约束双层优化的高效梯度近似方法

Siyuan Xu[1],**Minghui Zhu**[1*]

[1] 宾夕法尼亚州立大学电气工程与计算机科学学院,
美国大学公园市 {spx5032, muz16}@psu.edu

arXiv:2302.01970v1 [cs.LG] 3 Feb 2023

摘要

双层优化已应用于许多涉及大规模、高维数据的机器学习任务。本文研究一个约束双层优化问题，其中下层优化问题是带有等式约束和不等式约束的凸问题，而上层优化问题是非凸的。整体目标函数既非凸又不可微。为求解该问题，我们提出一种基于梯度的方法，称为梯度近似方法。该方法通过计算当前估计值邻域内目标函数的若干代表性梯度来确定下降方向。我们证明，该算法渐近收敛至 Clarke 驻点集合，并通过超参数优化和元学习实验展示了该算法的有效性。

1 引言

一般的带约束双层优化问题可表述如下：

$$\begin{aligned}&\min_{x\in\mathbb{R}^{d_x}} \ \Phi(x)=f\left(x,y^*(x)\right)\\&\text{s.t. } r\left(x,y^*(x)\right)\leq 0;\ s\left(x,y^*(x)\right)=0;\\&y^*(x)=\arg\min_{y\in\mathbb{R}^{d_y}}\{g(x,y):p\left(x,y\right)\leq 0;q\left(x,y\right)=0\}.\end{aligned}\tag{1}$$

该双层优化以 $\Phi(x)$ 为整体目标函数，并针对 x 进行最小化，其中 $y^*(x)$ 是下层优化问题的最优解，并且是上层决策变量 x 的参数化函数。本文假设，对于任意 $x\in\mathbb{R}^{d_x}$，$y^*(x)$ 都是唯一的。

求解问题 (1) 的现有方法可分为两类：单层化约简方法（Bard and Falk 1982；Bard and Moore 1990；Shi, Lu, and Zhang 2005）和下降方法（Savard and Gauvin 1994；Dempe 1998）。当下层优化问题为凸问题时，单层化约简方法使用KKT条件替代该问题。随后，它们将双层优化问题 (1) 改写为单层约束优化问题。下降方法旨在寻找下降方向，使新点满足可行性，同时降低目标函数值。Savard and Gauvin 1994 的论文通过求解一个二次规划来计算目标函数的下降方向。Dempe 1998 的论文采用 Kolstad and Lasdon 1990 和 Fiacco and McCormick1970 中计算得到的目标函数梯度来计算广义Clarke雅可比，并使用束方法（Schramm and Zowe 1992）进行优化。将双层优化应用于机器学习时，由于上层和下层问题的决策变量维数较高，因此会面临额外的挑战（Liu 等人，2021a）。

基于梯度的方法已被证明能够有效处理各种机器学习任务中的大规模和高维数据（Bottou 和 Bousquet 2008 年）。这类方法已扩展到求解下层优化不含约束的双层优化问题。这些方法可分为基于近似隐式微分（AID）的方法（Pedregosa 2016；Gould 等人 2016；Ghadimi 和 Wang 2018；Grazzi 等人 2020）和迭代微分（ITD）方法（Grazzi 等人 2020；Franceschi 等人 2017、2018；Shaban 等人 2019；Ji、Yang 和 Liang 2021）。基于 AID 的方法根据隐式微分（Bengio 2000）计算 $y^*(x)$ 和 $\Phi(x)$ 的梯度。基于 ITD 的方法将下层优化中的迭代优化步骤视为一个动力系统，将 $y^*(x)$ 设为其稳定点，并在每个迭代步骤计算 $\nabla y^*(x)$。基于梯度的算法已应用于多种机器学习任务，包括元学习（Franceschi 等人 2018；Rajeswaran 等人 2019；Ji 等人 2020）、超参数优化（Pedregosa 2016；Franceschi 等人 2017、2018）、强化学习（Hong 等人 2020；Konda 和 Tsitsiklis 2000）以及网络架构搜索（Liu、Simonyan 和 Yang 2018）。上述方法仅适用于无约束双层优化，并要求目标函数可微。当下层优化中存在约束时，由于目标函数不可微，这些方法无法直接应用。

贡献。 本文考虑问题 (1) 的一个特殊情形，其中不包含上层约束 r 和 s。一般而言，即使上层问题和下层问题都是凸的，且函数 f、g、p、q 均可微，目标函数 Φ 仍是非凸且不可微的（Hansen、Jaumard 和 Savard, 1992；Liu 等人，2021a）。大多数

*通讯作者。版权所有 © 2023，美国人工智能促进协会（www.aaai.org）。保留所有权利。

针对这一双层优化问题的方法通常十分复杂且计算成本高昂，尤其是在问题维度较大时（Liu 等人，2021a；Dempe 和 Franke，2016）。为应对这一挑战，我们通过计算若干个梯度来确定下降方向，这些梯度能够表示球内所有点处目标函数的梯度；同时，我们为带约束双层优化问题开发了一种计算高效且具有收敛保证的算法。总体贡献总结如下。(i) 首先，我们推导了下层最优解 $y^*(x)$ 连续可微或方向可微的条件。此外，我们给出了 $y^*(x)$ 连续可微时其梯度的解析表达式，以及 $y^*(x)$ 方向可微时其方向导数的解析表达式。(ii) 其次，我们提出梯度近似方法，将非凸且不可微目标函数 Φ 的 Clarke 次微分近似应用于线搜索法。具体而言，使用一组导数来近似当前估计点邻域内所有点处的梯度或方向导数。随后，根据这些导数近似 Clarke 次微分，并将近似 Clarke 次微分用作线搜索的下降方向。(iii) 我们证明，Clarke 次微分近似误差较小，线搜索始终可行，且算法渐近收敛到 Clarke 驻点集合。(iv) 我们通过在超参数优化和元学习任务上开展实验，实证验证了所提算法的有效性。

相关工作。长期以来，受约束优化问题最优解的微分一直受到研究。受约束优化的灵敏度分析（Fiacco，1983；Fiacco 和 Ishizuka 1990；Fiacco 和 McCormick 1970）表明，凸优化问题的最优解 $y^*(x)$ 具有方向可微性，但并非处处可微。这意味着问题 (1) 中的目标函数 $\Phi(x)$ 可能不可微。基于对 KKT条件的隐式微分，这些论文还会在 y^* 于 x 处可微时计算 $\nabla y^*(x)$。Optnet（Amos、Xu 和 Kolter，2017；Amos 和 Kolter 2017；Agrawal 等人 2019）将梯度计算应用于约束双层优化，其中上层优化问题包含一个深度神经网络。具体而言，最优解 $y^*(x)$ 作为深度神经网络中的一层，而 $\nabla y^*(x)$ 则用作反向传播梯度，以优化神经网络参数。然而，以上所有方法均未明确考虑 $y^*(x)$ 和 $\Phi(x)$ 的不可微性，也无法保证收敛。近期的论文（Liu 等人，2021b；Sow 等人，2022）考虑下层优化问题具有简单约束的情形，因此可以轻松计算向约束集的投影，并要求约束集有界。本文考虑等式约束和不等式约束，其一般性强于（Liu 等人，2021b；Sow 等人，2022）中的约束。

记号。对于向量 a，当$a_i > b_i$ 对所有 $1 \le i \le n$时，记 $a > b$ $b \in \mathbb{R}^n$，$a \ge b$。记号 $a = b$、$a \le b$、$a < b$ 和 以类似方式定义。记向量的l_2 范数记为 $\|\cdot\|$。函数 f 在 x 处沿方向 d 且 $\|d\| = 1$ 时的方向导数定义为 $\nabla_d f(x) \triangleq \lim_{h\to 0^+} {}^{f(}{}_{x+hd)-f(x)\,h}$。以 x 为球心、ϵ 为半径的球记为 $\mathcal{B}(x,\epsilon)$。集合 S 的补集记为 S^C。点 x与集合 S 之间的距离定义为 $d(x,S) \triangleq \inf\{\|x-a\| \mid a \in S\}$。$S$ 的凸包记为 $\operatorname{conv} S$。对于集合 S 和函数 f，定义像集 $f(S) \triangleq \{f(x) \mid x \in S\}$。对于有限正整数集 I 和向量函数 p，记子向量函数 $p_I \triangleq [p_{k_1}, \cdots, p_{k_j}, \cdots]^\top$，其中 $k_j \in I$。

2 问题陈述

考虑以下带约束双层优化问题：

$$\min_{x\in\mathbb{R}^{d_x}} \Phi(x) = f\left(x, y^*(x)\right) \tag{2}$$
$$\text{s.t. } y^*(x) = \arg\min_{y\in\mathbb{R}^{d_y}}\{g(x,y) : p\left(x,y\right) \le 0; q\left(x,y\right) = 0\},$$

其中 f，g：$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$；$p$：$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}^m$；$q$：$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}^n$。给定 $x \in \mathbb{R}^{d_x}$，将式 (2) 中的下层优化问题记为 $P(x)$。$P(x)$ 的可行集定义为 $K(x) \triangleq \{y \in \mathbb{R}^{d_y} : p(x, y) \le 0, q(x, y) = 0\}$。假设以下假设成立。

假设1。函数 f、g、p 和 q 二阶连续可微。

假设2。对于所有 $x \in \mathbb{R}^{d_x}$，函数 $g(x,y)$ 关于 y 是 μ-强凸的；对于每个 j，$p_j(x,y)$ 关于 y 是凸的；对于每个 i，$q_i(x,y)$ 关于 y 是仿射的。

注意，上层目标函数 $f(x, y)$ 和整体目标函数 $\Phi(x)$ 均为非凸。下层问题 $P(x)$ 是凸的，其拉格朗日函数为$\mathcal{L}(y, \lambda, \nu, x) \triangleq g(x, y) + \lambda^\top p(x, y) + \nu^\top q(x, y)$，其中 (λ, ν) 是拉格朗日乘子，且 $\lambda \ge 0$。

定义 *1.* 假设在 y 处，对于 $P(x)$，*KKT*条件成立，且拉格朗日乘子为 λ 和 ν。在 y 处，对于 $P(x)$，活跃不等式约束的集合定义为：$J(x,y) \triangleq \{j : 1 \le j \le m, p_j(x,y) = 0\}$。如果某个不等式约束不包含在 $J(x,y)$ 中，则称其为非活跃约束；非活跃约束的集合记为 $J(x,y)^C$。在 y 处，严格活跃不等式约束的集合定义为：$J^+(x,y,\lambda) \triangleq \{j : j \in J(x,y) \quad \lambda_j > 0\}$。在 y 处，非严格活跃不等式约束的集合定义为：$J^0(x,y,\lambda) \triangleq J(x,y) \setminus J^+(x,y,\lambda)$。注意，对于 $j \in J(x,y)$，$\lambda_j \ge 0$；对于 $j \in J^0(x,y,\lambda)$，$\lambda_j = 0$。

定义 2。若对于 $P(x)$，y 处满足线性无关约束资格条件（*LICQ*），则向量$\{\nabla_y p_j(x,y) \quad j \in J(x,y)$；$\nabla_y q_i(x,y) \quad 1 \le i \le n\}$ 线性无关。

假设3。假设对于所有 $x \in \mathbb{R}^{d_x}$，解 $y^*(x)$ 存在于 $P(x)$，并且对于 $P(x)$，$y^*(x)$ 处满足线性无关约束资格条件（*LICQ*）。

3 $y^*(x)$ 的可微性与梯度

本节给出下层最优解 $y^*(x)$ 连续可微或方向可微的充分条件。当 $y^*(x)$ 连续可微时，我们计算其梯度；

当 $y^*(x)$ 方向可微时，我们计算其方向导数。此外，我们给出 $y^*(x)$ 不可微的必要条件，并通过一个数值示例加以说明。

在问题 (2) 中，如果上层目标函数 f和下层问题 y^* 的解连续可微，那么 Φ也连续可微，并且根据复合函数的梯度计算规则，我们有

$$\nabla\Phi(x) = \nabla_x f(x, y^*(x)) + \nabla y^*(x)^\top \nabla_y f(x, y^*(x)). \quad (3)$$

(Domke 2012) 表明，当 p 和 q 不存在时，在某些假设下，y^* 和 Φ 可微。AID 方法（Pedregosa 2016；Gould 等人 2016；Ghadimi 和 Wang 2018；Grazzi 等人 2020；Domke 2012）利用 y^* 和 Φ 的可微性来近似∇y^*，并通过梯度下降最小化 Φ。然而，当下层问题 (2) 受到约束时，情况并非如此。

定理 1 给出了使 $y^*(x)$方向可微的条件。

定理 1. 假设假设*1*、*2*、*3*成立。对于任意 x，以下性质均成立。

(i) $P(x)$ 的全局最小值 $y^*(x)$ 存在且唯一。在 $y^*(x)$ 处，*KKT* 条件成立，且拉格朗日乘子 $\lambda(x)$ 和 $\nu(x)$ 唯一。*(ii)* 向量函数 $z(x) \triangleq [y^*(x)^\top, \lambda(x)^\top, \nu(x)^\top]^\top$ 连续且局部利普希茨连续。$z(x)$ 在任意方向上的方向导数均存在。

如定理 1 的 (i) 部分所示，$y^*(x)$、$\lambda(x)$ 和 $\nu(x)$ 由 x 唯一确定。因此，在本文余下部分，我们简化定义 1 中的记号：$J(x, y^*(x))$ 记为 $J(x)$，$J^+(x, y^*(x)\ \lambda(x))$ 记为 $J^+(x)$，而 $J^0(x, y^*(x)\ \lambda(x))$ 记为 $J^0(x)$。在 (ii) 部分，$z(x)$ 的方向导数计算见附录 C 中的定理 6。

定义 3. 假设对于 $P(x)$，在 y 处 *KKT*条件成立，且拉格朗日乘子为 λ 和 ν。若对于所有 $j \in J(x, y)$，$\lambda_j > 0$，则相对于 λ，严格互补松弛条件（*SCSC*）在 y 处对于 $P(x)$ 成立。

备注1. *KKT*条件包含互补松弛条件（*CSC*）。*SCSC*比*CSC*更强，而*CSC*仅要求对于所有 $j \in J(x, y)$，$\lambda_j \geq 0$。

定理 2 给出了 $y^*(x)$ 连续可微的条件，并推导出 $\nabla y^*(x)$。

定理 2. 假设*1*、*2*、*3*成立。若相对于 $\lambda(x)$，*SCSC* 在 $y^*(x)$ 处成立，则 $z(x)$ 在 x 处连续可微，且其梯度计算如下

$$\left[\nabla_x y^*(x)^\top, \nabla_x \lambda_{J(x)}^\top(x), \nabla_x \nu(x)^\top\right]^\top = -M_+^{-1}(x) N_+(x) \quad (4)$$

以及 $\nabla_x \lambda_{J(x)^C}(x) = 0$，其中 $M_+(x) \triangleq$

$$\begin{bmatrix} \nabla_y^2 \mathcal{L} & \nabla_y p_{J^+(x)}^\top & \nabla_y q^\top \\ \nabla_y p_{J^+(x)} & 0 & 0 \\ \nabla_y q & 0 & 0 \end{bmatrix} (x, y^*(x), \lambda(x), \nu(x))$$

是非奇异的，且 $N_+(x) \triangleq$

$$[\nabla_{xy}^2 \mathcal{L}^\top, \nabla_x p_{J^+(x)}^\top, \nabla_x q^\top]^\top (x, y^*(x), \lambda(x), \nu(x)).$$

定理 2 表明，若 $z(x)$ 不连续可微，则在 $y^*(x)$ 处相对于 $\lambda(x)$ 不满足 SCSC。定义 3 意味着，对于 $P(x)$，在 y 处相对于 λ 满足 SCSC，当且仅当 $J^0(x) = \emptyset$。由此可知，若 $y^*(x)$ 在 x 处不连续可微，则 $J^0(x) \neq \emptyset$，即 $y^*(x)$ 的不可微性出现在具有非严格活跃约束的点处。例1说明了这一结论。

例*1.* 考虑双层优化问题$\Phi(x) = y^*(x)$和下层问题 $P(x)$：$y^*(x) = \arg\min_y\{(y-x^2)^2 : p_1(x, y) = -x - y \leq 0\}$，其中 $x, y \in \mathbb{R}$。$z(x) = [y^*(x)\ \lambda(x)]$ 的解析解为：$y^*(x) = x^2$，$\lambda(x) = 0$ 当 $x \in (-\infty, -1] \cup [0, +\infty)$ 时；$y^*(x) = -x$，$\lambda(x) = -2x(1+x)$ 当 $x \in [-1, 0]$时。相应地，当 $x \in (-1, 0)$时，$J(x) = \{1\}$，$J^+(x) = \{1\}$，$J^0\ x = \emptyset$ $x \in -\infty\ -1 \cup 0\ +\infty$ $J\ x = \emptyset$ ( )；当 (,) (,) 时，( )，$J^+\ x = \emptyset$ $J^0\ x = \emptyset$ $x \in \{-1, 0\}$ $J\ x = \{1\}$ ( )，( )；当 时，( )，$J^+\ x = \emptyset$ $J^0\ x = \{1\}$ $1\ y^*\ x$ ( )，( )。如图所示，( ) 除在 $J^0(x) \neq \emptyset$ 时外处处连续可微。

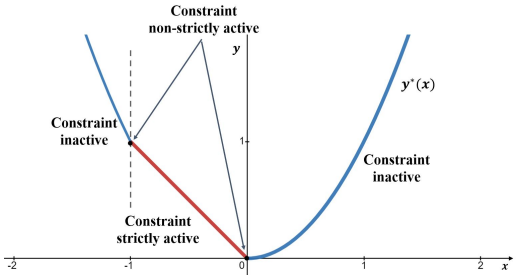


图 1：不可微性的出现。

(4) 中 $z(x)$ 的梯度计算是通过对问题 $P(x)$ 的 KKT条件进行隐式微分推导得到的，这一方法也用于（Fiacco 和 Ishizuka 1990；Amos 和 Kolter 2017；Agrawal 等人 2019）。与这些论文相比，定理 2 在 $z(x)$ 连续可微时，直接确定 $\nabla_x \lambda_{J(x)^C}(x) = 0$，并将 $\lambda_{J(x)^C}(x)$ 排除在 (4) 中逆矩阵的计算之外。附录 C 中的定理6推导了 $z(x)$ 不可微时的方向导数。

考虑下层优化问题 $P(x)$ 无约束的特殊情形。由于不再需要 SCSC，定理 2 中的假设可简化为：g 二阶连续可微，且对于 $x \in \mathbb{R}^{d_x}$，$g(x, y)$ 关于 y 是 μ-强凸的。根据定理 2，最优解 $y^*(x)$ 连续可微，矩阵 $\nabla_y^2 g(x, y)$ 非奇异，梯度可计算为 $\nabla y^*(x) = -[\nabla_y^2 g(x, y)]^{-1} \nabla_{xy}^2 g(x, y)$。这些结果在无约束双层优化的分析与应用中广为人知且被广泛使用（Pedregosa 2016；Franceschi 等人 2017、2018；Ji、Yang 和 Liang 2021）。

4 梯度近似方法

本节中，我们将开发梯度近似方法，以高效求解问题 (2)，其目标函数不可微且非凸。首先，我们定义

Clarke次微分（第4.1节），并高效近似目标函数 $\Phi(x)$ 的 Clarke 次微分（第 4.2 节）。接下来，我们提出梯度近似算法，给出其收敛保证（第4.3节），并介绍其实现细节（第4.4节）。

4.1 Φ的Clarke次微分

如第2节所示，且 Dempe 和 Franke（2016年）和 Liu 等人（2021a）也指出，问题（2）的目标函数 $\Phi(x)$ 通常不可微且非凸。为处理非光滑性和非凸性，我们引入Clarke次微分和Clarke驻点。

定义*4*（*Clarke*次微分和*Clarke*驻点（*Clarke*（*1975*年））） 对于局部利普希茨连续函数 $f: \mathbb{R}^n \to \mathbb{R}$，$f$ 在x 处的 *Clarke*次微分定义为函数 f 在收敛至 x的序列上的梯度极限的凸包，即 $\bar{\partial} f(x) \triangleq conv\{lim_{j\to\infty} \nabla f\left(y^j\right): \left\{y^j\right\} \to x$ 其中，对于所有 $j \in \mathbb{N}\}$，f 在 y^j 处可微。f在 x 处的*Clarke* ϵ-次微分定义为 $\bar{\partial}_\epsilon f(x) \triangleq conv\{\bar{\partial} f(x'): x' \in \mathcal{B}(x,\epsilon)\}$。若点 x 满足 $0 \in \partial f(x)$，则称其为 f 的*Clarke*驻点。

若 y^* 在 x处可微，则有 $\partial y^*(x) = \{\nabla y^*(x)\}$ 且 $\bar{\partial}\Phi(x) = \{\nabla_x f(x, y^*(x)) + \nabla y^*(x)^\top \nabla_y f(x, y^*(x))\}$；否则，$\bar{\partial}\Phi(x) = \{\nabla_x f(x, y^*(x)) + w^\top \nabla_y f(x, y^*(x)) : w \in \partial y^*(x)\}$。以例1和图 1 所示的函数为例，$\partial_\epsilon \Phi(-1) = \partial_\epsilon y^*(-1 = \{[-2 - 2\epsilon, -2] \cup \{-1\}\} = [-2 - 2\epsilon, -1]$，且 $\partial_\epsilon \Phi(0) = \partial_\epsilon y^*(0) = \text{conv}\{[0, 2\epsilon] \cup \{-1\}\} = [-1, 2\epsilon]$。

4.2 Clarke 次微分近似

基于梯度的方法已应用于凸和非凸优化问题（Hardt、Recht 和 Singer，2016年；Nemirovski 等，2009）。收敛性要求目标函数可微。如果存在目标函数不可微的点，算法访问这些点的概率非零，而这些点处的梯度未定义（Bagirov 等，2020）。此外，即使目标函数在所有被访问点处都可微，也可能发生振荡（Bagirov、Karmitsa 和 Mäkelä，2014）。

为处理不可微性，梯度采样法（Burke、Lewis 和 Overton，2005；Kiwiel，2007；Bagirov 等，2020）利用当前估计值邻域内的梯度来近似 Clarke次微分并确定下降方向。具体而言，该方法在邻域 $\mathcal{B}(x^0, \epsilon)$ 内采样一组点，选取目标函数可微的点，然后计算这些采样点处梯度的凸包。

然而，在问题 (2) 中，点采样的计算开销非常大。对于每个采样点 x^j，为了检查 Φ 的可微性，我们需要求解下层优化 $P(x^j)$，以获得 $y^*(x^j)$、$\lambda(x^j)$ 和 $\nu(x^j)$，并检查 SCSC。此外，采样完成后，每个点处的梯度通过式 (4) 计算。随着维度 d_x 的增加，为确保近似的准确率，采样数量也需要增加。更具体地，如图所示在 (Kiwiel2007) 中，若采样数量大于 $d_x + 1$，则该算法收敛。上述过程在每次优化迭代中执行。

为应对计算上的挑战，我们用少量梯度来近似 Clarke ϵ-次微分，这些梯度能够表示邻域内所有点处的梯度。以下命题区分两种情况：Φ 在 $\mathcal{B}(x^0, \epsilon)$ 上连续可微（命题1），以及它不可微（命题2）。

命题1。 假设假设*1*、*2*、*3*成立。考虑$x^0 \in R^{d_x}$。存在充分小的 $\epsilon > 0$，使得若对任意 $x \in \mathcal{B}(x^0, \epsilon)$，相对于 $\lambda(x)$，在 $y^*(x)$ 处满足 *SCSC*，则 $\nabla\Phi(x^0) \in \bar{\partial}_\epsilon \Phi(x^0)$ 并且

$$|\|\nabla\Phi(x^0)\| - d(0, \bar{\partial}_\epsilon \Phi(x^0))| < o(\epsilon).$$

命题1表明，单点处的梯度 $\nabla\Phi$ 可用于近似 Clarke ϵ-次微分 $\bar{\partial}_\epsilon \Phi(x^0)$，且近似误差的阶为 ϵ。回顾一下，梯度 $\nabla\Phi(x^0)$ 可通过式（3）和式（4）计算。图2展示了例1中问题的近似情况。若在 $y^*(x)$ 处满足 SCSC，且 $\Phi(x)$ 在 $\mathcal{B}(x^0, \epsilon)$ 上连续可微，则 $\bar{\partial}_\epsilon \Phi(x^0) = [2x^0 - 2\epsilon, 2x^0 + 2\epsilon]$ 可由 $\nabla\Phi(x^0) = 2x^0$近似，近似误差为 2ϵ。$\bar{\partial}_\epsilon \Phi(x^1)$ 和 $\bar{\partial}_\epsilon \Phi(x^2)$ 的近似也可用类似方式完成。

考虑到 $\Phi(x)$ 在 $\mathcal{B}(x^0, \epsilon)$ 中的某些点处不可连续可微。定义集合 $I^\epsilon(x^0)$，其中包含所有 j x' $x'' \in \mathcal{B}$ x^0 ϵ $j \in$，使 J^+ x' C $j \in J^+$ x'' I_+^ϵ x^0 得存在 , (,)，满足 ( ) 和 ( )。定义集合 ( )，其中包含所有 j，使得对于任意 $x \in \mathcal{B}(x^0, \epsilon)$，$j \in J^+(x)$。若 $I^\epsilon(x^0)$ 非空，则存在一点 $x \in \mathcal{B}(x^0, \epsilon)$，使得 SCSC 在 $y^*(x)$ 处不成立。$I^\epsilon(x^0)$ 的幂集将 $\mathcal{B}(x^0, \epsilon)$ 划分为若干子集，其中 $\Phi(x)$ 和 $y^*(x)$ 在每个子集中均连续可微。例1中的问题示意图见图3。点 $x' = -1$ 属于 $(x^0 - \epsilon, x^0 + \epsilon)$，且 SCSC 在 $y^*(x')$ 处不成立。注意，$I_+^\epsilon(x^0) = \emptyset$且 $I^\epsilon(x^0) = \{1\}$。于是，$I^\epsilon(x^0) = \{1\}$ 的幂集为 $\{S_{(1)}, S_{(2)}\}$，其中包含 $S_{(1)} = \emptyset$ 和 $S_{(2)} = \{1\}$。因此，$\mathcal{B}(x^0, \epsilon)$ 被划分为两个子集：约束 p_1 未激活的子集（球内的蓝色一侧），对应于 $S_{(1)}$；以及约束 p_1 严格激活的子集（球内的红色一侧），对应于 $S_{(2)}$。它们的边界是点 x'，此处约束 p_1 非严格激活。可以看出，$y^*(x)$ 在每个子集上连续可微；当 ϵ 较小时，子集内部的梯度变化较小。相比之下，两个子集之间的梯度变化较大。受命题1启发，我们计算一个代表性梯度，以近似 $\nabla y^*(x)$ 在 $\mathcal{B}(x^0, \epsilon)$ 的每个子集内的值。

现在我们继续推广上述思路。回顾一下，$\nabla\Phi(x)$ 根据式 (3) 计算，且 f 是二阶连续可微的。定义

$$\begin{aligned} G(x^0, \epsilon) \triangleq \{&\nabla_x f\left(x^0, y^*\left(x^0\right)\right) + w^S(x^0)^\top \\ &\nabla_y f\left(x^0, y^*\left(x^0\right)\right) : S \subseteq I^\epsilon(x^0)\}, \end{aligned} \tag{5}$$

其中，$w^S(x^0)$ 是通过提取前 d_x 行得到的

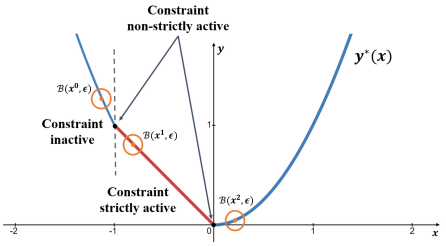


图2：SCSC 在 $\mathcal{B}(x^0,\ \epsilon)$ 上成立

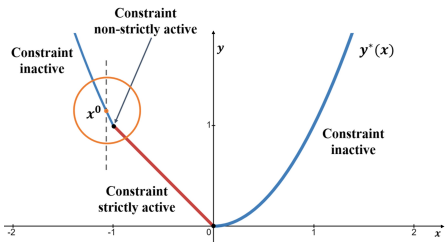


图3：$\mathcal{B}(x^0,\ \epsilon)$ 内的子集

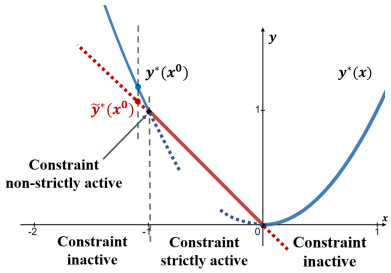


图4：近似的 $\bar{\partial}_\epsilon y^*(x^0)$

从矩阵 $-M_\epsilon^S(x^0,y^*(x^0))^{-1}N_\epsilon^S(x^0,y^*(x^0))$ 中得到，其中

$$M_\epsilon^S \triangleq \begin{bmatrix} \nabla_y^2\mathcal{L} & \nabla_y p_{I_+^\epsilon(x^0)}^\top & \nabla_y q^\top & \nabla_y p_S^\top \\ \nabla_y p_{I_+^\epsilon(x^0)} & 0 & 0 & 0 \\ \nabla_y q & 0 & 0 & 0 \\ \nabla_y p_S & 0 & 0 & 0 \end{bmatrix},$$

以及 $N_\epsilon^S \triangleq \left[\nabla_{xy}^2\mathcal{L}^\top, \nabla_x p_{I_+^\epsilon(x^0)}^\top\ \nabla_x q^\top,\ \nabla_x p_S^\top\right]^\top$。这里，$S$ 是 $I^\epsilon(x^0)$ 的子集，而 $w^S(x^0)$ 是代表性梯度，用于近似与 S 对应的 $\mathcal{B}(x^0,\ \epsilon)$ 子集内的 $\nabla y^*(x)$。命题2表明，Clarke ϵ-次微分 $\bar{\partial}_\epsilon y^*(x^0)$ 可以用代表性梯度集合 $G(x^0,\ \epsilon)$ 近似，且近似误差的阶为 ϵ。

命题2。 假设*1*、*2*、*3*成立。考虑$x^0 \in \mathbb{R}^{d_x}$，并假设存在一个充分小的 $\epsilon > 0$，使得存在 $x \in \mathcal{B}$（x^0、ϵ），使得 y^*（x）在 x处不是连续可微的。则对于任意$z \in \mathbb{R}^{d_x}$，以下不等式成立：

$$|d(z, \operatorname{conv} G(x^0,\epsilon)) - d(z, \bar{\partial}_\epsilon \Phi(x^0))| < o(\epsilon).$$

例1的代表性梯度 $w^S(x^0)$的计算如图4所示。由于 x^0 靠近两个子集的边界，命题2采用 $w^{S_{(1)}}(x^0) = \nabla y^*\ x^0$ ( ) 来近似具有非活跃约束的子集（蓝色一侧）的梯度，并采用 $w^{S_{(2)}}(x^0) = \nabla\tilde{y}^*(x^0)$来近似具有严格活跃约束的子集（红色一侧）中的梯度。二阶可微函数$\tilde{y}^*(x)$是 $y^*(x)$的扩展（参见（Dempe 1998）的(12.8)中$x^I(\cdot)$的定义）。梯度 $\nabla y^*(x^0)$和 $\nabla\tilde{y}^*(x^0)$分别在矩阵 $-M_\epsilon^{S_{(1)}{}^{-1}}N_\epsilon^{S_{(1)}}$和 $-M_\epsilon^{S_{(2)}{}^{-1}}N_\epsilon^{S_{(2)}}$中计算。然后，利用代表性梯度$w^{S_{(1)}}(x^0)$和$w^{S_{(2)}}(x^0)$来近似$\bar{\partial}_\epsilon y^*(x^0)$。接着，我们可以使用 $-1 \in [x^0-\epsilon, x^0+\epsilon]$计算 $G(x^0,\epsilon) = \{2x^0, -1\}$和$\bar{\partial}_\epsilon\Phi(x^0) = [2x^0 - 2\epsilon, -1]$。对于任意 z，近似误差 $|d(z, \operatorname{conv} G(x^0, \epsilon)) - d(z, \bar{\partial}_\epsilon\Phi(x^0))|$小于或等于 2ϵ。

4.3 梯度近似算法

我们提出的梯度近似算法总结于算法1，是一种线搜索算法。它使用Clarke次微分的近似作为下降方向

算法1：梯度近似方法

要求：初始点 x^0**；初始近似半径** $\epsilon_0 \in$**(0,** ∞**)；初始驻点性目标** $\nu_0 \in$ **(0,** ∞**)；线搜索参数** $(\beta,\gamma) \in$**(0, 1)** $\times$ **(0, 1)；终止容差** $(\epsilon_{\text{opt}};\nu_{\text{opt}}) \in [0,\infty) \times [0,\infty)$**；折扣因子** $(\theta_\epsilon,\theta_\nu) \in$**(0,1)** $\times$ **(0,1)**。**1：对** $k \in \mathbb{N}$ **执行 2：求解下层优化问题** $P(x^k)$**，并获得** $y^*(x^k)$**、** $\lambda(x^k)$ **和** $\nu(x^k)$ **3：根据(6)和(7)检查** y^* **在** $\mathcal{B}\left(x^k、\ \epsilon_k\right)$ **上的可微性 4：如果** y^* **在** $\mathcal{B}\left(x^k、\ \epsilon_k\right)$ **上连续可微，则 5：根据(3)计算** $g^k = \nabla\Phi(x^k)$ **6：否则 7：根据(5)计算** $G(x^k,\epsilon_k)$ **8：** $\bar{\partial}_{\epsilon_k}\Phi(x^k)$ **=conv** $G(x^k,\epsilon_k)$ **9：计算** $g^k =$**min** $\{\|g\|：\ g \in$**conv** $G(x^k,\epsilon_k)\}$**10：结束 if 11：如果** $\left\|g^k\right\| \le \nu_{\text{opt}}$ **和** $\epsilon_k \le \epsilon_{\text{opt}}$**，则 12：输出：** x^k **并终止 13：结束 if 14：如果** $\left\|g^k\right\| \le \nu_k$**，则 15：** $\nu_{k+1} \leftarrow \theta_\nu\nu_k$、$\epsilon_{k+1} \leftarrow \theta_\epsilon\epsilon_k$**和** $t_k \leftarrow 0$ **16：否则 17：通过线搜索计算**

$$t_k = \{t \in : \{\gamma,\gamma^2,\ldots\}\ \Phi\ x^k - tg^k\ < \Phi\ x^k - \beta t\|g^k\|^2\}$$

max：( ) ( ) 18： $\nu_{k+1} \leftarrow \nu_k$ **和** $\epsilon_{k+1} \leftarrow \epsilon_k$ **19：结束 if 20：** $x^{k+1} \leftarrow x^k - t_k g^k$**21：结束 for**

用于线搜索。在迭代 k 中，我们首先求解下层优化问题 $P(x^k)$，并得到 $y^*(x^k)$、$\lambda(x^k)$ 和 $\nu(x^k)$。为降低计算复杂度，迭代 k 中的解用作初始点，以求解迭代 $k+1$ 中的 $P(x^{k+1})$。其次，我们检查 y^* 在 $\mathcal{B}\left(x^k,\ \epsilon_k\right)$ 上的可微性，其实现细节见第4.4节。如果 y^* 在 $\mathcal{B}\left(x^k,\ \epsilon_k\right)$ 上连续可微，我们使用 $\nabla\Phi(x^k)$ 近似 $\bar{\partial}_\epsilon\Phi(x^k)$，对应命题1。否则，使用 $G(x^k,\ \epsilon_k)$，对应命题2。计算 $G(x^k,\ \epsilon_k)$ 的细节见式(5)和第4.4节。第三，线搜索方向 g^k 由一个向量确定，该向量的范数最小

，它取自 $G(x^k,\ \epsilon_k)$ 的凸包中的所有向量。在优化过程中，随着迭代次数 k 增加，近似半径 ϵ_k 减小。根据命题1和命题2，Clarke次微分的近似误差逐渐减小。接下来，我们刻画算法1的收敛性。

定理 3。 假设假设*1*、*2*、*3*成立，且 $\Phi(x)$ 在 $\mathbb{R}^{d_x}$上有下界。令 $\{x^k\}$ 为带有 $\nu_{\text{opt}}=\epsilon_{\text{opt}}=0$的算法*1*生成的序列。则，

(i) 对于每个 k，第*17*行中的线搜索都有解 t_k。*(ii)* 极限 $\lim_{k\to\infty}\nu_k=0$，极限$\lim_{k\to\infty}\epsilon_k=0$。*(iii)* 下极限$\liminf_{k\to\infty} d(0,\ \bar{\partial}\Phi(x^k))=0$。*(iv)* $\{x^k\}$ 的每个极限点都是 Φ 的 *Clarke* 驻点。

如果目标函数 Φ 是非凸但光滑的，则性质(iii)可化为$\liminf_{k\to\infty}\|\nabla\Phi(x^k)\|=0$，这是一种广泛用于光滑非凸优化的收敛判据（Nesterov 1998年；Jin等人 2021）。$\{x^k\}$ 存在极限点的一个充分条件是该序列有界。

4.4 实现细节

检查 y^* **在** $\mathcal{B}\left(x^0\right.$上的可微性，$\left.\epsilon_0\right)$ 我们提出命题 3，用于检查 y^* 在 $\mathcal{B}\left(x^0\right.$上的可微性，$\left.\epsilon_0\right)$这是算法1第3行所要求的。

命题 *3.* 考虑 $x^0\in\mathbb{R}^{d_x}$ 和 $\epsilon>0$。假设假设*1*、*2*、*3*成立。则函数 $\Phi(x)$、$\lambda_j(x)$ 和 $p_j(x,y^*(x))$ 在 $\mathcal{B}(x^0,\epsilon)$ 上的 *Lipschitz*常数存在，分别记为 $l_\Phi(x^0,\epsilon)$、$l_{\lambda_j}(x^0,\epsilon)$ 和 $l_{p_j}(x^0,\epsilon)$。此外，假设在 $y^*(x^0)$ 处相对于 $\lambda(x^0)$ 满足 *SCSC*。若存在 $\epsilon_1>0$ 使得

$$\begin{aligned}&\lambda_j(x^0)>l_{\lambda_j}(x^0,\epsilon_1)\epsilon_1\ \textit{for all}\ j\in J(x^0),\\&p_j(x^0,y^*(x^0))<-l_{p_j}(x^0,\epsilon_1)\epsilon_1\ \textit{for all}\ j\notin J(x^0),\end{aligned}\tag{6}$$

那么，y^* 在 $\mathcal{B}$（x^0、ϵ_1）上连续可微。

命题 3 表明，如果对于任意 j，以下条件之一成立，则 y^* 在 x^0的某个邻域上连续可微：(i) 当约束 $p_j(x^0,y^*(x^0))$ 处于激活状态时，λ_j以非平凡的幅度大于零；或 (ii) 当约束未激活时，满足 $p_j(x^0,y^*(x^0))$ 的程度是非平凡的。对于情形 (i)，$\lambda_j(x)>0$，并且对于所有 $x\in\mathcal{B}(x^0,\epsilon)$，该约束都严格激活；对于情形 (ii)，$p_j(x,y^*(x))<0$，并且对于所有 $x\in\mathcal{B}(x^0,\epsilon)$，该约束都未激活。以图2所示的例1中的问题为例，$y^*$ 在 $\mathcal{B}(x^0,\epsilon)$、$\mathcal{B}(x^1,\epsilon)$ 和 $\mathcal{B}(x^2,\epsilon)$ 上连续可微，并且在每个球中，该约束均处于未激活或严格激活状态。

根据命题 3，我们评估了 $y^*(x)$ 和 $\Phi(x)$ 在 $\mathcal{B}(x^0,\ \epsilon)$ 上的可微性。具体而言，如果满足 (6)，我们近似地认为 y^* 和 Φ 在 $\mathcal{B}(x^0,\ \epsilon)$ 上连续可微；否则，存在 $x\in\mathcal{B}(x^0,\ \epsilon)$，使得 y^* 和 Φ 在 x 处不连续可微。Lipschitz常数 $l_{\lambda_j}(x^0,\ \epsilon)$ 和 $l_{p_j}(x^0,\ \epsilon)$ 的计算结果如下

$$\begin{aligned}l_{\lambda_j}(x^0,\epsilon)&=\|\nabla\lambda_j(x^0)\|+\delta,\\l_{p_j}(x^0,\epsilon)&=\|\nabla_x p_j(x^0,y^*(x^0))+\\&\qquad\nabla y^*(x^0)^\top\nabla_y p_j(x^0,y^*(x^0))\|+\delta,\end{aligned}\tag{7}$$

其中，δ 是一个较小的参数，$\nabla_x p_j(x^0,y^*(x^0))$ 和 $\nabla\lambda_j(x^0)$ 在 (4) 中给出。这里，对于函数 f，我们近似计算其在 $\mathcal{B}(x^0,\epsilon)$ 上的 Lipschitz 常数，该常数定义为 $l_f(x^0,\epsilon)\triangleq\sup_x\{\|\nabla f(x)\|:\ x\in\mathcal{B}(x^0,\epsilon)\}$，并将其近似为 $l_f(x^0,\epsilon)\approx\|\nabla f(x^0)\|+\delta$。随着 ϵ 减小，$\mathcal{B}(x^0,\epsilon)$ 中的 f 趋近于一个仿射函数，从而 $l_f(x^0,\epsilon)$ 的近似误差也随之减小。

计算 $G(x^0,\ \epsilon)$ 为了在算法1的第7行计算 $G(x^0,\ \epsilon)$，我们需要计算命题2中定义的集合 $I_+^\epsilon(x^0)$ 和 $I^\epsilon(x^0)$。与命题3中的思路类似，我们将 $I_+^\epsilon(x^0)$ 和 $I^\epsilon(x^0)$ 计算为

$$\begin{aligned}I_+^\epsilon(x^0)&=\left\{j\in J(x^0):\lambda_j(x^0)>l_{\lambda_j}(x^0,\epsilon)\epsilon\right\},\\I_-^\epsilon(x^0)&=\left\{j\notin J(x^0):p_j(x^0,y^*(x^0))<-l_{p_j}(x^0,\epsilon)\epsilon\right\},\\I^\epsilon(x^0)&=\left\{j:j\notin I_+^\epsilon(x^0)\cup I_-^\epsilon(x^0)\right\}.\end{aligned}\tag{8}$$

回顾一下，对于问题 $P(x^0)$，KKT条件在 $y^*(x^0)$ 处成立，因此对于任意 $x\in\mathcal{B}(x^0,\ \epsilon)$，$p_j(x,\ y^*(x))=0$ 对于$j\in I_+^\epsilon(x^0)$，且 $\lambda_j(x)=0$ 对于 $j\in I_-^\epsilon(x^0)$。这里，我们还使用式（7）中给出的 l_{λ_j} 和 l_{p_j} 作为 Lipschitz常数。当 y^* 和 λ 在 x^0 处不可微时，我们在 x^0 附近采样一个点 x'，使得 y^* 和 λ 在 x' 处可微，然后将式（7）中的 $\nabla\lambda(x^0)$ 和 $\nabla y^*(x^0)$ 替换为 $\nabla\lambda(x')$ 和 $\nabla y^*(x')$。

5 实验

5.1 超参数优化

超参数优化已得到广泛研究（Pedregosa 2016；Franceschi 等人 2017；Ji、Yang 和 Liang 2021）。然而，现有方法无法处理约束学习问题的超参数优化，例如支持向量机（SVM）分类（Cortes 和 Vapnik 1995）、约束强化学习（Achiam 等人 2017；Chen、Dong 和 Wang 2021；Xu、Liang 和 Lan 2021）。我们将所提出的算法应用于约束学习的超参数优化。

SVM 超参数优化 我们优化 SVM 优化中的超参数，即针对分离违例的惩罚项。我们在 Dua 和 Graff 2017 中的糖尿病数据集上，对线性支持向量机和核化支持向量机进行了实验。这是首次解决 SVM 的超参数优化问题。我们在附录 B.1 中详细介绍了问题的形式化描述和实现设置。如图 5 所示，线性支持向量机和核化支持向量机的损失都几乎收敛，最终测试准确率远高于随机选择超参数所得的准确率；随机选择的超参数即为优化的初始点。

数据超清洗 数据超清洗（Franceschi 等人 2017；Shaban 等人 2019）旨在训练一个分类器，其中训练数据的标签以概率 p 被污染（即污染率）。我们将该问题建模为支持向量机的超参数优化，并在（Dua 和 Graff 2017）提供的乳腺癌数据集上进行实验。问题建模和实现设置见附录 B.1。我们将梯度近似方法与（Amos 和 Kolter 2017；Lee 等，2019年）中使用的直接梯度

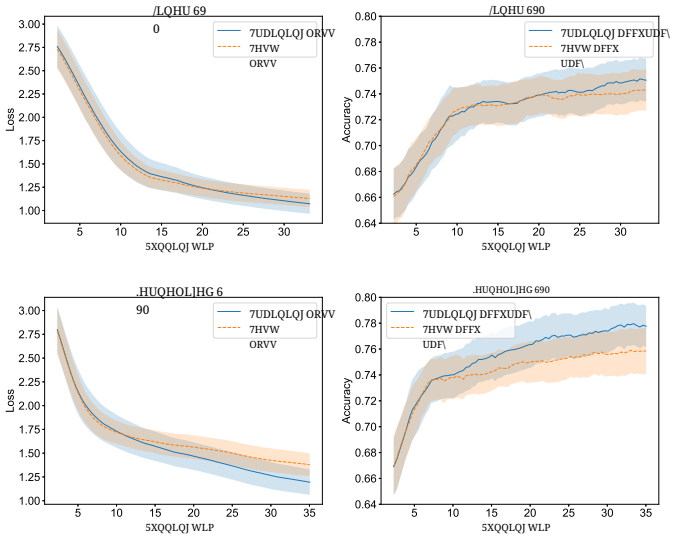


图 5：线性支持向量机和核化支持向量机超参数优化中的损失和准确率随运行时间的变化

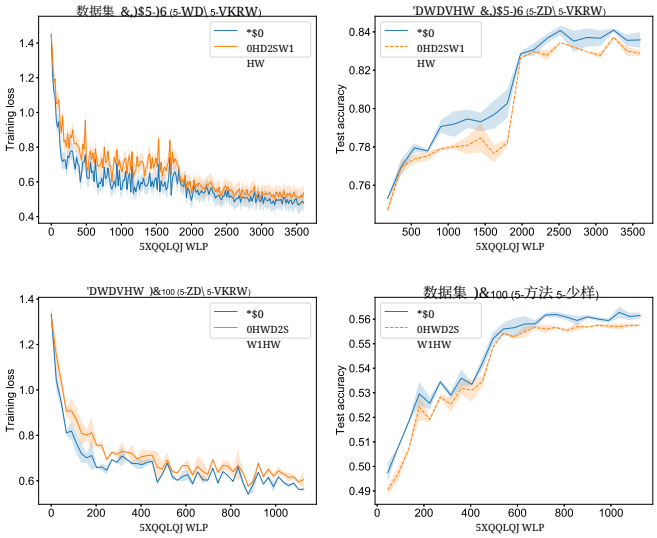


图 7：MetaOptNet 与梯度近似方法（GAM）的比较。对于每个数据集，左图：训练损失随运行时间的变化；右图：测试准确率随运行时间的变化。

下降进行比较。如图 6 所示，在训练和测试阶段，无论是损失还是准确率，我们的方法都比基准方法收敛得更快。此外，两种方法都在受污染数据（$p = 0.4$）上实现了 96.2% 的测试准确率。该准确率与数据未受污染时支持向量机模型的 96.5% 测试准确率相当。

5.2 元学习

少样本学习的元学习方法已在（Rajeswaran 等人 2019；Lee 等人 2019；Ji、Yang 和 Liang 2021）中被表述为双层优化问题。具体而言，MetaOptNet（Lee 等人 2019）中的问题具有带下层约束的问题（2）的形式。然而，其优化过程并未显式考虑

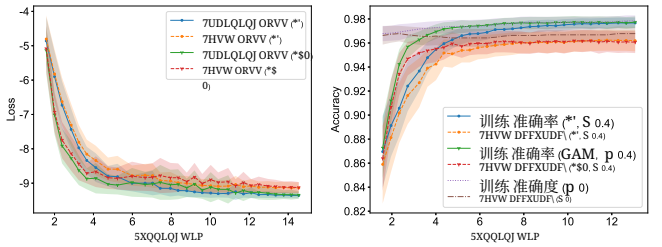


图 6：在污染率为 $p = 0.4$ 的数据超清洗中，梯度下降（GD）与梯度近似方法（GAM）的比较。左图：GD 和 GAM 的训练损失与测试损失随运行时间的变化；右图：使用 $p = 0.4$ 时 GD 和 GAM 的训练准确率与测试准确率，以及使用 $p = 0$ 时的训练准确率与测试准确率随运行时间的变化。

目标函数的不可微性，因此无法保证收敛。在实验中，我们将自己的算法与 MetaOptNet 中的优化方法进行比较，所用数据集为 CIFAR-FS（Bertinetto 等，2018）和 FC100（Oreshkin、Rodríguez López 和 Lacoste 2018），这两个数据集广泛用于少样本学习。附录 B.2 介绍了问题表述和实验设置的详情。

图 7 比较了我们的梯度近似方法与 MetaOptNet 中的直接梯度下降方法（Lee 等，2019年）。两种算法采用完全相同的训练配置，包括网络结构、每个训练轮次的学习率以及批量大小。对于 CIFAR-FS 和 FC100 数据集，与 MetaOptNet 中的优化相比，我们的方法在训练损失和测试准确率方面收敛得更快，并且最终测试准确率更高。请注意，在这项实验中，两种算法的唯一差别在于下降方向的计算方式。结果表明，我们算法中的 Clarke 次微分近似比将梯度用作下降方向效果更好。这与命题2一致：用一组代表性梯度而非单个梯度来近似 Clarke 次微分更为合适。附录 B.2 中给出了与其他元学习方法的更多比较结果。

6 结论

我们针对双层优化提出了一种梯度近似方法，其中下层优化问题为带有等式约束和不等式约束的凸优化问题，而上层优化为非凸优化。该方法能够高效近似非光滑目标函数的Clarke次微分，并在理论上保证收敛。实验验证了我们的理论分析，并展示了该算法的优越有效性。

7 致谢

本研究部分得到了美国国家科学基金会（NSF）项目资助，项目编号为电气、通信与网络系统（ECCS）1846706和电气、通信与网络系统（ECCS）2140175。

参考文献

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P. 2017. 约束策略优化. In 国际机器学习大会, 22–31. PMLR. Agrawal, A.; Amos, B.; Barratt, S.; Boyd, S.; Diamond, S.; and Kolter, J. Z. 2019. 可微凸优化层. 神经信息处理系统进展, 9562–9674. Amos, B.; and Kolter, J. Z. 2017. Optnet：将可微优化作为神经网络中的一层. In 国际机器学习大会, 136–145. PMLR. Amos, B.; Xu, L.; and Kolter, J. Z. 2017. 输入凸神经网络. In 国际机器学习大会, 146–155. PMLR. Bagirov, A.; Karmitsa, N.; and Mäkelä, M. M. 2014. 非光滑优化导论：理论、实践与软件, volume 12. 施普林格. Bagirov, A. M.; Gaudioso, M.; Karmitsa, N.; Mäkelä, M. M.; and Taheri, S. 2020. 非光滑优化的梯度采样方法. In数值非光滑优化：前沿算法. 施普林格. Bard, J.; and Moore, J. 1990. 一种用于双层规划问题的分支定界算法. *SIAM*科学与统计计算杂志, 11. Bard, J. F.; and Falk, J. E. 1982. 多层规划问题的显式解. 计算机与运筹学, 9(1): 77–100. Bengio, Y. 2000. 超参数的基于梯度的优化. 神经计算, 12(8): 1889–1900. Bertinetto, L.; Henriques, J. F.; Torr, P. H.; and Vedaldi, A. 2018. 使用可微闭式求解器进行元学习. *arXiv* 预印本 *arXiv:1805.08136*. Bottou, L.; and Bousquet, O. 2008. 大规模学习的权衡. 神经信息处理系统进展, 161–168. Burke, J. V.; Lewis, A. S.; and Overton, M. L. 2005. 一种用于非光滑非凸优化的鲁棒梯度采样算法. *SIAM* 优化杂志, 15(3): 751–779. Chen, Y.; Dong, J.; and Wang, Z. 2021. 一种求解约束马尔可夫决策过程的原始-对偶方法. *arXiv* 预印本 *arXiv:2101.10895*. Clarke, F. H. 1975. 广义梯度及其应用.美国数学会会刊, 205: 247–262. Cortes, C.; and Vapnik, V. 1995. 支持向量网络.机器学习, 20(3): 273–297. Dempe, S. 1998. 一种求解双层规划问题的隐函数方法. In 多层优化：算法与应用, 273–294. 施普林格. Dempe, S.; and Franke, S. 2016. 凸双层优化问题的求解. 计算优化与应用, 63(3): 685–703. Domke, J. 2012. 基于优化的建模通用方法。载于人工智能与统计学，318–326。PMLR。Dua, D.; and Graff, C. 2017. UCI 机器学习资料库。Fiacco, A. V. 1983. 非线性规划中的灵敏度与稳定性分析导论。Academic Press。Fiacco, A. V.; and Ishizuka, Y. 1990. 非线性规划的灵敏度与稳定性分析。《运筹学年鉴》，27(1): 215–235。Fiacco, A. V.; and McCormick, G. P. 1970. 非线性规划：序贯无约束极小化技术。《*SIAM*评论》，12(4): 593–594。Finn, C.; Abbeel, P.; and Levine, S. 2017. 用于深度网络快速适应的模型无关元学习。载于国际机器学习大会，1126–1135。PMLR。Franceschi, L.; Donini, M.; Frasconi, P.; and Pontil, M. 2017. 基于前向和反向梯度的超参数优化。载于国际机器学习大会，1165–1173。PMLR。Franceschi, L.; Frasconi, P.; Salzo, S.; Grazzi, R.; and Pontil, M. 2018. 用于超参数优化与元学习的双层规划。载于国际机器学习大会，1568–1577。PMLR。Ghadimi, S.; and Wang, M. 2018. 双层规划的近似方法。*arXiv* 预印本 *arXiv:1802.02246*。Giorgi, G.; and Zuccotti, C. 2018. 关于可微性假设下非线性规划与变分不等式中灵敏度和稳定性的教程。技术报告，DEM工作论文系列。Gould, S.; Fernando, B.; Cherian, A.; Anderson, P.; Cruz, R. S.; and Guo, E. 2016. 关于参数化的极小值与极大值问题的微分及其在双层优化中的应用。*arXiv* 预印本 *arXiv:1607.05447*。Grazzi, R.; Franceschi, L.; Pontil, M.; and Salzo, S. 2020. 关于超梯度计算的迭代复杂度。载于国际机器学习大会，3748–3758。PMLR。Hansen, P.; Jaumard, B.; and Savard, G. 1992. 线性双层规划的新分支定界规则。*SIAM* 科学计算杂志，13(5): 1194–1217。Hardt, M.; Recht, B.; and Singer, Y. 2016. 训练更快，泛化更好：随机梯度下降的稳定性。载于国际机器学习大会，1225–1234。PMLR。Hestenes, M. R.; and Stiefel, E. 1952. 求解问题的共轭梯度法。《美国国家标准局研究杂志》，49(6): 409。Hofmann, T.; Schölkopf, B.; and Smola, A. J. 2008. 机器学习中的核方法。《统计年鉴》，36(3): 1171 – 1220。Hong, M.; Wai, H.-T.; Wang, Z.; and Yang, Z. 2020. 双层优化的双时间尺度框架：复杂度分析及其在行动者-评论家中的应用。*arXiv* 预印本 *arXiv:2007.05170*。

Ji, K.; Lee, J. D.; Liang, Y.; and Poor, H. V. 2020. 具有任务特定参数适应的元学习收敛性。神经信息处理系统进展，11490–11500。Ji, K.; Yang, J.; and Liang, Y. 2021. 双层优化：收敛分析与改进设计。载于国际机器学习大会，4882–4892。PMLR。Jin, C.; Netrapalli, P.; Ge, R.; Kakade, S. M.; and J ordan, M. I. 2021. 机器学习的非凸优化：梯度、随机性与鞍点。《美国计算机学会杂志》，68(2): 1–29。Kiwiel, K. C. 2007. 非光滑非凸优化的梯度采样算法的收敛性。*SIAM* 优化杂志，18(2): 379–388。Koji ma, M. 1980. 非线性规划中的强稳定驻点解。载于不动点的分析与计算，93–138。Elsevier。Kolstad, C. D.; and Lasdon, L. S. 1990. 大型双层数学规划的导数求值与计算经验。《优化理论与应用杂志》，65(3): 485–499。Konda, V. R.; and Tsitsiklis, J. N. 2000. 行动者-评论家算法。载于神经信息处理系统进展，1008–1014。Kyparisis, J. 1985. 非线性规划中库恩-塔克乘子的唯一性。数学规划，32(2): 242–246。Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. 通过可微凸优化实现元学习。载于*2019 IEEE/CVF*计算机视觉与模式识别会议，10649–10657。Liu, H.; Simonyan, K.; and Yang, Y. 2018. DARTS：可微架构搜索。载于国际学习表征会议。Liu, R.; Gao, J.; Zhang, J.; Meng, D.; and Lin, Z. 2021a. 从统一视角探究用于学习和视觉的双层优化：综述及后续进展。*IEEE*模式分析与机器智能汇刊。Liu, R.; Liu, Y.; Zeng, S.; and Zhang, J. 2021b. 面向具有非凸跟随者的基于梯度的双层优化及后续研究。神经信息处理系统进展，34: 8662–8675。Lorraine, J.; Vicol, P.; and Duvenaud, D. 2020. 通过隐式微分优化数百万个超参数。载于国际人工智能与统计会议，1540–1552。PMLR。Malanowski, K. 1985. 凸规划问题解关于参数的可微性。数学规划，33(3): 352–361。Nemirovski, A.; Juditsky, A.; Lan, G.; and Shapiro, A. 2009. 随机规划的稳健随机近似方法。*SIAM* 优化杂志，19(4): 1574–1609。Nesterov, Y. 1998. 凸规划导论讲义。

Oreshkin, B.; Rodríguez López, P.; and Lacoste, A. 2018. Tadam：用于改进少样本学习的任务依赖型自适应度量。神经信息处理系统进展，719–729。Pedregosa, F. 2016. 使用近似梯度进行超参数优化。载于国际机器学习大会，737–746。PMLR。Rajeswaran, A.; Finn, C.; Kakade, S. M.; and Levine, S. 2019. 基于隐式梯度的元学习。载于神经信息处理系统进展，113–124。Ralph, D.; and De mpe, S. 1995. 参数化非线性规划解的方向导数。数学规划，70(1)：159–172。Savard, G.; and Gauvin, J. 1994. 非线性双层规划问题的最速下降方向。运筹学快报，15(5)：265–272。Schramm, H.; and Zowe, J. 1992. 一种用于最小化非光滑函数的束方法思想：概念构想、收敛性分析与数值结果。*SIAM* 优化杂志，2(1)：121–152。Shaban, A.; Cheng, C.-A.; Hatch, N.; and Boots, B. 2019. 用于双层优化的截断反向传播。载于第*22*届国际人工智能与统计会议，1723–1732。PMLR。Shi, C.; Lu, J.; and Zhang, G. 2005. 一种用于线性双层规划的扩展 Kuhn-Tucker 方法。应用数学与计算，162(1)：51–63。Snell, J.; Swersky, K.; and Zemel, R. 2017. 用于少样本学习的原型网络。神经信息处理系统进展，4080–4090。Sow, D.; Ji, K.; Guan, Z.; and Liang, Y. 2022. 一种针对具有多个内层极小值的双层优化的约束优化方法。*arXiv*预印本 *arXiv:2203.01123*。Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. 学习比较：用于少样本学习的关系网络。载于*IEEE*计算机视觉与模式识别会议论文集，1199–1208。Xu, T.; Liang, Y.; and Lan, G. 2021. Crpo：一种具有收敛保证的安全强化学习新方法。载于国际机器学习大会，11480–11491。PMLR。

补充材料

A 实现补充

A.1 梯度矩阵的计算

在算法1的第5行和第7行中，我们需要计算梯度矩阵 $\nabla\Phi(x^k)$ 和梯度集合 $G(x^k,\epsilon_k)$。注意，在式 (4) 中计算 $\nabla\Phi(x^k)$ 时的 $-M_+^{-1}(x)N_+(x)$，以及在式 (5) 中计算 $G(x^0,\epsilon)$ 时的 $-M_\epsilon^S(x^0,y^*(x^0))^{-1}N_\epsilon^S(x^0,y^*(x^0))$，都具有如下形式

$$-\begin{bmatrix}\nabla_y^2\mathcal{L} & \nabla_y r^\top\\ \nabla_y r & 0\end{bmatrix}^{-1}\begin{bmatrix}\nabla_{xy}^2\mathcal{L}\\ \nabla_x r\end{bmatrix}, \tag{9}$$

其中，对于任意 x， $\nabla_y^2\mathcal{L}^{-1}$ 在 $(y^*(x)\ \lambda(x)\ \nu(x)\ x)$ 处正定（见附录中定理6的证明(ii)）。我们可以按如下方式计算 (9)。首先，由于 $\nabla_y^2\mathcal{L}^{-1}$ 正定，我们可以使用共轭梯度（CG）法（Hestenes和Stiefel 1952）计算 $A=\nabla_y^2\mathcal{L}^{-1}\nabla_{xy}^2\mathcal{L}$ 和 $B=\nabla_y^2\mathcal{L}^{-1}\nabla_y r^\top$。其次，(9) 可以写为

$$\begin{bmatrix}-A+B(\nabla_y rB)^{-1}(\nabla_y rA-\nabla_x r)\\ -(\nabla_y rB)^{-1}(\nabla_y rA-\nabla_x r)\end{bmatrix}. \tag{10}$$

设在 $P(x)$ 中，严格活跃约束的数量为 m^+，非严格活跃约束的数量为 m^-，且 n 为等式约束的数量。那么， $r(x,\ y)$ 是一个维数至多为 $n+m^++m^-$ 的向量函数。$\nabla_y^2\mathcal{L}$ 的维数为 d_y。在机器学习应用中， d_x 通常较大，而 $n+m^++m^-$ 相对较小。Ji 、 Yang 和 Liang (2021) 证明了第一步中 A 和 B 的计算是可行的，因此 (10) 的计算也可行。

论文（Amos 和 Kolter 2017）提供了一种高效求解器，用于计算二次规划解的梯度。该求解器利用原始-对偶内点法中的快速图形处理器批量求解。下层优化问题为二次规划时，该工具还可用于计算算法1中的 $-M_+^{-1}(x)N_+(x)$ 和 $-M_\epsilon^S(x^0,\ y^*(x^0))^{-1}N_\epsilon^S(x^0,\ y^*(x^0))$。

A.2 假设的满足情况

这里重申一下假设3的两点。

注2。 对于 $P(x)$， y 处满足线性无关约束资格条件 一个充分条件 假设3的一个更容易检验的充分条件是，解 $y^*(x)$ 在 $P(x)$ 处存在，且 *t* *he* 所有 $x\in\mathbb{R}^{d_x}$ 和 $y\in\mathbb{R}^{d_y}$。

注释 3. 等式约束不能替换为两个不等式约束，即将 $p=0$ 替换为 $p\le 0$ 和 $-p\le 0$。否则，线性无关约束资格条件将不成立。

请注意，本文实验中的所有优化问题均满足假设1、2、3。具体内容见附录 B。

B 实验补充材料

所有实验均在一台配备 4.10 GHz 英特尔酷睿 i5 CPU 和 RTX 3080 图形处理器的计算机上执行。

B.1 超参数优化

在机器学习问题中，给定超参数 Λ，学习器需要最小化训练误差，其最优参数记为 $w^*(\Lambda)$。超参数优化旨在为学习问题搜索最佳超参数 Λ^*，这可以表述为一个双层优化问题。具体而言，上层优化旨在最小化学习器参数 $w^*(\Lambda)$ 的验证误差，其中 $w^*(\Lambda)$ 是在超参数 Λ 下，下层优化中训练误差的最小化解。

超参数优化已得到广泛研究（Pedregosa 2016；Franceschi 等人 2017、2018；Lorraine、Vicol 和 Duvenaud 2020；Ji、Yang 和 Liang 2021）。然而，这些方法无法处理约束学习问题的超参数优化，例如支持向量机（SVM）分类（Cortes 和 Vapnik 1995）、安全强化学习（Achiam 等人 2017；Chen、Dong 和 Wang 2021；Xu、Liang 和 Lan 2021）。我们将所提出的算法应用于约束学习问题的超参数优化，该问题表述为

$$\begin{aligned}&\min_{\Lambda}\ \Phi(\Lambda)=\mathcal{L}_{\mathcal{D}_{\text{val}}}(w^*)=\frac{1}{|\mathcal{D}_{\text{val}}|}\sum_{z\in\mathcal{D}_{\text{val}}}\mathcal{L}(w^*;z)\\ &\text{s.t. } w^*=\arg\min_{w}\{\mathcal{F}_{\mathcal{D}_{\text{tr}}}(\Lambda,w):p(\Lambda,w)\le 0;q(\Lambda,w)=0\},\end{aligned}$$

其中， $\mathcal{D}_{\text{val}}$ 和 $\mathcal{D}_{\text{tr}}$ 分别是验证数据和训练数据， $\mathcal{L}_{\mathcal{D}_{\text{val}}}$ 是模型参数 ω 在数据 $\mathcal{D}_{\text{val}}$ 上的损失函数。下层优化是对模型参数 w的训练，其中 $\mathcal{F}_{\mathcal{D}_{\text{val}}}$ 是关于 $\mathcal{D}_{\text{tr}}$ 的训练目标， p和 q 是约束。

SVM 超参数优化 SVM 的优化问题为：

$$(w^*, b^*, \xi^*) = \arg\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + \frac{1}{2}\sum_{i=1}^{N} e^{c_i}\xi_i^2$$
$$\text{s.t. } l_i\left(w^\top \phi(z_i) + b\right) \geq 1 - \xi_i,\ i = 1, 2, \ldots N, \tag{11}$$

其中，z_i是数据点，y_i是标签，并且对于所有 $1 \leq i \leq N$，$(z_i;\ l_i) \in \mathcal{D}_{\text{tr}}$。对于所有 $1 \leq i \leq N$，$(z_i;\ y_i) \in \mathcal{D}_{\text{tr}}$。向量函数 $\phi(z_i)$ 是点 z_i 的高维特征。核函数定义为 $K(z_i,\ z_j) = \phi(z_i)^T \phi(z_j)$。SVM 超参数优化问题表述为

$$\min_c \Phi(c) = \mathcal{L}_{\mathcal{D}_{\text{val}}}(w^*, b^*), \tag{12}$$

其中，w^*，b^* 在式（11）中给出，优化后的超参数为 $c = [c_1, \ldots, c_N]$。为满足假设2，我们将式（11）的目标函数设为 $\frac{1}{2}\|w\|^2 + \frac{1}{2}\sum_{i=1}^{N} e^{c_i}\xi_i^2 + \frac{1}{2}\mu b^2$，其中 μ是一个较小的正数。于是，目标函数关于 $(w,\ b,\ \xi)$ 是强凸的。容易验证，假设3中的线性无关约束资格条件成立。

上层目标函数定义为 $\mathcal{L}_{\mathcal{D}_{\text{val}}}(w^*, b^*) = \frac{1}{|\mathcal{D}_{\text{val}}|}\sum_{(z,l)\in\mathcal{D}_{\text{val}}} \mathcal{L}(w^*, b^*; z, l)$，其中 $\mathcal{L}(w^*, b^*; \mathcal{D}_{\text{val}})$ 定义为 $\mathcal{L}(w^*, b^*;\ z, l) = \sigma(-l(z^\top w^* + b)\|w^*\|)$，且 $\sigma(x) = \frac{1-e^{-x}}{1+e^{-x}}$。这里，$l(z^\top w^* + b)\|w^*\|$ 是点 z 与决策超平面 $z^\top w^* + b = 0$之间的有符号距离，其中预测正确时 $\frac{l(z^\top w^* + b)}{\|w^*\|} > 0$，预测错误时 $\frac{l(z^\top w^* + b)}{\|w^*\|} < 0$。因此，$\mathcal{L}_{\mathcal{D}_{\text{val}}}(w^*, b^*)$ 是验证准确率的可微替代函数。

当特征函数 ϕ 不可计算时，无法直接优化式 (11) 的超参数 c。例如，在核化支持向量机中（Hofmann、Schölkopf 和 Smola，2008年），对于大多数核函数，如高斯核和多项式核，ϕ 未知或非常复杂。因此，根据定理 2，很难计算 $\nabla w^*(c)$。因此，在核化支持向量机中，我们求解式 (11) 的对偶问题：

$$\min_\alpha \frac{1}{2}\alpha^\top (Q + C^{-1})\alpha - \sum_{i=1}^{n}\alpha_i$$
$$\text{s.t. } \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{13}$$
$$\alpha_i \geq 0,\ i = 1, 2, \ldots N,$$

其中，Q 是一个 $n \times n$ 的半正定矩阵，满足 $Q_{ij} \equiv y_i y_j K(z_i,\ z_j)$，且 $C = \text{diag}(e^{c_1},\ \ldots,\ e^{c_n})$。由于问题 (11) 满足强对偶性，因此有

$$w^* = \sum_{i=1}^{n}\alpha_i^* y_i \phi(z_i)$$

and

$$b^* = y_i(1 - e^{-c_i}\alpha_i) - \sum_{j=1}^{n}\alpha_j^* y_j \phi(z_j)^\top \phi(z_i) = y_i(1 - e^{-c_i}\alpha_i) - \sum_{j=1}^{n}\alpha_j^* y_j K(z_j, z_i)$$

对于任意满足 $\alpha_i^* > 0$ 的支持向量 z_i。根据核方法，在模型训练和预测这两个过程中，都无需计算 ϕ。z^{new} 的预测为

$$\text{sign}\{w^{*\top}\phi(z^{new}) + b^*\} = \text{sign}\{\sum_{i=1}^{n}\alpha_i^* y_i K(z_i, z_{new}) + b^*\}.$$

假设2成立，因为目标函数 $\frac{1}{2}\alpha^\top(Q + C^{-1})\alpha - \sum_{i=1}^{n}\alpha_i$，其中 Q 是一个半正定矩阵，且 $C^{-1} = \text{diag}(e^{-c_1},\ \ldots,\ e^{-c_n})$ 是正定的，因此目标函数是强凸的。由于存在 i 使得 $\alpha_i^* > 0$，因此很容易证明假设3中的线性无关约束资格条件成立。

在实验中，我们考虑线性支持向量机模型和核化支持向量机模型的超参数优化。**线性支持向量机：** 特征函数为 $\phi(x) = x$。下层问题 (11) 和 (13) 均适用于线性支持向量机的超参数优化。这里，我们求解双层问题 (11)。**核化支持向量机：** 我们采用多项式核，即 $K(\phi\ z\ ^T\phi\ z' = \gamma z^\top z' + r\ ^d$ $\gamma = 1$ $d = 3$ 2017 $, z')$ ( ) ( ) ( )，其中 和 。我们在 (Dua 和 Graff) 中的糖尿病数据集上测试算法。对于算法1，我们设置 $\gamma = 0.3, \epsilon_0 = 0.3, \beta = 0.5$，并将总迭代次数固定为 60。

数据超清洗 我们将数据超清洗建模为支持向量机的超参数优化问题，其中上层优化问题如式（12）所示，下层优化问题如式（11）所示。优化超参数 c 后，与污染数据 $(z_i,\ y_i)$ 对应的惩罚项 e^{c_i} 将接近 0。因此，污染数据 $(z_i,\ y_i)$ 得以检测，且几乎不会影响分类器模型的训练和预测。我们在 Dua 和 Graff 2017 提供的乳腺癌数据集上进行了实验。对于算法1，我们设置 $\gamma = 0.3$、$\epsilon_0 = 0.3$、$\beta = 0.5$，并将总迭代次数固定为 30。

B.2 元学习

少样本学习中的元学习旨在从任务分布中学习一个共享的先验参数，使得基于该先验，利用少样本数据进行简单的学习步骤，就能很好地适应该分布中的任务。具体而言，训练任务 $\mathcal{T}_i$ 从分布 $P_{\mathcal{T}}$ 中采样。每个任务 $\mathcal{T}_i$ 由其训练数据 $\mathcal{D}_i^{tr}$ 和测试数据 $\mathcal{D}_i^{test}$ 表征。元学习的目标是找到一个良好的参数 ϕ 和一个基学习器 $\mathcal{A}$，使得任务特定参数 $w^i = \mathcal{A}(\phi, \mathcal{D}_i^{tr})$ 的测试损失 $\mathcal{L}(w^i, \phi, \mathcal{D}_i^{test})$ 较小。

元学习的训练可以表述为一个带约束双层优化问题（Lee 等，2019年）。上层优化旨在从输入数据中提取特征。多类支持向量机作为下层优化中的基学习器 $\mathcal{A}$，用于根据提取的特征对数据进行分类。具体而言，特征提取模型 f_ϕ 将图像 x_n映射为其特征，记作 $f_\phi(x_n)$。下层优化中的多类支持向量机是一个约束优化问题:

$$
\begin{aligned}
w^i = \mathcal{A}\left(\mathcal{D}_i^{\mathrm{tr}}; \phi\right) = & \underset{\{w_k\},\{\xi_n\}}{\arg\min} \frac{1}{2} \sum_k \|w_k\|_2^2 + C \sum_n \xi_n \\
\text{s.t. } & w_{y_n} \cdot f_\phi\left(x_n\right) - w_k \cdot f_\phi\left(x_n\right) \geq 1 - \delta_{y_n,k} - \xi_n, \forall n, k
\end{aligned}
\tag{14}
$$

其中，$\mathcal{D}_i^{tr} = \{(x_n, y_n)\}$，其中图像为 x_n，其标签为 y_n，C 为正则化参数，$\delta_{\cdot,\cdot}$ 为克罗内克δ函数。这里，k 是特征 $f_\phi(x_n)$的索引，n 是数据的索引。上层优化为

$$
\min_\phi \sum_{\mathcal{T}_i \in P_{\mathcal{T}}} \mathcal{L}\left(w^i, \phi, \mathcal{D}_i^{test}\right). \tag{15}
$$

其中

$$
\mathcal{L}\left(w^i, \phi, \mathcal{D}_i^{test}\right) = \sum_{(x,y) \in \mathcal{D}_i^{test}} \left[-\gamma w_y^i \cdot f_\phi(x) + \log \sum_k \exp\left(\gamma w_k^i \cdot f_\phi(x)\right) \right]
$$

且 w^i 由式（14）给出。这里，$\mathcal{L}\left(w^i, \phi, \mathcal{D}_i^{test}\right)$ 是特征提取参数 ϕ 和支持向量机参数 w^i 在下层优化（14）中优化后对应的负对数似然损失。随后，通过上层优化（15）求得最佳特征提取参数 ϕ。

按照 MetaOptNet（Lee 等，2019年）中的实验设置，我们使用 ResNet-12 网络作为特征提取映射 $f_\phi()$。由于算法1中的线搜索不适用于元学习问题，我们计算式1中的下降方向，并使用 Nesterov动量为 0.9、权重衰减为 0.0005 的随机梯度下降法来求解该问题。每个小批量包含 8 个任务片段。模型经过 30 个训练轮次的元训练，每个训练轮次包含 1000 个任务片段。学习率初始设为 0.1，并分别在第 10、20 和 25 个训练轮次调整为 0.006、0.0012。我们对自己提出的方法和 MetaOptNet（Lee 等，2019年）中的方法均采用相同的配置。

表1展示了在预测准确率方面与先前工作在 CIFAR-FS 和 FC100 上的比较结果。结果表明，我们的优化算法最终测试准确率略高于 MetaOptNet-SVM（Lee 等，2019年）。

表1：**与 CIFAR-FS 和 FC100 上既有工作的比较**。CIFAR-FS 和 FC100 上小样本分类准确率（%）的平均值及其 95% 置信区间。*CIFAR-FS 的结果来自 (Bertinetto 等，2018)。FC100 的结果来自 †(Oreshkin、Rodríguez López 和 Lacoste 2018) 和 ‡(Ji、Yang 和 Liang 2021)。所有模型均使用 (Lee 等，2019年) 中 CIFAR-FS 和 FC100 的原始训练数据进行训练，且未使用验证数据。

	CIFAR-FS 5-way		**FC100 5-way**	
模型	**1-shot**	**5-shot**	**1-shot**	**5 次**
MAML*‡（Finn、Abbeel 和 Levine，2017）	58.9 ± 1.9	71.5 ± 1.0	-	47.2
原型网络*†（Snell、Swersky 和 Zemel，2017）	55.5 ± 0.7	72.0 ± 0.6	35.3 ± 0.6	48.6 ± 0.6
关系网络* (Sung 等，2018)	55.0 ± 1.0	69.3 ± 0.8	-	-
R2D2（Bertinetto 等，2018）	65.3 ± 0.2	79.4 ± 0.1	-	-
TADAM（Oreshkin、Rodríguez López 和 Lacoste 2018）	-	-	40.1 ± 0.4	56.1 ± 0.4
ProtoNets（Snell、Swersky 和 Zemel，2017）	**72.2 ± 0.7**	83.5 ± 0.5	37.5 ± 0.6	52.5 ± 0.6
MetaOptNet-RR（Lee 等，2019年）	**72.6 ± 0.7**	**84.3 ± 0.5**	40.5 ± 0.6	55.3 ± 0.6
MetaOptNet-SVM（Lee 等，2019年）	**72.0 ± 0.7**	**84.2 ± 0.5**	41.1 ± 0.6	55.5 ± 0.6
MetaOptNet-SVM-GAE（我们的方法）	**72.2 ± 0.7**	**84.6 ± 0.6**	**41.9 ± 0.6**	**56.4 ± 0.8**

C 证明与分析

首先，我们说明本节使用的记号。本节使用的所有记号均与论文正文中的记号相同，但第3节中的 $J(x)$、 $J^+(x)$ 和 $J^0(x)$ 除外。在第3节中，我们简化了式 (1) 中的记号，并将 $J(x,\ y^*(x))$ 记为 $J(x)$，将 $J^+(x,\ y^*(x)\ \lambda(x))$ 记为 $J^+(x)$，并将 $J^0(x,\ y^*(x)\ \lambda(x))$ 记为 $J^0(x)$，因为最优解 $y^*(x)$ 和拉格朗日乘子 λ 与 ν 均由 x 唯一确定。

本节保留定义 1 中的所有定义，并将 $J^+(x,\ y,\ \lambda)$ 简化为 $J^+(x,\ y)$，将 $J^0(x,\ y,\ \lambda)$ 简化为 $J^0(x,\ y)$，其中 KKT条件在 y 处对 $P(x)$ 成立。我们可以这样做，因为当 KKT条件在 y 处对 $P(x)$ 成立且线性无关约束资格条件成立（假设3）时，拉格朗日乘子 λ 和 μ 唯一，并由 x 和 y 确定（Kyparisis，1985）。

C.1 定理1和定理2的证明

我们首先列出定义5、定理4和定理5，这些结果见（Fiacco，1983；Giorgi 和 Zuccotti，2018；Kojima（1980））， 然后介绍引理1、2和3。最后，我们证明定理6，它是定理1和定理2组合后的完整版本。

ˆ定义5。 假设 *KKT*条件在 $\hat{}y$ 处对 $P(x)$ 成立，且拉格朗日乘子为 λ 和 $\hat{}\nu$。若严格二阶充分条件（*SSOSC*）在 $\hat{}y$ 处成立，则

$$z^{\top}\nabla_y^2\mathcal{L}\left(\hat{y},\hat{\lambda},\hat{\nu},x\right)z>0$$

对所有 $z\neq 0, z\in Z(\hat{}y,x)$均成立，其中 $\mathcal{L}$ 是与 $P(x)$相关联的拉格朗日函数，而 $Z(\hat{}y,x)$ 定义为

$$Z\left(\hat{y},x\right)\triangleq\{z\in\mathbb{R}^{d_y}:\nabla_y p_j\left(x,\hat{y}\right)z=0,\ \hat{\lambda}_j>0;\nabla_y q_i\left(x,\hat{y}\right)z=0,\ 1\leq i\leq n\}.$$

定理 4 （（Fiacco，1983；Giorgi 和 Zuccotti（2018））） 。 考虑问题 $P\left(x^0\right)$。假设假设*1*成立。假设 $y^0\in K\left(x^0\right)$ 和 *KKT*条件在 y^0 处成立，拉格朗日乘子为 λ^0,ν^0。此外，假设线性无关约束资格条件在 y^0处成立，关于 λ^0，严格互补松弛在 y^0 处成立，且严格二阶充分条件在 y^0处成立，其中 $\left(\lambda^0,\nu^0\right)$。那么，

(i) y^0 是 $P\left(x^0\right)$的局部唯一局部极小值，即存在 $\delta>0$，使得对于所有 $y\in\mathcal{B}(y^0,\delta)$， y^0 是 $P\left(x^0\right)$的唯一局部极小值。相应的拉格朗日乘子 λ^0 和 ν^0 是唯一的。*(ii)* 存在 $\epsilon>0$，使得存在一个唯一的连续可微向量函数

$$z(x)\triangleq[y(x)^{\top},\lambda(x)^{\top},\nu(x)^{\top}]^{\top},$$

该函数定义在 $\mathcal{B}(x^0,\epsilon)$上，并且 $y(x)$ 是 $P(x)$的局部唯一局部极小值。在 $y(x)$ 处，*KKT*条件成立，且相应的拉格朗日乘子 $\lambda(x)$ 和 $\nu(x)$ 唯一。

(iii) 对于所有 $x\in\mathcal{B}(x^0,\epsilon)$，在 $y(x)$ 处， $P(x)$ 满足线性无关约束资格条件和*SCSC*。

(iv) $z(x)$的梯度为

$$\begin{bmatrix}\nabla_x y(x^0)\\ \nabla_x\lambda(x^0)\\ \nabla_x\nu(x^0)\end{bmatrix}=-M(x^0,y^0,\lambda^0,\nu^0)^{-1}N(x^0,y^0,\lambda^0,\nu^0),$$

其中

$$M\triangleq\begin{bmatrix}\nabla_y^2\mathcal{L} & (\nabla_y p_1)^{\top} & \cdots & (\nabla_y p_m)^{\top} & (\nabla_y q)^{\top}\\ \lambda_1\nabla_y p_1 & p_1 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & 0 & \vdots\\ \lambda_m\nabla_y p_m & 0 & \cdots & p_m & 0\\ \nabla_y q & 0 & 0 & 0 & 0\end{bmatrix}$$

其中 $M(x^0,y^0)$ 是非奇异的，并且

$$N\triangleq\left[\nabla_{xy}^2\mathcal{L}^{\top},\lambda_1\left(\nabla_x p_1\right)^{\top},\cdots,\lambda_m\left(\nabla_x p_m\right)^{\top},\nabla_x q^{\top}\right]^{\top}.$$

注 4. 如果下层优化问题 $P(x^0)$ 是无约束的，那么定理 *4* 中的要求可简化为$\nabla_y^2 g\left(y^0,x^0\right)$ 正定。

定理5 （（Giorgi 和 Zuccotti（2018）； Kojima（1980））） 。 假设在 $P\left(x^0\right)$ 处，除定理 *4* 中的 *SCSC* 外，所有要求均在(y^0,λ^0,ν^0) 处成立。那么，

(i) y^0 是 $P\left(x^0\right)$的局部唯一局部极小值。相应的拉格朗日乘子 λ^0 和 ν^0 是唯一的。

(ii) 存在 $\epsilon > 0$，使得存在唯一的 *Lipschitz* 连续且一次方向可微的向量函数

$$z(x) \triangleq [y(x)^\top, \lambda(x)^\top, \nu(x)^\top]^\top,$$

其定义于 $\mathcal{B}(x^0, \epsilon)$ 上，且 $y(x)$ 是 $P(x)$ 的局部唯一局部极小值，其对应的拉格朗日乘子 $\lambda(x)$ 和 $\nu(x)$唯一。

(iii) 对于所有 $x \in \mathcal{B}(x^0, \epsilon)$，$P(x)$ 处的 $y(x)$ 均满足线性无关约束资格条件。

引理1. 假设除定理 *4* 中的 *SCSC* 外，所有条件均在 (y^0, λ^0, ν^0) 处对 $P\left(x^0\right)$ 成立（定理*5*中的所有条件均满足）。令 $y(x)$ 为定理 *5* 中所示的 $P(x)$ 的局部唯一局部极小值。则，

(i) 存在 $\beta > 0$，使得对于所有 $x \in \mathcal{B}(x^0, \beta)$，$p_j(x, y(x)) < 0$ 对所有 $j \notin J(x^0, y^0)$均成立。(ii) 存在 和，使得对于所有 (x^0, ξ) 和 $y' \in \mathcal{B}(y(x')\ \delta)$，$p_j(x', y') < 0$ *fo* 对所有 $j \notin J(x^0, y^0)$。

证明。(i) 对任意 $j \notin J(x^0, y^0)$，有 $p_j\left(x^0, y^0\right) < 0$。于是，存在 $\epsilon_1 > 0$，使得 $p_j\left(x^0, y^0\right) \leq -\epsilon_1$。由于函数 p_j 在 (x^0, y^0) 处连续，且 y 在 x^0 处连续，因此 $p_j(x, y(x))$ 在 x^0 处连续。于是，存在 $\beta_1 > 0$，使得对于所有 $x \in \mathcal{B}(x^0, \beta_1)$，都有 $|p_j(x, y(x)) - p_j\left(x^0, y^0\right)| \leq \frac{1}{2}\epsilon_1$，进而 $p_j(x, y(x)) \leq -\frac{1}{2}\epsilon_1$。取所有 $j \notin J(x^0, y^0)$ 对应的最小 β_1 作为 β，即可证明 (i)。

(ii) 由于 p_j 在 (x^0, y^0) 处连续，因此存在 $\delta_1 > 0$，使得对于所有满足 $\in \mathcal{B}((x^0, y^0)\ \delta_1)$ 的 (x', y')，都有 $|p_j(x', y') - p_j\left(x^0, y^0\right)| \leq \frac{1}{2}\epsilon_1$。由于 $y(x)$ 在 x^0 处连续，因此存在 $\xi_1 > 0$，对于所有满足 $\|y(x) - y(x^0)\| < \delta_1/4$ 的 $x \in \mathcal{B}(x^0, \xi_1)$，都有该关系。令 $\xi_2 = \min\{\delta_1/4, \xi_1\}$。那么，对于所有满足 $y' \in \mathcal{B}(y(x')\ \delta_1/2)$ 的 $x' \in \mathcal{B}(x^0, \xi_2)$，都有 $\|(x', y') - (x_0, y_0)\| < \delta_1$。于是 $|p_j(x', y') - p_j\left(x^0, y^0\right)| \leq \frac{1}{2}\epsilon_1$，且 $p_j(x', y') \leq -\frac{1}{2}\epsilon_1$。

通过在所有 $j \notin J(x^0,\ y^0)$ 中选取最小的 ξ_2 作为 ξ，并在所有 $j \notin J(x^0,\ y^0)$ 中选取最小的 $\delta_1/2$ 作为 δ，我们有：对于所有 $x' \in \mathcal{B}(x^0,\ \xi)$，都可以找到 $\delta > 0$，使得当 $||y' - y(x)|| \leq \delta$ 时，对于所有 $j \notin J(x^0,\ y^0)$，$p_j(x,\ y') \leq -\frac{1}{2}\epsilon_1$。 □

Rem引理 5。引理*1*表明，在 (x^0, y^0) 处的非活跃约束在 (x^0, y^0)附近仍然是非活跃的。

引理2。假设　假设定理 *4* 中的所有条件在 (y^0, λ^0, ν^0) 处对 $P\left(x^0\right)$ 均成立。定义问题 $\hat{}P(x)$ *a* *s:*

$$\begin{aligned} \arg\min_y\ & g(x, y): \\ \textit{s.t.}\ & p_j(x, y) \leq 0, j \in J(x^0, y^0), \\ & q(x, y) = 0. \end{aligned}$$

则以下性质成立：

（*i*） 定理 *4* 中的所有要求和所有结论在 (y^0, λ^0, ν^0) 处对 $P\left(x^0\right)$ 成立，并且在 $(y^0, \lambda^0_{J(x^0, y^0)}\ \nu^0)$ 处对$\hat{}P\left(x^0\right)$ 也成立。

令 $z(x) \triangleq [y(x)^\top、\ \lambda(x)^\top、\ \nu(x)^\top]^\top$ 为 x^0 的某个邻域内唯一的连续可微向量函数，使得 $y(x)$ 是 $P(x)$ 的局部唯一局部极小值，并具有唯一的相关拉格朗日乘子 $\lambda(x)$ 和 $\nu(x)$。令 $\hat{}z(x) \triangleq [\hat{}y(x)^\top、\ \hat{}\lambda(x)^\top、\ \hat{}\nu(x)^\top]^\top$ 为 x^0 的某个邻域内唯一的连续可微向量函数，使得 $\hat{}\hat{}\hat{}y(x)$ 是 $P(x)$ 的局部唯一局部极小值，并具有唯一的相关拉格朗日乘子 $\lambda(x)$ 和 $\hat{}\nu(x)$。

(ii) 我们有

$$\begin{aligned} &\nabla_x y(x^0) = \nabla_x \hat{y}(x^0), \\ &\nabla_x \nu(x^0) = \nabla_x \hat{\nu}(x^0), \\ &\nabla_x \lambda_j(x^0) = \nabla_x \hat{\lambda}_j(x^0)\ \textit{when}\ j \in J(x^0, y^0), \\ &\nabla_x \lambda_j(x^0) = 0\ \textit{when}\ j \notin J(x^0, y^0). \end{aligned}$$

证明。(i) 问题 $\hat{}\ P\left(x^0\right)$ 与问题 $P\left(x^0\right)$具有相同的目标函数和相同的等式约束。$P\left(x^0\right)$ 的 不等式约束是从 $P\left(x^0\right)$ 中移除非活跃约束后得到的。在 $P\left(x^0\right)$的 (y^0, λ^0, ν^0) 处，LICQ、SCSC、SSOSC 和 KKT条件均成立。因此，容易证明，在 $\hat{}P\left(x^0\right)$的 $(y^0, \lambda^0_{J(x^0, y^0)}\nu^0)$ 处，LICQ 和 SCSC 成立。由 $P\left(x^0\right)$在 (y^0, λ^0, ν^0) 处满足的 KKT条件可得，对于 $j \notin J(x^0, y^0)$，有 $\lambda_j = 0$，即 $p_j\left(x^0, y^0\right) < 0$。于是，在 $\hat{}\ P\left(x^0\right)$的 $(y^0, \lambda^0_{J(x^0, y^0)}\nu^0)$ 处，SSOSC 和 KKT条件成立。由定理 4，(i) 成立。因此，y^0 是 $\hat{}\ P\left(x^0\right)$的局部唯一局部极小值，并且存在 $\hat{}z(x) = [\hat{}y(x)^\top、\ \hat{}\lambda(x)^\top、\ \hat{}\nu(x)^\top]^\top$，它们是在 x^0的某个邻域内唯一的连续可微向量函数，使得 $\hat{}y(x)$ 是 $\hat{}P(x)$ 的局部唯一局部极小值，且具有唯一的拉格朗日乘子 $\hat{}\lambda(x)$ 和 $\hat{}\nu(x)$。

(ii) 定理 4 的所有结论在 $P(x^0)$的 $(y^0,\ \lambda^0,\ \nu^0)$ 处均成立，因此存在 $z(x) = [y(x)^\top,\ \lambda(x)^\top,\ \nu(x)^\top]^\top$，它们是在 x^0的某个邻域内唯一的连续可微向量函数，使得 $y(x)$ 是 $P(x)$ 的局部唯一局部极小值，且具有唯一的拉格朗日乘子 $\lambda(x)$ 和 $\nu(x)$。接下来，我们将证明，在 x^0的某个邻域内，

$$
\begin{aligned}
&y(x) = \hat{y}(x),\\
&\nu(x) = \hat{\nu}(x),\\
&\lambda_j(x) = \hat{\lambda}_j(x) \text{ when } j \in J(x^0, y^0),\\
&\lambda_j(x) = 0 \text{ when } j \notin J(x^0, y^0).
\end{aligned}
$$

由于 $y(x)$ 是 $P(x)$ 的局部唯一局部极小值，因此存在 $\beta_3 > 0$，使得对于任意 $||x - x^0|| \le \beta_3$，$y(x)$ 都是 $P(x)$ 的局部极小值，即存在 $\delta_3 > 0$，使得当 $||y' - y(x)|| \le \delta_3$、$p(x, y') \le 0$ 且 $q(x, y') = 0$ 时，$g(x, y(x)) \le g(x, y')$。令 β_2 为引理1中所示的 ξ，δ_3 为引理1中所示的 δ。令 $\beta_4 =$ 为 $\{\beta_2, \beta_3\}$ 的最小值，$\delta_4 =$ 为 $\{\delta_2, \delta_3\}$ 的最小值。那么，对于所有 $||x - x^0|| \le \beta_4$，存在 δ_4，使得以下两个陈述成立:

(a) 当 $||y' - y(x)|| \le \delta_4$、$p(x, y') \le 0$ 且 $q(x, y') = 0$ 时，$g(x, y(x)) \le g(x, y')$。

(b) 当 $||y' - y(x)|| \le \delta_4$ 时，对于所有 $j \notin J(x^0, y^0)$，$p_j(x, y') \le -\frac{1}{2}\epsilon_1 < 0$。

命题 (b) 表明，对于所有 $j \notin J(x^0, y^0)\}$，集合 $\{y' :\ ||y' - y(x)|| \le \delta_4\} \subset \{y' :\ p_j(x, y') \le -\frac{1}{2}\epsilon_1 < 0$。因此，当 $||y' - y(x)|| \le \delta_4$ 时，$g(x, y(x)) \le g(x, y')$，$q(x, y') = 0$，且 $p(x, y') \le 0, j \in J(x^0, y^0)$。这意味着，对于所有 $||x - x^0|| \le \beta_4$，$y(x)$ 是 $\hat{}P(x)$ 的一个局部极小值。

对于 $||x - x^0|| \le \beta_4$，$y(x)$ 是 $P(x)$ 的一个局部唯一局部极小值。假设 $y(x)$ 不是 $\hat{}P(x)$ 的局部唯一局部极小值，也就是说，对于任意 $\phi > 0$，存在 $||y'(\phi) - y(x)|| \le \phi$，使得 y' 是 $\hat{}P(x)$ 的一个局部极小值。我们可以令 $0 < \phi < \delta_4$，于是对于所有 $j \notin J(x^0, y^0)$，$p_j(x, y'(\phi)) < 0$。因此，对于任意 ϕ，$y'(\phi)$ 是 $P(x)$ 的一个局部极小值，这与 $y(x)$ 是 $P(x)$ 的局部唯一局部极小值相矛盾。因此，$y(x)$ 是 $\hat{}P(x)$ 的一个局部唯一局部极小值。令 $z_1(x) \triangleq [y(x)^\top,\ \lambda'(x)^\top,\ \nu(x)^\top]^\top$ 定义在 $\{x :\ ||x - x^0|| \le \beta_4\}$ 上，其中 $\lambda'(x)^\top$ 是一个向量函数，包含当 $j \in J(x^0, y^0)$ 时的所有 $\lambda_j(x)$。于是 $z_1(x)$ 是一个连续可微向量函数，并且容易验证，对于 $\hat{}P(x)$，在 $y(x)$ 处，KKT条件成立，其拉格朗日乘子为 $\lambda'(x)$ 和 $\nu(x)$。

由于 $\hat{}z(x) = [\hat{}y(x)^\top, \lambda(x)^\top, \nu(x)^\top]^\top$ 是在 x^0的某个邻域内唯一的连续可微向量函数，使得 $\hat{}y(x)$ 是 $\hat{}P(x)$ 的局部唯一局部极小值。那么对于 $||x - x^0|| \le \beta_4$，我们有 $\hat{}z(x) = z_1(x)$，并且

$$
\begin{aligned}
&y(x) = \hat{y}(x),\\
&\nu(x) = \hat{\nu}(x),\\
&\lambda_j(x) = \hat{\lambda}_j(x) \text{ when } j \in J(x^0, y^0).
\end{aligned}
$$

由于当 $j \notin J(x^0,\ y^0)$ 时，$p_j(x,\ y(x)) < 0$，因此根据 KKT条件，$\lambda_j(x) = 0$。然后，

$$
\begin{aligned}
&\nabla_x y(x^0) = \nabla_x \hat{y}(x^0),\\
&\nabla_x \nu(x^0) = \nabla_x \hat{\nu}(x^0),\\
&\nabla_x \lambda_j(x^0) = \nabla_x \hat{\lambda}_j(x^0) \text{ when } j \in J(x^0, y^0),\\
&\nabla_x \lambda_j(x^0) = 0 \text{ when } j \notin J(x^0, y^0).
\end{aligned}
$$

□

注意，对于 $P(x)$，$\hat{}y$ 关于 $\hat{}\lambda$ 满足 SCSC 等价于 $J^0(x,\ \hat{}y) = \emptyset$，于是 $J(x,\ \hat{}y) = J^+(x,\ \hat{}y)$。我们定义矩阵函数

$$
M_+(x^0, y^0) \triangleq \begin{bmatrix} \nabla_y^2 \mathcal{L} & \nabla_y p_{J^+(x^0,y^0)}^\top & \nabla_y q^\top \\ \nabla_y p_{J^+(x^0,y^0)} & 0 & 0 \\ \nabla_y q & 0 & 0 \end{bmatrix} (x^0, y^0, \lambda^0, \nu^0),
$$

and

$$
N_+(x^0, y^0) \triangleq [\nabla_{xy}^2 \mathcal{L}^\top, \nabla_x p_{J^+(x^0,y^0)}^\top, \nabla_x q^\top]^\top (x^0, y^0, \lambda^0, \nu^0).
$$

按照定理 4 所示，计算 $\hat{}z(x)$ 的梯度。由于对于所有 $j \in J^+(x^0, y^0)$，$\lambda_j > 0$，我们可以消去 M 和 N 中的所有 λ_j。然后我们可以得到

$$
\begin{bmatrix} \nabla_x \hat{y}(x^0) \\ \nabla_x \hat{\lambda}(x^0) \\ \nabla_x \hat{\nu}(x^0) \end{bmatrix} = -M_+^{-1}(x^0, y^0) N_+(x^0, y^0).
$$

根据引理2， $z(x)$ 的梯度计算如下：

$$\begin{aligned}\begin{bmatrix}\nabla_x y(x^0)\\ \nabla_x \lambda_{J(x^0,y^0)}(x^0)\\ \nabla_x \nu(x^0)\end{bmatrix} &= -M_+(x^0,y^0)^{-1}N_+(x^0,y^0),\\ \nabla_x \lambda_{J(x^0,y^0)^C}(x^0) &= 0.\end{aligned} \tag{16}$$

引理3。 假设 在 $P(x^0)$ 处，除定理 *4* 中的 *SCSC* 外，所有要求均在 (y^0,λ^0,ν^0) 处得到满足。那么 ，

(i) y^0 是 P的局部唯一局部极小值 (x^0).

(ii) 存在 $\epsilon>0$ ，使得存在一个唯一的 *Lipschitz* 连续向量函数

$$z(x) \triangleq [y(x)^\top, \lambda(x)^\top, \nu(x)^\top]^\top,$$

该函数定义在 $\mathcal{B}(x^0,\epsilon)$ 上，并且 $y(x)$ 是 $P(x)$ 的局部唯一局部极小值，且对应的拉格朗日乘子 $\lambda(x)$ 和 $\nu(x)$唯一。

对于方向 $d\in\mathbb{R}^{d_x}$，定义 $J_+^0(x^0,y^0,d)$ 为一个集合，其中包含所有满足以下条件的 $j\in J^0(x^0,y^0)$：存在 $\epsilon_0>0$，对于任意$0<\epsilon<\epsilon_0$， $p_j(x^0+\epsilon d, y(x^0+\epsilon d))=0$ 且 $\lambda_j(x^0+\epsilon d)>0$。记 $J_-^0(x^0,y^0,d)\triangleq J^0(x^0,y^0)\setminus J_+^0(x^0,y^0,d)$，

$$M_D(x^0,y^0,d)\triangleq\begin{bmatrix}\nabla_y^2\mathcal{L} & \nabla_y p_{J^+(x^0,y^0)}^\top & \nabla_y q^\top & \nabla_y p_{J_+^0(x^0,y^0,d)}^\top\\ \nabla_y p_{J^+(x^0,y^0)} & 0 & 0 & 0\\ \nabla_y q & 0 & 0 & 0\\ \nabla_y p_{J_+^0(x^0,y^0,d)} & 0 & 0 & 0\end{bmatrix}(x^0,y^0,\lambda^0,\nu^0) \tag{17}$$

and

$$N_D(x^0,y^0,d)\triangleq\left[\nabla_{xy}^2\mathcal{L}^\top, \nabla_x p_{J^+(x^0,y^0)}^\top, \nabla_x q^\top, \nabla_x p_{J_+^0(x^0,y^0,d)}^\top\right]^\top(x^0,y^0,\lambda^0,\nu^0). \tag{18}$$

(iii) $z(x)$ 在 x^0 处沿任意满足 $\|d\|=1$ 的方向 $d\in\mathbb{R}^{d_x}$ 的方向导数存在，且由下式给出

$$\begin{aligned}\begin{bmatrix}\nabla_d y(x^0)\\ \nabla_d \lambda_{J^+(x^0,y^0)}(x^0)\\ \nabla_d \nu(x^0)\\ \nabla_d \lambda_{J_+^0(x^0,y^0,d)}(x^0)\end{bmatrix} &= -M_D^{-1}(x^0,y^0,d)N_D(x^0,y^0,d)d,\\ \nabla_d \lambda_{J_-^0(x^0,y^0,d)}(x^0) &= 0,\\ \nabla_d \lambda_{J(x^0,y^0)^C}(x^0) &= 0,\end{aligned} \tag{19}$$

其中 $M_D(x^0,y^0,d)$ 是非奇异的。

证明。 当 y^0 关于 λ^0 满足 SCSC 时，我们有 $J(x^0,y^0)=J^+(x^0,y^0)$。该定理等价于引理2。当不满足 SCSC 时， $J^0(x^0,y^0)\neq\emptyset$。

(i) 根据定理5的第(i)部分，对于 P，在 (y^0,λ^0,ν^0) 处，线性无关约束资格条件、严格二阶充分条件和 KKT条件均成立 (x^0)，(i)成立。

(ii) 根据定理5的第(ii)部分，对于 $P(x^0)$，在 (y^0,λ^0,ν^0) 处，线性无关约束资格条件、严格二阶充分条件和 KKT条件均成立，(ii)成立，且 $z(x)$ 在 x^0 处沿任意方向的方向导数均存在。

(iii) 根据 (ii)，在 $\mathcal{B}(x^0,\epsilon)$ 上定义的向量函数 $z(x)\triangleq[y(x)^\top,\lambda(x)^\top,\nu(x)^\top]^\top$ 是 $P(x)$ 的局部解。根据定理 5 的 (iii)，对于所有 $x\in\mathcal{B}(x^0,\epsilon)$， $y(x)$ 处关于 $P(x)$ 的线性无关约束资格条件成立。然后，根据 Giorgi 和 Zuccotti（2018）中的定理 3，对于任意 $x\in\mathcal{B}(x^0,\epsilon)$， $z(x)\triangleq[y(x)^\top,\lambda(x)^\top,\nu(x)^\top]^\top$ 满足问题 $P(x)$ 的 KKT条件，并且 $\lambda(x)$ $\nu(x)$ 是唯一的拉格朗日乘子。 $z(x)$ 在 x^0 处沿任意方向的方向导数均存在。

(a) 考虑 $j\notin J(x^0,y^0)$。由引理1的(ii)部分可知，在 (x^0,y^0) 的一个小邻域内，对于所有 $j\notin J(x^0,y^0)$， $p_j(x,y)<0$。也就是说，(x^0,y^0) 处的非活跃约束在 (x^0,y^0) 的一个邻域内仍然是非活跃的。然后，对于 $j\notin J(x^0,y^0)$，存在 β_0，使得 $p_j(x^0+\beta d,y(x^0+\beta d))<0$，并且对于所有 $\beta<\beta_0$， $\lambda_j(x^0+\beta d)=0$。于是，$\nabla_d\lambda_{J(x^0,y^0)^C}(x^0)=0$。

(b) 考虑 $j\in J_-^0(x^0,y^0,d)$，即不存在 $\beta_0>0$，对于任意 $\beta<\beta_0$, $p_j(x^0+\beta d,y(x^0+\beta d))=0$ 和$\lambda_j(x^0+\beta d)>0$。方向 d 有两种可能的情况。第一种情况是，存在 $\beta_0>0$，对于任意 $\beta<\beta_0$,$p_j(x^0+\beta d,y(x^0+\beta d))\le 0$ 且 $\lambda_j(x^0+\beta d)=0$。第二种情况是，对于任意 $\beta_0>0$，总能找到 $\beta,\beta_1<\beta_0$，使得 $p_j(x^0+\beta d,y(x^0+\beta d))=0,\lambda_j(x^0+\beta d)>0$，并且 $p_j(x^0+\beta d,y(x^0+\beta d))\le 0,\lambda_j(x^0+\beta d)=0$。对于第一种情况，我们

有 $\nabla_d \lambda_{J^0_-(x^0,y^0,d)}(x^0) = 0$。对于第二种情况，由于 $y(x)$ 和 $\lambda(x)$ 在 x^0 沿该方向的方向导数存在，因此有 $\nabla_\beta p_j(x^0 + \beta d, y(x^0 + \beta d)) = 0$ 且 $\nabla_d \lambda_{J^0_-(x^0,y^0,d)}(x^0) = 0$。因此，对于 $j \in J^0_- (x^0, y^0, d)$, $\nabla_d \lambda_{J^0_-(x^0,y^0,d)}(x^0) = 0$。(c) 考虑 $j \in J^+(x^0, y^0)$，我们有 $\lambda^0_j > 0$。当 β 足够小时，有 $p_j(x^0 + \beta d, y(x^0 + \beta d)) = 0$ 且 $\lambda_j(x^0 + \beta d) > 0$。

(d) 考虑 $j \in J^0_+ (x^0, y^0, d)$, $p_j(x^0 + \beta d, y(x^0 + \beta d)) = 0$ 和 $\lambda_j(x^0 + \beta d) > 0$。
(e) 考虑 $z(x)$处的 KKT条件，我们有

$$\begin{cases} \nabla_y \mathcal{L}(y(x), \lambda(x), \nu(x), x) = 0 \\ \lambda_j(x) p_j(y(x), x) = 0,\ j = 1, \ldots, m \\ q_j(y(x), x) = 0,\ j = 1, \ldots, n \end{cases}$$

对于任意 $x \in \mathcal{B}(x^0, \epsilon)$。于是，对于任意充分小的 $\beta < \epsilon$，我们有

$$\begin{cases} \nabla_y \mathcal{L}(y(x^0), \lambda(x^0), \nu(x^0), x^0) = 0 \\ \lambda_j(x^0) p_j(y(x^0), x^0) = 0,\ j = 1, \ldots, m \\ q_j(y(x^0), x^0) = 0,\ j = 1, \ldots, n \end{cases} \text{ and } \begin{cases} \nabla_y \mathcal{L}(y(x^0 + \beta d), \lambda(x^0 + \beta d), \nu(x^0 + \beta d), x^0 + \beta d) = 0 \\ \lambda_j(x^0 + \beta d) p_j(y(x^0 + \beta d), x^0 + \beta d) = 0,\ j = 1, \ldots, m \\ q_j(y(x^0 + \beta d), x^0 + \beta d) = 0,\ j = 1, \ldots, n. \end{cases}$$

于是，我们有

$$\left. \frac{\partial \nabla_y \mathcal{L}\left(y\left(x^0 + \beta d\right), \lambda\left(x^0 + \beta d\right), \nu\left(x^0 + \beta d\right), x^0 + \beta d\right)}{\partial \beta} \right|_{\beta = 0} = 0,$$

$$\left. \frac{\partial \lambda_i(x^0 + \beta d) p_j(y(x^0 + \beta d), x^0 + \beta d)}{\partial \beta} \right|_{\beta = 0} = 0,\ j = 1, \ldots, m,$$

$$\left. \frac{\partial q_j(y(x^0 + \beta d), x^0 + \beta d)}{\partial \beta} \right|_{\beta = 0} = 0,\ j = 1, \ldots, n.$$

于是，我们得到

$$\begin{bmatrix} \nabla^2_y \mathcal{L} & (\nabla_y p_1)^\top & \cdots & (\nabla_y p_m)^\top & (\nabla_y q)^\top \\ \lambda_1 \nabla_y p_1 & p_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & \vdots \\ \lambda_m \nabla_y p_m & 0 & \cdots & p_m & 0 \\ \nabla_y q & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \nabla_d y \\ \nabla_d \lambda_1 \\ \vdots \\ \nabla_d \lambda_m \\ \nabla_d \nu \end{bmatrix} + \begin{bmatrix} \nabla^2_{xy} \mathcal{L} \\ \lambda_1 \nabla_x p_1 \\ \vdots \\ \lambda_m \nabla_x p_m \\ \nabla_x q \end{bmatrix} d = 0.$$

由 (a)(b)，我们有 $\nabla_d \lambda_{J(x^0,y^0)^C}(x^0) = 0$ 和 $\nabla_d \lambda_{J^0_-(x^0,y^0,d)}(x^0) = 0$。于是，该方程化简为 to

$$\begin{bmatrix} \nabla^2_y \mathcal{L} & \nabla_y p^\top_{J^+(x^0,y^0)} & \nabla_y p^\top_{J^0_+(x^0,y^0,d)} & \nabla_y q^\top \\ \lambda_{J^+(x^0,y^0)} \nabla_y p_{J^+(x^0,y^0)} & p_{J^+(x^0,y^0)} & 0 & 0 \\ \lambda_{J^0_+(x^0,y^0,d)} \nabla_y p_{J^0_+(x^0,y^0,d)} & 0 & p_{J^0_+(x^0,y^0,d)} & 0 \\ \nabla_y q & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \nabla_d y \\ \nabla_d \lambda_{J^+(x^0,y^0)} \\ \nabla_d \lambda_{J^0_+(x^0,y^0,d)} \\ \nabla_d \nu \end{bmatrix}$$

$$+ \begin{bmatrix} \nabla^2_{xy} \mathcal{L} \\ \lambda_{J^+(x^0,y^0)} \nabla_x p_{J^+(x^0,y^0)} \\ \lambda_{J^0_+(x^0,y^0,d)} \nabla_x p_{J^0_+(x^0,y^0,d)} \\ \nabla_x q \end{bmatrix} d = 0.$$

由 (c)(d)，$p_{J^+(x^0,y^0)} = 0, p_{J^0_+(x^0,y^0,d)} = 0, \lambda_{J^+(x^0,y^0)} > 0$，以及 $\lambda_{J^0_+(x^0,y^0,d)} > 0$。于是，$\lambda_{J^+(x^0,y^0)}$ 和 $\lambda_{J^0_+(x^0,y^0,d)}$ 被消去。我们得到

$$\begin{bmatrix} \nabla_d y(x^0) \\ \nabla_d \lambda_{J^+(x^0,y^0)}(x^0) \\ \nabla_d \nu(x^0) \\ \nabla_d \lambda_{J^0_+(x^0,y^0,d)}(x^0) \end{bmatrix} = -M_D^{-1}(x^0, y^0, d) N_D(x^0, y^0, d) d$$

其中， $M_D(x^0, y^0, d)$ 和 $N_D(x^0, y^0, d)$ 定义见 (17) 和 (18)。由于线性无关约束资格条件成立， $M_D(x^0, y^0, d)$ 非奇异。

□

假设1、2、3成立， $y^*(x)$ 是 $P(x)$ 的最优解。定义矩阵函数

$$M_+(x) \triangleq \begin{bmatrix} \nabla_y^2 \mathcal{L} & \nabla_y p_{J^+(x,y^*(x))}^\top & \nabla_y q^\top \\ \nabla_y p_{J^+(x,y^*(x))} & 0 & 0 \\ \nabla_y q & 0 & 0 \end{bmatrix} (x, y^*(x), \lambda(x), \nu(x))$$

and

$$N_+(x) \triangleq = [\nabla_{xy}^2 \mathcal{L}^\top, \nabla_x p_{J^+(x,y^*(x))}^\top, \nabla_x q^\top]^\top (x, y^*(x)).$$

对于方向 $d \in \mathbb{R}^{d_x}$，定义 $J_+^0(x,\ y^*(x)\ d)$ 为一个集合，其中包含所有满足以下条件的 $j \in J^0(x,\ y^*(x))$：存在 $\epsilon_0 > 0$，对于任意 $0 < \epsilon < \epsilon_0$，$p_j(x + \epsilon d,\ y^*(x + \epsilon d)) = 0$ 且 $\lambda_j(x + \epsilon d) > 0$。记 $J_-^0(x,\ y^*(x)\ d) \triangleq J^0(x,\ y^*(x)) \setminus J_+^0(x,\ y^*(x)\ d)$。定义

$$M_D(x, d) \triangleq \begin{bmatrix} & & \nabla_y p_{J_+^0(x,y^*(x),d)}^\top \\ & M_+ & 0 \\ & & 0 \\ \nabla_y p_{J_+^0(x,y^*(x),d)} & 0 \quad 0 & 0 \end{bmatrix} (x, y^*(x))$$

and

$$N_D(x, d) \triangleq \left[N_+^\top(x), \nabla_x p_{J_+^0(x,y^*(x),d)}^\top (x, y^*(x)) \right]^\top.$$

定理6 (F　　定理1和定理2的组合的完整版本)。 假设假设*1*、*2*、*3*成立。那么　　，

(i) $y^*（x）$ 的全局最小值存在且唯一。*KKT*条件在 $y^*（x）$ 处成立，且拉格朗日乘子 $\lambda（x）$ 和 $\nu（x）$ 唯一。

(ii) 向量函数 $z（x）\ \triangleq [y^*（x）\ \ \lambda（x）\ \ \nu（x）]$ 连续且局部利普希茨连续。

(iii) z 在 x 处沿任意满足 $\|d\| = 1$的方向 $d \in \mathbb{R}^{d_x}$ 的方向导数存在，且由下式给出

$$\begin{bmatrix} \nabla_d y^*(x) \\ \nabla_d \lambda_{J^+(x,y^*(x))}(x) \\ \nabla_d \nu(x) \\ \nabla_d \lambda_{J_+^0(x,y^*(x),d)}(x) \end{bmatrix} = -M_D^{-1}(x, d) N_D(x, d) d, \tag{20}$$

$$\nabla_d \lambda_{J_-^0(x,y^*(x),d)}(x) = 0,$$

$$\nabla_d \lambda_{J(x,y^*(x))^C}(x) = 0,$$

其中 $M_D（x,\ d）$ 非奇异。

(iv) 如果严格互补松弛 C 在 $y^*（x）$ 处相对于 $\lambda（x）$ 成立，则 z 在 x 处连续可微，且梯度计算为 b　　y

$$\begin{bmatrix} \nabla_x y^*(x) \\ \nabla_x \lambda_{J(x,y^*(x))}(x) \\ \nabla_x \nu(x) \end{bmatrix} = -M_+^{-1}(x) N_+(x),$$

$$\nabla_x \lambda_{J(x,y^*(x))^C}(x) = 0,$$

其中 $M_+(x)$是非奇异的。

证明。 (i) 首先，我们证明 $y^*(x)$ 存在且唯一。函数 $g(x, y)$ 关于 y 是 μ-强凸的，并且对于所有 $x \in \mathbb{R}^{d_x}$，可行集 $K(x)$ 是凸且闭的。情况一： $K(x)$ 有界。于是， $K(x)$ 是一个紧集，且 $g(x, y)$ 连续，这意味着最小值 $y^*(x)$ 存在。情况二： $K(x)$ 无界。函数 $g(x, y)$ 关于 y 是 μ- 强凸的，那么

$$g(x, y) \geq g(x, y^0) + \nabla_y g(x, y^0)^\top (y - y^0) + \frac{\mu}{2} \|y - y^0\|^2.$$

于是，$\lim_{\|y\| \to \infty} g(x, y) = +\infty$。接着，对于任意实数 α，集合 $\{x \mid g(x, y) \leq \alpha\}$ 是闭且有界的，并且存在α 使得 $\{y \mid g(x, y) \leq \alpha\}$ 非空。因此， $\{x \mid g(x, y) \leq \alpha\}$ 是紧的。于是， $\{y \mid g(x, y) \leq \alpha\}$的紧致性

以及 $g(x,y)$ 的连续性意味着，优化问题 $\min g(x,y)$ s.t. $y \in \{t \mid g(x,t) \leq \alpha\}$ 可行且存在一个最小值，该最小值也是问题 $P(x)$ 的最小值。因此，全局最小值 $y^*(x)$ 存在。

$y^*(x)$ 是 $P(x)$ 的全局最小值，这意味着它也是局部极小值。假设 $y^*(x)$ 不是 $P(x)$ 的唯一局部极小值，且还存在另一个局部极小值 $y' \in K(x)$，那么存在一个 $0<\alpha<1$，使得 $g(x,y^*(x)) \leq g(x,\alpha y^*(x)+(1-\alpha)y')$ 且 $g(x,y' \leq g(x,\alpha y^*(x)+(1-\alpha)y')$，那么 $\alpha g(x,y^*(x))+(1-\alpha)g(x,y') \leq g(x,\alpha y^*(x)+(1-\alpha)y')$，这与 $g(x,y)$ 关于 y 是 μ-强凸的事实相矛盾。因此，$y^*(x)$ 是唯一的局部极小值，进而也是唯一的全局最小值。

解 $y^*(x)$ 是 $P(x)$ 的全局最小值，这意味着 $y^*(x)$ 是一个局部极小值。根据假设3，对于 $P(x)$，在 $y^*(x)$ 处满足线性无关约束资格条件。然后，根据定理5的 (i)，对于 $P(x)$，在 $y^*(x)$ 处满足 KKT 条件，且存在唯一的拉格朗日乘子 $\lambda(x)$ 和 $\nu(x)$。

(ii)(iii) 首先证明，对于所有 $x \in \mathbb{R}^{d_x}$，对于 $P(x)$，在 $y^*(x)$ 处且拉格朗日乘子为 $\lambda(x)$ 和 $\nu(x)$ 时，满足严格二阶充分条件。根据假设2，$g(x,y)$ 关于 y 是 μ-强凸的，这意味着 $\nabla_y^2 g(x,y^*(x))$ 是正定的；$p_j(x,y)$ 是凸的，这意味着 $\nabla_y^2 p_j(x,y^*(x))$ 是半正定的；$q_j(x,y)$ 是仿射的，这意味着 $\nabla_y^2 q(x,y^*(x)) = 0$。(i) 中的 KKT 条件意味着 $\lambda(x) \geq 0$。然后，

$$\nabla_y^2 \mathcal{L}\left(y^*(x),\lambda(x),\hat{\nu}(x),x\right) = \nabla_y^2 g(x,y^*(x)) + \lambda(x)^\top \nabla_y^2 p(x,y^*(x)) + \nu(x)^\top \nabla_y^2 q(x,y^*(x))$$

为正定。因此，严格二阶充分条件成立。

在 $y^*(x)$ 处，KKT条件、线性无关约束资格条件和严格二阶充分条件均成立，且对于 $P(x)$，其拉格朗日乘子为 $\lambda(x)$ 和 $\nu(x)$。根据引理3，对于所有 $x^0 \in \mathbb{R}^{d_x}$，在 x^0 的某个邻域内，存在唯一的 Lipschitz 连续向量函数 $\hat{}z(x) \triangleq [\hat{}y^\top(x), \hat{}\lambda^\top(x), \hat{}\nu^\top(x)]^\top$，并且 $\hat{}y(x)$ 是 $P(x)$ 的局部唯一局部极小值。由于对于 $P(x)$，具有唯一拉格朗日乘子 $\lambda(x)$ 和 $\nu(x)$ 的局部极小值 $y^*(x)$ 是唯一的，因此 $\hat{}z(x) = z(x)$。

对于任意 $x^0 \in \mathbb{R}^{d_x}$，向量函数 $z(x) \triangleq [y^*(x)^\top, \lambda(x)^\top, \nu(x)^\top]^\top$ 在 x^0 的某个邻域内是 Lipschitz 连续的，这意味着 $z(x)$ 在 $\mathbb{R}^{d_x}$ 上局部利普希茨连续。梯度的计算见引理3。

(iv) 对于 $P(x)$，在 $y^*(x)$ 处，KKT条件、线性无关约束资格条件和严格二阶充分条件均成立，且对应的拉格朗日乘子为 $\lambda(x)$ 和 $\nu(x)$。根据引理2，若关于 $\lambda(x^0)$ 的严格互补松弛在 $y^*(x^0)$ 处成立，则对于所有 $x^0 \in \mathbb{R}^{d_x}$，在 x^0 的某个邻域内，存在唯一的连续可微向量函数 $\hat{}z(x) \triangleq [\hat{}y^\top(x), \hat{}\lambda^\top(x), \hat{}\nu^\top(x)]^\top$，且 $\hat{}y(x)$ 是 $P(x)$ 的局部唯一局部极小值。由于对于 $P(x)$，具有唯一拉格朗日乘子 $\lambda(x)$ 和 $\nu(x)$ 的局部极小值 $y^*(x)$ 是唯一的，因此 $\hat{}z(x) = z(x)$。梯度的计算见引理2。

□

对于带约束的下层优化 $P(x)$，对每个方向 d，Ralph和Dempe（1995）通过求解一个二次规划问题来计算 $\nabla_d y^*(x)$。在定理6中，我们通过式（20）计算 $\nabla_d y^*(x)$，而集合 $J_+^0(x,\ y^*(x)\ d)$ 可以通过在小邻域内进行采样 ϵ_0 来确定。

C.2 命题1和命题2的证明

下面的引理给出了在不满足 SCSC 的情况下计算 y^* 的 Clarke次微分的方法。

引理 4。 假设定理 *2* 中的所有假设均成立。那么，

$$\bar{\partial} y^*\left(x^0\right) = \operatorname{conv}\left\{w^S(x^0) : S \subseteq J^0(x^0, y^*(x^0))\right\}. \tag{21}$$

这里，$w^S(x^0)$ 是从矩阵 $-M_+^S(x^0,y^*(x^0))^{-1}N_+^S(x^0,y^*(x^0))$ 中提取前 d_x 行得到的。当 S 不为空时，

$$M_+^S(x,y) \triangleq \begin{bmatrix} \nabla_y^2 \mathcal{L} & \nabla_y p_{J^+(x^0,y^*(x^0))}^\top & \nabla_y q^\top & \nabla_y p_S^\top \\ \nabla_y p_{J^+(x^0,y^*(x^0))} & 0 & 0 & 0 \\ \nabla_y q & 0 & 0 & 0 \\ \nabla_y p_S & 0 & 0 & 0 \end{bmatrix}(x,y),$$

and

$$N_+^S(x,y) \triangleq \left[\nabla_{xy}^2 \mathcal{L}^\top, \nabla_x p_{J^+(x^0,y^*(x^0))}^\top, \nabla_x q^\top, \nabla_x p_S^\top\right]^\top (x,y).$$

当 S 为空，$M_+^S(x^0,y^*(x^0)) = M_+(x^0,y^*(x^0))$ 和 $N_+^S(x^0,y^*(x^0)) = N_+(x^0,y^*(x^0))$，其中

$$M_+(x,y) = \begin{bmatrix} \nabla_y^2 \mathcal{L} & \nabla_y p_{J^+(x^0,y^*(x^0))}^\top & \nabla_y q^\top \\ \nabla_y p_{J^+(x^0,y^*(x^0))} & 0 & 0 \\ \nabla_y q & 0 & 0 \end{bmatrix}(x,y),$$

and

$$N_+(x,y) = [\nabla_{xy}^2 \mathcal{L}^\top, \nabla_x p_{J^+(x^0,y^*(x^0))}^\top, \nabla_x q^\top]^\top (x,y).$$

引理 4 已见于相关论文（Dempe 1998；Malanowski 1985），也可以由定理 2 第 (iii) 部分的方向导数推导得出。

引理 5。 设 $f\colon\ \mathbb{R}^{d_x} \longrightarrow \mathbb{R}$ 为一个局部*Lipschitz*函数。则 f 在任意紧集 $S \in \mathbb{R}^{d_x}$上都是*Lipschitz*连续的。

证明。函数 $f\colon\ \mathbb{R}^{d_x} \longrightarrow \mathbb{R}$ 局部Lipschitz，因此对于任意 $x \in \mathbb{R}^{d_x}$，存在一个Lipschitz常数 $l(x)$ 和 $\epsilon(x) > 0$，使得 $l(x)$ 是 f 在 $\mathcal{B}(x, \epsilon(x))$ 上的Lipschitz常数。对于任意紧集 $S \in \mathbb{R}^{d_x}$，我们有 $S \subset \cup_{x\in S}\mathcal{B}(x, \epsilon(x))$，于是$\cup_{x\in S}\mathcal{B}(x, \epsilon(x))$ 是 S的一个开覆盖。集合 S为紧集，这意味着存在一个有限集 F，使得 $S \subset \cup_{x\in F}\mathcal{B}(x, \epsilon(x))$。于是，$\max_{x\in F} l(x) < \infty$ 是 f在 $\cup_{x\in F}\mathcal{B}(x, \epsilon(x))$ 上的Lipschitz常数，因而也是 f在 S上的Lipschitz常数。 □

L引理 6。 假设定理*6*中的所有假设均成立。给定一个开球 B，若对于每个 j，要么 r

$$j \in J^+(x, y^*(x))^C = J(x, y^*(x))^C \cup J^0(x, y^*(x)) \text{ for all } x \in B,$$

or

$$j \in J^+(x, y^*(x)) \text{ for all } x \in B.$$

则 $y^*(x)$ 在 B上连续可微。

证明。 对于任意 j，要么对于所有 $x \in B$，都有 $p_j(x, y^*(x)) \le 0, \lambda_j(x) = 0$，要么对于所有 $x \in B$都有 $p_j(x, y^*(x)) = 0, \lambda_j(x) > 0$。考虑集合 B'

$\triangleq \{x' \in B : \text{for all } j, \text{ either } p_j(x', y^*(x')) < 0, \lambda_j(x') = 0 \text{ or } p_j(x', y^*(x')) = 0, \lambda_j(x') > 0\}.$

(i）根据定理6的第（iii）部分，对于任意点 $x' \in B'$，SCSC在 $y^*(x')$ 处成立，并且 y^* 在 处连续可微 x'.

(ii) C onsider the poi数 $x \in B \setminus B'$, th非空 ist j such that $p_j(x, y^*(x)) = 0, \lambda_j(x) = 0$, i.e., $J^0(x, y^*(x))$ i集。

(ii.a) 考虑满足以下条件的 j：对所有 $x \in B$，$p_j(x, y^*(x)) \le 0, \lambda_j(x) = 0$。根据定理6中的定义，对于任意方向 d，$j \in J^0_-(x, y^*(x); d)$ 不会被纳入式 (19) 中 $-M_D^{-1}N_D$ 的计算。

(ii.b) 考虑满足以下条件的 j：对所有 $x \in B$，$p_j(x, y^*(x)) = 0, \lambda_j(x) > 0$。根据定理6中的定义，对于任意方向 d，$j \in J^0_+(x, y^*(x); d)$ 会被纳入式 (19) 中 $-M_D^{-1}N_D$ 的计算。

由 (ii.a) 和 (ii.b) 可知，对于点 $x \in B \setminus B'$，任意给定 d 时，式 (19) 中 $-M_D^{-1}N_D$ 的计算完全相同，即 li $\mathrm{m}_{h\to 0} \frac{\|y^*(x+h) - y^*(x) + M_D^{-1}(\cdots x\ d\ N_D\ x\ d\ h\|}{|h|} = 0$.)(,)。因此，$y^*(x)$ 在 B 上可微。于是，$y^*(x)$ 在 $B \setminus B'$ 上可微。此外，对于任意 x 和 d，导数 $M_D^{-1}(x, d)N_D(x, d)$ 连续。因此，$y^*(x)$ 在 B 上连续可微。 □

命题1的证明

命题*1*的证明。 根据定理6第(iii)部分中 y^* 的梯度计算，$\Phi(x)$和 y^* 在$\mathcal{B}(x^0, \epsilon)$上可微，因此它们在 $\mathcal{B}(x^0, \epsilon)$上二次可微。当 ϵ 足够小时，对于任意$x' \in \mathcal{B}(x^0, \epsilon)$，$|\|\nabla\Phi(x^0)\| - \|\nabla\Phi(x')\|| < o(\epsilon)$。于是，$|\|\nabla\Phi(x^0)\| - d(0, \bar\partial_\epsilon\Phi(x^0))| < o(\epsilon)$。 □

命题2的证明

命题 4 (命题2的完整版本)。 假设假设*1*、*2*、*3*成立。考虑 $x^0 \in \mathbb{R}^{d_x}$ 和 $\epsilon > 0$，存在$x \in \mathcal{B}(x^0, \epsilon)$ 使得 y^* 在 x处不可连续可微。于是，至少存在一个 j，使得存在 x'、

$x'' \in \mathcal{B}\ x^0\ \epsilon$ (,) 满足

$$j \in J^+(x', y^*(x'))^C = J(x', y^*(x'))^C \cup J^0(x', y^*(x')) \text{ and } j \in J^+(x'', y^*(x'')); \tag{22}$$

定义集合 $I^\epsilon(x^0)$，它包含所有满足 (22) 的 j。定义集合 $I^\epsilon_+(x^0)$，它包含所有满足以下条件的 j：对于任意$x \in \mathcal{B}(x^0, \epsilon)$，

$$j \in J^+(x, y^*(x));$$

定义集合 $I^\epsilon_-(x^0)$，它包含所有满足以下条件的 j：对于任意 $x \in \mathcal{B}(x^0, \epsilon)$，

$$j \in J(x, y^*(x))^C \cup J^0(x, y^*(x)).$$

集合 $G(x^0, \epsilon)$ 定义为

$$G(x^0, \epsilon) \triangleq \{\nabla_x f\left(x^0, y^*\left(x^0\right)\right) + w^S(x^0)^\top \nabla_y f\left(x^0, y^*\left(x^0\right)\right) : S \subseteq I^\epsilon(x^0)\}.$$

这里，$w^S(x^0)$是从矩阵 $-M^S_\epsilon(x^0, y^*(x^0))^{-1}N^S_\epsilon(x^0, y^*(x^0))$中提取前 d_x 行得到的，其中

$$M^S_\epsilon \triangleq \begin{bmatrix} \nabla^2_y\mathcal{L} & \nabla_y p^\top_{I^\epsilon_+(x^0)} & \nabla_y q^\top & \nabla_y p^\top_S \\ \nabla_y p_{I^\epsilon_+(x^0)} & 0 & 0 & 0 \\ \nabla_y q & 0 & 0 & 0 \\ \nabla_y p_S & 0 & 0 & 0 \end{bmatrix},$$

and

$$N_\epsilon^S \triangleq \left[\nabla_{xy}^2 \mathcal{L}^\top, \nabla_x p_{I_+^\epsilon(x^0)}^\top, \nabla_x q^\top, \nabla_x p_S^\top\right]^\top.$$

考虑 $x^0 \in \mathbb{R}^{d_x}$，并假设存在一个足够小的 $\epsilon > 0$，使得存在 $x \in \mathcal{B}(x^0, \epsilon)$，满足 $y^*(x)$ 在 x 处不是连续可微的。那么，有如下结论

(i) 对于任意 $g \in G\left(x^0, \epsilon\right)$，存在 $x' \in \mathcal{B}(x^0, \epsilon)$ 使得 $\|g - \nabla\Phi(x')\| < o(\epsilon)$。对于任意 $x'' \in \mathcal{B}(x^0, \epsilon)$，若 $y^*(x)$ 在 x'' 处可微，则存在 $g \in G\left(x^0, \epsilon\right)$ 使得 $\|g - \nabla\Phi(x'')\| < o(\epsilon)$。

(ii) 对于任意 $z \in \mathbb{R}^{d_x}$,

$$|d(z, \operatorname{conv} G(x^0, \epsilon)) - d(z, \bar{\partial}_\epsilon \Phi(x^0))| < o(\epsilon).$$

命题 *4* 的证明。若对于 $x^0 \in \mathbb{R}^{d_x}$ 和 $\epsilon > 0$，$y^*(x)$ 在开球 $\mathcal{B}(x^0, \epsilon)$ 上不连续可微。假设对于所有 j，都不存在满足 (22) 的点 $x', x'' \in \mathcal{B}(x^0, \epsilon)$。根据引理 6，$y^*(x)$ 在 $\mathcal{B}(x^0, \epsilon)$ 上连续可微，这将导致矛盾。因此，至少存在一个 j，使得存在满足 (22) 的 $x', x'' \in \mathcal{B}(x^0, \epsilon)$。

(i) 首先，与命题1中的证明类似，如果

$$M_\epsilon^S = \begin{bmatrix} \nabla_y^2 \mathcal{L} & \nabla_y p_{I_+^\epsilon(x^0)}^\top & \nabla_y q^\top & \nabla_y p_S^\top \\ \nabla_y p_{I_+^\epsilon(x^0)} & 0 & 0 & 0 \\ \nabla_y q & 0 & 0 & 0 \\ \nabla_y p_S & 0 & 0 & 0 \end{bmatrix},$$

and

$$N_\epsilon^S = \left[\nabla_{xy}^2 \mathcal{L}^\top, \nabla_x p_{I_+^\epsilon(x^0)}^\top, \nabla_x q^\top, \nabla_x p_S^\top\right]^\top.$$

的分量与式（20）中所示的 $M_D(x^0, d)$ 和 $N_D(x^0, d)$ 的分量相同，即矩阵计算中涉及的集合:

$$I_+^\epsilon(x') \cup S = J_+^0(x'', y^*(x''), d) \cup J^+(x'', y^*(x'')),$$

那么，对于所有 x' 和 $x'' \in \mathcal{B}(x^0, \epsilon)$，差值 $\|M_\epsilon^S(x', y^*(x'))^{-1} N_\epsilon^S(x', y^*(x')) - M_D(x'', d))^{-1} N_D(x'', d)\| < o(\epsilon)$ 为零，因为函数 $(M_\epsilon^S)^{-1} N_\epsilon^S$ 可微（参见命题1中的证明）。

(a) 假设对于任意 $x' \in \mathcal{B}(x^0, \epsilon)$ 和任意 $d \in \mathbb{R}^{d_x}$，都存在 $g^S \in G\left(x^0, \epsilon\right)$，使得

$$I_+^\epsilon(x^0) \cup S = J_+^0(x', y^*(x'), d) \cup J^+(x', y^*(x'));$$

那么，对于任意满足 y^* 在 x' 处可微的 $x' \in \mathcal{B}(x^0, \epsilon)$，都存在 $g^S \in G\left(x^0, \epsilon\right)$，使得对于方向 $d \in \mathbb{R}^{d_x}$, $\|M_\epsilon^S(x^0, y^*(x^0))^{-1} N_\epsilon^S(x^0, y^*(x^0)) - M_D(x', d)^{-1} N_D(x', d)\| < o(\epsilon)$。然后，$\|w^S(x^0) - y^*(x')\| < o(\epsilon)$ 且 $\|g^S - \nabla\Phi(x')\| < o(\epsilon)$。

(b) 假设对于任意 $g^S \in G\left(x^0, \epsilon\right)$，都存在 $x' \in \mathcal{B}(x^0, \epsilon)$ 和 $d \in \mathbb{R}^{d_x}$，使得

$$I_+^\epsilon(x^0) \cup S = J_+^0(x', y^*(x'), d) \cup J^+(x', y^*(x')).$$

由于集合 $J_+^0(x', y^*(x'), d) \subseteq J^0(x', y^*(x'))$，根据引理 4 及其证明（Dempe 1998年；Malanowski 1985），$g^S \in \bar{\partial} y^*(x')$，并且存在 $\{x^j\}$，使得 $g^S = \lim_{j\to\infty} \nabla y^*\left(x^j\right)$：$\{x^j\} \to x'$，其中 y^* 对所有 j 在 y^j 处可微。于是，在 x' 的任意小邻域内，存在 x''，使得 $J_+^0(x', y^*(x'), d) \cup J^+(x', y^*(x')) = J^+(x'', y^*(x''))$，且 y^* 在 x'' 处可微。于是，

$$I_+^\epsilon(x^0) \cup S = J^+(x'', y^*(x'')),$$

并且我们有 $\|w^S(x^0) - y^*(x'')\| < o(\epsilon)$ 和 $\|g^S - \nabla\Phi(x'')\| < o(\epsilon)$，其中 $x'' \in \mathcal{B}(x^0, \epsilon)$。于是，命题的 (i) 部分得证。

现在，我们可以通过证明 (a) 和 (b) 的假设成立来证明 (i)。定义

$$J^N(x, y^*, x) \triangleq \{j \mid j \notin (,)) : (,)) (,)) 。$$

$J^+(x, y^*, x) \cup J^0(x, y^*, x)\}$ $x^* \in \mathcal{B}(x^0, \epsilon)$ $d \in \mathbb{R}^{d_x}$ $I^\epsilon(x^0) \cup I_+^\epsilon(x^0) \cup I_-^\epsilon(x^0) = J^0(x^*, y^*, x^*) \cup$

然后，对于任意 (,) 和任意，()()()(,()) $J^+(x^*, y^*(x^*)) \cup J^N(x^*, y^*(x^*)) = J_-^0(x^*, y^*(x^*), d) \cup J_+^0(x^*, y^*(x^*), d) \cup J^+(x^*, y^*(x^*)) \cup J^N(x^*, y^*(x^*))$，因为它们都包含所有约束 p_j。注意，在 $I^\epsilon(x^0)$、$I_+^\epsilon(x^0)$ 和 $I_-^\epsilon(x^0)$ 中任取两个的交集为空，在 $J_-^0(x^*, y^*(x^*), d)$、$J_+^0(x^*, y^*(x^*), d)$、$J^+(x^*, y^*(x^*))$ 和 $J^N(x^*, y^*(x^*))$ 中任取两个的交集为空，并且 $J^0(x^*, y^*(x^*)) = J_+^0(x^*, y^*(x^*), d) \cup J_-^0(x^*, y^*(x^*), d)$。

(1) 考虑 $j \in I_+^\epsilon(x^0)$，即 $p_j(x, y^*(x)) = 0$ 且 $\lambda_j(x) > 0$ 对所有 $x \in \mathcal{B}(x^0, \epsilon)$ 均成立。那么，对于任意 $x' \in \mathcal{B}(x^0, \epsilon)$，都有 $j \in J^+(x', y^*(x'))$。于是，对于任意 $d \in \mathbb{R}^{d_x}$ 以及任意 $x' \in \mathcal{B}(x^0, \epsilon)$，都有 $j \in J_+^0(x', y^*(x'), d) \cup J^+(x', y^*(x'))$。因此，

$$I_+^\epsilon(x^0) \subseteq J_+^0(x', y^*(x'), d) \cup J^+(x', y^*(x'))$$

对于任意 $d \in \mathbb{R}^{d_x}$ 以及任意 $x' \in \mathcal{B}(x^0, \epsilon)$。

(2) 考虑 $j \in I_{-}^{\epsilon}(x^0)$。若给定 $x \in \mathcal{B}(x^0,\epsilon)$，我们有 $p_j(x,y^*(x)) < 0$ 且 $\lambda_j(x) = 0$，则 $j \notin J^+(x,y^*(x)) \cup J^0(x,y^*(x))$，即 $j \in J^N(x,y^*(x))$。若给定 $x \in \mathcal{B}(x^0,\epsilon)$，我们有 $p_j(x,y^*(x)) = 0, \lambda_j(x) = 0$，并且这对任意 $x \in \mathcal{B}(x^0,\epsilon)$ 都成立，则对任意 $d \in \mathbb{R}^{d_x}$，有 $j \in J_{-}^0(x,y^*(x)\ d)$。若给定 $x \in \mathcal{B}(x^0,\epsilon)$，我们有 $p_j(x,y^*(x)) = 0, \lambda_j(x) = 0$，但这并非对所有集合 $\mathcal{B}(x^0,\epsilon)$ 都成立，则对任意 $d \in \mathbb{R}^{d_x}$，有 $j \in J_{-}^0(x,y^*(x)\ d)$。然后，

$$I_{-}^{\epsilon}(x^0) \subseteq J_{-}^0(x,y^*(x),d) \cup J^N(x,y^*(x)),$$

对于任意 $x \in \mathcal{B}(x^0,\epsilon)$ 以及任意 $d \in \mathbb{R}^{d_x}$。

(3) 考虑 $j \in I^{\epsilon}(x^0)$。当 $\epsilon > 0$ 足够小时，存在 $x^* \in \mathcal{B}(x^0,\epsilon)$，使得对于所有 $j \in I^{\epsilon}(x^0)$，$p_j(x^*,y^*(x^*)) = 0$ 且$\lambda_j(x^*) = 0$。于是，我们有

$$I^{\epsilon}(x^0) \subseteq J^0(x^*,y^*(x^*)).$$

于是，对于任意 $S \subseteq I^{\epsilon}(x^0)$，$S \subseteq J^0(x^*,y^*(x^*))$。此外，

$$J^+(x^*,y^*(x^*)) \subseteq I_{+}^{\epsilon}(x^0).$$

这是因为，假设 $j \in J^+(x^*,y^*(x^*))$ 且 $j \notin I_{+}^{\epsilon}(x^0)$，那么 $j \in I^{\epsilon}(x^0)$，进而 $p_j(x^*,y^*(x^*)) = 0$ 且$\lambda_j(x^*) = 0$，并且 $j \in J^0(x^*,y^*(x^*))$，这与 $j \in J^+(x^*,y^*(x^*))$ 矛盾。

由 (1) 和 (3)，我们有：对于任意 $d \in \mathbb{R}^{d_x}$，存在 $x^* \in \mathcal{B}(x^0,\epsilon)$，

$$J^+(x^*,y^*(x^*)) \subseteq I_{+}^{\epsilon}(x^0) \subseteq J^+(x^*,y^*(x^*)) \cup J_{+}^0(x^*,y^*(x^*),d). \tag{23}$$

由引理 4 以及 (Dempe 1998; Malanowski 1985) 中给出的引理证明，对于任意集合 $A \subseteq J^0(x^*,y^*(x^*))$，我们可以找到一个方向 d，使得 $A = J_{+}^0(x^*,y^*(x^*)\ d)$。这里，我们可以令 $A = S \cup I_{+}^{\epsilon}(x^0) \setminus J^+(x^*,y^*(x^*))$。由 (23)，$A \subseteq J^0(x^*,y^*(x^*))$。然后，$S \cup I_{+}^{\epsilon}(x^0) \setminus J^+(x^*,y^*(x^*)) = J_{+}^0(x^*,y^*(x^*)\ d)$。接着，对于任意 $g^S \in G\left(x^0,\epsilon\right)$，存在$x^* \in \mathcal{B}(x^0,\epsilon)$ 和 $d \in \mathbb{R}^{d_x}$，使得

$$I_{+}^{\epsilon}(x^0) \cup S = J_{+}^0(x^*,y^*(x^*),d) \cup J^+(x^*,y^*(x^*)).$$

假设 (b) 得证。

由 (2)，若 $j \in I_{-}^{\epsilon}(x^0)$，则对于任意 $x \in \mathcal{B}(x^0,\epsilon)$ 以及任意 $d \in \mathbb{R}^{d_x}$, $j \in J_{-}^0(x,y^*(x)\ d)$，有 $\cup J^N(x,y^*(x))$。于是，$j \notin J_{+}^0(x,y^*(x)\ d) \cup J^+(x,y^*(x))$。于是，对于任意集合 $C \subseteq I_{-}^{\epsilon}(x^0)$，$C \cap (J_{+}^0(x,y^*(x)\ d) \cup J^+(x,y^*(x)))$ 为空。此外，由 (1)，$I_{+}^{\epsilon}(x^0) \subseteq J_{+}^0(x,y^*(x)\ d) \cup J^+(x,y^*(x))$。于是，对于任意 $x \in \mathcal{B}(x^0,\epsilon)$ 以及任意 $d \in \mathbb{R}^{d_x}$，存在 $g^S \in G\left(x^0,\epsilon\right)$ 使得

$$I_{+}^{\epsilon}(x^0) \cup S = J_{+}^0(x,y^*(x),d) \cup J^+(x,y^*(x));$$

已证明 (a) 的假设成立。至此，命题 (i) 部分的证明完成。

(ii) 根据 Clarke 广义微分的定义，不难看出 $\bar{\partial}_{\epsilon}\Phi(x^0) \subseteq \operatorname{conv}\{\nabla\Phi(x') : x' \in \mathcal{B}(x^0,\epsilon)\ y^* \text{在}\ x' \text{处可微}\}$。于是，对于任意 $g_0 \in \bar{\partial}_{\epsilon}\Phi(x^0)$，存在 $g_1, g_2 \in \{\nabla\Phi(x') : x' \in \mathcal{B}(x^0,\epsilon)\ y^* \text{在}\ x' \text{处可微}\}$，使得 $g_0 = \theta g_1 + (1-\theta)g_2$ 且 $0 \le \theta \le 1$。根据 (i)，我们可以找到 g_1'，$g_2' \in G\left(x^0,\epsilon\right)$，使得 $\|g_1 - g_1'\| < o(\epsilon)$ 且$\|g_2 - g_2'\| < o(\epsilon)$。于是，

$$g_0' = \theta g_1' + (1-\theta)g_2' \in \operatorname{conv} G(x^0,\epsilon),$$

且对于任意 $z \in \mathbb{R}^{d_x}$，

$$\begin{aligned}
|\|z-g_0\| - \|z-g_0'\|| &= |\|z-(\theta g_1 + (1-\theta)g_2)\| - \|z-(\theta g_1' + (1-\theta)g_2')\|| \\
&\le |\|\theta(z-g_1) + (1-\theta)(z-g_2)\| - \|(\theta(z-g_1')) + (1-\theta)(z-g_2')\|| \\
&\le \|\theta(z-g_1) + (1-\theta)(z-g_2) - (\theta(z-g_1')) - (1-\theta)(z-g_2')\| \\
&\le \|\theta(z-g_1) - (\theta(z-g_1'))\| + \|(1-\theta)(z-g_2) - (1-\theta)(z-g_2')\| \\
&< o(\epsilon).
\end{aligned}$$

令 $l = d(z,\bar{\partial}_{\epsilon}\Phi(x^0))$。那么，对于任意 $\sigma > 0$，存在 $g_0 \in \bar{\partial}_{\epsilon}\Phi(x^0)$ 使得 $l \le \|z-g_0\| < l+\sigma$。于是，存在 $g_0' \in G\ x^0\ \epsilon$ $|\|z-g_0\| - \|z-g_0'\|| < o\ \epsilon$ $l - o\ \epsilon \le \|z-g_0'\| < l + \sigma + o\ \epsilon$ $\|z - \text{conv}(,))$，满足 ()。因此，() ()。由于 (, c $g_0'\| \ge d\ z$ $G\ x^0\ \epsilon$ $d\ z$ $G\ x^0\ \epsilon = \{\|z-a\| \mid a \in G\ x^0\ \epsilon\} < l + \sigma + o\ \epsilon$ onv (,)), 我们有 (,conv (,)) inf conv (,) ()。于是，对于任意 $\sigma > 0$，$d(z,\text{conv } G(x^0,\epsilon)) < d(z,\bar{\partial}_{\epsilon}\Phi(x^0)) + \sigma + o(\epsilon)$，即 $d(z,\text{conv } G(x^0,\epsilon)) \le d(z,\bar{\partial}_{\epsilon}\Phi(x^0)) + o(\epsilon)$。

与 $d\ z$ $G\ x^0\ \epsilon + o\ \epsilon$ $\le$ $|d\ z$ $G\ x^0\ \epsilon - d\ z\ \bar{\partial}_{\epsilon}\Phi\ x^0\ | < o\ \epsilon$ 比较可得 $d(z,\ \text{conv}\ G(x^0,\ \epsilon)) \le d(z,\ \bar{\partial}_{\epsilon}\Phi(x^0)) + o(\epsilon)$，由 (i) 可得 $d(z,\ \partial_{\epsilon}\Phi\ x^0\ ^-) \le ()) (,$ conv (,)) ()。然后，(, conv (,)) (, ( )) ()。然后，命题 4 的第 (ii) 部分的证明就完成了。

□

注意到命题2包含在命题 4 中，因此命题2得证。

C.3 定理 3 的证明

我们首先 pa 定理 3 的证明中的第 (i) 部分由引理 7 得出，并由引理 8 证明第 (ii) 部分。

引理 7. 设 $\emptyset \neq C \subset \mathbb{R}^n$ 是紧凸集，且 $\beta \in (0,1)$ 并且 $0 \notin C$。若 u、$v \in C$ 和 $\|u\| = d(0, C)$，则 $\langle v, u\rangle \geq \|u\|^2$。

证明。假设 $\|u\| = d(0, C)$ 且 $\langle v, u\rangle < \|u\|^2$。那么，存在 $0 < \theta < 1$ 使得 $w = \theta v + (1-\theta)u$ 且 $\|w\| < \|u\|$。于是，由于 C 是凸的，$w \in C$。接着 $\|u\| \neq d(0, C)$，这就产生了矛盾。 □

定理 3 (i) 部分的证明。首先，考虑 y^* 在 x^k 处可微。我们有 $g^k = \min\{\|g\| : g \in \text{conv } G(x^k, \epsilon_k)\}$ 和 $\|g^k\| > 0$。我们有 conv $G(x^k, \epsilon_k)$ 是闭、有界且凸的，因此是凸且紧的。由于 conv $G(x^k, \epsilon_k)$ 是闭的，$g^k \in \text{conv } G(x^k, \epsilon_k)$ 和 $\|g^k\| = d(0, \text{conv } G(x^k, \epsilon_k))$。由引理 7，$\langle v, g_k\rangle \geq \|g^k\|^2$ 对任意 $v \in \text{conv } G(x^k, \epsilon_k)$ 都成立。由于 y^* 在 x^k 处可微，我们有 $\nabla\Phi(x^k) \in \text{conv } G(x^k, \epsilon_k)$。然后，$\langle \nabla\Phi(x^k)\ g_k\rangle \geq \|g^k\|^2$。由于 y^* 在 x^k 处可微，且 $\nabla\Phi$ 在 x^k 处可微，因此 Φ 在 x^k 处二阶可微。于是，当 t 足够小时，

$$\begin{aligned}\Phi(x^k - tg^k) =& \Phi(x^k) - t\langle \nabla\Phi(x^k), g^k\rangle + o(t^2)\\ \leq& \Phi(x^k) - t\|g^k\|^2 + o(t^2)\\ <& \Phi(x^k) - \beta t\|g^k\|^2,\end{aligned}$$

对任意的 $0 < \beta < 1$。那么，线搜索存在一个正解 t_k。

其次，考虑 y^x 在 x^k 处不可微。由引理 4、命题2及命题2的证明（亦见于（Dempe 1998年；Malanowski 1985））， $\Phi(x)$ 是由有限个二阶可微函数 $\{\Phi_i(x)\}_{1\leq i\leq m}$ 构成的分段函数，并且对于每个 $1 \leq i \leq m$， $\nabla\Phi_i(x^k) \in \text{conv } G(x^k,\ \epsilon_k)$。那么，当 t 足够小时，对所有 $1 \leq i \leq m$,

$$\begin{aligned}\Phi_i(x^k - tg^k) =& \Phi_i(x^k) - t\langle \nabla\Phi_i(x^k), g^k\rangle + o(t^2)\\ \leq& \Phi_i(x^k) - t\|g^k\|^2 + o(t^2)\\ <& \Phi_i(x^k) - \beta t\|g^k\|^2\\ =& \Phi(x^k) - \beta t\|g^k\|^2,\end{aligned}$$

对任意的 $0 < \beta < 1$。那么，我们有 $\Phi(x^k - tg^k) < \Phi(x^k) - \beta t\|g^k\|^2$。那么，线搜索存在一个正解 t_k。 □

引理 8。若对于充分小的 ϵ_k，下极限 $k\to\infty$ 最大值 $\left\{\|x^k - x\|, \|g^k\|, \epsilon_k\right\} = 0$ 满足 $|\|g^k\| - d(0,\ \bar\partial_{\epsilon_k}\Phi(x^k))| < o(\epsilon_k)$，则 $0 \in \partial\Phi(x)$。

证明。若 $\liminf_{k\to\infty} \max\left\{\|x^k - x\|, \|g^k\|, \epsilon_k\right\} = 0$，则序列 $\{x^k\}$ 存在一个子序列 $\{x^{k_n}\}$，使得 $x^{k_n} \to x$、$\|g^k\| \to 0$ 和 $\epsilon_k \to 0$。由于 $|\|g^k\| - d(0, \bar\partial_\epsilon\Phi(x^k))| < o(\epsilon_k)$ 且 $\epsilon_k \to 0$，我们有 $d(0, \bar\partial_{\epsilon_k}\Phi(x^k)) \to 0$，即存在一个序列 $\{h^k\}$，满足 $h^k \in \bar\partial_{\epsilon_k}\Phi(x^k)$ 且 $\|h^k\| \to 0$，这意味着 $0 \in \bar\partial\Phi(x)$。 □

定理 *3* 的 *(ii)-(iv)* 部分的证明。(ii.a) 根据定理 3 的 (i) 部分，线搜索存在一个非零解 t_k，我们有 $\Phi(x^{k+1}) < \Phi(x^k) - \beta t_k\|g^k\|^2$，于是，

$$\sum_{k=1}^{\infty} \beta t_k\|g^k\|^2 < \sum_{k=1}^{\infty} \Phi(x^k) - \Phi(x^{k+1}) \leq \Phi(x^1) - a < \infty,$$

其中 a 是 $\mathbb{R}^{d_x}$ 上 $\Phi(x)$ 的下界。由于 $x^{k+1} = x^k - tg^k$，我们有

$$\sum_{k=1}^{\infty} \|x^{k+1} - x^k\|\|g^k\| < \infty. \tag{24}$$

(ii.b) 现在我们证明 $\lim_{k\to\infty}\nu_k = 0$ 和 $\lim_{k\to\infty}\epsilon_k = 0$。假设结论不成立。则存在 k_1、$\hat\nu$ 和 $\hat\epsilon$，使得对所有 $k > k_1$，都有 $\nu_k = \hat\nu$、$\epsilon_k = \hat\epsilon$ 和 $\|g^k\| > \hat\nu$。于是，由 (24) 可得 $\sum_{k=1}^{\infty}\|x^{k+1} - x^k\| < \infty$，进而 x^k 收敛到一点 $\bar x$ 和 $t_k\|g^k\| \to 0$，这意味着 $t_k \to 0$。

与定理 3 的第 (i) 部分的证明类似，当 t 充分小时，

$$\begin{aligned}\Phi(x^k - tg^k) =& \Phi(x^k) - t\langle \nabla\Phi_i(x^k), g^k\rangle + \alpha t^2\\ \leq& \Phi(x^k) - t\|g^k\|^2 + \alpha t^2\\ <& \Phi(x^k) - \beta t\|g^k\|^2,\end{aligned} \tag{25}$$

其中，$\Phi(x)$是由有限个二阶可微函数 $\{\Phi_i(x)\}_{1\le i\le m}$构成的分段函数，$\alpha$ 是集合 $\mathrm{conv}\{x^k\}$上 $\{\|\nabla^2\Phi_i(x)\|\}_{1\le i\le m}$ 的上界。由于 $x^k\to\bar{x}$，存在一个有界闭集 A，使得 $\mathrm{conv}\{x^k\}\subset A$。由于 $\nabla^2\Phi_i(x)$在紧集 A上连续，$\|\nabla^2\Phi_i(x)\|$ 有界。因此，上界 α 存在。

由 (25) 可知，存在 t_0，使得对于所有 $t<t_0$ 和 $-t\|g^k\|^2+\alpha t^2<-\beta t\|g^k\|^2$，即 $t<(1-\beta)\frac{\|g^k\|^2}{\alpha}$，都有 t 满足不等式 $\Phi(x^k-tg^k)<\Phi(x^k)-\beta t\|g^k\|^2$。由于对于所有 $k>k_1$，$\|g^k\|>\hat{\nu}$，因此若 $t<t_0$ 且 $t<(1-\beta)\hat{\alpha}\frac{\nu^2}{}$，则对于所有 $k>k_1$，都有$t<(1-\beta)\frac{\|g^k\|^2}{\alpha}$。于是，线搜索 $t_k=\sup\{t\in\{\gamma,\gamma^2,\ldots\}:\Phi(x^k-tg^k)<\Phi(x^k)-\beta t\|g^k\|^2\}$ 存在解$t_k\ge\gamma^N$，且对于所有 $k>k_1$，$\gamma^N\le\min\{t_0,(1-\beta)\hat{\alpha}\nu^2\}<\gamma^{N-1}$。因此，$t_k$ 不会收敛到 0，这与 (24) 矛盾。于是，$\lim_{k\to\infty}\nu_k=0$ 且 $\lim_{k\to\infty}\epsilon_k=0$。

(iii) 由于 $\lim_{k\to\infty}\nu_k=0$，因此对于任意小的 $\nu>0$ 和任意 k_0，存在 $k>k_0$ 使得 $\|g_k\|<\nu$，于是 $\liminf_{k\to\infty}\|g^k\|=0$。由于 $|\|g^k\|-d(0,\bar{\partial}_\epsilon\Phi(x^k))|<o(\epsilon_k)$ 且 $\epsilon_k\to0$，我们有 $\liminf_{k\to\infty}d(0,\bar{\partial}\Phi(x^k))=0$。

(iv) 现在我们有 $\lim_{k\to\infty}\nu_k=0$、$\lim_{k\to\infty}\epsilon_k=0$ 和 $\liminf_{k\to\infty}\|g^k\|=0$。设$\overset{k}{x}\overset{\infty}{\to}$是 $\{x^k\}$ 的一个极限点。若 x^k 收敛到 $\bar{x}$，则 $\|x^k-x\|\to0$。因此我们有 $\liminf_{k\to\infty}\max\{\|x^k-x\|,\|g^k\|,\epsilon_k\}=0$。根据命题2，若 Φ 在球 $\mathcal{B}(x^0,\epsilon)$ 上不可微，则对于充分小的 ϵ_k，有 $|\|g^k\|-d(0,\bar{\partial}_{\epsilon_k}\Phi(x^k))|=|d(0,\mathrm{conv}\ G(x^k,\epsilon_k))-d(0,\bar{\partial}_{\epsilon_k}\Phi(x^k))|<o(\epsilon_k)$。根据命题 3，若 Φ 在球 $\mathcal{B}(x^0,\epsilon)$、$g^k=\bar{\partial}\Phi(x^k))$ 上不可微，则 $|\|g^k\|-d(0,\bar{\partial}_{\epsilon_k}\Phi(x^k))|<o(\epsilon_k)$。然后，根据引理 8，我们有 $0\in\partial\Phi(x)$。

考虑 x^k 不收敛于 $\bar{x}$。注意到 $\bar{x}$ 是 $\{x^k\}$的极限点。由于 $\lim_{k\to\infty}\epsilon_k=0$，我们只需证明 $\liminf_{k\to\infty}\max\{\|x^k-x\|,\|g^k\|\}=0$，于是由引理 8 可得 $0\in\bar{\partial}\Phi(x)$。假设 $\liminf_{k\to\infty}\max\{\|x^k-x\|,\|g^k\|\}>0$。由于 $\bar{x}$ 是 $\{x^k\}$的极限点，对于任意 $\hat{v}>0$ 和 $\hat{k}$，存在一个无限集 $K(\hat{k},\hat{v})\triangleq\{k:\ k\ge\hat{k},\|x^k-\bar{x}\|<\hat{v}\}$。由于 $\liminf_{k\to\infty}\max\{\|x^k-x\|,\|g^k\|\}>0$，存在 $\hat{v}>0$ 和 $\hat{k}$，使得对于所有 $k\in K(\hat{k},\hat{v})$，均有 $\|g^k\|>\hat{v}$。由 (24) 可得 $\sum_{k\in K(\hat{k},\hat{v})}\|x^{k+1}-x^k\|\|g^k\|<\infty$，于是 $\sum_{k\in K(\hat{k},\hat{v})}\|x^{k+1}-x^k\|<\infty$。由于 x^k 不收敛于 $\bar{x}$，存在 $\epsilon>0$，使得对于每个满足 $\|x^k-\bar{x}\|\le\hat{v}/2$的 $k\in K(\hat{k},\hat{v})$，都存在满足 $\|x^k-x^{k'}\|>\epsilon$ 且对所有 $k\le i<k'$均有 $x^i\in K(\hat{k},\hat{v})$ 的 $k'>k$。于是，我们有 $\epsilon<\|x^{k'}-x^k\|\le\sum_{i=k}^{k'-1}\|x^{i+1}-x^i\|$。由于 $\sum_{i\in K(\hat{k},\hat{v})}\|x^{i+1}-x^i\|<\infty$，我们可以选取充分大的 k，使得 $\sum_{i=k}^{\infty}\|x^{i+1}-x^i\|<\epsilon$。于是得到矛盾。因此，$\liminf_{k\to\infty}\max\{\|x^k-x\|,\|g^k\|\}=0$。

因此，无论 x^k 是否收敛到极限点 $\bar{x}$，我们都有 $0\in\bar{\partial}\Phi(x)$。定理 3 的第 (iii) 部分得证。

□

C.4 命题 3 的证明

命题 3 的证明。(i) 根据定理6的第 (ii) 部分，向量函数 $z(x)\triangleq[y^*(x)^\top,\lambda(x)^\top,\nu(x)^\top]^\top$ 是局部利普希茨连续的。引理 5 表明，$z(x)$ 在任意紧集 $S\in\mathbb{R}^{d_x}$ 上都是利普希茨连续函数。因此，对于所有j，$y^*(x)$、$\lambda_j(x)$ 和 $p_j(x)$ 在任意紧集上都是利普希茨连续函数。由于 f 和 p 连续可微，因此 $\Phi(x)=f(x,y^*(x))$ 和 $p(x,y^*(x)$ 都是局部利普希茨连续的。因此，$\Phi(x)$ 和 $f(x,y^*(x))$ 在任意紧集上都是利普希茨连续函数。至此，第 (i) 部分得证。

对于所有 $j\in J(x^0,\hat{y})$，要么 $\lambda_j(x^0)>l_{\lambda_j}(x^0,\epsilon)\epsilon$，要么 $p_j(x^0,y^*(x^0))<-l_{p_j}(x^0,\epsilon)\epsilon$。首先，若 $\lambda_j(x^0)>l_{\lambda_j}(x^0,\epsilon)\epsilon$，则对于任意 $x\in\mathcal{B}(x^0,\epsilon)$，$|\lambda_j(x)-\lambda_j(x^0)|\le l_{\lambda_j}(x^0,\epsilon)\|x-x^0\|\le l_{\lambda_j}(x^0,\epsilon)\epsilon$。于是，对于任意 $x\in\mathcal{B}(x^0,\epsilon)$，$\lambda_j(x)>0$，并且$p_j(x^0,y^*(x^0))=0$，这是由 KKT条件得到的。然后 $j\in J^+(x,y^*(x))$。其次，若 $p_j(x^0,y^*(x^0))<-l_{p_j}(x^0,\epsilon)\epsilon$，则对于任意 $x\in\mathcal{B}(x^0,\epsilon)$，$|p_j(x^0,y^*(x^0))-p_j(x,y^*(x))|\le l_{p_j}(x^0,\epsilon)\|x-x^0\|\le l_{p_j}(x^0,\epsilon)\epsilon$。于是，对于任意$x\in\mathcal{B}(x^0,\epsilon)$，$p_j(x,y^*(x))<0$，且 $j\notin J(x,y^*(x))$。因此，对于任意 $x\in\mathcal{B}(x^0,\epsilon)$，$J^0(x,y^*(x))$ 为空，这意味着对于任意 $x\in\mathcal{B}(x^0,\epsilon)$，SCSC 在 $y^*(x)$ 相对于 $\lambda(x)$ 成立。根据定理6的 (iii) 部分，$y^*(x)$ 在 $\mathcal{B}(x^0,\epsilon)$ 上可微。于是 $\Phi(x)$ 在 $\mathcal{B}(x^0,\epsilon)$ 上可微。

□