义，标量积 $\langle \mathbf{g},\ \mathbf{y}_j\rangle$ 随 j 单调递增，因此我们有

$$\langle \mathbf{g}, \tilde{\mathbf{w}}_k - \mathbf{w}_k\rangle = \langle \mathbf{g}, \mathbf{y}_l - \mathbf{w}_k\rangle \le \langle \mathbf{g}, \mathbf{y}_T - \mathbf{w}_k\rangle = \langle \mathbf{g}, \mathbf{w}_{k+1} - \mathbf{w}_k\rangle\,.$$

根据 SSC 的终止方式，我们区分四种情况。在前两种情况下，我们证明可以选择最后一步 $\tilde{\mathbf{w}} = \mathbf{y}_T$；在第三种情况下，倒数第二个选择 $\tilde{\mathbf{w}} = \mathbf{y}_{T-1}$ 满足所有条件；而在第四种情况下，某个中间步骤是合适的选择。我们将$B_j = B_{\langle \mathbf{g},\ \hat{\mathbf{d}}_j\rangle/L}(\mathbf{w}_k)$记作缩写。

情形 1： $T = 0$ 或 $\mathbf{d}_T = \mathbf{o}$。由于不存在下降方向，$\mathbf{w}_{k+1} = \mathbf{y}_T$ 必定是梯度 $-\mathbf{g}$ 的驻点。等价地，$\tilde{p}_T = \|\pi(T_{\mathcal{C}_{(i)}}(\mathbf{w}_{k+1}), \mathbf{g})\| = 0$。最后，显然，若 $T = 0$，则$\mathbf{d}_0 = \mathbf{o}$，因为$\mathbf{y}_0$ 必定是 $-\mathbf{g}$ 的驻点。因此，取 $\tilde{\mathbf{w}}_k = \mathbf{y}_T$ 即可得到所需性质。

在考察其余情形之前，我们指出，若 SSC 在第二阶段终止，则$\alpha_{T-1} = \beta_{T-1}$必须关于条件$\mathbf{y}_T \in B_{T-1}$ 或$\mathbf{y}_T \in B$达到极大。若 $\alpha_{T-1} = 0$，则$\mathbf{y}_{T-1} = \mathbf{y}_T$，且在此情形下不可能有$\mathbf{y}_{T-1} \in \partial B$，否则 SSC 会在前一轮的第二阶段终止。因此，必有$\mathbf{y}_T = \mathbf{y}_{T-1} \in \mathrm{int}(B_{T-1})^c$ （情形 2）。若 $\beta_{T-1} = \alpha_{T-1} > 0$，则必有$\mathbf{y}_{T-1} \in \mathcal{C}_{T-1} = B_{T-1} \cap B$，且$\mathbf{y}_T \in \partial B_{T-1}$（情形 3）或$\mathbf{y}_T \in \partial B$（情形 4）。

情形 2： $\mathbf{y}_{T-1} = \mathbf{y}_T \in \mathrm{int}(B_{T-1})^c$。我们可以将条件改写为

$$\langle \mathbf{g}, \hat{\mathbf{d}}_{T-1}\rangle \le L\|\mathbf{y}_{T-1} - \mathbf{w}_k\| = L\|\mathbf{y}_T - \mathbf{w}_k\|\,, \tag{36}$$

这正是 (35)。因此，$\tilde{\mathbf{w}}_k = \mathbf{w}_{k+1} = \mathbf{y}_T$ 满足所需条件。**情形 3：** $\mathbf{y}_T = \mathbf{y}_{T-1} + \beta_{T-1}\mathbf{d}_{T-1}$且 $\mathbf{y}_T \in \partial B_{T-1}$。于是由 $\mathbf{y}_{T-1} \in B_{T-1}$可得

$$L\|\mathbf{y}_{T-1} - \mathbf{w}_k\| \le \langle \mathbf{g}, \hat{\mathbf{d}}_{T-1}\rangle\,, \tag{37}$$

且 $\mathbf{y}_T \in \partial B_{T-1}$ 意味着

$$\langle \mathbf{g}, \hat{\mathbf{d}}_{T-1}\rangle = L\|\mathbf{y}_T - \mathbf{w}_k\|\,, \tag{38}$$

这就是对 $l = T-1$ 而言的 (35)。结合 (37) 和 (38)，我们还可以得到

$$L\|\mathbf{y}_{T-1} - \mathbf{w}_k\| \le L\|\mathbf{y}_T - \mathbf{w}_k\|\,, \tag{39}$$

因此，特别地，我们可以取 $\tilde{\mathbf{w}}_k = \mathbf{y}_{T-1}$。

情形 4： $\mathbf{y}_T = \mathbf{y}_{T-1} + \beta_{T-1}\mathbf{d}_{T-1}$ 且 $\mathbf{y}_T \in \partial B$。

条件 $\mathbf{w}_{k+1} = \mathbf{y}_T \in \partial B$ 可改写为

$$L\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 - \langle \mathbf{g}, \mathbf{w}_{k+1} - \mathbf{w}_k\rangle = 0\,. \tag{40}$$

对于每个 $j \in [0\colon T]$，我们有

$$\mathbf{w}_{k+1} = \mathbf{y}_j + \sum_{i=j}^{T-1} \alpha_i \mathbf{d}_i \,. \tag{41}$$

现在我们要证明，对于每个 $j \in [0\colon T]$

$$\|\mathbf{w}_{k+1} - \mathbf{w}_k\| \geq \|\mathbf{y}_j - \mathbf{w}_k\| \,. \tag{42}$$

事实上，我们有

$$\begin{aligned} L\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 = \langle \mathbf{g}, \mathbf{w}_{k+1} - \mathbf{w}_k \rangle &= \langle \mathbf{g}, \mathbf{y}_j - \mathbf{w}_k \rangle + \sum_{i=j}^{T-1} \alpha_i \langle \mathbf{g}, \mathbf{d}_i \rangle \\ &\geq \langle \mathbf{g}, \mathbf{y}_j - \mathbf{w}_k \rangle \geq L\|\mathbf{y}_j - \mathbf{w}_k\|^2 \,, \end{aligned}$$

其中，第一个等式用到了 (40)，第二个等式用到了 (41)，第一个不等式对每个 j 都用到了 $\langle g, \mathbf{d}_j \rangle \geq 0$，第二个等式用到了 $\mathbf{y}_j \in B$，由此证明了 (42)。我们还有

$$\frac{\langle \mathbf{g}, \mathbf{w}_{k+1} - \mathbf{w}_k \rangle}{\|\mathbf{w}_{k+1} - \mathbf{w}_k\|} = \frac{\langle \mathbf{g}, \sum_{j=0}^{T-1} \alpha_j \mathbf{d}_j \rangle}{\|\sum_{j=0}^{T-1} \alpha_j \mathbf{d}_j\|} \geq \frac{\langle \mathbf{g}, \sum_{j=0}^{T-1} \alpha_j \mathbf{d}_j \rangle}{\sum_{j=0}^{T-1} \alpha_j \|\mathbf{d}_j\|} \geq \min \left\{ \frac{\langle \mathbf{g}, \mathbf{d}_j \rangle}{\|\mathbf{d}_j\|} : j \in [0\colon T-1] \right\} . \tag{43}$$

因此，对于 $\tilde{T} \in \operatorname{argmin} \left\{ \frac{\langle \mathbf{g}, \mathbf{d}_j \rangle}{\|\mathbf{d}_j\|} : j \in [0\colon T-1] \right\}$，我们有

$$\langle g, \hat{\mathbf{d}}_{\tilde{T}} \rangle \leq \frac{\langle \mathbf{g}, \mathbf{w}_{k+1} - \mathbf{w}_k \rangle}{\|\mathbf{w}_{k+1} - \mathbf{w}_k\|} = L\|\mathbf{w}_{k+1} - \mathbf{w}_k\| \,, \tag{44}$$

其中，第一个不等式用到了 (43)，第二个（等式）用到了 (40)。

特别地，$\tilde{\mathbf{w}}_k = \mathbf{y}_{\tilde{T}}$ 满足所需性质，其中 $\|\tilde{\mathbf{w}}_k - \mathbf{w}_k\| \leq \|\mathbf{w}_{k+1} - \mathbf{w}_k\|$ 由 (44) 得出，而 (35) 由 (44) 得出。(42)

我们用 $\mathrm{SSC}(\overline{\mathbf{w}_k, \mathbf{g}})$ 表示具有上述引理所述性质的点 $\tilde{\mathbf{w}}_k$。定义 U_0 为 $\{\mathbf{x} \in \mathcal{C}$ 的连通分支也很有用：$f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$，其中包含 $\mathbf{x}_0$。接下来的结果说明，在我们的块坐标设置中，定理 1 对 U_0 的假设使我们能够为 SSC 生成的点得出目标函数的下界。该下界类似于 [39, 定理 4.2] 中所要求的下界。

引理 2 设 $\{x_k\}$ 是由算法 1 生成的序列。设 $\bar{x}_k = [SSC(x_k^{(i)}, -\nabla f(x_k)^{(i)})]_{i=1}^m$。另假设 $f(x_*)$ 是 U_0 中的一个极小值。那么，$\{f(x_k)\}$ 是递减的，并且对每个 k，$\{x_k, \bar{x}_k\} \subset U_0$，均有 $f(y) \in [f(x_*)\ f(x_0)]$，其中 $y \in \{x_k, \bar{x}_k\}$。

证明 令 U_k 为包含 $\mathbf{x}_k$ 的 $\{\mathbf{x} \in \mathcal{C}\colon f(\mathbf{x}) \leq f(\mathbf{x}_k)\}$ 的最小连通分支，令 $\mathbf{g} = -\nabla f(\mathbf{x}_k)$，并令 $\bar{B}_k^{\mathcal{C}} = \mathcal{C} \cap \prod_i \bar{B}_{\|\mathbf{g}^{(i)}\|}(\mathbf{x}_k^{(i)} + \mathbf{g}_{2L}^{(i)})$。对于 $\mathbf{y} \in \bar{B}_k^{\mathcal{C}}$，我们有 $\mathbf{x}_k \in \bar{B}_k^{\mathcal{C}}$ 和 $\frac{}{2L}$

$$\begin{aligned} f(\mathbf{y}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{y} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_k\|^2 \\ &= f(\mathbf{x}_k) + \sum_{i=1}^m \langle \nabla f(\mathbf{x}_k)^{(i)}, \mathbf{y}^{(i)} - \mathbf{x}_k^{(i)} \rangle + \frac{L}{2}\|\mathbf{y}^{(i)} - \mathbf{x}_k^{(i)}\|^2 \leq f(\mathbf{x}_k) - \frac{L}{2}\sum_{i=1}^m \|\mathbf{y}^{(i)} - \mathbf{x}_k^{(i)}\|^2 \\ &= f(\mathbf{x}_k) - \frac{L}{2}\|\mathbf{x}_k - \mathbf{y}\|^2 \end{aligned} \tag{45}$$

其中第一个不等式用到了标准的下降引理，第二个不等式则由 $\bar{B}_k^{\mathcal{C}}$ 的定义得出。由 (45) 可知，$\{f(\mathrm{x}_k)\}$ 是递减的，并且 $\bar{B}_k^{\mathcal{C}} \subset \{\mathrm{x} \in \mathcal{C}\colon f(\mathrm{x}) \leq f(\mathrm{x}_k)\}$。此外，由于 $\bar{B}_k^{\mathcal{C}}$ 连通且包含 x_k，更强的包含关系 $\bar{B}_k^{\mathcal{C}} \subset U_k$ 也成立。因此 $\{\mathrm{x}^{k+1}, \bar{\mathrm{x}}_k\} \subset \bar{B}_k^{\mathcal{C}} \subset U_k$，特别是，由于 $f(\mathrm{x}^{k+1}) \leq f(\mathrm{x}_k)$，有 $U_{k+1} \subset U_k$；再由归纳法可得 $\{\mathrm{x}^{k+1}, \bar{\mathrm{x}}_k\} \subset U_0$。最后，对于 $\mathrm{y} \in \{\mathrm{x}^{k+1}, \bar{\mathrm{x}}_k\}$，有 $f(\mathrm{y}) \in [f(\mathrm{x}_*)\ f(\mathrm{x}_k)]$，其中下界来自假设 $f(\mathrm{x}_*)$ 是 U_0 中的最小值，而上界则由 (45) 得出。□

在下述引理中，我们将引理 1 中针对单个块证明的 SSC 性质结合起来，得到其在整个块的乘积上的类似性质，然后利用 KL 条件，以目标函数的最优性间隙为合适的改进度量给出下界。我们想要强调的是，与单块情形不同，此处的最优性间隙是相对于一个辅助点来衡量的，而该点不一定是算法生成的点。因此，在这种情形下证明线性收敛速率需要妥善处理。

引理 *3* 设 $\{x_k\}$ 是由算法 *1* 生成的序列，并假设对于所有 $i \in [1\colon m]$，方法 $\mathcal{A}^{(i)}$ 满足角度条件，且具有相同的 τ。设 $\bar{x}_k = [SSC(x_k^{(i)}, -\nabla f(x_k)^{(i)})]_{i=1}^m$，且 $\tilde{x}_{k+1} = [\overline{SSC}(x_k^{(i)}, -\nabla f(x_k)^{(i)})]_{i=1}^m$。若 *(14)* 在 $\bar{x}_k$ 处成立，则记 $g = -\nabla f(x_k)$ 为：

$$\|\tilde{\mathbf{x}}_{k+1} - \mathbf{x}_k\|^2 \geq \frac{\tau^2}{2(1+\tau^2)L^2}\|\pi(T_{\mathcal{C}}(\bar{\mathbf{x}}_k), -\nabla f(\bar{\mathbf{x}}_k))\|^2 \geq \frac{\tau^2\mu}{L^2(1+\tau^2)}[f(\bar{\mathbf{x}}_k) - f^*]\,, \tag{46}$$

$$\frac{1}{2}\langle \mathbf{g}, \tilde{\mathbf{x}}_{k+1} - \mathbf{x}_k\rangle \geq \frac{1}{3}[f(\mathbf{x}_k) - f(\bar{\mathbf{x}}_k)]\,. \tag{47}$$

证明 令 $\bar{\mathbf{g}} = -\nabla f(\bar{\mathbf{x}}_k)$、$\bar{q}_{(i)} = \|\pi(T_{\mathcal{C}_{(i)}}(\bar{\mathbf{x}}_k^{(i)}) - \bar{\mathbf{g}}^{(i)})\|$ 以及 $q_{(i)} = \|\pi(T_{\mathcal{C}_{(i)}}(\bar{\mathbf{x}}_k^{(i)}), \mathbf{g}^{(i)})\|$。注意到，根据梯度的利普希茨连续性，我们有不等式

$$\bar{q}_{(i)} \leq q_{(i)} + L\|\bar{\mathbf{x}}_k^{(i)} - \mathbf{x}_k^{(i)}\| \tag{48}$$

因此

$$\bar{q}_{(i)}^2 \leq 2q_{(i)}^2 + 2L^2\|\bar{\mathbf{x}}_k^{(i)} - \mathbf{x}_k^{(i)}\|^2 \leq \frac{2L^2(1+\tau^2)}{\tau^2}\|\tilde{\mathbf{x}}_{k+1}^{(i)} - \mathbf{x}_k^{(i)}\|^2\,, \tag{49}$$

其中，在第一个不等式中我们对（48）应用了詹森不等式，在第二个不等式中应用了（30）和（31）。

因此我们可以写成

$$\begin{aligned}\|\tilde{\mathbf{x}}_{k+1} - \mathbf{x}_k\|^2 &= \sum_{i=1}^m \|\tilde{\mathbf{x}}_{k+1}^{(i)} - \mathbf{x}_k^{(i)}\|^2 \geq \frac{\tau^2}{2L^2(1+\tau^2)}\sum_{i=1}^m \bar{q}_{(i)}^2 \\ &= \frac{\tau^2}{2L^2(1+\tau^2)}\|\pi(T_{\mathcal{C}}(\bar{\mathbf{x}}_k), -\nabla f(\bar{\mathbf{x}}_k))\|^2 \geq \frac{\tau^2\mu}{L^2(1+\tau^2)}[f(\bar{\mathbf{x}}_k) - f^*]\,,\end{aligned} \tag{50}$$

其中，在第一个不等式中我们使用了（49），在第二个不等式中使用了KL性质。这就证明了（46）。利用标准的下降引理，我们可以给出上界

$$f(\mathbf{x}_k) - f(\bar{\mathbf{x}}_k) \leq \langle \mathbf{g}, \bar{\mathbf{x}}_k - \mathbf{x}_k\rangle + \tfrac{L}{2}\|\bar{\mathbf{x}}_k - \mathbf{x}_k\|^2 \leq \tfrac{3}{2}\langle \mathbf{g}, \bar{\mathbf{x}}_k - \mathbf{x}_k\rangle = \tfrac{3}{2}\sum_{i=1}^m \langle \mathbf{g}^{(i)}, \bar{\mathbf{x}}_k^{(i)} - \mathbf{x}_k^{(i)}\rangle\,, \tag{51}$$

其中，在第二个不等式中我们使用了（33）。最后，我们可以证明（47）：

$$\tfrac{1}{2}\langle \mathbf{g}, \tilde{\mathbf{x}}_{k+1} - \mathbf{x}_k\rangle \geq \tfrac{1}{2}\langle \mathbf{g}, \bar{\mathbf{x}}_k - \mathbf{x}_k\rangle = \tfrac{1}{2}\sum_{i=1}^m \langle \mathbf{g}^{(i)}, \bar{\mathbf{x}}_k^{(i)} - \mathbf{x}_k^{(i)}\rangle \geq \tfrac{1}{3}[f(\mathbf{x}_k) - f(\bar{\mathbf{x}}_k)]\,, \tag{52}$$

其中，第一个不等式中使用了（32），第二个不等式中使用了（51）。 □

下一个结果直接由前述引理推出，明确给出了利用上文引入的最优性间隙对目标函数改进的下界。

引理 4 设 $\{\mathbf{x}_k\}$ 是由算法 *1* 生成的序列，并假设对于所有 $i \in [1\colon m]$，方法 $\mathcal{A}^{(i)}$ 在相同的 τ 下满足角度条件。设 $\bar{\mathbf{x}}_k = (\mathrm{SSC}(\mathbf{x}_k^{(i)}, -\nabla f(\mathbf{x}_k)^{(i)}))_{i=1}^m$。那么，如果 *KL* 性质(14) 在 $\bar{\mathbf{x}}_k$ 处成立，则对于并行更新

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{\tau^2 \mu}{2L(1+\tau^2)} \left(f(\bar{\mathbf{x}}_k) - f^*\right), \tag{53}$$

对于 *GS* 更新

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{\tau^2 \mu}{2L(1+\tau^2)} \frac{1}{m} \left(f(\bar{\mathbf{x}}_k) - f^*\right), \tag{54}$$

对于随机更新

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})] \geq \frac{\tau^2 \mu}{2L(1+\tau^2)} \frac{1}{m} \mathbb{E}[f(\bar{\mathbf{x}}_k) - f^*]. \tag{55}$$

证明 我们先证明并行更新时的不等式。我们有

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) &\geq \frac{L}{2} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \geq \frac{L}{2} \frac{\tau^2}{2L^2(1+\tau^2)} \|\pi(T_{\mathcal{C}}(\bar{\mathbf{x}}_k), -\nabla f(\bar{\mathbf{x}}_k))\|^2 \\ &= \frac{\tau^2}{4L(1+\tau^2)} \|\pi(T_{\mathcal{C}}(\bar{\mathbf{x}}_k), -\nabla f(\bar{\mathbf{x}}_k))\|^2 \ \geq \frac{\tau^2 \mu}{2L^2(1+\tau^2)} \left[f(\bar{\mathbf{x}}_k) - f^*\right], \end{aligned} \tag{56}$$

其中，第一个不等式由 (45) 得出，第二个不等式由 (46) 得出。采用引理 3 中引入的记号，根据定义有 $\mathbf{x}_{k+1} = \tilde{\mathbf{x}}^{k+1}$。对于 GS 更新，我们有

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) &\geq \langle g, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle - \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \geq \frac{1}{2} \langle \mathbf{g}, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle \\ &= \frac{1}{2} \max_{i \in [1:m]} \langle \mathbf{g}^{(i)}, \tilde{\mathbf{x}}_{k+1}^{(i)} - \mathbf{x}_k \rangle \geq \frac{1}{2m} \langle \mathbf{g}, \tilde{\mathbf{x}}_{k+1} - \mathbf{x}_k \rangle \geq \frac{L}{2m} \|\tilde{\mathbf{x}}_{k+1} - \mathbf{x}_k\|^2 \\ &\geq \frac{\tau^2 \mu}{2mL(1+\tau^2)} \left[f(\bar{\mathbf{x}}_k) - f^*\right], \end{aligned} \tag{57}$$

其中，第一个不等式使用了标准的下降引理，第二个不等式使用了（33）；等式由 GS 更新的定义得到，第四个不等式再次应用了（33），最后一个不等式则应用了（46）。

最后，对于随机更新，记 $i(k) = j$ 为在步骤 k 选择的索引是 j 这一事件:

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})] &\geq \frac{L}{2} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2] = \frac{L}{2} \sum_{j=1}^m \mathbb{P}(i(k) = j) \mathbb{E}[\|\tilde{\mathbf{x}}_{k+1}^{(j)} - \mathbf{x}_k^{(j)}\|^2] \\ &= \frac{L}{2m} \sum_{j=1}^m \mathbb{E}[\|\tilde{\mathbf{x}}_{k+1}^{(j)} - \mathbf{x}_k^{(j)}\|^2] = \frac{L}{2m} \mathbb{E}[\|\tilde{\mathbf{x}}_{k+1} - \mathbf{x}_k\|^2] \geq \frac{\tau^2 \mu}{2L(1+\tau^2)} \mathbb{E}[f(\bar{\mathbf{x}}_k) - f^*], \end{aligned} \tag{58}$$

其中，第一个不等式由（45）得出；第二个等式中，我们使用了 $\mathbb{P}(\{i(k) = j\}) = \frac{1}{m}$，最后一个不等式则由（46）得出。□

在接下来的两个引理中，我们将相对于辅助点衡量的改进与目标函数的实际改进联系起来，从而将[39,引理 4.3] 中的线性收敛速率推广到块坐标设置。

引理 5 设 $\{\mathbf{x}_k\}$ 是由算法 *1* 生成的序列，并假设对于所有 $i \in [1\!:\!m]$，方法 $\mathcal{A}^{(i)}$ 满足角度条件，且采用相同的 τ。设 $\bar{\ }\mathbf{x}_k = (\mathrm{SSC}(\mathbf{x}_k^{(i)}, -\nabla f(\mathbf{x}_k)^{(i)}))_{i=1}^m$。那么，对于并行更新

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{3}[f(\mathbf{x}_k) - f(\bar{\mathbf{x}}_k)]\,, \tag{59}$$

对于 *GS* 更新

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{3m}[f(\mathbf{x}_k) - f(\bar{\mathbf{x}}_k)]\,, \tag{60}$$

对于随机更新

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})] \geq \frac{1}{3m}\mathbb{E}[f(\mathbf{x}_k) - f(\bar{\mathbf{x}}_k)] \tag{61}$$

证明 对于并行更新，我们有

$$\begin{aligned} &f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \langle \mathbf{g}, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle - \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \geq \frac{1}{2}\langle \mathbf{g}, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle \\ &= \frac{1}{2}\langle \mathbf{g}, \tilde{\mathbf{x}}_{k+1} - \mathbf{x}_k\rangle \geq \frac{1}{3}[f(\mathbf{x}_k) - f(\bar{\mathbf{x}}_k)]\,, \end{aligned} \tag{62}$$

其中，第一处不等式使用了标准下降引理，第二处不等式使用了 (33)，最后一处不等式使用了 (47)。

注意到.....之后，GS 更新的证明也可类似地得出

$$\langle \mathbf{g}, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle \geq \frac{1}{m}\langle \mathbf{g}, \tilde{\mathbf{x}}_{k+1} - \mathbf{x}_k\rangle\,, \tag{63}$$

如 (57) 所示；对于随机更新，则使用

$$\mathbb{E}[\langle \mathbf{g}, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle] = \frac{1}{m}\mathbb{E}[\langle \mathbf{g}, \tilde{\mathbf{x}}_{k+1} - \mathbf{x}_k\rangle]\,, \tag{64}$$

分别。 □

引理 6 设 $\{\mathbf{x}_k\}$ 是由算法 *1* 生成的序列，并假设对于所有 $i \in [1\!:\!m]$，方法 $\mathcal{A}^{(i)}$ 满足角度条件，且具有相同的 τ。那么，若 *KL*性质 (14) 在$\mathbf{x}_k$处成立，对于并行更新

$$f(\mathbf{x}_{k+1}) - f^* \leq q_P(f(\mathbf{x}_k) - f^*)\,, \tag{65}$$

对于 *GS* 更新

$$f(\mathbf{x}_{k+1}) - f^* \leq q_{GS}(f(\mathbf{x}_k) - f^*)\,, \tag{66}$$

以及对于随机更新

$$\mathbb{E}[f(\mathbf{x}_{k+1}) - f^*] \leq q_R\mathbb{E}[f(\mathbf{x}_k) - f^*]\,. \tag{67}$$

证明 首先注意，由于 $\tau \in [0,1]$ 且 $\mu \le L$，因此我们有

$$\frac{\tau^2\mu}{2L(1+\tau^2)} \le \frac{\mu}{3L} \le \frac{1}{3}. \tag{68}$$

然后，结合 (53) 和 (59)，可得

$$\begin{aligned} 2[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})] &\ge \frac{1}{3}[f(\mathbf{x}_k) - f(\bar{\mathbf{x}}_k)] + \frac{\tau^2\mu}{2L(1+\tau^2)}[f(\bar{\mathbf{x}}_k) - f^*] \\ &\ge \frac{\tau^2\mu}{2L(1+\tau^2)}[f(\mathbf{x}_k) - f(\bar{\mathbf{x}}_k)] + \frac{\tau^2\mu}{2L(1+\tau^2)}[f(\bar{\mathbf{x}}_k) - f^*] \\ &= \frac{\tau^2\mu}{2L(1+\tau^2)}[f(\mathbf{x}_k) - f^*)], \end{aligned} \tag{69}$$

重新整理后

$$f(\mathbf{x}_{k+1}) - f^* \le q_P[f(\mathbf{x}_k) - f^*]. \tag{70}$$

GS 和随机更新的待证命题也可类似地得出。 □

根据块坐标设置下的先前结果，证明的剩余部分只需直接改编 [39, 定理 4.2] 证明中使用的论证。

（定理 *1* 的证明）我们需要证明 KL 性质 (14) 在 $\{\bar{}\mathbf{x}_k\}$ 中成立。由引理 6 通过归纳法可立即得到关于 $f(\mathbf{x}_k) - f(\mathbf{x}_*)$ 的界，进而，关于 $\|\mathbf{x}_k - \mathbf{x}_*\|$ 的界也可按照 [39, 引理 4.3] 的证明得出。

对于随机更新，我们取 $\tilde{}f(\mathbf{x}_0) < \tilde{}f(\mathbf{x}_*) + \eta$。由构造可知，KL性质 (14) 在 U_0 中成立；又因为根据引理 2，$\{\bar{}\mathbf{x}_k\}$ 包含于 U_0 中，所以 (14) 特别在 $\{\bar{}\mathbf{x}_k\}$ 中成立。

对于并行更新，根据引理 2，我们有 $\{f(\mathbf{x}_k)\}$ 是递减的，并且 $f(\bar{}\mathbf{x}_k) \ge f(\mathbf{x}_k) \ge f(\mathbf{x}_*)$。随后可以用类似于定理 4.2] 证明的论证证明：当 δ 足够小时，(14) 在 $\tilde{\{\bar{}_k\}}$ 中成立。为完整起见，我们在此给出该论证。令 $f(\mathbf{x}_k) - f(\mathbf{x}_*)$，并令 $\delta < \delta/2$ 按照定理 4.2] 的证明中所述定义，使得

$$\tilde{\delta} < \frac{\delta}{2} < \delta - \frac{\sqrt{2f_0(1-q)}}{L(1-\sqrt{q})} - \sqrt{\frac{2}{L}}\sqrt{f_0}, \tag{71}$$

这里的 $q = q_P$。现在我们要证明，对于每个 $k \in \mathbb{N}$，通过对 k 进行归纳法可得 $\bigcup_{[0:\ k-1]}\{\mathbf{x}_i, \bar{}\mathbf{x}_i\} \cup \{\mathbf{x}_k\} \subset B_\delta(\mathbf{x}_*)$。注意，根据构造，$\mathbf{x}_0 \in B_\delta(\mathbf{x}_*)$。首先来看归纳步骤，

$$\sum_{i=0}^{k-1}\|\mathbf{x}_i - \mathbf{x}_{i+1}\| \le \sqrt{\frac{2}{L}}\sum_{i=0}^{k-1}\sqrt{f_i - f_{i+1}} \le \frac{\sqrt{2f_0(1-q)}}{\sqrt{L}(1-\sqrt{q})} \tag{72}$$

其中，第一个不等式中用到了 (45)，而第二个不等式可仿照 [39, 定理 4.2]的证明，由 [39, 引理 8.1] 推出。但接下来

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}_*\| &\le \|\mathbf{x}_0 - \mathbf{x}_*\| + \left(\sum_{i=0}^{k-1}\|\mathbf{x}_i - \mathbf{x}_{i+1}\|\right) + \|\mathbf{x}_k - \mathbf{x}_{k+1}\| \\ &\le \tilde{\delta} + \frac{\sqrt{2f_0(1-q)}}{L(1-\sqrt{q})} + \sqrt{\frac{2}{L}}\sqrt{f_k - f_{k+1}} \\ &< \tilde{\delta} + \frac{\sqrt{2f_0(1-q)}}{L(1-\sqrt{q})} + \sqrt{\frac{2}{L}}\sqrt{f_k} < \delta, \end{aligned} \tag{73}$$

其中，第二个不等式中结合使用了 (72) 和 (45)，第三个不等式中使用了 $f_{k+1} = f(\mathbf{x}^{k+1}) - f(\mathbf{x}_*) \geq 0$，而最后一个不等式中结合使用了 (71) 和 $f_0 \geq f_k$。现在我们有

$$\begin{aligned} \|\tilde{\mathbf{x}}_k - \mathbf{x}_*\| &\leq \|\mathbf{x}_0 - \mathbf{x}_*\| + \left(\sum_{i=0}^{k-1} \|\mathbf{x}_i - \mathbf{x}_{i+1}\|\right) + \|\mathbf{x}_k - \tilde{\mathbf{x}}_k\| \\ &\leq \|\mathbf{x}_0 - \mathbf{x}_*\| + \left(\sum_{i=0}^{k-1} \|\mathbf{x}_i - \mathbf{x}_{i+1}\|\right) + \|\mathbf{x}_k - \mathbf{x}_{k+1}\| < \delta\,, \end{aligned} \tag{74}$$

其中，第二个不等式中我们使用了 $\|\tilde{\mathbf{x}}_k - \mathbf{x}_k\| \leq \|\mathbf{x}^{k+1} - \mathbf{x}_k\|$，最后一个不等式则可由 (73) 同理推出。因此，$\tilde{\mathbf{x}}_k \in B_\delta(\mathbf{x}_*)$ 也成立，归纳法至此完成。对于 GS 更新，证明 $\{\tilde{\mathbf{x}}_k\} \subset B_\delta(\mathbf{x}_*)$ 的方法类似。□

7.2 活跃集识别准则

本节证明定理 2，并针对可行集 $\mathcal{C}$ 为单纯形乘积的特殊情形，为算法 1 提出一个通用的活跃集识别准则。采用第 4 节引入的记号，令 $\mathcal{C}_* = \{\mathbf{x} \in \mathcal{C} : \mathrm{supp}(\mathbf{x}) = \mathrm{supp}(\mathbf{x}_*)\}$，且令 $S_* = \{\mathbf{x} \in \mathbb{R}^n : \mathrm{supp}(\mathbf{x}) = \mathrm{supp}(\mathbf{x}_*)\}$，分别表示 $\mathcal{C}$ 中与 $\mathbf{x}_*$ 具有相同支撑集的点的子集和方向子空间。

定义 4 我们称方法 A 在 $\mathbf{x}_*$ 处具有与活跃集相关的方向，V 如果它可以执行有限步连续最大步，并且对于 $\mathbf{x}_*$ 的某个邻域，$\mathbf{x} \to \mathbf{x}_*$、$\mathbf{g} \to -\nabla f(\mathbf{x}_*)$ 和 $\mathbf{d} \in A(\mathbf{x},\mathbf{g})$ 满足：

- 若 $\mathbf{x} \in \mathcal{C}_*$，则 $\mathbf{d} \in S_*$ 且 $\alpha_{\max}(\mathbf{x},\mathbf{d}) = \Theta(1)$，
- 若 $\mathbf{x} \in \mathcal{C} \setminus \mathcal{C}_*$，则 $\langle \mathbf{g},\mathbf{d}\rangle = \Theta(1)$。

引理 7 在定义 *4* 的假设下：

- 若 $\mathbf{x} \in \mathcal{C} \setminus \mathcal{C}_*$，则有 $\alpha_{\max}(\mathbf{x},\mathbf{d}) = o(1)$，
- 若 $\mathbf{x} \in \mathcal{C}_*$，则 $\langle \mathbf{g},\mathbf{d}\rangle = o(1)$。

证明 注意到

$$\begin{aligned} 0 \leq \alpha_{\max}(\mathbf{x},\hat{\mathbf{d}})\langle \mathbf{g},\hat{\mathbf{d}}\rangle &= \langle \mathbf{g}, (\mathbf{x} + \alpha_{\max}(\mathbf{x},\hat{\mathbf{d}})\hat{\mathbf{d}}) - \mathbf{x}\rangle \\ &= \langle -\nabla f(\mathbf{x}), (\mathbf{x} + \alpha_{\max}(\mathbf{x},\hat{\mathbf{d}})\hat{\mathbf{d}}) - \mathbf{x}\rangle + o(1) \leq \langle -\nabla f(\mathbf{x}), \mathbf{x}_* - \mathbf{x}\rangle + o(1) = o(1)\,. \end{aligned} \tag{75}$$

Thus

$$\alpha_{\max}(\mathbf{x},\hat{\mathbf{d}}) \leq \frac{o(1)}{\langle \mathbf{g},\hat{\mathbf{d}}\rangle} = o(1)\,, \tag{76}$$

其中，在等式中我们根据假设使用了 $\langle \mathbf{g},\mathbf{d}\rangle = \Theta(1)$。这证明了该结论的第一部分。至于第二部分，我们有

$$\langle \mathbf{g},\hat{\mathbf{d}}\rangle = \langle -\nabla f(\mathbf{x}_*),\hat{\mathbf{d}}\rangle + o(1) = o(1)\,, \tag{77}$$

其中，在第二个等式中我们使用了 $\langle -\nabla f(\mathbf{x}_*),\ \mathbf{d}\rangle = 0$，这由平稳性条件保证，因为 $\mathbf{d} \in S_*$。□

命题1 设将算法 *1* 应用于一种在$\mathbf{x}_*$ 处具有与活跃集相关的方向的方法，如定义 *4* 所述。那么存在$\mathbf{x}_*$ 的一个邻域 U，使得若$\mathbf{x}_k \in U$，则 $supp(\mathbf{x}^{k+1}) = supp(\mathbf{x}_*)$。

证明：令 $\{\mathbf{y}_i : i \in [0:j]\}$ 为由 SSC$(\mathbf{x}_k, -\nabla f(\mathbf{x}_k))$ 生成的点集，T为连续最大步数的上界，使得 $T \le T+1$，并令 B 和 B_j 如引理 1 的证明所示。不失一般性，假设对于 $j \in [0:T]$，有 $\|\mathbf{d}_j\| = 1$。

我们将证明，当 $\mathbf{x}_k$ 充分接近 $\mathbf{x}_*$ 时，某些不等式（即 (79)、(80) 和 (83)）成立，从而可以推出识别性质。令

$$T^* = \max\{j \in [0:T] : \{\mathbf{y}_i : i \in [0:j]\} \subset \mathcal{C} \setminus \mathcal{C}_*\}$$

当 $\mathbf{y}_0 \notin \mathcal{C}_*$时，否则为 $T^* = -1$。我们首先断言 $T \ge T^* + 1$。根据 T^*的定义，当 $T^* = -1$时这一点显然成立。否则，对于 $j \in [0:T^*]$，令 $\tilde{\mathbf{y}}^{j+1} = \mathbf{y}_j + \alpha_{\max}^{(j)} \mathbf{d}_j$。现在证明 $\tilde{\mathbf{y}}_{j+1} \in \mathcal{C}_j$。首先，检查 $\tilde{\mathbf{y}}_{j+1} \in$int$(B)$。一方面，我们有

$$\begin{aligned} L\|\tilde{\mathbf{y}}_{j+1} - \mathbf{y}_0\|^2 &= L\|\alpha_{\max}^{(j)} \mathbf{d}_j + \sum_{i=0}^{j-1} \alpha_i \mathbf{d}_i\|^2 \le L\left(\sum_{i=0}^{j-1} \alpha_{\max}^{(i)} \|\mathbf{d}_i\|\right)^2 \\ &= L\left(\sum_{i=0}^{j-1} \alpha_{\max}^{(i)}\right)^2 = O(\max_{i\in[0:j]} (\alpha_{\max}^{(i)})^2)\,, \end{aligned} \tag{78}$$

其中，在第二个等式中，我们根据假设使用了 $\|\mathbf{d}_i\| = 1$。另一方面

$$\langle \mathbf{g}, \tilde{\mathbf{y}}_{j+1} - \mathbf{y}_0 \rangle = \alpha_{\max}^{(j)} \mathbf{d}_j + \sum_{i=0}^{j-1} \alpha_i \langle \mathbf{g}_i, \mathbf{d}_i \rangle = O(\max_{i\in[0:j]} \alpha_{\max}^{(i)})\,. \tag{79}$$

由于 $\mathbf{y}_{(i)} \in \mathcal{C} \setminus \mathcal{C}_*$，根据与活跃集相关的性质 $\alpha_{\max}^{(i)} = o(1)$，对于$\mathbf{x}_k \to \mathbf{x}_*$。现在令 M_1 和 M_2分别为 (78) 和 (79) 中的隐式常数。对于足够接近$\mathbf{x}_*$ 的$\mathbf{y}_0 = \mathbf{x}_k$，我们得到

$$L\|\tilde{\mathbf{y}}_{j+1} - \mathbf{y}_0\|^2 \le M_1 \max_{i\in[0:j]} (\alpha_{\max}^{(i)})^2 < M_2 \max_{i\in[0:j]} \alpha_{\max}^{(i)} \le \langle \mathbf{g}, \tilde{\mathbf{y}}_{j+1} - \mathbf{y}_0 \rangle\,, \tag{80}$$

其中，第一个不等式中使用了 (78)，第二个不等式中使用了 $\max_{i\in[0:j]} \alpha_{\max}^{(i)} = o(1)$，最后一个不等式中使用了 (79)。由 (80) 可轻易得到 $\tilde{\mathbf{y}}_{j+1} \in$int$B$，正如所需。
现在需要验证 $\tilde{\mathbf{y}}_{j+1} \in B_j$。与上文推理相同，一方面，对于$\mathbf{x}_k \to \mathbf{x}$ </style}, 有 $\mathbf{g}\frac{\|}{2L} = \Theta(1)$（忽略 $-\nabla f(\mathbf{x}_*) = 0$ 的平凡情形）；另一方面，由 (78) 有 $\|\tilde{\mathbf{y}}^{j+1} - \mathbf{y}_0\| = o(1)$，因此

$$\|\tilde{\mathbf{y}}_{j+1} - \mathbf{y}_0\| < \frac{\|\mathbf{g}\|}{2L} \tag{81}$$

使得对于$\mathbf{y}_0$，其足够接近$\mathbf{x}_*$，并且 $\tilde{\mathbf{y}}_{j+1} \in B_j$ 可按要求达到。于是 $\tilde{\mathbf{y}}_{j+1} \in$ int$(\mathcal{C}_j)$ 对于 $j \in [0 : T^*]$，或者等价地，$\beta_j > \alpha_{\max}^{(j)}$ SSC 在前 T^*+1 次迭代中始终执行最大步长。特别地，它生成点$\mathbf{y}_{T^*+1} \in \mathcal{C}_* \setminus \{\mathbf{y}_{T^*}\}$。命题得证。
如果$\mathbf{y}^{T^*+1}$ 是 g 的驻点，SSC 会在第 4 步终止，输出为 $\mathbf{y}_{T^*+1} \in \mathcal{C}_*$，待证命题得证。否则，我们声称 SSC 会终止并输出$\mathbf{y}_{T^*+2} \in \mathcal{C}_*$ 以及 $\beta_{T^*+1} < \alpha^{(T^*+1)_{\max}}$。首先，注意到根据假设，必有$\mathbf{d}_{T^*+1} \in S_*$，因此$\mathbf{y}_{T^*+2} = \mathbf{y}_{T^*+1} + \alpha_{T^*+1}\mathbf{d}_{T^*+1} \in \mathcal{C}_*$。其次，我们有 $\alpha^{(T^*+1)_{\max}} = \Theta(1)$，同时

$$\beta_{T^*+1} \le \operatorname{diam}(\mathcal{C}_{T^*+1}) \le \operatorname{diam}(B_{T^*+1}) \le 2\langle \mathbf{g}_{T^*+1}, \mathbf{d}_{T^*+1} \rangle = o(1)\,. \tag{82}$$

因此，当$\mathbf{y}_0$ 足够接近$\mathbf{x}_*$ 时，必有

$$\beta_{T^*+1} < \alpha_{\max}^{(T^*+1)}\,, \tag{83}$$

且该断言得证。由于 SSC 以 $\mathbf{y}^{T^*+1}$ 或$\mathbf{y}^{T^*+2}$ 终止，且这两个点都属于 $\mathcal{C}_*$，因此待证命题成立。□

引理 8 对于$\mathbf{x} \to \mathbf{x}_*$，若$\mathbf{g} \to -\nabla f(\mathbf{x}_*)$，且$\mathbf{x} \in \mathcal{C}_*$、$\mathbf{d} \in S_*$，并且 $\alpha_{\max}(\mathbf{x}, \mathbf{d})$ 与最大可行步长一致，则 $\alpha_{\max}(\mathbf{x},\mathbf{d}) = \Theta(1)$。

证明我们有

$$\alpha_{\max}(\mathbf{x}, \hat{\mathbf{d}}) = \min_{i:\hat{d}_i<0} \frac{x_i}{|\hat{d}_i|} \geq \min_{i:\hat{d}_i<0} x_i \geq \min_{i\in\mathrm{supp}(\mathbf{x}_*)} x_i = \Theta(1)\,, \tag{84}$$

其中，第一处不等式中我们使用了 $|d_i| \leq \|\mathbf{d}\| \leq 1$，第二处不等式中使用了 $\mathrm{supp}(\mathbf{d}) \subseteq \mathrm{supp}(\mathbf{x}_*)$，第三处不等式中使用了$x_i \to x_{*,i} > 0$。□

对于$\mathbf{x} \in \mathcal{C}$，我们定义如下表达式

$$\lambda^{(i)}(\mathbf{x}, \mathbf{g}) = \langle \mathbf{g}^{(i)}, \mathbf{x}^{(i)} \rangle\, \mathbf{e}^{(i)} - \mathbf{g}^{(i)}\,, \quad i \in [1\!:\!m]\,,$$

以及拉格朗日乘子向量

$$\lambda^{(i)}(\mathbf{x}) = \lambda^{(i)}(\mathbf{x}, -\nabla f(\mathbf{x})) = \nabla f(\mathbf{x})^{(i)} - \langle \nabla f(\mathbf{x})^{(i)}, \mathbf{x}^{(i)} \rangle\, \mathbf{e}^{(i)}\,, \quad i \in [1\!:\!m]\,. \tag{85}$$

我们注意到，在驻点$\mathbf{x}_* \in \mathcal{C}$ 处，对于 $\nabla f(\mathbf{x}_*)$，严格互补性成立，当且仅当对于每个 $i \in [1\!:\!\ m]$，在$\mathbf{x}_*^{(i)} \in \mathcal{C}^{(i)}$处以及 $\nabla f(\mathbf{x}_*)^{(i)}$，严格互补性都成立。

引理 9 假设严格互补性在$\mathbf{x}_*$处成立。那么，将 *AFW* 应用于单纯形时，在$\mathbf{x}_*$ 处具有定义 *4* 中所述的与活跃集相关的方向。

证明 对于 $\mathbf{x} \to \mathbf{x}_*$ 以及 $\mathbf{g} \to -\nabla f(\mathbf{x}_*)$，我们有 $\lambda(\mathbf{x}, \mathbf{g}) \to \lambda(\mathbf{x}_*)$，因此特别地，对于 $i \in \mathrm{supp}(\mathbf{x}_*)$，有 $\lambda_i(\mathbf{x}, \mathbf{g}) \to 0$；而对于 $i \in [1\!:\!\ n] \setminus \mathrm{supp}(\mathbf{x}_*)$，有 $\lambda_i(\mathbf{x}, \mathbf{g}) \to \lambda_i(\mathbf{x}_*) > 0$。因此，当 $\mathbf{x} \in \mathcal{C} \setminus \mathcal{C}_*$足够接近 $\mathbf{x}_*$时，我们必有 $\max\{\lambda_i(\mathbf{x}, \mathbf{g}) :\ i \in \mathrm{supp}(\mathbf{x})\} > \max\{-\lambda_i(\mathbf{x}, \mathbf{g}) :\ i \in [1\!:\!n]\}$，故由 [13, 引理 3.2(a)] 可知，AFW 选取的下降方向满足 $\mathbf{d} = \mathbf{x} - \mathbf{e}_{\hat{i}}$，其中某个 $\hat{i} \in \mathrm{argmax}\{\lambda_i(\mathbf{x}, \mathbf{g}) :\ i \in \mathrm{supp}(\mathbf{x})\} \subset [1\!:\!\ n] \setminus \mathrm{supp}(\mathbf{x}_*)$。因此

$$\langle \mathbf{g}, \hat{\mathbf{d}} \rangle = \langle \mathbf{g}, \frac{\mathbf{x} - \mathbf{e}_{\hat{i}}}{\|\mathbf{x} - \mathbf{e}_{\hat{i}}\|} \rangle = \frac{\lambda_{\hat{i}}(\mathbf{x}, \mathbf{g})}{\|\mathbf{x} - \mathbf{e}_{\hat{i}}\|} = \Theta(1) \tag{86}$$

对于 $\mathbf{x} \to \mathbf{x}_*$ 和 $\mathbf{g} \to -\nabla f(\mathbf{x}_*)$。

对于 $\mathbf{x} \in \mathcal{C}_*$ 这种情况，如果 $\mathbf{x}$、$\mathbf{g}$ 与 $\mathbf{x}_*$ 足够接近，那么对于 [1 中的每个 i，我们必有 $\lambda_i(\mathbf{x}, \mathbf{g}) > 0\ n] \setminus \mathrm{supp}(\mathbf{x}_*)$。因此，根据 [13, 引理 3.2(b)]，如果 y 是通过对 $\mathbf{x}$ 执行一次 Frank-Wolfe 更新得到的，那么对于 $i \in [1$，我们必有 $\mathbf{y}_i = 0\ n] \setminus \mathrm{supp}(\mathbf{x}_*)$；这等价于说，更新方向必须属于 S_*。由引理 8 可得性质 $\alpha_{\max}(\mathbf{x},\mathbf{d}) = \Theta(1)$。□

证明（定理 *2*）对该方法选取的每个块应用引理 9 中证明的性质即可得出。□

8 声明

8.1 资金支持和/或利益冲突/竞争性利益

所有作者均无需要申报的事项。

8.2 遵守伦理标准

所有作者均严格遵守广泛认可的道德规范，完成了本文的全部研究工作和撰写工作。

参考文献

1. Ala´ız, C.M., Barbero, A., Dorronsoro, J.R.：群组融合套索。载于：国际人工神经网络会议，第 66–73 页。Springer（2013） 2. Alon, N., Spencer, J.H.：概率方法。约翰·威立父子出版公司（2016） 3. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.：针对非凸问题的近端交替最小化与投影方法：一种基于 Kurdyka-Łojasiewicz 不等式的方法。运筹学数学**35**(2), 438–457（2010） 4. Beck, A.：优化中的一阶方法。工业与应用数学学会（2017） 5. Bertsekas, D., Tsitsiklis, J.：并行与分布式计算：数值方法。Athena Scientific（2015） 6. Birgin, E.G., Mart´ınez, J.M., Raydan, M.：凸集上的非单调谱投影梯度方法。SIAM 优化期刊**10**(4), 1196–1211（2000） 7. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.：可分层函数的 Clarke 次梯度。SIAM 优化期刊**18**(2), 556–572（2007） 8. Bolte, J., Daniilidis, A., Ley, O., Mazet, L.： -Lojasiewicz 不等式的刻画：次梯度流、谷径与凸性。美国数学会汇刊**362**(6), 3319–3363（2010） 9. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.：从误差界到凸函数一阶下降方法的复杂度。数学规划**165**(2), 471–507（2017） 10. Bomze, I.M., Budinich, M., Pardalos, P.M., Pelillo, M.：最大团问题。载于：组合优化手册，第 1–74 页。Springer（1999） 11. Bomze, I.M., Gabl, M., Maggioni, F., Pflug, G.：两阶段随机标准二次优化。欧洲运筹学杂志 **299**(1), 21–34（2022） 12. Bomze, I.M., Rinaldi, F., Rota Bul`o, S.：面向急性子的人的一阶方法：利用收敛的 Frank-Wolfe 变体在有限时间内识别支撑集。SIAM 优化期刊**29**(3), 2211–2226（2019） 13. Bomze, I.M., Rinaldi, F., Zeffiro, D.：远离步 Frank-Wolfe 算法的活跃集复杂度。SIAM 优化期刊**30**(3), 2470–2500（2020） 14. Bomze, I.M., Rinaldi, F., Zeffiro, D.：Frank-Wolfe 及其变体：无投影一阶优化方法探索之旅。4OR **19**(3), 313–345（2021） 15. Bomze, I.M., Rinaldi, F., Zeffiro, D.：利用一阶优化快速检测网络中的簇。SIAM 数据科学数学期刊**4**(1), 285–305（2022） 16. Bomze, I.M., Schachinger, W.：多标准二次优化：内点法与锥规划重构。计算优化与应用 **45**(2), 237–256 （ 2010 ） 17. Bomze, I.M., Schachinger, W., Ullrich, R.：简单模型的复杂度——标准二次优化问题最坏与典型困难情形研究。运筹学数学**43**(2), 347–692（2017） 18. Boumal, N.：光滑流形上的优化导论。在线发布，5 月**3** 日（2020） 19. Calamai, P.H., Mor´e, J.J.：线性约束问题的投影梯度方法。数学规划**39**(1), 93–116（1987） 20. Combettes, C.W., Pokutta, S.：若干集合上线性最小化与投影的复杂度。运筹学快报**49**(4), 565–571（2021） 21. Di Serafino, D., Toraldo, G., Viola, M., Barlow, J.：一种用于具有单个线性约束及变量界的二次规划问题的两阶段梯度法。SIAM 优化期刊**28**(4), 2809–2838（2018） 22. Foygel, R., Horrell, M., Drton, M., Lafferty, J.：非参数降秩回归。神经信息处理系统进展**25** （2012） 23. Fukunaga, T., Kasai, H.：用于半松弛最优传输的快速分块坐标 Frank-Wolfe 算法。arXiv 预印本 arXiv:2103.05857（2021） 24. Garber, D.：重新审视多胞体上的 Frank-Wolfe：严格互补性与稀疏性。arXiv 预印本 arXiv:2006.00558（2020） 25. Jaggi, M.：重新审视 Frank-Wolfe：无投影稀疏凸优化。载于：国际机器学习会议（第 1 卷），第 427–435 页（2013） 26. Jegelka, S., Bach, F., Sra, S.：面向易用次模优化的反射方法。神经信息处理系统进展**26** （2013）

27. Karimi, H., Nutini, J., Schmidt, M.：Polyak-Łojasiewicz 条件下梯度方法与近端梯度方法的线性收敛。载于：欧洲机器学习与数据库知识发现联合会议，第 795–811 页。Springer（2016） 28. Lacoste-Julien, S., Jaggi, M., Schmidt, M., Pletscher, P.：结构化支持向量机的块坐标 Frank-Wolfe 优化。载于：S. Dasgupta, D. McAllester（编），第 30 届国际机器学习会议论文集， *Proceedings of Machine Learning Research*，第 28(1) 卷，第 53–61 页。PMLR，亚特兰大，佐治亚州，美国（2013）。网址：http://proceedings.mlr.press/v28/lacoste-julien13.html。Lan, G.：机器学习的一阶与随机优化方法。Springer（2020） 30. Leary, R.H.：漏斗景观上的全局优化。全球优化期刊**18**(4), 367–383（2000） 31. LeBlanc, L.J., Morlok, E.K., Pierskalla, W.P.：求解道路网络均衡交通分配问题的有效方法。交通研究**9**(5), 309–318（1975） 32. Liu, J., Musialski, P., Wonka, P., Ye, J.：用于估计视觉数据缺失值的张量补全。IEEE 模式分析与机器智能汇刊**35**(1), 208–220（2012） 33. Locatelli, M., Schoen, F.：全局优化：理论、算法与应用。工业与应用数学学会（2013） 34. Luo, Z.Q., Tseng, P.：凸可微最小化中坐标下降方法的收敛性。优化理论与应用期刊**72**(1), 7–35（1992） 35. Nesterov, Y.：坐标下降方法在超大规模优化问题中的效率。SIAM 优化期刊**22**(2), 341–362（2012） 36. Osokin, A., Alayrac, J.B., Lukasewitz, I., Dokania, P., Lacoste-Julien, S.：关注差距：结构化支持向量机的块 Frank-Wolfe 优化。载于：国际机器学习会议，第 593–602 页。PMLR（2016） 37. Richt´arik, P., Tak´aˇc, M.：用于复合函数最小化的随机块坐标下降方法的迭代复杂度。数学规划**144**(1), 1–38（2014） 38. Rinaldi, F., Zeffiro, D.：在充分斜率条件下分析无投影一阶方法的统一框架。arXiv 预印本 arXiv:2008.09781（2020） 39. Rinaldi, F., Zeffiro, D.：避免 Frank-Wolfe 变体中的不良步。计算优化与应用**84**, 225–264（2023） 40. di Serafino, D., Hager, W.W., Toraldo, G., Viola, M.：具有多面体约束的非线性优化问题的驻点性。数学规划，第 1–28 页（2023） 41. Wang, Y.X., Sadhanala, V., Dai, W., Neiswanger, W., Sra, S., Xing, E.：并行与分布式块坐标 Frank-Wolfe 算法。载于：国际机器学习会议，第 1548–1557 页。PMLR（2016）