双层优化的一阶罚函数方法

Zhaosong Lu * Sanyou Mei *

January 4, 2023（修订：2023年4月18日；2023年9月27日；2023年12月6日 ）

arXiv:2301.01716v2 [math.OC] 7 Mar 2024

摘要

本文研究一类无约束和有约束双层优化问题，其中下层是可能非光滑的凸优化问题，而上层是可能非凸的优化问题。我们为这类问题引入了 ε-KKT 解的概念，并证明在适当假设下，一个 ε-KKT 解会导出一个基于 $\mathcal{O}(\sqrt{\varepsilon})$-或 $\mathcal{O}(\varepsilon)$-超梯度的驻点。我们还提出了一阶罚函数方法，以求解这类问题的 ε-KKT 解。其子问题最终可表示为结构化极小极大问题，并可通过作者近期开发的一阶方法进行求解。在适当假设下，针对所提出的罚函数方法，分别建立了求解无约束和有约束双层优化问题的 ε-KKT 解所需的基本运算复杂度，为 $operationcomplexity$ of $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$ 和 $\mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$。我们给出了初步数值结果，以展示所提方法的性能。据我们所知，本文首次证明双层优化可以近似地作为极小极大优化求解；此外，本文还为这类复杂的双层优化提供了首个具有复杂度保证的可实现方法。

关键词：双层优化、极小极大优化、罚函数法、一阶方法、运算复杂度

数学主题分类：90C26, 90C30, 90C47, 90C99, 65K05

1 引言

双层优化是一种两层的层次优化，其中上层的决策变量也涉及下层。一般而言，它可以写成以下形式：

$$\begin{array}{ll}\min\limits_{x,y} & f(x,y)\\ \text{s.t.} & g(x,y)\le 0,\quad y\in \operatorname{argmin}\limits_{z}\{\tilde{f}(x,z)|\tilde{g}(x,z)\le 0\}.^{1}\end{array}\tag{1}$$

双层优化已在许多重要领域得到应用，包括对抗训练 [45,46, 57], 持续学习 [40], 超参数调优 [3, 17], 图像重建 [9], 元学习[4, 28, 52], 神经架构搜索 [15, 38], 强化学习 [23, 31], 和斯塔克尔伯格博弈 [59]。关于双层优化的更多应用可见 [2, 8, 10, 11, 12, 54] 及其中的参考文献。文献已广泛研究了问题 (1) 的理论性质，包括其最优性条件（例如，参见[12, 13, 43, 58, 62]）。

目前已有许多方法用于求解（1）的一些特殊情形。例如，已有研究提出了基于约束的方法 [22, 53], 确定性梯度方法 [16, 17, 20, 24, 44, 51, 52], 和随机梯度方法 [6, 18, 21, 23, 25, 26, 29, 30, 34, 35, 61] ，用于求解满足以下条件的（1）：~ ~$g\equiv 0$、$\tilde{g}\equiv 0$、f、f 均为光滑函数，且 f 关于 y强凸。对于与此类似的情形~，但 f 关于 y为凸函数，近期[5], 提出了一种零阶方法；此外，[36, 37, 56] 还通过将（1）作为单个或一系列~光滑约束优化问题来求解，开发了数值方法。此外，当（1）中的所有函数均为光滑函数，且 f、$\tilde{g}$关于 y为凸函数时，研究者通过求解一个带均衡约束的数学规划，提出了梯度型方法；该规划是将（1）的下层优化问题替换为其一阶最优性条件后得到的（例如，参见 [1, 42, 50]）。近期，~在 [63] 中开发了差分凸（DC）算法，用于求解满足以下条件的（1）：$g\equiv 0$、f 为差分凸函数，且 f、$\tilde{g}$ 为凸函数

*美国明尼苏达大学工业与系统工程系（电子邮件： zhaosong@umn.edu,mei00035@umn.edu）。本研究部分得到了美国国家科学基金会 IIS-2211491号资助项目的支持。[1]为便于阅读，本文中特别使用波浪号表示与下层优化问题相关的函数。此外，“使目标函数取最小值的自变量”表示相关问题的最优解集合。

函数。此外，针对（1），[27] 提出了一种双重惩罚法，该方法求解一系列如下形式的双层优化问题

$$\begin{array}{ll}\min\limits_{x,y} & f(x,y)+\rho_k\Psi(x,y)\\ \text{s.t.} & y\in\operatorname{argmin}\limits_z \tilde{f}(x,z)+\rho_k\tilde{\Psi}(x,z),\end{array}\tag{2}$$

其中，$\{\rho_k\}$ 是一个惩罚参数序列，Ψ 和 $\tilde{\Psi}$ 分别是与集合 $\{(x,y)|g(x,y)\le 0\}$ 以及 $\{(x,z)|\tilde{g}(x,z)\le 0\}$相关的罚函数。尽管问题 (2) 看起来比问题 (1) 简单，但一般而言，目前没有可用于求解问题 (2) 的近似解的方法。因此，双重惩罚法 [27] 通常无法实施。关于双层优化的算法发展，更多讨论可见 [2, 8, 12, 39, 55, 58]及其中的参考文献。

众所周知，双层优化问题 (1) 的棘手难题主要来自下层部分，因为它要求变量 y 是另一个优化问题的解。因此，为了简化起见，我们只考虑不包含约束$g(x,y)\le 0$ 的一类双层优化问题，即

$$\begin{array}{ll}\min\limits_{x,y} & f(x,y)\\ \text{s.t.} & y\in\operatorname{argmin}\limits_z\{\tilde{f}(x,z)|\tilde{g}(x,z)\le 0\}.\end{array}\tag{3}$$

尽管如此，本文中的结果可能可以扩展到问题 (1)。

本文的主要目标是开发一种可实现的一阶方法，并为求解问题 (3) 提供复杂度保证。我们在此方面的关键见解是：(i) 问题 (3) 可以通过一种新颖的罚函数方法所得到的结构化极小极大问题来近似求解；(ii) 可以采用 [41,算法 2] 中提出的一阶方法，适当地求解由此得到的结构化极小极大问题。因此，这些见解促成了一种新颖的一阶罚函数法，用于求解 (3)，并具有以下优点。

- 它仅使用问题的一阶信息。具体而言，其基本运算仅包括对 $\tilde{g}$ 和 f 的光滑部分以及 f 和的非光滑部分进行梯度计算 f 和 f近端算子计算。因此，它适用于求解大规模问题（见第 2 节和第 3 节）。

- 它在运算复杂度方面具有理论保证；运算复杂度由前述基本运算衡量，用于求得问题 (3) 的 ε-KKT 解。具体而言，当 $\tilde{g}\equiv 0$时，其运算复杂度为 $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$。否则，其运算复杂度为$\mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$（见定理3和定理5）。

- 与现有方法相比，该方法适用于更广泛的问题类别。例如，它可以应用于（3），其中 f, f 是非光滑的，且 $f, \tilde{g}$ 关于 x 是非凸的；然而，现有方法并不适用于这种情形。

据我们所知，本文首次证明了双层优化可以通过极小极大优化近似求解，并且首次为复杂的双层优化问题（3）提供了具有复杂度保证的可实现方法。

本文其余部分安排如下。小节 1.1 介绍一些记号和术语。第 2 节和第 3 节分别提出用于无约束和有约束双层优化的一阶罚函数方法，并研究其复杂度。初步数值结果和主要结果的证明分别见第 4 节和第 5 节。最后，我们在第 6 节作总结。

1.1 记号和术语

本文将始终使用以下记号。令 $\mathbb{R}^n$ 表示 n 维欧几里得空间，令 $\mathbb{R}^n_+$ 表示 $\mathbb{R}^n$ 中的非负正交象限。除非另有说明，标准内积和欧几里得范数分别记为 $\langle\cdot,\cdot\rangle$ 和 $\|\cdot\|$。对于任意 $v\in\mathbb{R}^n$，令 v_+ 表示 v 的非负部分，即对于所有 i，$(v_+)_i=\max\{v_i,0\}$。对于任意两个向量 u 和 v，$(u;v)$ 表示将 v 叠放在 u 下方所得的向量。给定 x 中的一个点 S 和一个闭集 $\mathbb{R}^n$，令 $\operatorname{dist}(x,S)=\min_{x'\in S}\|x'-x\|$，并令 $\mathscr{I}_S$ 表示与 S 相关联的示性函数。

函数或映射 ϕ 若在集合 S 上满足 $\|\phi(x)-\phi(x')\|\le L_\phi\|x-x'\|$，则称其为 L_ϕ-Lipschitz 连续，对于所有 $x,x'\in S$。此外，如果 $\|\nabla\phi(x)-\nabla\phi(x')\|\le L_{\nabla\phi}\|x-x'\|$，则称其在 S 上是 $L_{\nabla\phi}$-光滑的。

所有 x、$x' \in S$。[2] 对于闭凸函数 $p: \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$、[3]，与 p 相关联的近端算子记作 prox_p，即

$$\mathrm{prox}_p(x) = \underset{x' \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2}\|x' - x\|^2 + p(x') \right\} \quad \forall x \in \mathbb{R}^n.$$

鉴于计算 $\mathrm{prox}_{\gamma p}(x)$ 的开销通常与计算 $\mathrm{prox}_p(x)$ 一样低，对于任意 $\gamma > 0$ 和 $x \in \mathbb{R}^n$，我们将计算 $\mathrm{prox}_{\gamma p}(x)$ 计作对 p 的近端算子的一次计算。

给定下半连续函数 $\phi: \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$，其定义域为集合 $\mathrm{dom}\,\phi := \{x | \phi(x) < \infty\}$。$\phi$ 在 $x \in \mathrm{dom}\,\phi$ 处沿方向 $d \in \mathbb{R}^n$ 的上次导数定义为

$$\phi'(x; d) = \limsup_{x' \xrightarrow{\phi} x,\, t \downarrow 0} \inf_{d' \to d} \frac{\phi(x' + td') - \phi(x')}{t},$$

其中，$t \downarrow 0$ 表示 $t > 0$ 和 $t \to 0$，而 $x' \xrightarrow{\phi} x$ 表示 $x' \to x$ 和 $\phi(x') \to \phi(x)$。ϕ 在 $x \in \mathrm{dom}\,\phi$ 处的次微分为集合

$$\partial\phi(x) = \{s \in \mathbb{R}^n \,|\, s^T d \le \phi'(x; d) \ \ \forall d \in \mathbb{R}^n\}.$$

我们用 $\partial_{x_i}\phi(x)$ 表示关于 x_i 的次微分。此外，对于上半连续函数 ϕ，其次微分定义为 $\partial\phi = -\partial(-\phi)$。如果 ϕ 在局部 Lipschitz 连续，则上述次微分定义与 Clarke次微分一致。此外，如果 ϕ 是凸的，则它与凸函数的通常次微分一致。另外，如果 ϕ 在 x 处连续可微，则有 $\partial\phi(x) = \{\nabla\phi(x)\}$，其中 $\nabla\phi(x)$ 是 ϕ 在 x 处的梯度。此外，不难验证，如果 ϕ_1 在 x 处连续可微，且 ϕ_2 在 x 处下半连续或上半连续，则 $\partial(\phi_1 + \phi_2)(x) = \nabla\phi_1(x) + \partial\phi_2(x)$。更多详情请参见 [7, 60] 。

最后，我们为一般的极小极大问题引入两类近似解

$$\Psi^* = \min_x \max_y \Psi(x, y), \tag{4}$$

其中，$\Psi(\cdot, y): \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ 是下半连续函数，$\Psi(x, \cdot): \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ 是上半连续函数，且 Ψ^* 有限。

定义1. 若点$(x_\varepsilon, y_\varepsilon)$满足以下条件，则称其为极小极大问题(4)的 ϵ-最优解

$$\max_y \Psi(x_\epsilon, y) - \Psi(x_\varepsilon, y_\varepsilon) \le \epsilon, \quad \Psi(x_\varepsilon, y_\varepsilon) - \Psi^* \le \epsilon.$$

定义2. 若点(x, y)满足以下条件，则称其为极小极大问题(4)的驻点

$$0 \in \partial_x \Psi(x, y), \quad 0 \in \partial_y \Psi(x, y).$$

此外，对于任意 $\epsilon > 0$，若点$(x_\varepsilon, y_\varepsilon)$满足以下条件，则称其为极小极大问题(4)的 ϵ-驻点 *if*

$$\mathrm{dist}\,(0, \partial_x \Psi(x_\varepsilon, y_\varepsilon)) \le \epsilon, \quad \mathrm{dist}\,(0, \partial_y \Psi(x_\varepsilon, y_\varepsilon)) \le \epsilon.$$

2 无约束双层优化

本节中，我们考虑一个无约束双层优化问题[4]

$$\begin{aligned} f^* = \min \quad & f(x, y) \\ \text{s.t.} \quad & y \in \underset{z}{\operatorname{argmin}}\, \tilde{f}(x, z). \end{aligned} \tag{5}$$

˜假设问题（5）至少存在一个最优解。此外，f 和 f 满足以下假设。

[2]当 ϕ 是一个映射时，$\|\nabla\phi(x) - \nabla\phi(x')\|$ 中使用的范数是弗罗贝尼乌斯范数。[3]为方便起见，∞ 表示 $+\infty$。[4]为方便起见，我们将问题（5）称为无约束双层优化问题，因为其下层部分没有显式约束。严格来说，它也可以是一个有约束双层优化问题。例如，˜当 f 和/或 f 的一部分是闭凸集的示性函数时，（5）本质上是一个有约束双层优化问题。

假设1。（i） $f(x,y) = f_1(x,y) + f_2(x)$ 以及 $f(x,y) = f_1(x,y) + f_2(y)$ 在 $\mathcal{X}\times\mathcal{Y}$上连续，其中 f_2：$\mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ 和 f_2： $\mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ 是真闭凸函数， $f_1(x,\cdot)$ 对于任意给定的 $x \in \mathcal{X}$都是凸的，并且 f_1、f_1分别在 $\mathcal{X}\times\mathcal{Y}$上是 $L_{\nabla f_1}$-光滑和 $L_{\nabla \tilde{f}_1}$-光滑的，其中$\mathcal{X}$：$=$dom f_2且 $\mathcal{Y}$：$=$dom f_2。

(ii) 与 f_2 和 f_2 相关联的近端算子可以精确求值。

(iii) 集合 $\mathcal{X}$和 $\mathcal{Y}$（即 dom f_2和 dom f_2）是紧集。

为方便记号，我们定义

$$D_{\mathbf{x}} := \max\{\|u-v\| \big| u,v \in \mathcal{X}\}, \quad D_{\mathbf{y}} := \max\{\|u-v\| \big| u,v \in \mathcal{Y}\}, \tag{6}$$

$$\tilde{f}_{\text{hi}} := \max\{\tilde{f}(x,y) | (x,y) \in \mathcal{X}\times\mathcal{Y}\}, \quad \tilde{f}_{\text{low}} := \min\{\tilde{f}(x,y) | (x,y) \in \mathcal{X}\times\mathcal{Y}\}, \tag{7}$$

$$f_{\text{low}} := \min\{f(x,y) | (x,y) \in \mathcal{X}\times\mathcal{Y}\}. \tag{8}$$

根据假设1，可知 $D_{\mathbf{x}}$、 $D_{\mathbf{y}}$、 f_{hi}、 f_{low} 和 f_{low} 均为有限值。

本节的目标是提出求解问题（5）的一阶罚函数方法。为此，我们首先注意到，问题（5）可以看作

$$\min_{x,y}\{f(x,y) | \tilde{f}(x,y) \le \min_z \tilde{f}(x,z)\}. \tag{9}$$

注意到，对于所有 x、 y， $f(x,y) - \min_z f(x,z) \ge 0$。因此，与（9）相关联的一个自然的罚函数问题是

$$\min_{x,y} f(x,y) + \rho(\tilde{f}(x,y) - \min_z \tilde{f}(x,z)), \tag{10}$$

其中， $\rho > 0$ 是一个罚参数。我们进一步注意到，问题（10）等价于极小极大问题

$$\min_{x,y}\max_z P_\rho(x,y,z), \quad \text{where} \quad P_\rho(x,y,z) := f(x,y) + \rho(\tilde{f}(x,y) - \tilde{f}(x,z)). \tag{11}$$

鉴于假设1(i)， P_ρ 可改写为

$$P_\rho(x,y,z) = \big(f_1(x,y) + \rho\tilde{f}_1(x,y) - \rho\tilde{f}_1(x,z)\big) + \big(f_2(x) + \rho\tilde{f}_2(y) - \rho\tilde{f}_2(z)\big). \tag{12}$$

由此以及假设1可知， P_ρ 具有以下良好性质。

- P_ρ是光滑函数 $f_1(x,y)+\rho f_1(x,y) - \rho f_1(x,z)$与具有Lipschitz 连续梯度的可能非光滑函数 $f_2(x)+\rho f_2(y)-\rho f_2(z)$之和，其中后者的近端算子可以精确计算。
- P_ρ 关于(x,y)是非凸的，但关于 z是凹的。

得益于 P_ρ的这一良好结构，可以通过 [41, 提出的一阶方法找到极小极大问题（11）的近似驻点（见附录A中的算法 6）。

基于上述观察，我们现在可以通过求解一系列极小极大问题，或求解一个形式如（11）的极小极大问题，来为无约束双层优化问题（5）提出罚函数法。具体而言，我们首先通过求解一系列极小极大问题，为（5）提出一种理想的罚函数法（见算法 1）。然后，我们通过求得一个单一极小极大问题的近似驻点，为（5）提出一种实用罚函数法（见算法 2）。

算法 1 问题（5）的一种理想罚函数法

输入：正数列 $\{\rho_k\}$和 $\{\epsilon_k\}$，满足 $\lim_{k\to\infty}(\rho_k,\epsilon_k) = (\infty, 0)$。1: 对于 $k = 0$、1、2、.....，执行2: 求解问题（11），得到其一个 ϵ_k-最优解 (x^k, y^k, z^k)，其中 $\rho = \rho_k$。3: 结束循环

以下 定理给出了算法 1 的收敛结果，其证明推迟到第 n 5.1.

定理1（算法 1 的收敛性）。假设假设1成立，且 $\{(x^k, y^k, z^k)\}$由算法 1 生成。那么， $\{(x^k, y^k)\}$ 的任意聚点都是问题（5）的最优解。

注意，(11) 是一个非凸-凹极小极大问题。对于任意 $\epsilon > 0$，通常很难找到 (11) 的 ϵ-最优解。因此，算法 1 通常无法实现。接下来，我们提出一种求解问题（5）的实用罚函数法：通过对单个极小极大问题 (11) 采用合适的 ρ，应用算法 6（见附录A）求得一个近似驻点。

算法 2 问题（5）的实用罚函数法

输入： $\varepsilon \in (0, 1/4]$, $\rho = \varepsilon^{-1}$, $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$，其中 $\tilde{f}(x^0, y^0) \le \min_y \tilde{f}(x^0, y) + \varepsilon$。1: 调用附录A中的算法 6，使用 $\epsilon \leftarrow \varepsilon$, $\epsilon_0 \leftarrow \varepsilon^{3/2}$, $\hat{x}^0 \leftarrow (x^0, y^0)$, $\hat{y}^0 \leftarrow y^0$ 和 $L_{\nabla h} \leftarrow L_{\nabla f_1} + 2\varepsilon^{-1} L_{\nabla \tilde{f}_1}$，求得问题（11）的一个 ε-驻点 $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$，其中 $\rho = \varepsilon^{-1}$。2: 输出：$(x_\varepsilon, y_\varepsilon)$。

注释1。(i) 算法 2 的初始点 (x^0, y^0) 可通过额外的过程求得。具体而言，可以先任意选择一个 $x^0 \in \mathcal{X}$，然后对问题 $\min_y \tilde{f}(x^0, y)$ 应用加速近端梯度法 [47]，以求得满足 $\tilde{f}(x^0, y^0) \le \min_y \tilde{f}(x^0, y) + \varepsilon$ 的 $y^0 \in \mathcal{Y}$；(ii) 如定理 6（见附录A）所示，通过对问题（11）应用算法 6，可以在算法 2 的步骤 1 中成功求得一个 (11) 的 ϵ-驻点；(iii) 为简便起见，算法 2 求解一个具有静态罚参数和容差参数的、形式如（11）的子问题。不过，算法 2 也可以改进为一种或许在实践中效率更高的算法，即求解一系列具有动态罚参数和容差参数的、形式如（11）的子问题。

为了刻画算法 2 找到的近似解，下面我们引入问题（5）的 ε-KKT 解这一概念。

回顾一下，问题（5）可以视为问题（9），后者是一个约束优化问题。按照经典约束优化的思路，人们自然会关注问题（9）或等价的问题（5）的 KKT 解 $\tilde{(x, y)}$，即 (x, y) 满足 $f(x, y) \le \min_z f(x, z)$，并且 (x, y) 是问题

$$\min_{x', y'} f(x', y') + \rho\big(\tilde{f}(x', y') - \min_{z'} \tilde{f}(x', z')\big) \tag{13}$$

对某个 $\rho \ge 0$ 而言的驻点。[5] 然而，由于复杂的问题结构，刻画问题（13）的驻点通常十分困难。另一方面，注意到问题（13）等价于极小极大问题

$$\min_{x', y'} \max_{z'} f(x', y') + \rho(\tilde{f}(x', y') - \tilde{f}(x', z')),$$

根据定义2和假设1，其驻点 (x, y, z) 满足

$$0 \in \partial f(x, y) + \rho \partial \tilde{f}(x, y) - (\rho \nabla_x \tilde{f}(x, z); 0), \quad 0 \in \rho \partial_z \tilde{f}(x, z). \tag{14}$$

基于这一观察，我们转而关注问题（5）的（弱）KKT 解及其非精确对应问题，具体定义如下。

定义3. 若存在 $\tilde{(x, y)}$，使得（14）和 $f(x, y) \le \min_{z'} f(x, z')$ 成立，则称数对 $\tilde{(x, y)}$ 为问题（5）的 KKT 解。此外，对于任意 $\tilde{\varepsilon} > 0$，若存在 $(z, \rho) \in \mathbb{R}^m \times \mathbb{R}_+$，使得

$$\begin{aligned} &\operatorname{dist}\Big(0, \partial f(x, y) + \rho \partial \tilde{f}(x, y) - (\rho \nabla_x \tilde{f}(x, z); 0)\Big) \le \varepsilon, \quad \operatorname{dist}\big(0, \rho \partial_z \tilde{f}(x, z)\big) \le \varepsilon, \\ &\tilde{f}(x, y) - \min_{z'} \tilde{f}(x, z') \le \varepsilon. \end{aligned}$$

近期，文献中考虑了问题（5）的基于超梯度的驻点（例如，[18, 51]）$\tilde{}$，其假设是 f 和 $\tilde{f}$ 关于 $\mathbb{R}^n \times \mathbb{R}^m$ 和 $\tilde{}$ 二阶连续可微，且对于任意 $x \in \mathbb{R}^n$，$f_x(\cdot)$ 为强凸函数。在此假设下，问题（5）的超目标函数 Φ 定义为

$$\Phi(x) := f(x, y^*(x)), \text{ where } y^*(x) = \operatorname*{argmin}_z \tilde{f}(x, z), \tag{15}$$

是连续可微的。此外，根据 [18, 公式（2.8）], ，问题（5）的超梯度，即 Φ 的梯度，为

$$\nabla \Phi(x) = \nabla_x f(x, y^*(x)) - \nabla^2_{xy} \tilde{f}(x, y^*(x)) [\nabla^2_{yy} \tilde{f}(x, y^*(x))]^{-1} \nabla_y f(x, y^*(x)) \quad \forall x \in \mathbb{R}^n. \tag{16}$$

[5] 关系 $\tilde{f}(x, y) \le \min_z \tilde{f}(x, z)$ 意味着 $\tilde{f}(x, y) = \min_z \tilde{f}(x, z)$，因此互补松弛条件 $\rho(\tilde{f}(x, y) - \min_z \tilde{f}(x, z)) = 0$ 成立。

此外，不难看出，问题（5）等价于

$$\min_x \Phi(x). \tag{17}$$

鉴于此，文献（例如 [18, 51]）针对问题（5）引入了基于超梯度的驻点及其近似形式，它们分别基于问题（17）的经典驻点及其近似形式。更具体地说，如果 $x \in \mathbb{R}^n$ $\nabla\Phi(x) = 0$，则称其为问题（5）的基于超梯度的驻点；如果对于任意 $\varepsilon > 0$， $\|\nabla\Phi(x)\| \le \varepsilon$，则称其为问题（5）的 ε-基于超梯度的驻点。

现在，我们研究 ε-KKT 解与问题（5）的近似基于超梯度的驻点之间的关系。具体而言，在一些适当的假设下，下面的定理表明，如果 (x, y) 是问题（5）的 ε-KKT 解，那么 x 是其 $\mathcal{O}(\sqrt{\varepsilon})$-或 $\mathcal{O}(\varepsilon)$-基于超梯度的驻点。该定理的证明推迟到第 5.1 小节。

˜定理2. 设给定 $\varepsilon_0, \rho_0 > 0$，且 $\Omega \subset \mathbb{R}^n$ 为非空紧集。假设 f 和 f˜分别关于 $\mathbb{R}^n \times \mathbb{R}^m$ 连续可微和二阶连续可微， $f(x', \cdot)$ 对所有属于包含 Ω 的开集 $\mathcal{N}$的 x'均为模数为 $\sigma > 0$的强凸函数， $\nabla f(x', \cdot)$ 对所有 $x' \in \Omega$ 均为 L_1-Lipschitz 连续，并且 $\nabla^2 \tilde{f}(x', \cdot)$ 对所有 $x' \in \Omega$ 均为 L_2-Lipschitz 连续。假设 $(x, y) \in \Omega \times \mathbb{R}^m$ 是问题（5）的一个 ε-KKT 解，其关联的 $\rho \ge \rho_0$ 对应某个 $0 < \varepsilon \le \varepsilon_0$。令 $y^*(x')$ 如（15）所定义，并

$$\bar{C} = \max\big\{\|\nabla_y f(x', y')\| : x' \in \Omega, \|y' - y^*(x')\| \le \sqrt{2\sigma^{-1}\varepsilon}\big\}, \tag{18}$$

$$\theta = \min\big\{(\rho\sigma)^{-1}(\varepsilon + \bar{C}),\ \sqrt{2\sigma^{-1}\varepsilon}\big\}, \quad C = \max_{x' \in \Omega} \|\nabla^2_{xy} \tilde{f}(x', y^*(x'))[\nabla^2_{yy} \tilde{f}(x', y^*(x'))]^{-1}\|. \tag{19}$$

那么我们有

$$\|\nabla\Phi(x)\| \le (2C+1)\varepsilon + (C+1)\Big(L_1\theta + \frac{L_2\rho\theta^2}{2} + \frac{L_2\varepsilon^2}{2\rho\sigma^2}\Big) \tag{20}$$

$$\le (2C+1)\varepsilon + (C+1)\Big(L_1\sqrt{2\sigma^{-1}} + \frac{L_2\sigma^{-3/2}(\varepsilon_0 + \bar{C})}{\sqrt{2}} + \frac{\varepsilon_0^{3/2} L_2}{2\rho_0\sigma^2}\Big)\sqrt{\varepsilon}. \tag{21}$$

˜注2. 根据定理2中的假设，不难看出 C, C 和 θ 均为有限值。由此根据（21）可知，一般而言 $\nabla\Phi(x) = \mathcal{O}(\sqrt{\varepsilon})$ 成立。然而，当 ρ 至少为 ε^{-1}的阶，即对于某个与 ε无关的常数 $c > 0$，有 $\rho \ge c\varepsilon^{-1}$ 时，这一结果可改进为$\nabla\Phi(x) = \mathcal{O}(\varepsilon)$；这一点可由（20）看出。因此，在定理2所述假设下，若 (x, y) 是问题（5）的一个 ε-KKT 解，则 x是该问题的一个 $\mathcal{O}(\sqrt{\varepsilon})$-或 $\mathcal{O}(\varepsilon)$-基于超梯度的驻点。

接下来，我们给出一个关于算法 2 的运算复杂度的定理。该复杂度以求解（5）的一个 $\mathcal{O}$（{v7）-KKT 解所需的 ∇f_1、 ∇f_1 和 f_2 的近端算子求值次数衡量；其证明将在第 5.1 小节中给出。

˜定理 3（算法 2 的复杂度）。假设假设1成立。设 f^*、 f、 f、 $D_{\mathbf{x}}$、 $D_{\mathbf{y}}$、 $\tilde{f}_{\text{hi}}$、 $\tilde{f}_{\text{low}}$ 和 f_{low} 由式（5）、（6）、（7）和（8）定义， $L_{\nabla f_1}$ 和 $L_{\nabla \tilde{f}_1}$ 由假设1给出， ε、 ρ、 x^0、y^0 和 z_ε 由算法 2 给出，并且

$$\begin{aligned}
&\widehat{L} = L_{\nabla f_1} + 2\varepsilon^{-1} L_{\nabla \tilde{f}_1},\ \hat{\alpha} = \min\Big\{1, \sqrt{4\varepsilon/(D_{\mathbf{y}}\widehat{L})}\Big\}, \\
&\hat{\delta} = (2 + \hat{\alpha}^{-1})(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\widehat{L} + \max\Big\{\varepsilon/D_{\mathbf{y}}, \hat{\alpha}\widehat{L}/4\Big\} D_{\mathbf{y}}^2, \\
&\widehat{C} = \frac{4\max\Big\{\frac{1}{2\widehat{L}}, \min\Big\{\frac{D_{\mathbf{y}}}{\varepsilon}, \frac{4}{\hat{\alpha}\widehat{L}}\Big\}\Big\}\Big[\hat{\delta} + 2\hat{\alpha}^{-1}(f^* - f_{\text{low}} + \varepsilon^{-1}(\tilde{f}_{\text{hi}} - \tilde{f}_{\text{low}}) + \varepsilon D_{\mathbf{y}}/4 + \widehat{L}(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2))\Big]}{\Big[(3\widehat{L} + \varepsilon/(2D_{\mathbf{y}}))^2/\min\{\widehat{L}, \varepsilon/(2D_{\mathbf{y}})\} + 3\widehat{L} + \varepsilon/(2D_{\mathbf{y}})\Big]^{-2}\varepsilon^3}, \\
&\widehat{K} = \Big\lceil 16(1 + f(x^0, y^0) - f_{\text{low}} + \varepsilon D_{\mathbf{y}}/4)\widehat{L}\varepsilon^{-2} + 32(1 + 4D_{\mathbf{y}}^2\widehat{L}^2\varepsilon^{-2})\varepsilon - 1\Big\rceil_+, \\
&\widehat{N} = \Big(\Big\lceil 96\sqrt{2}(1 + (24\widehat{L} + 4\varepsilon/D_{\mathbf{y}})\widehat{L}^{-1})\Big\rceil + 2\Big)\max\Big\{2, \sqrt{D_{\mathbf{y}}\widehat{L}\varepsilon^{-1}}\Big\} \\
&\quad \times ((\widehat{K} + 1)(\log\widehat{C})_+ + \widehat{K} + 1 + 2\widehat{K}\log(\widehat{K} + 1)).
\end{aligned} \tag{22}$$

那么，算法 2 输出（5）的一个近似解（x_ε、 y_ε），满足

$$\operatorname{dist}\left(0, \partial f(x_\varepsilon, y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon, y_\varepsilon) - (\rho\nabla_x\tilde{f}(x_\varepsilon, z_\varepsilon); 0)\right) \le \varepsilon, \quad \operatorname{dist}\left(0, \rho\partial_z\tilde{f}(x_\varepsilon, z_\varepsilon)\right) \le \varepsilon, \tag{23}$$

$$\tilde{f}(x_\varepsilon, y_\varepsilon) \le \min_z \tilde{f}(x_\varepsilon, z) + \varepsilon\left(1 + f(x^0, y^0) - f_{\text{low}} + 2\varepsilon^3(\widehat{L}^{-1} + 4D_{\mathbf{y}}^2\widehat{L}\varepsilon^{-2}) + D_{\mathbf{y}}\varepsilon/4\right), \tag{24}$$

在至多经过 N 次对 ∇f_1、 ∇f_1 以及 f_2 和 f_2 的近端算子的求值后

.

注 3. 从定理 3 可以看出， $\widehat{L} = \mathcal{O}(\varepsilon^{-1})$、$\hat{}\alpha = \mathcal{O}(\varepsilon)$、$\hat{}\delta = \mathcal{O}(\varepsilon^{-2})$、 $\widehat{C} = \mathcal{O}(\varepsilon^{-11})$、$\widehat{K} = \mathcal{O}(\varepsilon^{-3})$ 和 $\widehat{N} = \mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$。因此，算法 2 的运算复杂度为 $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$，按对 ∇f_1、 $\nabla\tilde{f}_1$ 以及 f_2 和 $\tilde{f}_2$ 的近端算子的求值次数衡量，用于求解满足以下条件的 (5) 的 $\mathcal{O}(\varepsilon)$-KKT 解 $(x_\varepsilon, y_\varepsilon)$:

$$\operatorname{dist}\left(0, \partial f(x_\varepsilon, y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon, y_\varepsilon) - (\rho\nabla_x\tilde{f}(x_\varepsilon, z_\varepsilon); 0)\right) \le \varepsilon, \quad \operatorname{dist}\left(0, \rho\partial_z\tilde{f}(x_\varepsilon, z_\varepsilon)\right) \le \varepsilon,$$
$$\tilde{f}(x_\varepsilon, y_\varepsilon) - \min_z \tilde{f}(x_\varepsilon, z) = \mathcal{O}(\varepsilon),$$

其中， z_ε 在算法 2 中给出，且 $\rho = \varepsilon^{-1}$。

3 约束双层优化

在本节中，我们考虑一个有约束双层优化问题[6]

$$\begin{aligned} f^* = \min \quad & f(x, y) \\ \text{s.t.} \quad & y \in \operatorname*{argmin}_z\{\tilde{f}(x, z) | \tilde{g}(x, z) \le 0\}, \end{aligned} \tag{25}$$

~ ~其中 f和 f 满足假设1。回顾假设1可知， $\mathcal{X} = \text{dom } f_2$且 $\mathcal{Y} = \text{dom } f_2$。现在，我们对问题 (25) 作出一些额外假设。

~假设 2。 (i) f和 f分别在 $\mathcal{X} \times \mathcal{Y}$上是 L_f-Lipschitz 连续和 $L_{\tilde{f}}$-Lipschitz 连续。

(ii) $\tilde{g}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^l$ 在 $\mathcal{X} \times \mathcal{Y}$上是 $L_{\nabla\tilde{g}}$-光滑且 $L_{\tilde{g}}$-Lipschitz 连续。

(iii) $\tilde{g}_i(x, \cdot)$ 是凸的，并且对每个 $x \in \mathcal{X}$，都存在 $\hat{}z_x \in \mathcal{Y}$，使得对所有 $i = 1, 2, \ldots, l$以及 $G := \min\{-\tilde{g}_i(x, \hat{}z_x) | x \in \mathcal{X},\ i = 1, \ldots, l\} > 0$，均有 $\tilde{g}_i(x, \hat{}z_x) < 0$。[7]

为方便记号，我们定义

$$\tilde{f}^*(x) := \min_z\{\tilde{f}(x, z) | \tilde{g}(x, z) \le 0\}, \tag{26}$$

$$\tilde{f}^*_{\text{hi}} := \sup\{\tilde{f}^*(x) | x \in \mathcal{X}\}, \tag{27}$$

$$\tilde{g}_{\text{hi}} := \max\{\|\tilde{g}(x, y)\| \big| (x, y) \in \mathcal{X} \times \mathcal{Y}\}. \tag{28}$$

由假设 2(ii)可知

$$\|\nabla\tilde{g}(x, y)\| \le L_{\tilde{g}} \qquad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{29}$$

此外，根据假设1和假设2以及 $\mathcal{X}$ 和 $\mathcal{Y}$的紧致性，可以看出 $\tilde{g}_{\text{hi}}$ 和 G~均为有限值。此外，正如引理3(ii)将要证明的， f^*_{hi} 是有限值。

本节旨在提出用于求解问题 (25)的一阶罚函数方法。为此，~我们先为下层优化问题 $y \in$使目标函数取最小值的自变量$\{f(x, z) | \tilde{g}(x, z) \le_z$引入一个罚函数
$0\}$ 问题 (25)的，其定义为

$$\widetilde{P}_\mu(x, z) = \tilde{f}(x, z) + \mu\left\|[\tilde{g}(x, z)]_+\right\|^2 \tag{30}$$

对于一个罚 参数 $\mu > 0$。注意，问题 (25) 可以近似求解为一个无约束 n-
双层优化问题 层优化问题

$$f^*_\mu = \min_{x,y}\left\{f(x, y) | y \in \operatorname*{argmin}_z \widetilde{P}_\mu(x, z)\right\}. \tag{31}$$

[6]为方便起见，问题 (25) 被称为约束双层优化问题，因为其下层部分至少包含一个显式约束。[7]该假设的后半部分可以弱化为：问题 (25) 的下层部分满足逐点 Slater 条件，即，对每个 $x \in \mathcal{X}$，存在$\hat{z}_x \in \mathcal{Y}$ 使得 $\tilde{g}(x, \hat{z}_x) < 0$。事实上，如果 $G > 0$，则假设2(iii)显然成立。否则，可以改为求解问题 (25) 的扰动对应问题，其中将 $\tilde{g}(x, z)$ 替换为 $\tilde{g}(x, z) - \epsilon$，其中 $\epsilon > 0$ 取适当值，从而满足假设2(iii)。

此外，根据第2节的研究，问题（31）可以近似求解为罚函数问题

$$\min_{x,y} f(x,y) + \rho\Big(\widetilde{P}_\mu(x,y) - \min_z \widetilde{P}_\mu(x,z)\Big) \tag{32}$$

对于某个适当的 $\rho > 0$。还可以看出，问题 (32) 等价于极小极大问题

$$\min_{x,y}\max_z P_{\rho,\mu}(x,y,z), \quad \text{where} \quad P_{\rho,\mu}(x,y,z) \coloneqq f(x,y) + \rho(\widetilde{P}_\mu(x,y) - \widetilde{P}_\mu(x,z)). \tag{33}$$

鉴于 (30)、(33) 和假设1(i)，$P_{\rho,\ \mu}$ 可改写为

$$\begin{aligned} P_{\rho,\mu}(x,y,z) = &\big(f_1(x,y) + \rho\tilde{f}_1(x,y) + \rho\mu\,\|[\tilde{g}(x,y)]_+\|^2 - \rho\tilde{f}_1(x,z) - \rho\mu\,\|[\tilde{g}(x,z)]_+\|^2\big) \\ &+ \big(f_2(x) + \rho\tilde{f}_2(y) - \rho\tilde{f}_2(z)\big). \end{aligned} \tag{34}$$

由此以及假设1和假设2，可以看出，$P_{\rho,\ \mu}$ 具有以下良好性质。

- $P_{\rho,\ \mu}$ 是光滑函数 $f_1(x,y)+\rho\tilde{f}_1(x,y)+\rho\mu\,\|[\tilde{g}(x,y)]_+\|^2-\rho\tilde{f}_1(x,z)-\rho\mu\,\|[\tilde{g}(x,z)]_+\|^2$ ~与可能非光滑函数 $f_2(x) + \rho f_2(y) - \rho f_2(z)$ 之和，前者具有 Lipschitz 连续梯度，后者的近端算子可精确计算；

- $P_{\rho,\ \mu}$ 关于 (x,y) 是非凸的，但关于 z 是凹的。

由于 $P_{\rho,\ \mu}$具有这种良好结构，可以通过 [41, 算法 2] 提出的一阶方法找到极小极大问题 (33) 的近似驻点（参见附录A中的算法 6）。

基于上述观察，我们现在可以通过求解一系列极小极大问题，或求解一个形式如 (33) 的极小极大问题，来为约束双层优化问题 (25) 提出罚函数法。具体而言，我们首先通过求解一系列极小极大问题，为 (25) 提出一种理想的罚函数法（见算法 3）。然后，我们通过求出单个极小极大问题的一个近似驻点，为 (25) 提出一种实用罚函数法（见算法4）。

算法 3 问题 (25) 的一种理想罚函数法

输入：正数序列 $\{\rho_k\}, \{\mu_k\}$ 和 $\{\epsilon_k\}$，满足 $\lim_{k\to\infty}(\rho_k, \mu_k, \epsilon_k) = (\infty, \infty, 0)$。1: 对于 $k = 0$，1，2，.....，执行2: 求解问题 (33)，得到其一个 ϵ_k-最优解 (x^k, y^k, z^k)，其中 $(\rho, \mu) = (\rho_k, \mu_k)$。3: 结束循环

为了研究算法 3 的收敛性，我们对问题 (25) 的下层优化问题的解集作出如下误差界假设。文献中已在集值映射的背景下研究过这类误差界（例如，参见 [14]）。

~假设 3. 存在 $\theta > 0$ 和一个非递减函数 $\omega: \mathbb{R}_+ \to \mathbb{R}_+$，满足 $\lim_{\theta\downarrow 0}\omega(\theta) = 0$~，使得对所有 $x \in \mathcal{X}$, $z \in \mathcal{S}_0(x)$ 和 $\theta \in [0,\theta]$, 均有 $\mathrm{dist}(z, \mathcal{S}_\theta(x)) \le \omega(\theta)$，其中

$$\mathcal{S}_\theta(x) \coloneqq \operatorname*{argmin}_z\{\tilde{f}(x,z) : \|[\tilde{g}(x,z)]_+\| \le \theta\}.$$

现在我们可以陈述算法 3 的一个收敛性结果，其证明将在第 5.2 节给出。

定理 4（算法 3 的收敛性）。假设假设 1-3 成立，且 $\{(x^k, y^k, z^k)\}$由算法 3 生成。那么，$\{(x^k, y^k)\}$ 的任意聚点都是问题 (25) 的最优解。

注意到，(33) 是一个非凸-凹极小极大问题。对于任意的 $\epsilon > 0$，通常很难找到 (33) 的 ϵ-最优解。因此，算法 3 通常无法实施。接下来，我们通过应用算法 6（见附录A），并对 ρ 和 μ进行适当选择，提出一种针对问题 (25) 的实用罚函数法，以求得 (33) 的近似驻点。

算法4 针对问题 (25) 的实用罚函数法

输入：$\varepsilon \in (0, 1/4]$, $\rho = \varepsilon^{-1}$, $\mu = \varepsilon^{-2}$, $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$，其中 $\widetilde{P}_\mu(x^0, y^0) \le \min_y \widetilde{P}_\mu(x^0, y) + \varepsilon$。1: 调用附录A中的算法 6，输入 $\epsilon \leftarrow \varepsilon$, $\epsilon_0 \leftarrow \varepsilon^{5/2}$, $\hat{}x^0 \leftarrow (x^0, y^0)$, $\hat{}y^0 \leftarrow y^0$ 和 $L_{\nabla h} \leftarrow L_{\nabla f_1} + 2\rho L_{\nabla\tilde{f}_1} + 4\rho\mu(\tilde{g}_{\mathrm{hi}}L_{\nabla\tilde{g}} + L_{\tilde{g}}^2)$，以求得问题 (33) 的一个 ε-驻点 $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$，其中$\rho = \varepsilon^{-1}$ 和 $\mu = \varepsilon^{-2}$。2: 输出：$(x_\varepsilon, y_\varepsilon)$。

注释4. (i) 算法4的初始点 (x^0, y^0) 可通过类似于注释1中所述的过程求得，只需将 $\tilde{f}$ 替换为 $\widetilde{P}_\mu$；(ii) $\rho = \varepsilon^{-1}$ 和 $\mu = \varepsilon^{-2}$ 的选择对于算法4求得问题 (25) 的一个 $\mathcal{O}(\varepsilon)$-KKT 解时的 ε 阶运算复杂度至关重要。与所有可能的 ρ 和 μ的选择相比，该选择可获得最佳运算复杂度，这一点可从定理5的证明中看出。直观而言，下层罚参数 μ 的阶必须大于上层罚参数 ρ 的阶，这样下层约束违反所引入的误差才不会被放大（见引理8中的 (92)）。(iii) 如定理 6（见附录A）所示，通过对 (33) 应用算法 6，可在算法4的步骤1中成功求得 (33) 的一个 ϵ-驻点；(iv) 为简便起见，算法4只求解一个具有静态罚参数和容差参数的 (33) 形式子问题。不过，算法4也可修改为一个可能在实践中更高效的算法，通过求解一系列具有动态罚参数和容差参数的 (33) 形式子问题来实现。

为了刻画算法4找到的近似解，我们接下来引入问题 (25) 的 ε-KKT 解这一概念。

根据.....的定义，$\tilde{f}^*$ 在 (26) 中的定义，问题 (25) 可视为

$$\min_{x,y}\{f(x,y)|\tilde{f}(x,y) \le \tilde{f}^*(x),\ \tilde{g}(x,y) \le 0\}. \tag{35}$$

其对应的拉格朗日函数为

$$\mathcal{L}(x,y,\rho,\lambda) = f(x,y) + \rho(\tilde{f}(x,y) - \tilde{f}^*(x)) + \langle\lambda, \tilde{g}(x,y)\rangle. \tag{36}$$

按照经典约束优化的思路，人们自然会关注 (35)（或等价地，问题 (25)）的一个 KKT 解 (x, y)，即 (x, y) 满足

$$\tilde{f}(x,y) \le \tilde{f}^*(x), \quad \tilde{g}(x,y) \le 0, \quad \rho(\tilde{f}(x,y) - \tilde{f}^*(x)) = 0, \quad \langle\lambda, \tilde{g}(x,y)\rangle = 0, \tag{37}$$

并且 (x, y) 是问题的一个驻点

$$\min_{x',y'} \mathcal{L}(x', y', \rho, \lambda) \tag{38}$$

对于某些 $\rho \ge 0$ 和 $\lambda \in \mathbb{R}^l_+$。然而，由于问题结构复杂，通常很难刻画 (38) 的驻点。另一方面，由引理 3 和 (36) 可知，问题 (38) 等价于以下极小极大问题

$$\min_{x',y',\tilde{\lambda}'} \max_{z'} \left\{f(x',y') + \rho\big(\tilde{f}(x',y') - \tilde{f}(x',z') - \langle\tilde{\lambda}', \tilde{g}(x',z')\rangle\big) + \langle\lambda, \tilde{g}(x',y')\rangle + \mathscr{I}_{\mathbb{R}^l_+}(\tilde{\lambda}')\right\},$$

˜其驻点 (x, y, λ, z) 根据定义2和假设1和假设2满足

$$0 \in \partial f(x,y) + \rho\partial\tilde{f}(x,y) - \rho(\nabla_x\tilde{f}(x,z) + \nabla_x\tilde{g}(x,z)\tilde{\lambda}; 0) + \nabla\tilde{g}(x,y)\lambda, \tag{39}$$

$$0 \in \rho(\partial_z\tilde{f}(x,z) + \nabla_z\tilde{g}(x,z)\tilde{\lambda}), \tag{40}$$

$$\tilde{\lambda} \in \mathbb{R}^l_+, \quad \tilde{g}(x,z) \le 0, \quad \langle\tilde{\lambda}, \tilde{g}(x,z)\rangle = 0.^8 \tag{41}$$

基于这一观察以及 (37) 等价于以下问题这一事实

$$\tilde{f}(x,y) = \tilde{f}^*(x), \quad \tilde{g}(x,y) \le 0, \quad \langle\lambda, \tilde{g}(x,y)\rangle = 0, \tag{42}$$

因此，我们转而关注问题 (25) 的（弱）KKT 解及其非精确对应问题，定义如下。

˜定义 4. 若存在 $(z, \rho, \lambda, \lambda) \in \mathbb{R}^m \times \mathbb{R}_+ \times \mathbb{R}^l_+ \times \mathbb{R}^l_+$ 使得 (39)-(42) 成立，则称这对 (x, y) 是问题 (25) 的 KKT 解。此外，对于任意 $\varepsilon > 0$，若存在 $(z, \rho, \lambda, \tilde{\lambda}) \in \mathbb{R}^m \times \mathbb{R}_+ \times \mathbb{R}^l_+ \times \mathbb{R}^l_+$ 使得，则称 (x, y) 是问题 (25) 的 ε-KKT 解。

$$\begin{aligned}
&\operatorname{dist}\left(0, \partial f(x,y) + \rho\partial\tilde{f}(x,y) - \rho(\nabla_x\tilde{f}(x,z) + \nabla_x\tilde{g}(x,z)\tilde{\lambda}; 0) + \nabla\tilde{g}(x,y)\lambda\right) \le \varepsilon,\\
&\operatorname{dist}\left(0, \rho(\partial_z\tilde{f}(x,z) + \nabla_z\tilde{g}(x,z)\tilde{\lambda})\right) \le \varepsilon,\\
&\|[\tilde{g}(x,z)]_+\| \le \varepsilon, \quad |\langle\tilde{\lambda}, \tilde{g}(x,z)\rangle| \le \varepsilon,\\
&|\tilde{f}(x,y) - \tilde{f}^*(x)| \le \varepsilon, \quad \|[\tilde{g}(x,y)]_+\| \le \varepsilon, \quad |\langle\lambda, \tilde{g}(x,y)\rangle| \le \varepsilon,
\end{aligned}$$

˜其中 f^* 在式 (26) 中定义。

[8] (41) 中的关系等价于 $0 \in -\tilde{g}(x,z) + \partial\mathscr{I}_{\mathbb{R}^l_+}$ ({v6)。

现在，我们可以给出算法4的运算复杂度，该复杂度以求值 ∇f_1、 ∇f_1、 $\nabla \tilde{g}$ 以及 f_2 和 f_2的近端算子的次数来衡量，用于求解式 (25) 的一个 $\mathcal{O}(\varepsilon)$-KKT 解。其证明将在第5.2小节给出。

~定理5（算法4的复杂度）。假设假设1和假设2成立。设 f^*, f, f, $\tilde{g}$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $f_{\rm hi}$, $f_{\rm low}$, $f_{\rm low}$, f^*, ~ ~ ~

4.1 线性上层与二次下层的无约束双层优化

在本小节中，我们考虑如下形式的线性上层、二次下层无约束双层优化问题：

$$\begin{aligned} \min \quad & c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x) \\ \text{s.t.} \quad & y \in \operatorname*{argmin}_z x^T \widetilde{A} z + z^T \widetilde{B} z + \tilde{d}^T z + \mathscr{I}_{[-1,1]^m}(z), \end{aligned} \tag{52}$$

其中，$\widetilde{A} \in \mathbb{R}^{n\times m}$、$\widetilde{B} \in \mathbb{R}^{m\times m}$、$c \in \mathbb{R}^n$、$d$、$\tilde{d} \in \mathbb{R}^m$ 和 $\mathscr{I}_{[-1,1]^n}(\cdot)$ 与 $\mathscr{I}_{[-1,1]^m}(\cdot)$ 分别是 $[-1,1]^n$ 和 $[-1,1]^m$ 的示性函数。[9]

对于每一对 (n, m)，我们随机生成问题(52)的10个实例。具体而言，我们首先随机生成 c、d，其所有元素均独立地取自标准正态分布；$\widetilde{B}$ 和 A 的所有元素均独立地取自均值为0、标准差为0.01的正态分布。然后，我们通过执行 $U = \mathrm{orth}(\mathrm{randn}(m))$ 随机生成一个正交矩阵 U，再生成一个 $m \times m$ 对角矩阵 D，其对角元素独立地取自均值为0、标准差为0.01的正态分布，并将其投影到 $\mathbb{R}_+$，随后设定 $\widetilde{B} = UDU^T$。此外，我们随机生成 $\hat{y} \in [-1,1]^m$，其所有元素均独立地取自均值为0、标准差为0.1的正态分布，然后将其投影到 $[-1,1]^m$；并选取 d，使得在 $x = 0$ 条件下，$\hat{y}$ 是问题(52)下层优化的最优解。

请注意，(52) 是 (5) 的一个特例，其中 $f(x,y)$ 可通过算法 2 进行适当求解。为提高效率，我们实现了算法 2 的一个变体，对惩罚参数和容差参数进行动态更新。具体而言，我们设定 $\rho_k = 5^{k-1}$，$\varepsilon_k = \rho_k^{-1}$ 和 $x_{\varepsilon_{-1}} = 0$。对于每个 $k \geq 0$，我们以 $(\varepsilon, \rho) = (\varepsilon_k, \rho_k)$ 和 $(x_{\varepsilon_{k-1}}, \tilde{y}_{\varepsilon_{k-1}})$ 为初始点运行算法 2，生成 $(x_{\varepsilon_k}, y_{\varepsilon_k})$，其中 $\tilde{y}_{\varepsilon_{k-1}} \in \mathrm{argmin}_z f(x_{\varepsilon_{k-1}}, z)$ 由 CVX [19] 求得。一旦 $\varepsilon_{\bar{k}} \leq 10^{-4}$ 且 $(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ 满足以下条件，我们便终止该过程

$$\tilde{f}(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}}) - \min_z \tilde{f}(x_{\varepsilon_{\bar{k}}}, z) \leq 10^{-4}$$

对于某个 k，并将 $(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ 作为问题 (52) 的近似解输出，其中 $\min_z f(x_{\varepsilon_{\bar{k}}}, z)$ 的值由 CVX 计算。

上述针对随机生成实例的算法 2 变体的计算结果列于表1。具体而言，n 和 m 的值列在前两列中。对于每一对 (n, m)，其余各列给出了 10 个随机实例上的平均初始目标值 $f(x_{\varepsilon_{-1}}, \tilde{y}_{\varepsilon_{-1}})$ 和平均最终目标值 $f(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$。可以观察到，与初始点 $(x_{\varepsilon_{-1}}, \tilde{y}_{\varepsilon_{-1}})$ 相比，该方法找到的近似解 $(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ 显著降低了目标函数值。

n	m	初始目标值	最终目标值
100	100	-0.35	-101.67
200	200	-0.53	-194.91
300	300	-0.48	-307.43
400	400	-0.44	-401.71
500	500	-0.05	-527.45
600	600	0.99	-644.53
700	700	0.49	-759.54
800	800	-1.23	-872.77
900	900	-2.07	-1004.27
1000	1000	-1.06	-1107.61

表1：问题 (52) 的数值结果

4.2 约束双层线性优化

在本小节中，我们考虑如下形式的约束双层线性优化

$$\begin{aligned} \min \quad & c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x) \\ \text{s.t.} \quad & y \in \operatorname*{argmin}_z \left\{ \tilde{d}^T z + \mathscr{I}_{[-1,1]^m}(z) \middle| \widetilde{A} x + \widetilde{B} z - \tilde{b} \leq 0 \right\}, \end{aligned} \tag{53}$$

[9]记号 $[-1,1]^n$ 表示集合 $\{x \in \mathbb{R}^n | x_i \in [-1,1],\ i = 1, \ldots, n.\}$。

其中，$c \in \mathbb{R}^n$、d、$\tilde{d} \in \mathbb{R}^m$、$\tilde{b} \in \mathbb{R}^l$、$\widetilde{A} \in \mathbb{R}^{l\times n}$、$\widetilde{B} \in \mathbb{R}^{l\times m}$、$\mathscr{I}_{[-1,1]^n}(\cdot)$ 和 $\mathscr{I}_{[-1,1]^m}(\cdot)$ 分别是 $[-1,1]^n$ 和 $[-1,1]^m$ 的示性函数。

对于每组参数 (n,m,l)，我们随机生成问题（53）的 10 个实例。具体而言，我们首先随机生成 c 和 d，其所有元素均独立地从标准正态分布 ~ ~中抽取。然后，我们随机生成 A 和 B，其所有元素均独立地从均值为 0、标准差为 0.01 的正态分布中抽取。此外，我们随机生成 $\hat{y} \in [-1,1]^m$，其所有元素均独立地从均值为 0、标准差为~ ~0.1 的正态分布中抽取，随后将其投影到 $[-1,1]^m$，并选择 d 和 b，使得 $\hat{y}$ 是问题（53）在 $x=0$下层优化的最优解。

注意，问题（53）是问题（25）的一个特例，其中 $f(x,y)=c^Tx+d^Ty+\mathscr{I}_{[-1,1]^n}(x)$，$\tilde{f}(x,z)=\tilde{d}^Tz+\sim \sim \mathscr{I}_{[-1,1]^m}(z)$，且 $\tilde{g}(x,z)=Ax+Bz-b$，因此可以用算法4妥善求解。为提高效率，我们实现了算法4的一个变体，对惩罚参数和容差参数进行动态更新。具体而言，我们设定 $\rho_k=5^{k-1}$，$\mu_k=\rho_k^2$，$\varepsilon_k=\rho_k^{-1}$ 和 $x_{\varepsilon_{-1}}=0$。对于每个 $k\geq 0$，我们以 $(\varepsilon,\rho,\mu)=(\varepsilon_k,\rho_k,\mu_k)$ 和 $(x_{\varepsilon_{k-1}},\tilde{y}_{\varepsilon_{k-1}})$ 为初始点运行算法4，生成 $(x_{\varepsilon_k},y_{\varepsilon_k})$，~ ~ ~其中 $\tilde{y}_{\varepsilon_{k-1}}$ 满足 $P_{\mu_k}(x_{\varepsilon_{k-1}},\tilde{y}_{\varepsilon_{k-1}})\leq \min_z P_{\mu_k}(x_{\varepsilon_{k-1}},z)+\varepsilon_k$，其中 P_{μ_k} 由式（30）给出，可通过加速近端梯度法 [48]求得。一旦 $\varepsilon_{\bar{k}}\leq 10^{-4}$且 $(x_{\varepsilon_{\bar{k}}},y_{\varepsilon_{\bar{k}}})$ 满足下式，我们便终止该过程

$$\|[\tilde{g}(x_{\varepsilon_{\bar{k}}},y_{\varepsilon_{\bar{k}}})]_+\|\leq 10^{-4},\quad \tilde{f}(x_{\varepsilon_{\bar{k}}},y_{\varepsilon_{\bar{k}}})-\tilde{f}^*(x_{\varepsilon_{\bar{k}}})\leq 10^{-4}$$

– ~对于某个 k，并输出 $(x_{\varepsilon_{\bar{k}}},y_{\varepsilon_{\bar{k}}})$ 作为问题 (53) 的近似解，其中 f^*在 (26) 中定义，且~值 $f^*(x_{\varepsilon_{\bar{k}}})$ 由 CVX [19]计算。

上述针对随机生成实例的算法4变体的计算结果见表2。具体而言，n、m 和 l 的值列于前三列。对于每个三元组 (n,m,l)，其余各列给出了 10 个随机实例上的平均初始目标值 $f(x_{\varepsilon_{-1}},\hat{y})$（其中 $\hat{y}$由上述方式生成[10]）以及平均最终目标值 $f(x_{\varepsilon_{\bar{k}}},y_{\varepsilon_{\bar{k}}})$。可以看出，与初始点 $(x_{\varepsilon_{-1}},\hat{y})$ 相比，该方法得到的近似解 $(x_{\varepsilon_{\bar{k}}},y_{\varepsilon_{\bar{k}}})$ 显著降低了目标函数值。

n	m	l	初始目标值	最终目标值
100	100	5	-0.51	-34.83
200	200	10	-0.15	-121.41
300	300	15	1.56	-208.44
400	400	20	-0.04	-298.25
500	500	25	1.45	-384.77
600	600	30	0.75	-470.31
700	700	35	0.09	-568.26
800	800	40	-0.98	-629.61
900	900	45	1.21	-689.00
1000	1000	50	1.44	-781.79

表2：问题（53）的数值结果

5 主要结果的证明

本节将证明我们在第 2 节和第 3 节中提出的主要结果，尤其是定理1–5。

5.1 第 2 节主要结果的证明

在本小节中，我们证明定理1、2和3。首先，我们在下文建立一个引理，随后将用它来证明定理1。

引理1. 假设假设1成立，且 $(x_\varepsilon,y_\varepsilon,z_\varepsilon)$ 是问题（11）的 ϵ-最优解~，其中 $\epsilon>0$。设 f、f、f^*、$f_{\rm low}$ 和 ρ 分别由 (5)、(8) 和 (11) 给出。则有

$$\tilde{f}(x_\varepsilon,y_\varepsilon)\leq \min_z \tilde{f}(x_\varepsilon,z)+\rho^{-1}(f^*-f_{\rm low}+2\epsilon),\quad f(x_\varepsilon,y_\varepsilon)\leq f^*+2\epsilon.$$

[10]注意，$(x_{\varepsilon_{-1}},\tilde{y}_{\varepsilon_{-1}})$可能不是(53)的可行点。尽管如此，由于$x_{\varepsilon_{-1}}=0$以及(53)实例的特定生成方式，$(x_{\varepsilon_{-1}},\hat{y})$是(53)的可行点。此外，(53)可以看作关于变量 x的隐式优化问题。因此，将 $f(x_{\varepsilon_{-1}},\hat{y})$作为用于比较的初始目标值是合理的。

证明. 由于 $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ 是 (11) 的 ϵ-最优解，根据定义1可得

$$\max_z P_\rho(x_\varepsilon, y_\varepsilon, z) \le P_\rho(x_\varepsilon, y_\varepsilon, z_\varepsilon) + \epsilon, \quad P_\rho(x_\varepsilon, y_\varepsilon, z_\varepsilon) \le \min_{x,y} \max_z P_\rho(x, y, z) + \epsilon.$$

将这些不等式相加可得

$$\max_z P_\rho(x_\varepsilon, y_\varepsilon, z) \le \min_{x,y} \max_z P_\rho(x, y, z) + 2\epsilon. \tag{54}$$

设 (x^*, y^*) 是 (5) 的一个最优解。由此可得 $f(x^*, y^*) = f^*$ 且 $f(x^*, y^*) = \min_z f(x^*, z)$。根据这些结果以及 (11) 中 P_ρ 的定义，可得

$$\max_z P_\rho(x^*, y^*, z) = f(x^*, y^*) + \rho(\tilde{f}(x^*, y^*) - \min_z \tilde{f}(x^*, z)) = f(x^*, y^*) = f^*,$$

这意味着

$$\min_{x,y} \max_z P_\rho(x, y, z) \le \max_z P_\rho(x^*, y^*, z) = f^*. \tag{55}$$

由 (11)、(54) 和 (55) 可得

$$f(x_\varepsilon, y_\varepsilon) + \rho(\tilde{f}(x_\varepsilon, y_\varepsilon) - \min_z \tilde{f}(x_\varepsilon, z)) \overset{(11)}{=} \max_z P_\rho(x_\varepsilon, y_\varepsilon, z) \overset{(54)(55)}{\le} f^* + 2\epsilon,$$

结合 $f(x_\varepsilon, y_\varepsilon) - \min_z f(x_\varepsilon, z) \ge 0$，可得

$$f(x_\varepsilon, y_\varepsilon) \le f^* + 2\epsilon, \quad \tilde{f}(x_\varepsilon, y_\varepsilon) \le \min_z \tilde{f}(x_\varepsilon, z) + \rho^{-1}\left(f^* - f(x_\varepsilon, y_\varepsilon) + 2\epsilon\right).$$

由此及 (8) 可直接得到本引理的结论。 □

现在我们可以证明定理1了。

定理1的证明。设 $\{(x^k, y^k, z^k)\}$ 由算法 1 生成，且 $\lim_{k\to\infty}(\rho_k, \epsilon_k) = (\infty, 0)$。如有必要，考虑一个收敛子列，不失一般性地假设 $\lim_{k\to\infty}(x^k, y^k) = (x^*, y^*)$。现在证明 (x^*, y^*) 是问题（5）的最优解。事实上，由于 (x^k, y^k, z^k) 是满足 $\rho = \rho_k$ 的 (11) 的 ϵ_k- 最优解，根据引理 1，取 $(\rho, \epsilon) = (\rho_k, \epsilon_k)$ 和 $(x_\varepsilon, y_\varepsilon) = (x^k, y^k)$，可得

$$\tilde{f}(x^k, y^k) \le \min_z \tilde{f}(x^k, z) + \rho_k^{-1}(f^* - f_{\text{low}} + 2\epsilon_k), \quad f(x^k, y^k) \le f^* + 2\epsilon_k.$$

由 f 和 $\tilde{f}$ 的连续性，$\lim_{k\to\infty}(x^k, y^k) = (x^*, y^*)$，$\lim_{k\to\infty}(\rho_k, \epsilon_k) = (\infty, 0)$，并对上述关系两边取 $k \to \infty$ 的极限，可得 $f(x^*, y^*) \le \min_z f(x^*, z)$ 且 $f(x^*, y^*) \le f^*$，这显然意味着 $y^* \in \operatorname{argmin}_z f(x^*, z)$ 且 $f(x^*, y^*) = f^*$。因此，(x^*, y^*) 是 (5) 的最优解，证毕。 □

下面我们证明定理2。

定理2的证明。由于 (x, y) 是问题（5）的一个 ε-KKT 解，且与之关联的为 $\rho \ge \rho_0$，由定义3可知，存在 $z \in \mathbb{R}^m$ 使得

$$\|\nabla_x f(x, y) + \rho\nabla_x \tilde{f}(x, y) - \rho\nabla_x \tilde{f}(x, z)\| \le \varepsilon, \tag{56}$$
$$\|\nabla_y f(x, y) + \rho\nabla_y \tilde{f}(x, y)\| \le \varepsilon, \tag{57}$$
$$\rho\|\nabla_y \tilde{f}(x, z)\| \le \varepsilon, \quad \tilde{f}(x, y) - \min_{z'} \tilde{f}(x, z') \le \varepsilon. \tag{58}$$

利用 (56)、三角不等式，以及对于所有 $x' \in \Omega$，$x \in \Omega$、$\nabla f(x', \cdot)$ 是 L_1-Lipschitz 连续且 $\nabla^2 \tilde{f}(x', \cdot)$ 是 L_2-Lipschitz 连续的假设，我们得到

$$\begin{aligned}
&\|\nabla_x f(x, y^*(x)) + \rho\nabla^2_{xy} \tilde{f}(x, y^*(x))(y - z)\| \\
&\le \|\nabla_x f(x, y) + \rho\nabla_x \tilde{f}(x, y) - \rho\nabla_x \tilde{f}(x, z)\| + \|\nabla_x f(x, y^*(x)) - \nabla_x f(x, y)\| \\
&\quad + \rho\|\nabla_x \tilde{f}(x, y^*(x)) + \nabla^2_{xy} \tilde{f}(x, y^*(x))(y - y^*(x)) - \nabla_x \tilde{f}(x, y)\| \\
&\quad + \rho\|\nabla_x \tilde{f}(x, z) - \nabla_x \tilde{f}(x, y^*(x)) - \nabla^2_{xy} \tilde{f}(x, y^*(x))(z - y^*(x))\| \\
&\le \varepsilon + L_1\|y - y^*(x)\| + \frac{\rho L_2}{2}\|y - y^*(x)\|^2 + \frac{\rho L_2}{2}\|z - y^*(x)\|^2.
\end{aligned} \tag{59}$$

由 (57)、(58) 以及与推导 (59) 类似的论证可得

$$\begin{aligned}
&\|\nabla_y f(x,y^*(x)) + \rho\nabla^2_{yy}\tilde{f}(x,y^*(x))(y-z)\| \\
&\le \|\nabla_y f(x,y^*(x)) - \nabla_y f(x,y)\| + \|\nabla_y f(x,y) + \rho\nabla_y\tilde{f}(x,y)\| + \rho\|\nabla_y\tilde{f}(x,z)\| \\
&\quad + \rho\|\nabla_y\tilde{f}(x,y^*(x)) + \nabla^2_{yy}\tilde{f}(x,y^*(x))(y-y^*(x)) - \nabla_y\tilde{f}(x,y)\| \\
&\quad + \rho\|\nabla_y\tilde{f}(x,z) - \nabla_y\tilde{f}(x,y^*(x)) - \nabla^2_{yy}\tilde{f}(x,y^*(x))(z-y^*(x))\| \\
&\le L_1\|y-y^*(x)\| + 2\varepsilon + \frac{\rho L_2}{2}\|y-y^*(x)\|^2 + \frac{\rho L_2}{2}\|z-y^*(x)\|^2.
\end{aligned}$$

利用该不等式、(16)、(19) 和 (59)，我们得到

$$\begin{aligned}
&\|\nabla\Phi(x)\| = \|\nabla_x f(x,y^*(x)) - \nabla^2_{xy}\tilde{f}(x,y^*(x))[\nabla^2_{yy}\tilde{f}(x,y^*(x))]^{-1}\nabla_y f(x,y^*(x))\| \\
&= \|\nabla_x f(x,y^*(x)) + \rho\nabla^2_{xy}\tilde{f}(x,y^*(x))(y-z) \\
&\quad - \nabla^2_{xy}\tilde{f}(x,y^*(x))[\nabla^2_{yy}\tilde{f}(x,y^*(x))]^{-1}[\nabla_y f(x,y^*(x)) + \rho\nabla^2_{yy}\tilde{f}(x,y^*(x))(y-z)]\| \\
&\le \|\nabla_x f(x,y^*(x)) + \rho\nabla^2_{xy}\tilde{f}(x,y^*(x))(y-z)\| \\
&\quad + \|\nabla^2_{xy}\tilde{f}(x,y^*(x))[\nabla^2_{yy}\tilde{f}(x,y^*(x))]^{-1}\| \cdot \|\nabla_y f(x,y^*(x)) + \rho\nabla^2_{yy}\tilde{f}(x,y^*(x))(y-z)\| \\
&\le (2C+1)\varepsilon + (C+1)\Big(L_1\|y-y^*(x)\| + \frac{\rho L_2}{2}\|y-y^*(x)\|^2 + \frac{\rho L_2}{2}\|z-y^*(x)\|^2\Big). \qquad (60)
\end{aligned}$$

根据假设，$x\in\Omega\subset\mathcal{N}$，且对于所有 $x'\in\mathcal{N}$，$f(x',\cdot)$ 关于模数 $\sigma>0$ 是强凸的。由此以及 (58) 和 (15) 中 $y^*(x)$ 的定义可得

$$\|y-y^*(x)\|^2 \le 2\sigma^{-1}\big(\tilde{f}(x,y) - \min_{z'}\tilde{f}(x,z')\big) \le 2\sigma^{-1}\varepsilon, \qquad (61)$$

结合 $x\in\Omega$ 和 (18)，可推出 $\|\nabla_y f(x,y)\|\le\bar{C}$。利用这一结论、(57)、(58)、$x\in\Omega\subset\mathcal{N}$、$\nabla_y f(x,y^*(x))=0$，以及对于所有 $x'\in\mathcal{N}$，$f(x',\cdot)$ 关于模数 $\sigma>0$ 是强凸这一假设，可得

$$\begin{aligned}
\|y-y^*(x)\| &\le \sigma^{-1}\|\nabla_y\tilde{f}(x,y) - \nabla_y\tilde{f}(x,y^*(x))\| = \sigma^{-1}\|\nabla_y\tilde{f}(x,y)\| \\
&\le (\rho\sigma)^{-1}(\|\nabla_y f(x,y) + \rho\nabla_y\tilde{f}(x,y)\| + \|\nabla_y f(x,y)\|) \le (\rho\sigma)^{-1}(\varepsilon+\bar{C}), \qquad (62) \\
\|z-y^*(x)\| &\le \sigma^{-1}\|\nabla_y\tilde{f}(x,z) - \nabla_y\tilde{f}(x,y^*(x))\| = \sigma^{-1}\|\nabla_y\tilde{f}(x,z)\| \le (\rho\sigma)^{-1}\varepsilon. \qquad (63)
\end{aligned}$$

由 (61)、(62) 和 (19) 中 θ 的定义可得 $\|y-y^*(x)\|\le\theta$。由此结合 (60) 和 (63)，可得 (20) 成立。此外，根据 (19)，有 $\theta\le\sqrt{2\sigma^{-1}\varepsilon}$，并且

$$\rho\theta^2 = \min\big\{\rho^{-1}\sigma^{-2}(\varepsilon+\bar{C})^2,\ 2\rho\sigma^{-1}\varepsilon\big\} \le \min\big\{\rho^{-1}\sigma^{-2}(\varepsilon_0+\bar{C})^2,\ 2\rho\sigma^{-1}\varepsilon\big\} \le \sqrt{2}\sigma^{-3/2}(\varepsilon_0+\bar{C})\sqrt{\varepsilon}.$$

利用这些不等式、(20)、$\varepsilon\le\varepsilon_0$ 和 $\rho\ge\rho_0$，可得(21)成立。 □

接下来我们证明定理 3。在继续之前，先建立下面的一个引理，随后将用它来证明定理 3。

引理 2. 假设假设1成立，且$(x_\varepsilon, y_\varepsilon, z_\varepsilon)$是(11)的一个 ϵ-驻点。设 $D_{\mathbf{y}}$,$f_{\rm low}$, f, ρ和 P_ρ 分别由(6)、(8)和(11)给出。则有

$$\begin{aligned}
&\mathrm{dist}\Big(0, \partial f(x_\varepsilon,y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon,y_\varepsilon) - (\rho\nabla_x\tilde{f}(x_\varepsilon,z_\varepsilon);0)\Big) \le \epsilon, \quad \mathrm{dist}\big(0,\rho\partial_z\tilde{f}(x_\varepsilon,z_\varepsilon)\big) \le \epsilon, \\
&\tilde{f}(x_\varepsilon,y_\varepsilon) \le \min_z\tilde{f}(x_\varepsilon,z) + \rho^{-1}(\max_z P_\rho(x_\varepsilon,y_\varepsilon,z) - f_{\rm low}).
\end{aligned}$$

证明. 由于$(x_\varepsilon, y_\varepsilon, z_\varepsilon)$是(11)的一个 ϵ-驻点，由定义2可得

$$\mathrm{dist}\big(0,\partial_{x,y}P_\rho(x_\varepsilon,y_\varepsilon,z_\varepsilon)\big) \le \epsilon, \quad \mathrm{dist}\big(0,\partial_z P_\rho(x_\varepsilon,y_\varepsilon,z_\varepsilon)\big) \le \epsilon.$$

利用这些结果以及(11)中 P_ρ 的定义，可得

$$\mathrm{dist}\Big(0, \partial f(x_\varepsilon,y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon,y_\varepsilon) - (\rho\nabla_x\tilde{f}(x_\varepsilon,z_\varepsilon);0)\Big) \le \epsilon, \quad \mathrm{dist}\big(0,\rho\partial_z\tilde{f}(x_\varepsilon,z_\varepsilon)\big) \le \varepsilon.$$

此外，由式 (11) 可得

$$f(x_\varepsilon,y_\varepsilon) + \rho(\tilde{f}(x_\varepsilon,y_\varepsilon) - \min_z\tilde{f}(x_\varepsilon,z)) = \max_z P_\rho(x_\varepsilon,y_\varepsilon,z),$$

结合式 (8)，可得

$$\tilde{f}(x_\varepsilon, y_\varepsilon) - \min_z \tilde{f}(x_\varepsilon, z) \le \rho^{-1}(\max_z P_\rho(x_\varepsilon, y_\varepsilon, z) - f_{\text{low}}).$$

引理的证明至此完成。 □

现在我们可以证明定理 3。

定理 3 的证明。由（12）可知，问题（11）可以视为

$$\min_{x,y} \max_z \{P_\rho(x, y, z) = h(x, y, z) + p(x, y) - q(z)\},$$

其中，$h(x, y, z) = f_1(x, y) + \rho f_1(x, y) - \rho f_1(x, z)$，$p(x, y) = f_2(x) + \rho f_2(y)$，且 $q(z) = \rho f_2(z)$。因此，问题（11）具有（99）的形式，其中 $H = P_\rho$。根据假设1和 $\rho = \varepsilon^{-1}$，可知 h 在其定义域上是$\widehat{L}$-光滑的，其中 $\widehat{L}$ 由（22）给出。此外，注意到根据算法 2，$\epsilon_0 = \varepsilon^{3/2} \le \varepsilon/2$，这是由于 $\varepsilon \in (0, 1/4]$。因此，可以将算法 6 适当地应用于问题（11），令 $\rho = \varepsilon^{-1}$，以求得其一个 ϵ-驻点（$x_\varepsilon, y_\varepsilon, z_\varepsilon$）。

此外，注意到根据算法 2，$\tilde{f}(x^0, y^0) \le \min_y \tilde{f}(x^0, y) + \varepsilon$。利用这一点以及 (11) 和 $\rho = \varepsilon^{-1}$，我们得到

$$\max_z P_\rho(x^0, y^0, z) = f(x^0, y^0) + \rho(\tilde{f}(x^0, y^0) - \min_z \tilde{f}(x^0, z)) \le f(x^0, y^0) + \rho\varepsilon = f(x^0, y^0) + 1. \tag{64}$$

由此以及令 $H = P_\rho$, $\epsilon = \varepsilon$, $\epsilon_0 = \varepsilon^{3/2}$, $\hat{x}^0 = (x^0, y^0)$, $D_q = D_{\mathbf{y}}$ 和 $L_{\nabla h} = \widehat{L}$ 的 (104)，可得

$$\begin{aligned}\max_z P_\rho(x_\varepsilon, y_\varepsilon, z) \le & \max_z P_\rho(x^0, y^0, z) + \varepsilon D_{\mathbf{y}}/4 + 2\varepsilon^3(\widehat{L}^{-1} + 4D_{\mathbf{y}}^2\widehat{L}\varepsilon^{-2}) \\ \overset{(64)}{\le} & 1 + f(x^0, y^0) + \varepsilon D_{\mathbf{y}}/4 + 2\varepsilon^3(\widehat{L}^{-1} + 4D_{\mathbf{y}}^2\widehat{L}\varepsilon^{-2}).\end{aligned}$$

由此可知，结合引理 2 以及 $\epsilon = \varepsilon$ 和 $\rho = \varepsilon^{-1}$，$(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ 满足 (23) 和 (24)。

接下来我们证明，在算法 2 中，∇f_1、∇f_1 和 f_2 与 f_2 的近端算子分别至多执行 N 次求值。事实上，由 (7)、(8) 和 (11) 可得

$$\min_{x,y} \max_z P_\rho(x, y, z) \overset{(11)}{=} \min_{x,y}\{f(x, y) + \rho(\tilde{f}(x, y) - \min_z \tilde{f}(x, z))\} \ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x, y) \overset{(8)}{=} f_{\text{low}}, \tag{65}$$

$$\min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} P_\rho(x, y, z) \overset{(11)}{=} \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}}\{f(x, y) + \rho(\tilde{f}(x, y) - \tilde{f}(x, z))\} \overset{(7)(8)}{\ge} f_{\text{low}} + \rho(\tilde{f}_{\text{low}} - \tilde{f}_{\text{hi}}). \tag{66}$$

为便于后续证明，令

$$H = P_\rho, \quad H^* = \min_{x,y} \max_z P_\rho(x, y, z), \quad H_{\text{low}} = \min\{P_\rho(x, y, z) | (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}\}. \tag{67}$$

鉴于上述结果、(55)、(64)、(65)、(66) 和 $\rho = \varepsilon^{-1}$，我们得到

$$\max_z H(x^0, y^0, z) \overset{(64)}{\le} f(x^0, y^0) + 1, \qquad f_{\text{low}} \overset{(65)}{\le} H^* \overset{(55)}{\le} f^*,$$

$$H_{\text{low}} \overset{(66)}{\ge} f_{\text{low}} + \rho(\tilde{f}_{\text{low}} - \tilde{f}_{\text{hi}}) = f_{\text{low}} + \varepsilon^{-1}(\tilde{f}_{\text{low}} - \tilde{f}_{\text{hi}}).$$

利用上述结果以及定理 6，并结合 (67) 中给出的 $\epsilon = \varepsilon$、$\hat{x}^0 = (x^0, y^0)$、$D_p = \sqrt{D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2}$、$D_q = D_{\mathbf{y}}$、$\epsilon_0 = \varepsilon^{3/2}$、$L_{\nabla h} = \widehat{L}$、$\hat{\alpha} = \hat{\alpha}$、$\delta = \delta$ 和 H、H^*、H_{low}，我们可以得出，算法 2 至多进行 $\widehat{N}$ 次 ∇f_1、∇f_1 和 f_2 的近端算子求值，分别用于求解问题（5）的一个近似解 $(x_\varepsilon, y_\varepsilon)$，使其满足 (23) 和 (24)。 □

5.2 第3节主要结果的证明

本小节证明定理4和定理5。在继续之前，我们先定义

$$r = G^{-1}D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}), \quad \mathbb{B}_r^+ = \{\lambda \in \mathbb{R}_+^l : \|\lambda\| \le r\}, \tag{68}$$

其中，$D_{\mathbf{y}}$ 在式 (6) 中定义，G 由假设2(iii)给出，ϵ 和 ρ 由算法4给出。此外，由式 (26) 和 (30) 可知，

$$\min_z \widetilde{P}_\mu(x, z) \le \tilde{f}^*(x) \qquad \forall x \in \mathcal{X}, \tag{69}$$

这一点将在后文中经常用到。

接下来，我们将建立若干技术引理，以便随后证明定理 4。

~ ~引理 3. 假设假设1和假设2成立。设 $D_{\mathbf{y}}$、 $L_{\tilde{f}}$、 G、 f^*、 f^*_{hi}和 $\mathbb{B}_r^+$分别由式 (6)、(26)、(27)、(68) 以及假设 2 给出。则以下结论成立。

(i) 对所有 $\lambda^* \in \Lambda^*(x)$ 和 $x \in \mathcal{X}$， $\|\lambda^*\| \le G^{-1}L_{\tilde{f}}D_{\mathbf{y}}$和 $\lambda^* \in \mathbb{B}_r^+$，其中 $\Lambda^*(x)$ 表示任意 $x \in \mathcal{X}$下问题 (26) 的最优拉格朗日乘子集合。

~ ~(ii) 函数 f^*在 $\mathcal{X}$上 Lipschitz 连续，且 f^*嗨有限。

(iii) 有

$$\tilde{f}^*(x) = \max_{\lambda} \min_{z} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle - \mathscr{I}_{\mathbb{R}^l_+}(\lambda) \qquad \forall x \in \mathcal{X}, \tag{70}$$

其中 $\mathscr{I}_{\mathbb{R}^l_+}(\cdot)$ 是与 $\mathbb{R}^l_+$相关的示性函数。

证明。(i) 任取 $x \in \mathcal{X}$和 $\lambda^* \in \Lambda^*(x)$，并令 $z^* \in \mathcal{Y}$满足（z^*， λ^*）是 (26) 的一对原始-对偶最优解。于是可得

$$z^* \in \operatorname*{argmin}_z \tilde{f}(x,z) + \langle \lambda^*, \tilde{g}(x,z)\rangle, \quad \langle \lambda^*, \tilde{g}(x,z^*)\rangle = 0, \quad \tilde{g}(x,z^*) \le 0, \quad \lambda^* \ge 0.$$

上面的第一个关系给出

$$\tilde{f}(x,z^*) + \langle \lambda^*, \tilde{g}(x,z^*)\rangle \le \tilde{f}(x,\hat{z}_x) + \langle \lambda^*, \tilde{g}(x,\hat{z}_x)\rangle,$$

其中，$\hat{z}_x$ 由假设2(iii)给出。由此以及 $\langle \lambda^*, \tilde{g}(x,z^*)\rangle = 0$，可得

$$\langle \lambda^*, -\tilde{g}(x,\hat{z}_x)\rangle \le \tilde{f}(x,\hat{z}_x) - \tilde{f}(x,z^*),$$

结合 $\lambda^* \ge 0$、(6) 和假设 2，可推出

$$G\sum_{i=1}^{l} \lambda_i^* \le \langle \lambda^*, -\tilde{g}(x,\hat{z}_x)\rangle \le \tilde{f}(x,\hat{z}_x) - \tilde{f}(x,z^*) \le L_{\tilde{f}}\|\hat{z}_x - z^*\| \le L_{\tilde{f}}D_{\mathbf{y}}, \tag{71}$$

其中，第一个不等式由假设2(iii)得出，第三个不等式则由 (6) 以及 $\tilde{f}$ 的 $L_{\tilde{f}}$-利普希茨连续性得出（见假设2(i)）。由 (68)、(71) 和 $\lambda^* \ge 0$，可得 $\|\lambda^*\| \le \sum_{i=1}^{l} \lambda_i^* \le G^{-1}L_{\tilde{f}}D_{\mathbf{y}}$和 $\lambda^* \in \mathbb{B}_r^+$。

~(ii) 回顾假设 1(i) 和假设2(iii)，可知对于任意给定的 $x \in \mathcal{X}$， $f(x,\cdot)$ 和 $\tilde{g}_i(x,\cdot)$ $i = 1,\ldots,l$ 均为凸函数。利用这一点、(26) 以及本引理的第一条结论，可得

$$\tilde{f}^*(x) = \min_{z} \max_{\lambda \in \mathbb{B}_r^+} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle \qquad \forall x \in \mathcal{X}. \tag{72}$$

~注意到，根据假设 2， f 和 $\tilde{g}$ 在其定义域上 Lipschitz 连续。因此不难~看出，$\max\{f(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle | \lambda \in \mathbb{B}_r^+\}$是关于 (x,z) 的 Lipschitz 连续函数，其定义域为相应区域。~由此以及式 (72)，很容易验证 f^*在 $\mathcal{X}$上 Lipschitz 连续。此外，由式 (27)、 f^*的连续性以及 $\mathcal{X}$的紧致性可知， f^*_{hi}是有限的~ ~。

(iii) 从式 (26) 可以看出，对所有 $x \in \mathcal{X}$，

$$\tilde{f}^*(x) = \min_{z} \max_{\lambda} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle - \mathscr{I}_{\mathbb{R}^l_+}(\lambda) \ge \max_{\lambda} \min_{z} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle - \mathscr{I}_{\mathbb{R}^l_+}(\lambda)$$

其中，不等式由弱对偶性得出。此外，由假设1可知，对所有 $x \in \mathcal{X}$， $f(x,\cdot)$的~定义域是紧集。由此、式 (72) 以及强对偶性可得

$$\tilde{f}^*(x) = \max_{\lambda \in \mathbb{B}_r^+} \min_{z} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle - \mathscr{I}_{\mathbb{R}^l_+}(\lambda) \qquad \forall x \in \mathcal{X},$$

结合上述不等式可得（70）成立。 □

引理4. 假设假设1和假设2成立，且 $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ 是问题 (33) 的一个 ϵ-最优解，其中 $\epsilon > 0$ 为某个值。设 f_{low}、 f、 P_μ、 f^*_μ、 ρ 和 μ 分别由 (8)、(25)、(30)、(31) 和 (33) 给出。则有

$$\widetilde{P}_\mu(x_\varepsilon, y_\varepsilon) \le \min_{z} \widetilde{P}_\mu(x_\varepsilon, z) + \rho^{-1}(f^*_\mu - f_{\mathrm{low}} + 2\epsilon), \qquad f(x_\varepsilon, y_\varepsilon) \le f^*_\mu + 2\epsilon. \tag{73}$$

~证明。证明沿用引理1的论证，只需分别将 f^* 和 f 替换为 f^*_μ 和 P_μ。

□

~ ~引理 5。假设假设 1-3 成立。设 f_{low}、 f^*、 f^*_{hi}、 f^*_μ 分别由式（7）、（25）、（27）和（31）定义，且 L_f、 ω和 $\bar{\theta}$由假设 2 和 3 给出。假设 $\mu \geq (\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})/\bar{\theta}^2$。则有

$$f^*_\mu \leq f^* + L_f \omega\Big(\sqrt{\mu^{-1}(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})}\Big). \tag{74}$$

~ ~证明。任取 $x \in \mathcal{X}$, $y \in$使目标函数取最小值的自变量$_z\{f(x,z)|\tilde{g}(x,z) \leq 0\}$ 以及 $z^* \in$使目标函数取最小值的自变量$_z P_\mu(x,z)$。由 (30) 和 (69) 可知， $\tilde{f}(x,z^*) + \mu \|[\tilde{g}(x,z^*)]_+\|^2 \leq \tilde{f}^*(x)$，结合 (7) 和 (27) 可得

$$\|[\tilde{g}(x,z^*)]_+\|^2 \leq \mu^{-1}(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}}). \tag{75}$$

由于 $\mu \geq (\tilde{f}^*$嗨$-\tilde{f}_{\text{low}})/\bar{\theta}^2$，由 (75) 可得 $\|[\tilde{g}(x,z^*)]_+\| \leq \bar{\theta}$。根据这一关系、 $y \in$使目标函数取最小值的自变量$\{\tilde{f}(x,z)|\tilde{g}(x,z) \leq 0\}$ 以及假设 3，存在某个 $\hat{}z^*$ 使得

$$\|y - \hat{z}^*\| \leq \omega(\|[\tilde{g}(x,z^*)]_+\|), \qquad \hat{z}^* \in \operatorname*{argmin}_z \left\{ \tilde{f}(x,z) \middle| \|[\tilde{g}(x,z)]_+\| \leq \|[\tilde{g}(x,z^*)]_+\| \right\}. \tag{76}$$

~根据 (30)、 $z^* \in$使目标函数取最小值的自变量$_z P_\mu(x,z)$ 以及 (76) 中的第二个关系，可以看出 $\hat{}z^* \in$~使目标函数取最小值的自变量$_z P_\mu(x,z)$。结合 (31)，可得 $f(x,\hat{}z^*) \geq f^*_\mu$。此外，利用 (76) 以及 L_f-利普希茨连续性 f (见假设 2)，我们有

$$f(x,y) - f(x,\hat{z}^*) \geq -L_f \|y - \hat{z}^*\| \overset{(76)}{\geq} -L_f \omega(\|[\tilde{g}(x,z^*)]_+\|).$$

~在该关系两边分别对 $x \in \mathcal{X}$ 和 $y \in$使目标函数取最小值的自变量$_z\{f(x,z)|\tilde{g}(x,z) \leq 0\}$ 取最小值，并利用 (25)、(75)、 $f(x,\hat{}z^*) \geq f^*_\mu$ 以及 ω 的单调性，可以得出 (74) 成立。 □

~ ~ ~ ~引理 6. 假设假设 1-3 成立。设 f_{low}, f_{low}, f, f, f^*, f^*, f^*嗨、 ρ 和 μ 分别由 (7)、¯(8)、(25)、(26)、(27) 和 (33) 给出，而 L_f、 ω 和 θ 分别由假设 2 和 3 给出。假设 $\mu \geq (\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})/\bar{\theta}^2$，且 $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ 是某个 $\epsilon > 0$ 下问题 (33) 的 ϵ-最优解。那么我们有

$$\begin{aligned}
&f(x_\varepsilon, y_\varepsilon) \leq f^* + L_f \omega\Big(\sqrt{\mu^{-1}(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})}\Big) + 2\epsilon,\\
&\tilde{f}(x_\varepsilon, y_\varepsilon) \leq \tilde{f}^*(x_\varepsilon) + \rho^{-1}\Big(f^* - f_{\text{low}} + L_f \omega\Big(\sqrt{\mu^{-1}(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})}\Big) + 2\epsilon\Big),\\
&\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|^2 \leq \mu^{-1}\Big(\tilde{f}^*(x_\varepsilon) - \tilde{f}_{\text{low}} + \rho^{-1}\Big(f^* - f_{\text{low}} + L_f \omega\Big(\sqrt{\mu^{-1}(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})}\Big) + 2\epsilon\Big)\Big).
\end{aligned}$$

证明。由 (30)、(69) 以及 (73) 中的第一个关系可得

$$\tilde{f}(x_\varepsilon, y_\varepsilon) + \mu \|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|^2 \overset{(30)}{=} \widetilde{P}_\mu(x_\varepsilon, y_\varepsilon) \overset{(69)(73)}{\leq} \tilde{f}^*(x_\varepsilon) + \rho^{-1}(f^*_\mu - f_{\text{low}} + 2\epsilon).$$

由此以及 (7) 可得

$$\text{-}f(x_\varepsilon, y_\varepsilon) \leq \tilde{f}^*(x_\varepsilon) + \rho^{-1}(f^*_\mu - f_{\text{low}} + 2\epsilon), \quad \|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|^2 \leq \mu^{-1}(\tilde{f}^*(x_\varepsilon) - \tilde{f}_{\text{low}} + \rho^{-1}(f^*_\mu - f_{\text{low}} + 2\epsilon)).$$

此外，回顾 (73) 可知 $f(x_\varepsilon, y_\varepsilon) \leq f^*_\mu + 2\epsilon$。由这三个关系以及 (74) 即可得出本引理的结论。 □

现在我们可以证明定理 4。

定理 4 的证明。设 $\{(x^k, y^k, z^k)\}$ 由算法 3 生成，且 $\lim_{k\to\infty}(\rho_k, \mu_k, \epsilon_k) = (\infty, \infty, 0)$。如有必要，考虑一个收敛子列；不失一般性地，假设 $\lim_{k\to\infty}(x^k, y^k) = (x^*, y^*)$。现在证明 (x^*, y^*) 是问题 (25) 的最优解。事实上，由于 (x^k, y^k, z^k) 是 (33) 的一个 ϵ_k-最优解，且 $(\rho, \mu) = (\rho_k, \mu_k)$ 并且 $\lim_{k\to\infty} \mu_k = \infty$，由引理 6 可知，当 $(\rho, \mu, \epsilon) = (\rho_k, \mu_k, \epsilon_k)$ 且 $(x_\varepsilon, y_\varepsilon) = (x^k, y^k)$ 时，对所有充分大的 k，有

$$\begin{aligned}
&f(x^k, y^k) \leq f^* + L_f \omega\Big(\sqrt{\mu_k^{-1}(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})}\Big) + 2\epsilon_k,\\
&\tilde{f}(x^k, y^k) \leq \tilde{f}^*(x^k) + \rho_k^{-1}\Big(f^* - f_{\text{low}} + L_f \omega\Big(\sqrt{\mu_k^{-1}(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})}\Big) + 2\epsilon_k\Big),\\
&\big\|[\tilde{g}(x^k, y^k)]_+\big\|^2 \leq \mu_k^{-1}\Big(\tilde{f}^*(x^k) - \tilde{f}_{\text{low}} + \rho_k^{-1}\Big(f^* - f_{\text{low}} + L_f \omega\Big(\sqrt{\mu_k^{-1}(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})}\Big) + 2\epsilon_k\Big)\Big).
\end{aligned}$$

由 f, $\tilde{f}$ 的连续性以及 $\tilde{f}^*$ 的连续性（见假设1(i)和引理3(ii)），$\lim_{k\to\infty}(x^k, y^k) = (x^*, y^*)$，$\lim_{k\to\infty}(\rho_k, \mu_k, \epsilon_k) = (\infty, \infty, 0)$，$\lim_{\theta\downarrow 0}\omega(\theta) = 0$；再对上述~关系两边取 $k \to \infty$ 的极限，得到 $f(x^*, y^*) \le f^*$，$f(x^*, y^*) \le f^*(x^*)$ 且 $[\tilde{g}(x^*, y^*)]_+ = 0$。结合~(25) 和 (26)，可知 $f(x^*, y^*) = f^*$，并且 $y^* \in$使目标函数取最小值的自变量$_z\{f(x^*, z)|\tilde{g}(x^*, z) \le 0\}$。因此，$(x^*, y^*)$ 是 (25) 的最优解，证毕。 □

接下来证明定理5。在继续之前，我们先建立以下几个技术引理，随后将用它们来证明定理5。

引理7. 假设假设1和假设2成立，且 $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ 是问题 (33) 的一个 ϵ-驻点，其中 $\epsilon > 0$ 为某个给定值。设 (6)、(25)、(33) 和假设 2 中分别给出了 $D_{\mathbf{y}}$, $\tilde{g}$, ρ, μ, L_f, $L_{\tilde{f}}$ 和 G。则有

$$\|[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+\| \le (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}), \tag{77}$$

$$\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \le (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}}). \tag{78}$$

证明. 我们先证明 (77)。由于 $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ 是 (33) 的一个 ϵ-驻点，根据定义2可得 dist(0,$\partial_z P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z_\varepsilon)$) $\le \epsilon$。此外，由 (30) 和 (33) 可得

$$P_{\rho,\mu}(x, y, z) = f(x, y) + \rho(\tilde{f}(x, y) + \mu\,\|[\tilde{g}(x, y)]_+\|^2) - \rho(\tilde{f}(x, z) + \mu\,\|[\tilde{g}(x, z)]_+\|^2). \tag{79}$$

利用这些关系，我们得到

$$\operatorname{dist}\Big(0, \partial_z \tilde{f}(x_\varepsilon, z_\varepsilon) + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \nabla_z \tilde{g}_i(x_\varepsilon, z_\varepsilon)\Big) \le \rho^{-1}\epsilon.$$

~因此，存在 $s \in \partial_z f(x_\varepsilon, z_\varepsilon)$ 使得

$$\Big\| s + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \nabla_z \tilde{g}_i(x_\varepsilon, z_\varepsilon)\Big\| \le \rho^{-1}\epsilon. \tag{80}$$

设假设2(iii)中给定了 $\hat{}z_{x_\varepsilon}$ 和 G。于是可得，对于所有 i，$\hat{}z_{x_\varepsilon} \in \mathcal{Y}$ 和 $-\tilde{g}_i(x_\varepsilon, \hat{}z_{x_\varepsilon}) \ge G > 0$。注意，由于式 (6)，对于所有 i 和 $\|z_\varepsilon - \hat{}z_{x_\varepsilon}\| \le D_{\mathbf{y}}$，$[\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \tilde{g}_i(x_\varepsilon, z_\varepsilon) \ge 0$。利用这些结果、式 (80)、~以及 $f(x_\varepsilon, \cdot)$ 和 $\tilde{g}_i(x_\varepsilon, \cdot)$ 对所有 i 的凸性，我们得到

$$\begin{aligned}
&\tilde{f}(x_\varepsilon, z_\varepsilon) - \tilde{f}(x_\varepsilon, \hat{z}_{x_\varepsilon}) + 2\mu G \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \le \tilde{f}(x_\varepsilon, z_\varepsilon) - \tilde{f}(x_\varepsilon, \hat{z}_{x_\varepsilon}) - 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon}) \\
&\le \tilde{f}(x_\varepsilon, z_\varepsilon) - \tilde{f}(x_\varepsilon, \hat{z}_{x_\varepsilon}) + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ (\tilde{g}_i(x_\varepsilon, z_\varepsilon) - \tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon})) \\
&\le \langle s, z_\varepsilon - \hat{z}_{x_\varepsilon}\rangle + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \langle \nabla_z \tilde{g}_i(x_\varepsilon, z_\varepsilon), z_\varepsilon - \hat{z}_{x_\varepsilon}\rangle \\
&= \langle s + 2\mu \sum_{i=1}^{l} [\tilde{g}(x_\varepsilon, z_\varepsilon)]_+ \nabla_z \tilde{g}_i(x_\varepsilon, z_\varepsilon), z_\varepsilon - \hat{z}_{x_\varepsilon}\rangle \le \rho^{-1} D_{\mathbf{y}} \epsilon,
\end{aligned} \tag{81}$$

其中，第一个不等式是因为对于所有 i，$-\tilde{g}_i(x_\varepsilon, \hat{}z_{x_\varepsilon}) \ge G$；第二个不等式由对于所有 i，$-[\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \tilde{g}_i(x_\varepsilon, z_\varepsilon) \ge 0$ 得出；第三个不等式是因为 $s \in \partial_z f(x_\varepsilon, z_\varepsilon)$，以及对于所有 i，$f(x_\varepsilon, \cdot)$ 和 $\tilde{g}_i(x_\varepsilon, \cdot)$ 的凸性；最后一个不等式由式 (6) 和 (80) 得出。根据式 (6)、(81)、~以及 $f(x, y)$ 的 $L_{\tilde{f}}$-利普希茨连续性（见假设 2），可得

$$\begin{aligned}
\|[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+\| \le \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ &\overset{(81)}{\le} (2\mu G)^{-1}(\rho^{-1} D_{\mathbf{y}}\epsilon + \tilde{f}(x_\varepsilon, \hat{z}_{x_\varepsilon}) - \tilde{f}(x_\varepsilon, z_\varepsilon)) \\
&\le (2\mu G)^{-1}(\rho^{-1} D_{\mathbf{y}}\epsilon + L_{\tilde{f}}\|\hat{z}_{x_\varepsilon} - z_\varepsilon\|) \overset{(6)}{\le} (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}).
\end{aligned}$$

因此，(77) 成立。

接下来证明（78）。由于 $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ 是（33）的 ϵ-驻点，根据定义2可得 dist(0,$\partial_{x,\ y} P_{\rho,\ \mu}(x_\varepsilon, y_\varepsilon, z_\varepsilon)$) $\le \epsilon$。此外，由（34）可知，$P_{\rho,\ \mu}$ 是一个光滑

函数与一个可能非光滑且关于 x、 y 和 z 可分离的函数之和。因此，$\partial_{x,\ y}P_{\rho,\ \mu} = \partial_x P_{\rho,\ \mu} \times \partial_y P_{\rho,\ \mu}$，结合 $\mathrm{dist}(0, \partial_{x,\ y}P_{\rho,\ \mu}(x_\varepsilon, y_\varepsilon, z_\varepsilon)) \leq \epsilon$，可得 $\mathrm{dist}(0, \partial_y P_{\rho,\ \mu}(x_\varepsilon, y_\varepsilon, z_\varepsilon)) \leq \epsilon$。由此关系和（34）可得

$$\mathrm{dist}\big(0,\ \partial_y f(x_\varepsilon, y_\varepsilon) + \rho\partial_y \tilde{f}(x_\varepsilon, y_\varepsilon) + 2\rho\mu\nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\big) \leq \epsilon.$$

因此，存在 $s \in \partial_y f(x_\varepsilon, y_\varepsilon)$ 和 $\tilde{s} \in \partial_y f(x_\varepsilon, y_\varepsilon)$，使得

$$\|s + \rho\tilde{s} + 2\rho\mu\nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \leq \epsilon. \tag{82}$$

设假设2(iii)中的 $\mathcal{A}(x_\varepsilon, y_\varepsilon)$、 $= \{i|\tilde{g}_i(x_\varepsilon, y_\varepsilon)$、 > 0、 $1 \leq i \leq l\}$、$\hat{z}_{x_\varepsilon}$ 和 G 已给定。于是可得 $\hat{z}_{x_\varepsilon} \in \mathcal{Y}$ 和 $-\tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon}) \geq G > 0$ 对所有 i均成立。利用这些结果以及对所有 i， $\tilde{g}_i(x_\varepsilon, \cdot)$ 的凸性，可得

$$\begin{aligned}
&\langle \nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+, y_\varepsilon - \hat{z}_{x_\varepsilon}\rangle = \sum_{i\in\bar{\mathcal{A}}(x_\varepsilon, y_\varepsilon)} \langle \nabla_y \tilde{g}_i(x_\varepsilon, y_\varepsilon), y_\varepsilon - \hat{z}_{x_\varepsilon}\rangle [g_i(x_\varepsilon, y_\varepsilon)]_+ \\
&\geq \sum_{i\in\bar{\mathcal{A}}(x_\varepsilon, y_\varepsilon)} (\tilde{g}_i(x_\varepsilon, y_\varepsilon) - \tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon}))[\tilde{g}_i(x_\varepsilon, y_\varepsilon)]_+ \\
&\geq \sum_{i\in\bar{\mathcal{A}}(x_\varepsilon, y_\varepsilon)} G[\tilde{g}_i(x_\varepsilon, y_\varepsilon)]_+ = G\sum_{i=1}^{l}[\tilde{g}_i(x_\varepsilon, y_\varepsilon)]_+ \geq G\,\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|,
\end{aligned} \tag{83}$$

其中，第一个不等式由 $\tilde{g}(x_\varepsilon, \cdot)$ 的凸性得出，第二个不等式则是因为$-\tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon}) \geq G$。由此以及 (82) 和 (83) 可得

$$\begin{aligned}
D_{\mathbf{y}}\epsilon &\geq \|s + \rho\tilde{s} + 2\rho\mu\nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \cdot \|y_\varepsilon - \hat{z}_{x_\varepsilon}\| \\
&\geq \langle s + \rho\tilde{s} + 2\rho\mu\nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+, y_\varepsilon - \hat{z}_{x_\varepsilon}\rangle \\
&= \langle s + \rho\tilde{s}, y_\varepsilon - \hat{z}_{x_\varepsilon}\rangle + 2\rho\mu\langle \nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+, y_\varepsilon - \hat{z}_{x_\varepsilon}\rangle \\
&\overset{(83)}{\geq} -(\|s\| + \rho\|\tilde{s}\|)\,\|y_\varepsilon - \hat{z}_{x_\varepsilon}\| + 2\rho\mu G\,\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \\
&\geq -(L_f + \rho L_{\tilde{f}})D_{\mathbf{y}} + 2\rho\mu G\,\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|,
\end{aligned} \tag{84}$$

其中，最后一个不等式由 $\|y_\varepsilon - \hat{z}_{x_\varepsilon}\| \leq D_{\mathbf{y}}$ 以及 $\|s\| \leq L_f$和 $\|\tilde{s}\| \leq L_{\tilde{f}}$这一事实得出，而是由 (6)、$s \in \partial_y f(x_\varepsilon, y_\varepsilon)$、 $\tilde{s} \in \partial_y f(x_\varepsilon, y_\varepsilon)$ 和假设2(i)所致。根据 (84)，可立即看出 (78) 成立。

□

引理8. 假设假设1和假设2成立。设 f, f, $\tilde{g}$, $D_{\mathbf{y}}$, f_{low}, f^*和$P_{\rho,\mu}$分别由(5)、(6)、(8)、(26)和(33)给出， L_f, $L_{\tilde{f}}$和 G由假设1和假设2给出，$(x_\varepsilon, y_\varepsilon, z_\varepsilon)$是(33)的一个 ϵ-驻点，其中 $\epsilon > 0$为某个值，并且

$$\tilde{\lambda} = 2\mu[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+, \quad \hat{\lambda} = 2\rho\mu[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+. \tag{85}$$

则有

$$\mathrm{dist}\left(\partial f(x_\varepsilon, y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon, y_\varepsilon) - \rho(\nabla_x \tilde{f}(x_\varepsilon, z_\varepsilon) + \nabla_x \tilde{g}(x_\varepsilon, z_\varepsilon)\tilde{\lambda}; 0) + \nabla\tilde{g}(x_\varepsilon, y_\varepsilon)\hat{\lambda}\right) \leq \epsilon, \tag{86}$$

$$\mathrm{dist}\left(0, \rho(\partial_z \tilde{f}(x_\varepsilon, z_\varepsilon) + \nabla_z \tilde{g}(x_\varepsilon, z_\varepsilon)\tilde{\lambda})\right) \leq \epsilon, \tag{87}$$

$$\|[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+\| \leq (2\mu G)^{-1}D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}), \tag{88}$$

$$|\langle\tilde{\lambda}, \tilde{g}(x_\varepsilon, z_\varepsilon)\rangle| \leq (2\mu)^{-1}G^{-2}D_{\mathbf{y}}^2(\rho^{-1}\epsilon + L_{\tilde{f}})^2, \tag{89}$$

$$|\tilde{f}(x_\varepsilon, y_\varepsilon) - \tilde{f}^*(x_\varepsilon)| \leq \max\left\{\rho^{-1}(\max_z P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z) - f_{\mathrm{low}}), (2\mu)^{-1}G^{-2}D_{\mathbf{y}}^2 L_{\tilde{f}}(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}})\right\}, \tag{90}$$

$$\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \leq (2\mu G)^{-1}D_{\mathbf{y}}(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}}), \tag{91}$$

$$|\langle\hat{\lambda}, \tilde{g}(x_\varepsilon, y_\varepsilon)\rangle| \leq (2\mu)^{-1}\rho G^{-2}D_{\mathbf{y}}^2(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}})^2. \tag{92}$$

证明。由于 $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ 是问题 (33) 的 ϵ-驻点，由 (79)、(85) 和定义2可直接得出 (86) 和 (87) 成立。此外，由 (77) 和 (78) 可得 (88) 和 (91) 成立。进一步，根据 (85)、(88) 和 (91)，有

$$|\langle\tilde{\lambda}, \tilde{g}(x_\varepsilon, z_\varepsilon)\rangle| \overset{(85)}{=} 2\mu\,\|[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+\|^2 \overset{(88)}{\leq} (2\mu)^{-1}G^{-2}D_{\mathbf{y}}^2(\rho^{-1}\epsilon + L_{\tilde{f}})^2,$$

$$|\langle\hat{\lambda}, \tilde{g}(x_\varepsilon, y_\varepsilon)\rangle| \overset{(85)}{=} 2\rho\mu\,\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]\|_+\|^2 \overset{(91)}{\leq} (2\mu)^{-1}\rho G^{-2}D_{\mathbf{y}}^2(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}})^2,$$

因此 (89) 和 (92) 成立。此外，由 (33) 中 $P_{\rho,\ \mu}$ 的定义可知

$$\widetilde{P}_\mu(x_\varepsilon,y_\varepsilon)-\min_z \widetilde{P}_\mu(x_\varepsilon,z)=\rho^{-1}(\max_z P_{\rho,\mu}(x_\varepsilon,y_\varepsilon,z)-f(x_\varepsilon,y_\varepsilon)).$$

利用这一点以及 (8)、(30) 和 (69)，可得

$$\begin{aligned}\tilde{f}(x_\varepsilon,y_\varepsilon)+\mu\|[\tilde{g}(x_\varepsilon,y_\varepsilon)]_+\|^2 &\overset{(30)}{=} \widetilde{P}_\mu(x_\varepsilon,y_\varepsilon)= \min_z \widetilde{P}_\mu(x_\varepsilon,z)+\rho^{-1}(\max_z P_{\rho,\mu}(x_\varepsilon,y_\varepsilon,z)-f(x_\varepsilon,y_\varepsilon))\\ &\overset{(8)(69)}{\le} \tilde{f}^*(x_\varepsilon)+\rho^{-1}(\max_z P_{\rho,\mu}(x_\varepsilon,y_\varepsilon,z)-f_{\rm low}).\end{aligned}\tag{93}$$

另一方面，设 $\lambda^*\in\mathbb{R}^l_+$ 是问题 (26) 在 $x=x_\varepsilon$ 下的一个最优拉格朗日乘子。由引理 3(i) 可得 $\|\lambda^*\|\le G^{-1}L_{\tilde{f}}D_{\mathbf{y}}$。利用这些结果以及 (91)，有

$$\begin{aligned}\tilde{f}^*(x_\varepsilon)&=\min_y\left\{\tilde{f}(x_\varepsilon,y)+\langle\lambda^*,\tilde{g}(x_\varepsilon,y)\rangle\right\}\le \tilde{f}(x_\varepsilon,y_\varepsilon)+\langle\lambda^*,\tilde{g}(x_\varepsilon,y_\varepsilon)\rangle\\ &\le \tilde{f}(x_\varepsilon,y_\varepsilon)+\|\lambda^*\|\|[\tilde{g}(x_\varepsilon,y_\varepsilon)]_+\|\le \tilde{f}(x_\varepsilon,y_\varepsilon)+(2\mu)^{-1}G^{-2}D_{\mathbf{y}}^2L_{\tilde{f}}(\rho^{-1}\epsilon+\rho^{-1}L_f+L_{\tilde{f}}).\end{aligned}$$

由此以及 (93) 可知，(90) 成立。 □

现在我们可以证明定理5。

定理5的证明。由 (34) 可见，问题 (33) 可视为

$$\min_{x,y}\max_z\{P_{\rho,\mu}(x,y,z)=h(x,y,z)+p(x,y)-q(z)\},$$

其中 $h(x,y,z)=f_1(x,y)+\rho\tilde{f}_1(x,y)+\rho\mu\|[\tilde{g}(x,y)]_+\|^2-\rho\tilde{f}_1(x,z)-\rho\mu\|[\tilde{g}(x,z)]_+\|^2$，$p(x,y)=f_2(x)+(\tilde{f}_1(x,y)-\tilde{f}_1(x,z))+\rho\mu(\|[\tilde{g}(x,y)]_+\|^2-\|[\tilde{g}(x,z)]_+\|^2)$ 且 $q(z)=\rho f_2(y)$ 且 $q(z)=\rho f_2(z)$。因此，问题 (33) 具有 (99) 的形式，其中 $H=P_{\rho,\ \mu}$。由假设1、(28)、(29)、$\rho=\varepsilon^{-1}$ 和 $\mu=\varepsilon^{-2}$可知，h 在其定义域上是 $\widetilde{L}$-光滑的，其中 $\widetilde{L}$ 由 (44) 给出。此外，注意到根据算法4，$\epsilon_0=\varepsilon^{5/2}\le\varepsilon/2=\epsilon/2$，这是因为 $\varepsilon\in(0,1/4]$。因此，可将算法 6 适当地应用于问题 (33)，取 $\rho=\varepsilon^{-1}$ 和 $\mu=\varepsilon^{-2}$，以求得其一个 ϵ-驻点 $(x_\varepsilon,y_\varepsilon,z_\varepsilon)$。

此外，由算法4可知 $\widetilde{P}_\mu(x^0,y^0)\le\min_y\widetilde{P}_\mu(x^0,y)+\varepsilon$。利用这一点以及 (33) 和$\rho=\varepsilon^{-1}$，可得

$$\max_z P_{\rho,\mu}(x^0,y^0,z)\overset{(33)}{=}f(x^0,y^0)+\rho(\widetilde{P}_\mu(x^0,y^0)-\min_z\widetilde{P}_\mu(x^0,z))\le f(x^0,y^0)+\rho\varepsilon=f(x^0,y^0)+1.\tag{94}$$

由此以及 (104)，结合 $H=P_{\rho,\mu}$, $\epsilon=\varepsilon$, $\epsilon_0=\varepsilon^{5/2}$, $\hat{}x^0=(x^0,y^0)$, $D_q=D_{\mathbf{y}}$ 和 $L_{\nabla h}=\widetilde{L}$，可得

$$\begin{aligned}\max_z P_{\rho,\mu}(x_\varepsilon,y_\varepsilon,z)&\le \max_z P_{\rho,\mu}(x^0,y^0,z)+\varepsilon D_{\mathbf{y}}/4+2\varepsilon^5(\widetilde{L}^{-1}+4D_{\mathbf{y}}^2\widetilde{L}\varepsilon^{-2})\\ &\overset{(94)}{\le} 1+f(x^0,y^0)+\varepsilon D_{\mathbf{y}}/4+2\varepsilon^5(\widetilde{L}^{-1}+4D_{\mathbf{y}}^2\widetilde{L}\varepsilon^{-2}).\end{aligned}$$

由此可 以推出，结合引理8以及 $\epsilon=\varepsilon$, $\rho=\varepsilon^{-1}$ 和 $\mu=\varepsilon^{-2}$，$(x_\varepsilon,y_\varepsilon,z_\varepsilon)$ 满足以下 e 关系 (45)-(51)。

接下来，我们证明在算法4中，∇f_1、∇f_1、$\nabla\tilde{g}$ 以及 f_2 和 f_2 的近端算子分别至多进行 N 次求值。事实上，根据(7)、(8)、(28)、(30)和(33)，有

$$\min_{x,y}\max_z P_{\rho,\mu}(x,y,z)\overset{(33)}{=}\min_{x,y}\{f(x,y)+\rho(\widetilde{P}_\mu(x,y)-\min_z\widetilde{P}_\mu(x,z))\}\ge\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}}f(x,y)\overset{(8)}{=}f_{\rm low},\tag{95}$$

$$\begin{aligned}&\min\{P_{\rho,\mu}(x,y,z)|(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\}\overset{(33)}{=}\min\{f(x,y)+\rho(\widetilde{P}_\mu(x,y)-\widetilde{P}_\mu(x,z))|(x,y,z)\\ &\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\}\overset{(30)}{=}\min\{f(x,y)+\rho(\tilde{f}(x,y)+\mu\|[\tilde{g}(x,y)]_+\|^2-\tilde{f}(x,z)-\mu\|[\tilde{g}(x,z)]_+\|^2)|(x,y,z)\\ &\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\}\ge f_{\rm low}+\rho(\tilde{f}_{\rm low}-\tilde{f}_{\rm hi})-\rho\mu\tilde{g}_{\rm hi}^2,\end{aligned}\tag{96}$$

其中，最后一个不等式由 (7)、(8) 和 (28) 得出。此外，设 (x^*,y^*) 是 (25) 的一个最优解。由此可得 $f(x^*,y^*)=f^*$ and $[\tilde{g}(x^*,y^*)]_+=0$。结合这些结果以及 (7)、(30) 和 (33)，可得

$$\begin{aligned}\min_{x,y}\max_z P_{\rho,\mu}(x,y,z)\le\max_z P_{\rho,\mu}(x^*,y^*,z)&\overset{(33)}{=}f(x^*,y^*)+\rho\left(\widetilde{P}_\mu(x^*,y^*)-\min_z\widetilde{P}_\mu(x^*,z)\right)\\ &\overset{(30)}{=}f(x^*,y^*)+\rho(\tilde{f}(x^*,y^*)+\mu\|[\tilde{g}(x^*,y^*)]_+\|^2-\min_z\{\tilde{f}(x^*,z)+\mu\|[\tilde{g}(x^*,z)]_+\|^2\})\\ &\overset{(7)}{\le}f^*+\rho(\tilde{f}_{\rm hi}-\tilde{f}_{\rm low}).\end{aligned}\tag{97}$$

为方便后续证明，令

$$H = P_{\rho,\mu}, \quad H^* = \min_{x,y}\max_{z} P_{\rho,\mu}(x,y,z), \quad H_{\text{low}} = \min\{P_{\rho,\mu}(x,y,z)|(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\}. \tag{98}$$

鉴于这些结果、(94)、(95)、(96)、(97)、 $\rho=\varepsilon^{-1}$ and $\mu=\varepsilon^{-2}$，我们可得

$$\max_z H(x^0,y^0,z) \overset{(94)}{\leq} f(x^0,y^0)+1, \quad f_{\text{low}} \overset{(95)}{\leq} H^* \overset{(97)}{\leq} f^*+\rho(\tilde{f}_{\text{hi}}-\tilde{f}_{\text{low}}) = f^*+\varepsilon^{-1}(\tilde{f}_{\text{hi}}-\tilde{f}_{\text{low}}),$$

$$H_{\text{low}} \overset{(96)}{\geq} f_{\text{low}}+\rho(\tilde{f}_{\text{low}}-\tilde{f}_{\text{hi}})-\rho\mu\tilde{g}_{\text{hi}}^2 = f_{\text{low}}+\varepsilon^{-1}(\tilde{f}_{\text{low}}-\tilde{f}_{\text{hi}})-\varepsilon^{-3}\tilde{g}_{\text{hi}}^2.$$

利用这些结果以及定理 6，并取 ( ) 中给出的 $\epsilon=\varepsilon$、$\hat{}x^0=(x^0,y^0)$、 $D_p=\sqrt{D_{\mathbf{x}}^2+D_{\mathbf{y}}^2}$, $D_q=D_{\mathbf{y}}$, $\epsilon_0=\varepsilon^{5/2}$, $L_{\nabla h}=\widetilde{L}$, $\sim\alpha=\tilde{\alpha}$、 $\delta=\delta$ 和 H, H^*, H_{low} 98 4 N，可以得出，算法在寻找满足 (45)-(51) 的问题 (25) 的近似解 $(x_\varepsilon,y_\varepsilon)$ 时，至多需要对 ∇f_1, ∇f_1, $\nabla\tilde{g}$ 求值，并计算 f_2 和 f_2 的近端算子。

□

6 总结

为简单起见，本文仅针对问题 (3) 提出一阶罚函数方法。采用处理约束 $g(x,y)\leq 0$ 的标准技术（例如，参见 [49]），将其推广到问题 (1) 也会很有意思。这将留待未来研究。

参考文献

[1] G. B. Allende 和 G. Still。使用 KKT 方法求解双层规划。《数学规划》，138(1)：309–332，2013。[2] J. F. Bard。《实用双层优化：算法与应用》，第 30 卷。Springer Science & Business Media，2013。[3] K. P. Bennett、G. Kunapuli、J. Hu 和 J.-S. Pang。双层优化与机器学习。载于 IEEE 世界计算智能大会，页码 25–47。Springer，2008。[4] L. Bertinetto、J. F. Henriques、P. Torr 和 A. Vedaldi。使用可微闭式求解器进行元学习。载于国际学习表征会议，2018。[5] L. Chen、J. Xu 和 J. Zhang。关于不具备下层强凸性的双层优化。arXiv 预印本 arXiv:2301.00712，2023。[6] T. Chen、Y. Sun、Q. Xiao 和 W. Yin。随机双层优化的单时间尺度方法。载于国际人工智能与统计会议，页码 2466–2488，2022。[7] F. H. Clarke。《优化与非光滑分析》。SIAM，1990。[8] B. Colson、P. Marcotte 和 G. Savard。双层优化综述。《运筹学年鉴》，153(1)：235–256，2007。[9] C. Crockett、J. A. Fessler 等。图像重建的双层方法。《信号处理基础与趋势》，15(2-3)：121–289，2022。[10] S. Dempe。《双层规划基础》。Springer Science & Business Media，2002。[11] S. Dempe、V. Kalashnikov、G. A. P´erez-Vald´es 和 N. Kalashnykova。双层规划问题。《能源系统》。Springer，柏林，10:978–3，2015。[12] S. Dempe 和 A. Zemkoho。双层优化。载于《Springer 优化及其应用》，第 161 卷。Springer，2020。[13] S. Dempe 和 A. B. Zemkoho。双层规划问题：重构、约束资格条件与最优性条件。《数学规划》，138(1)：447–473，2013。[14] A. L. Dontchev 和 R. T. Rockafellar。《隐函数与解映射》，第 543 卷。Springer，2009。

[15] M. Feurer 和 F. Hutter。超参数优化。载于《自动化机器学习》，页码 3–33。Springer，沙姆，2019。[16] L. Franceschi、M. Donini、P. Frasconi 和 M. Pontil。基于前向与反向梯度的超参数优化。载于国际机器学习会议，页码 1165–1173，2017。[17] L. Franceschi、P. Frasconi、S. Salzo、R. Grazzi 和 M. Pontil。用于超参数优化和元学习的双层规划。载于国际机器学习会议，页码 1568–1577，2018。[18]S. Ghadimi 和 M. Wang。双层规划的近似方法。arXiv 预印本 arXiv:1802.02246，2018。[19] M. Grant 和 S. Boyd。CVX：用于规范凸规划的 Matlab 软件，2.1 版，2014。[20] R. Grazzi、L. Franceschi、M. Pontil 和 S. Salzo。关于超梯度计算的迭代复杂度。载于国际机器学习会议，页码 3748–3758，2020。[21] Z. Guo、Q. Hu、L. Zhang 和 T. Yang。用于多任务随机双层优化的随机化随机方差缩减方法。arXiv 预印本 arXiv:2105.02266，2021。[22] P. Hansen、B. Jaumard 和 G. Savard。线性双层规划的新分支定界规则。《SIAM科学与统计计算杂志》，13(5)：1194–1217，1992。[23] M. Hong、H.-T. Wai、Z. Wang 和 Z. Yang。双层优化的双时间尺度随机算法框架：复杂度分析及其在演员-评论家中的应用。《SIAM优化杂志》，33(1)：147–180，2023。[24] X. Hu、N. Xiao、X. Liu 和 K.-C. Toh。双层优化的改进无约束方法。《SIAM优化杂志》，33(4)：2801–2829，2023。[25] F. Huang 和 H. Huang。Biadam：快速自适应双层优化方法。arXiv 预印本 arXiv:2106.11396，2021。[26] M. Huang、K. Ji、S. Ma 和 L. Lai。高效逃离双层优化中的鞍点。arXiv 预印本 arXiv:2202.03684，2022。[27] Y. Ishizuka 和 E. Aiyoshi。双层优化问题的双重惩罚法。《运筹学年鉴》，34(1)：73–88，1992。[28] K. Ji、J. D. Lee、Y. Liang 和 H. V. Poor。元学习在部分参数上进行任务特定适应时的收敛性。《神经信息处理系统进展》，33：11490–11500，2020。[29] K. Ji、J. Yang 和 Y. Liang。双层优化：收敛性分析与改进设计。载于国际机器学习会议，页码 4882–4892，2021。[30] P. Khanduri、S. Zeng、M. Hong、H.-T. Wai、Z. Wang 和 Z. Yang。通过双动量实现随机双层优化的近最优算法。《神经信息处理系统进展》，34：30271–30283，2021。[31] V. Konda 和 J. Tsitsiklis。演员-评论家算法。《神经信息处理系统进展》，12，1999。[32] W. Kong 和 R. D. Monteiro。求解非凸-凹极小极大问题的加速非精确近端点方法。《SIAM优化杂志》，31(4)：2558–2585，2021。[33] D. Kovalev 和 A. Gasnikov。首个适用于光滑且强凸-强凹极小极大优化的最优算法。《神经信息处理系统进展》，35：14691–14703，2022。[34] J. Kwon、D. Kwon、S. Wright 和 R. D. Nowak。随机双层优化的完全一阶方法。载于国际机器学习会议，页码 18083–18113，2023。[35] J. Li、B. Gu 和 H. Huang。一种无需海森逆矩阵的双层优化完全单循环算法。载于AAAI人工智能会议论文集，页码 7426–7434，2022。

[36] Y. Li、G.-H. Lin、J. Zhang 和 X. Zhu。基于 Wolfe 对偶的一种新型双层规划方法。arXiv 预印本 arXiv:2302.06838，2023。[37] B. Liu、M. Ye、S. Wright、P. Stone 和 Q. Liu。Bome! 让双层优化变得简单：一种简单的一阶方法。《神经信息处理系统进展》，35：17248–17262，2022。[38] H. Liu、K. Simonyan 和 Y. Yang。Darts：可微架构搜索。发表于《国际学习表征会议》，2018。[39] R. Liu、J. Gao、J. Zhang、D. Meng 和 Z. Lin。从统一视角探究学习与视觉中的双层优化：综述及拓展。《IEEE模式分析与机器智能汇刊》，2021。[40] D. Lopez-Paz 和 M. Ranzato。用于持续学习的梯度情景记忆。《神经信息处理系统进展》，30，2017。[41] Z. Lu 和 S. Mei。一种用于约束极小极大优化的一阶增广拉格朗日法。arXiv 预印本 arXiv:2301.02060，2023。[42] Z.-Q. Luo、J.-S. Pang 和 D. Ralph。带均衡约束的数学规划。剑桥大学出版社，1996。[43] X. Ma、W. Yao、J. J. Ye 和 J. Zhang。结合二阶最优性条件求解双层规划问题的方法。2023。即将在《凸分析杂志》发表。[44] D. Maclaurin、D. Duvenaud 和 R. Adams。通过可逆学习实现基于梯度的超参数优化。发表于《国际机器学习会议》，页码 2113–2122，2015。[45] A. Madry、A. Makelov、L. Schmidt、D. Tsipras 和 A. Vladu。迈向能够抵御对抗攻击的深度学习模型。发表于《国际学习表征会议》，2018。[46] J. A. Mirrlees。道德风险与不可观测行为理论：第一部分。《经济研究评论》，66(1)：3–21，1999。[47] Y. Nesterov。用于最小化复合函数的梯度法。《数学规划》，140(1)：125–161，2013。[48] Y. E. Nesterov。用于最小化复合函数的梯度法。《数学规划》，140：125–161，2013。[49] J. Nocedal 和 S. J. Wright。数值优化。Springer，1999。[50] J. Outrata、M. Kocvara 和 J. Zowe。带均衡约束的优化问题的非光滑方法：理论、应用与数值结果，第28卷。Springer Science & Business Media，2013。[51] F. Pedregosa。利用近似梯度进行超参数优化。发表于《国际机器学习会议》，页码 737–746，2016。[52] A. Rajeswaran、C. Finn、S. M. Kakade 和 S. Levine。利用隐式梯度进行元学习。《神经信息处理系统进展》，32，2019。[53] C. Shi、J. Lu 和 G. Zhang。线性双层规划的扩展 Kuhn–Tucker 方法。《应用数学与计算》，162(1)：51–63，2005。[54] K. Shimizu、Y. Ishizuka 和 J. F. Bard。不可微与两层数学规划。Springer Science & Business Media，2012。[55] A. Sinha、P. Malo 和 K. Deb。双层优化综述：从经典方法到进化方法及其应用。《IEEE进化计算汇刊》，22(2)：276–295，2017。[56] D. Sow、K. Ji、Z. Guan 和 Y. Liang。求解具有多个内层极小值的双层优化问题的原始-对偶方法。arXiv 预印本 arXiv:2203.01123，2022。[57] C. Szegedy、W. Zaremba、I. Sutskever、J. Bruna、D. Erhan、I. Goodfellow 和 R. Fergus。神经网络的奇妙特性。arXiv 预印本 arXiv:1312.6199，2013。

[58] L. N. Vicente 和 P. H. Calamai。双层与多层规划：文献综述。《全球优化杂志》，5(3)：291–306，1994。[59] H. Von Stackelberg。市场结构与均衡。Springer Science & Business Media，2010。[60] D. Ward 和 J. M. Borwein。有限维空间中的非光滑微积分。《SIAM控制与优化杂志》，25(5)：1312–1340，1987。[61] J. Yang、K. Ji 和 Y. Liang。具有可证明更快速度的双层优化算法。《神经信息处理系统进展》，34：13670–13682，2021。[62] J. J. Ye。双层优化中的约束资格条件与最优性条件。载于《双层优化》，页码 227–251。Springer，2020。[63] J. J. Ye、X. Yuan、S. Zeng 和 J. Zhang。用于双层规划的差分凸算法及其在超参数选择中的应用。《数学规划》，页码 1–34，2022。[64] R. Zhao。用于最大结构非凸优化的原始-对偶平滑框架。《运筹学数学》，2023。

A 非凸-凹极小极大问题的一种一阶方法

在这一部分，我们旨在寻找非凸-凹极小极大问题的一个 ϵ-驻点

$$H^* = \min_x \max_y \{H(x,y) \coloneqq h(x,y) + p(x) - q(y)\}, \tag{99}$$

该问题至少存在一个最优解，并满足以下假设。

假设 4. (i) $p\colon \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ 和 $q\colon \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ 是真凸函数，且分别在 dom p和 dom q上连续；此外，dom p和 dom q是紧集。

(ii) 与 p 和 q 相关联的近端算子可以精确求值。

(iii) h 在 dom $p\times$ dom q上是 $L_{\nabla h}$-光滑的；此外，对于任意 $x\in$ dom p， $h(x,\cdot)$ 都是凹函数。

最近，文献 [32]提出了一种加速非精确近端点平滑（AIPP-S）方案，用于求解一类极小极大复合非凸优化问题的近似驻点，其中包括作为特例的（99）。将 AIPP-S 应用于（99）时，对于任意 $\lambda>0$、$x'\in\mathbb{R}^n$ 和 $y'\in\mathbb{R}^m$，都需要精确求解 $\max_y\{h(x',y) - q(y) - \frac{1}{2\lambda}\|y-y'\|^2\}$。然而， h 通常较为复杂，且这类问题通常无法精确求解。因此，AIPP-S 通常无法用于（99）。此外，文献 [64]提出了一种一阶方法，其一阶预言机复杂度为 $\mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$，用于求解满足以下条件的（99）的一个 ϵ-原始驻点 x'：

$$\left\|\lambda^{-1}(x' - \operatorname*{argmin}_x\left\{\max_y H(x,y) + \frac{1}{2\lambda}\|x-x'\|^2\right\}\right\| \le \epsilon$$

其中某个 $0<\lambda<L_{\nabla h}^{-1}$ 满足上述条件。然而，该方法不适用于我们的需求，因为我们的目标是求解定义2中引入的（99）的一个 ϵ-驻点。接下来，我们介绍文献[41, 算法 2] 中提出的一阶方法，用于求解（99）的此类 ϵ-驻点。

为便于表述，我们定义

$$D_p = \max\{\|u-v\|\big|u,v\in\operatorname{dom}p\},\quad D_q = \max\{\|u-v\|\big|u,v\in\operatorname{dom}q\}, \tag{100}$$

$$H_{\text{low}} = \min\{H(x,y)|(x,y)\in\operatorname{dom}p\times\operatorname{dom}q\}. \tag{101}$$

给定一个迭代点 (x^k,y^k)，一阶方法 [41, 算法 2] 通过将 [41, 算法 1], 应用于强凸强凹极小极大问题，求得下一个迭代点 (x^{k+1},y^{k+1})；该算法是对一种新颖的最优一阶方法 [33,算法 4], 的轻微修改

$$\min_x\max_y\left\{h_k(x,y) = h(x,y) - \epsilon\|y-y^0\|^2/(4D_q) + L_{\nabla h}\|x-x^k\|^2\right\}. \tag{102}$$

为便于参考，我们接下来给出一种改进的最优一阶方法 [41, 算法 1] ，见下文算法 5，用于求解强凸-强凹极小极大问题

$$\min_x\max_y\left\{\bar{h}(x,y) + p(x) - q(y)\right\}, \tag{103}$$

¯其中 $h(x,y)$ 是 σ_x-强凸-σ_y-强凹的，并且在 $\mathrm{dom}\, p \times \mathrm{dom}\ q$ 上是 $L_{\nabla \bar{h}}$-光滑的，其中$\sigma_x, \sigma_y > 0$ 为某些值。在算法 5 中，函数 $\hat{}h$, a_x^k 和 a_y^k 定义如下：

$$\hat{h}(x,y) = \bar{h}(x,y) - \sigma_x\|x\|^2/2 + \sigma_y\|y\|^2/2,$$
$$a_x^k(x,y) = \nabla_x \hat{h}(x,y) + \sigma_x(x - \sigma_x^{-1} z_g^k)/2, \quad a_y^k(x,y) = -\nabla_y \hat{h}(x,y) + \sigma_y y + \sigma_x(y - y_g^k)/8,$$

其中， y_g^k 和 z_g^k 由下文算法 5 的第 k 次迭代生成。

算法 5：问题 (103) 的改进的最优一阶方法 输入： $\tau > 0$, $\bar{z}^0 = z_f^0 \in -\sigma_x \mathrm{dom}\, p$,[11] $\bar{y}^0 = y_f^0 \in \mathrm{dom}\ q$, $(z^0, y^0) = (\bar{z}^0, \bar{y}^0)$, $\bar{\alpha} = \min\left\{1, \sqrt{8\sigma_y/\sigma_x}\right\}$, $\eta_z = \sigma_x/2$, $\eta_y = \min\{1/(2\sigma_y),\ 4/(\bar{\alpha}\sigma_x)\}$, $\beta_t = 2/(t+3)$, $\zeta = \left(2\sqrt{5}(1 + 8L_{\nabla \bar{h}}/\sigma_x)\right)^{-1}$, $\gamma_x = \gamma_y = 8\sigma_x^{-1}$, 以及 $\hat{}\zeta = \min\{\sigma_x, \sigma_y\}/L_{\nabla \bar{h}}^2$. 1: 对于 $k = 0, 1, 2, \ldots,$ 执行 2: $(z_g^k, y_g^k) = \bar{\alpha}(z^k, y^k) + (1 - \bar{\alpha})(z_f^k, y_f^k)$. 3: $(x^{k,-1}, y^{k,-1}) = (-\sigma_x^{-1} z_g^k, y_g^k)$. 4: $x^{k,0} =$ 近端算子$_{\zeta\gamma_x p}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}))$. 5: $y^{k,0} =$ 近端算子$_{\zeta\gamma_y q}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}))$. 6: $b^{k,}{}_x{}^0 = \frac{1}{\zeta\gamma_x}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}) - x^{k,0})$. 7: $b^{k,}{}_y{}^0 = \frac{1}{\zeta\gamma_y}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}) - y^{k,0})$. 8: $t = 0$. 9: 当$\gamma_x\|a_x^k(x^{k,t}, y^{k,t}) + b^{k,}{}_x{}^t\|^2 + \gamma_y\|a_y^k(x^{k,t}, y^{k,t}) + b^{k,}{}_y{}^t\|^2 > \gamma_x^{-1}\|x^{k,t} - x^{k,-1}\|^2 + \gamma_y^{-1}\|y^{k,t} - y^{k,-1}\|^2$时， 执行 10: $x^{k,t+1/2} = x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x(a_x^k(x^{k,t}, y^{k,t}) + b^{k,}{}_x{}^t)$. 11: $y^{k,t+1/2} = y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y(a_y^k(x^{k,t}, y^{k,t}) + b^{k,}{}_y{}^t)$. 12: $x^{k,t+1} =$ 近端算子$_{\zeta\gamma_x p}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}))$. 13: $y^{k,t+1} =$ 近端算子$_{\zeta\gamma_y q}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}))$. 14: $b^{k,}{}_x{}^{t+1} = \frac{1}{\zeta\gamma_x}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}) - x^{k,t+1})$. 15: $b^{k,}{}_y{}^{t+1} = \frac{1}{\zeta\gamma_y}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}) - y^{k,t+1})$. 16: $t \leftarrow t + 1$. 17: 结束 while 循环 18: $(x_f^{k+1}, y_f^{k+1}) = (x^{k,t}, y^{k,t})$. 19: $(z_f^{k+1}, w_f^{k+1}) = (\nabla_x \hat{}h(x_f^{k+1}, y_f^{k+1}) + b^{k,}{}_x{}^t, -\nabla_y \hat{}h(x_f^{k+1}, y_f^{k+1}) + b^{k,}{}_y{}^t)$. 20: $z^{k+1} = z^k + \eta_z\sigma_x^{-1}(z_f^{k+1} - z^k) - \eta_z(x_f^{k+1} + \sigma_x^{-1} z_f^{k+1})$. 21: $y^{k+1} = y^k + \eta_y\sigma_y(y_f^{k+1} - y^k) - \eta_y(w_f^{k+1} + \sigma_y y_f^{k+1})$. 22: $x^{k+1} = -\sigma_x^{-1} z^{k+1}$. 23: $\hat{}x^{k+1} =$ 近端算子$_{\hat{\zeta}p}(x^{k+1} - \hat{}\zeta\nabla_x \bar{h}(x^{k+1}, y^{k+1}))$. 24: $\hat{}y^{k+1} =$ 近端算子$_{\hat{\zeta}q}(y^{k+1} + \hat{}\zeta\nabla_y \bar{h}(x^{k+1}, y^{k+1}))$. 25: 若满足以下条件，则终止算法并输出 $(\hat{}x^{k+1}, \hat{}y^{k+1})$

$$\|\hat{\zeta}^{-1}(x^{k+1} - \hat{x}^{k+1}, \hat{y}^{k+1} - y^{k+1}) - (\nabla \bar{h}(x^{k+1}, y^{k+1}) - \nabla \bar{h}(\hat{x}^{k+1}, \hat{y}^{k+1}))\| \le \tau.$$

26: </style}结束循环

现在我们准备好在下文的算法 6 中给出一阶方法 [41, 算法 2]，用于求解 (99) 的 ϵ-驻点。

[11]为方便起见， $-\sigma_x \mathrm{dom}\, p$ 表示集合 $\{-\sigma_x u | u \in \mathrm{dom}\ p\}$。

算法 6 问题 (99) 的一阶方法

输入： $\epsilon > 0$, $\epsilon_0 \in (0, \epsilon/2]$, $(\hat{x}^0, \hat{y}^0) \in \mathrm{dom}\, p \times \mathrm{dom}\, q$, $(x^0, y^0) = (\hat{x}^0, \hat{y}^0)$, 以及 $\epsilon_k = \epsilon_0/(k+1)$。1: 对于 $k = 0, 1, 2, \ldots$, 执行_2: 调用算法 5，使用 $h \leftarrow h_k$, $\tau \leftarrow \epsilon_k$, $\sigma_x \leftarrow L_{\nabla h}$, $\sigma_y \leftarrow \epsilon/(2D_q)$, $L_{\nabla \bar{h}} \leftarrow 3L_{\nabla h} + \epsilon/(2D_q)$,$\bar{z}^0 = z_f^0 \leftarrow -\sigma_x x^k$, $\bar{y}^0 = y_f^0 \leftarrow y^k$，并将其输出记为 (x^{k+1}, y^{k+1})，其中 h_k 由 (102) 给出。

3: 若满足以下条件，则终止算法并输出 $(x_\varepsilon, y_\varepsilon) = (x^{k+1}, y^{k+1})$

$$\|x^{k+1} - x^k\| \le \epsilon/(4L_{\nabla h}).$$

4: 结束循环

以下定理给出了算法 6 的迭代复杂度，其证明见 [41,定理 2]。

定理 6（算法 6 的复杂度）。假设假设 4 成立。设 H^*, H D_p, D_q 和 H_{low} 由 (99)、(100) 和 (101) 定义， $L_{\nabla h}$ 由假设 4 给出， ϵ, ϵ_0 和 x^0 由算法 6 给出，并且

$$\begin{aligned}
\alpha =& \min\left\{1, \sqrt{4\epsilon/(D_q L_{\nabla h})}\right\},\\
\delta =& (2+\alpha^{-1}) L_{\nabla h} D_p^2 + \max\left\{\epsilon/D_q, \alpha L_{\nabla h}/4\right\} D_q^2,\\
K =& \left\lceil 16(\max_y H(x^0, y) - H^* + \epsilon D_q/4) L_{\nabla h}\epsilon^{-2} + 32\epsilon_0^2(1 + 4D_q^2 L_{\nabla h}^2 \epsilon^{-2})\epsilon^{-2} - 1 \right\rceil_+,\\
N =& \left(\left\lceil 96\sqrt{2}\left(1 + (24L_{\nabla h} + 4\epsilon/D_q) L_{\nabla h}^{-1}\right)\right\rceil + 2\right)\left\{2, \sqrt{D_q L_{\nabla h}\epsilon^{-1}}\right\}\\
&\times \left((K+1)\left(\log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{D_q}{\epsilon}, \frac{4}{\alpha L_{\nabla h}}\right\}\right\}\left(\delta + 2\alpha^{-1}(H^* - H_{\mathrm{low}} + \epsilon D_q/4 + L_{\nabla h} D_p^2)\right)}{\left[(3L_{\nabla h} + \epsilon/(2D_q))^2/\min\{L_{\nabla h}, \epsilon/(2D_q)\} + 3L_{\nabla h} + \epsilon/(2D_q)\right]^{-2}\epsilon_0^2} \right)_+ \right.\\
&\left. + K + 1 + 2K\log(K+1) \right).
\end{aligned}$$

则算法 6 将在至多 $K+1$ 次外层迭代内终止，并输出 (99) 的一个 ϵ-驻点 $(x_\varepsilon, y_\varepsilon)$，满足

$$\max_y H(x_\varepsilon, y) \le \max_y H(\hat{x}^0, y) + \epsilon D_q/4 + 2\epsilon_0^2\left(L_{\nabla h}^{-1} + 4D_q^2 L_{\nabla h}\epsilon^{-2}\right). \tag{104}$$

此外，算法 6 中执行的 ∇h 和 p 的近端算子以及 q 的求值次数分别不超过 N。