arXiv:2305.03208v1 [math.OC] 4 May 2023

从度量视角看一阶优化算法的复杂度

A.S. Lewis * Tonghua Tian †

2023年5月8日

摘要

理解一阶优化算法的一个核心工具是 Kurdyka-Lojasiewicz 不等式。此类方法的标准处理方式关键依赖于这一不等式，以利用涉及梯度或次梯度的充分下降条件。然而，KL 性质从根本上关注的并不是次梯度，而是“斜率”这一纯粹的度量概念。通过强调这一视角，并避免使用任何次梯度，我们针对度量空间上的一阶优化算法，给出了一种简明扼要的复杂度分析。这种无次梯度视角还为非光滑半代数函数的 KL 性质提供了一个简短而聚焦的证明框架。

关键词：非光滑优化、一阶算法、斜率、KL 性质、复杂度、半代数

AMS 学科分类：90C48, 49J52, 65Y20, 14P10

1 引言

过去二十年来，从 [1, 6]，等开创性研究开始，Kurdyka-Lojasiewicz 不等式已成为一阶优化算法收敛分析中的核心工具。若将其视为一种次梯度性质 [12, 13]，那么若干因素——该不等式为何对一大类非光滑函数（例如半代数函数）成立这一深层原因，以及将其应用于优化复杂度分析时有时较为繁复的论证——可能会掩盖这一方法本质上的简洁性。然而，正如 [14]，

*ORIE，康奈尔大学，纽约州伊萨卡。people.orie.cornell.edu/aslewis 研究部分获得美国国家科学基金会 DMS-2006990 号资助。

†ORIE，康奈尔大学，纽约州伊萨卡。tt543@cornell.edu

所指出的，该不等式本质上是度量空间的一种性质。这一视角既能在 [11, 26], 等非欧几里得应用中发挥强大而启发性的作用，也能——这正是我们在此秉持的理念——帮助我们理解和证明该不等式，并将其用于简明的复杂度分析。KL 性质最初旨在分析 [32], 更一般的光滑 [29], 和非光滑 [12] 情形下，界定连续时间下降轨迹的长度。这个主题不在本文讨论范围之内，但对于一般度量空间而言至关重要[3]。

在这篇自成一体的论述中，我们将介绍 KL 性质的一些基础且广泛适用的复杂度推论，并给出简短、初等的证明。部分要素已见于零散文献，但通过重新编排证明步骤，并避免使用次梯度或其他传统的非光滑分析工具，我们对相关论述进行了精简与统一。首先，在引入 KL 不等式后，我们立即证明一个关于目标函数值收敛的初步复杂度结果，该结果在任意度量空间中均成立。这一新颖的观察结论适用范围之广、形式之简洁，令人惊叹，也彰显了我们的两项核心主张。

- 当下，人们对流形等非欧几里得情形下的优化算法日益关注 [2, 17, 43]，这促使我们开展一种无需次梯度的复杂度分析。
- 基于 KL 的复杂度分析若完全从度量角度展开，带来的不仅是普适性，还有简洁性与透明度。

受此启发，我们为欧几里得空间上半代数（以及更一般的“可定义”）函数的非光滑 Kurdyka–Lojasiewicz 性质给出了新的证明。通过聚焦于该性质的度量本质，而非次微分本质，我们简化并阐明了先前的研究成果，同时未实质性地牺牲其适用性。

回到复杂度结果的研究，我们考察满足基本斜率下降条件的迭代点序列，该条件结合了传统的充分下降概念与度量斜率性质。我们证明，在适当的温和条件下，这一条件在度量空间中可以得到满足，并推导出由此产生的迭代点的局部和全局收敛结果。作为本框架的示例，我们最后分析黎曼流形及更一般测地空间上的优化主化-最小化方法。

2 Kurdyka–Łojasiewicz性质

我们的研究设定在度量空间 X中。实际应用中， X 可以是欧几里得空间或希尔伯特空间，也可以是黎曼流形，但我们的论述完全基于度量。为在第4节证明 Kurdyka-Lojasiewicz 不等式，我们暂且不考虑半代数情形；直到第 7节之前，我们都不作任何类似有限维性的假设；而且直到最后的示例中，我们才会用到内积。因此，我们的方法完全是原始方法。

¯我们希望最小化一个下半连续函数 $f\colon X \to \mathrm{R} = [-\infty, +\infty]$，该函数是适当的，也就是说，它不会取值 $-\infty$，且其定义域非空

$$\operatorname{dom} f = \{x \in X : f(x) < +\infty\}.$$

实践中，f 通常在某种意义上具有正则性——例如光滑、凸，或是这类函数的某种复合函数——但本理论不需要任何此类假设。

将空间 X 中的度量记为 $d(\cdot,\cdot)$，将到任意点 $x \in X$ 的距离函数记为 $d_x(\cdot)$。按照 [20], ，函数 f 在任意取有限值 $f(x)$的点 x 处的斜率 $|\nabla f|(x)$，是所有满足以下条件的值 $\epsilon > 0$ 的下确界：x 是扰动函数 $f + \epsilon d_x$的局部极小点。在满足 $f(x) = +\infty$ 的点 x 处，斜率定义为 $+\infty$。这种记法的由来可由以下标准例子说明。

例 2.1（欧几里得空间中的子流形）

在欧几里得空间 R^n 中，考虑任意一个 C^1 光滑子流形 $\mathcal{M}$，赋予欧几里得度量，[1]并取任意可微函数 $f\colon \mathcal{M} \to \mathrm{R}$。在任意点$x \in \mathcal{M}$处，$f$ 的协变梯度属于 R^n 中在 x处与$\mathcal{M}$ 相切的子空间，其范数即为 f 的斜率。特别地，当 $\mathcal{M} = \mathrm{R}^n$时，有如下公式

$$|\nabla f|(x) = |\nabla f(x)|.$$

定义2.2 去奇异函数是一个连续函数 $\phi\colon \mathrm{R}_+ \to \mathrm{R}_+$，其在 $\phi(0) = 0$ 处取值，并且在 $(0, +\infty)$ 上具有连续且严格为正的导数。若去奇异函数 ϕ 的导数满足 $\phi'(\tau) = o(\frac{1}{\tau})$，当 $\tau \downarrow 0$ 时，则称其具有适度增长。

以下定义沿袭了 [14]的思路。其策略是利用去奇异函数对间隙进行缩放，从而诱导出一种“尖锐性”：

$$g(\cdot) = f(\cdot) - \inf f.$$

¯定义2.3 设 $f\colon X \to \mathrm{R}$ 是一个下有界的下半连续真函数。若存在一个去奇异函数 ϕ，使得所有具有较小且严格为正的间隙 $g(x)$ 的点 $x \in X$ 都满足 $|\nabla(\phi \circ g)|(x) \geq 1$，则称 f 具有 Kurdyka–Łojasiewicz性质（对于极小化）。

We 可以利用一个简单的链式法则重写基本不等式 [9, 引理 4.1] :

$$|\nabla(\phi \circ g)|(x) = \phi'\big(g(x)\big) \cdot |\nabla g|(x)$$

每当 $0 < g(x) < +\infty$ 时。证明只需简单应用经典中值定理。

比斜率这一概念更稳健的是极限斜率，它在任意点$\bar{x} \in \operatorname{dom} f$处的取值为

$$|\overline{\nabla} f|(\bar{x}) = \liminf_{\substack{x \to \bar{x} \\ f(x) \to f(\bar{x})}} |\nabla f|(x).$$

很容易验证，上述 KL 性质的定义可等价地用极限斜率替代斜率来表述，后续的链式法则也同样如此。当度量空间 X 为欧氏空间时，以下重要恒等式（例如 [27, 命题 8。5], ）总是将极限斜率与次微分联系起来∂f [40]:

$$\operatorname{dist}\big(0, \partial f(x)\big) \;=\; |\overline{\nabla} f|(x),$$

从而得到 KL 性质的一个等价的次微分形式。然而，我们强调，在本文的论述中并未使用次梯度。

当度量空间 X 是完备的，KL 性质等价于一种“误差界”：对于所有具有较小间隙 $g(x)$ 的点 x，以及所有满足 $0 < r < g(x)$ 的值 r，从 x 到水平集 $g^{-1}[0, r]$ 的距离不超过 $\phi\big(g(x)\big) - \phi(r)$。我们不会利用这一观察结果（参见定理 2.1][8,），但它为理解该不等式的强大作用提供了一些颇具说服力的直观解释。

[12,13] 中的结果保证：在欧几里得空间上，任何具有有界极小点集合的下半连续半代数（或更一般地说，“全局次解析”）函数，在最小化问题中都具有 KL 性质，并且某个相应的去奇异函数具有适度增长。事实上，存在一个形如凹函数 $\phi(\tau) \;=\; \kappa\tau^{1-\theta}$的此类去奇异函数，其中指数 θ属于区间 $[0, 1)$ 和 $\kappa > 0$。例如，这类去奇异函数就出现在 Lojasiewicz 最初关于实解析函数 KL 性质的证明 [32], 及其在 [29]中的光滑次解析扩展中。在收敛性证明中，我们通常会假设去奇异函数是凹函数：相关讨论见 [5]。

关于 [12,13], 中非光滑 KL 性质的原始论证相当复杂；此外，[27] （由 Drusvyatskiy 给出的）后续证明更贴近 [29], 所铺设的路径，两者都高度依赖次梯度几何。秉持本文的理念，并受到另一项近期证明的启发[41, 定理 2。9], 在第 4 节中，我们仅用一个简单的基于斜率的论证，便从 [29]中的光滑版本推导出非光滑结果。

3 目标函数值的收敛

接下来，我们将沿用一个源自[33], 、由 [1], 提出并在[6,15,16,38]中得到推广的主题。许多经典优化算法—— [15] 中所考察的一阶方法，例如梯度下降、近端点方法、近端梯度方法、变度量法和拟牛顿法、信赖域方法、交替投影法、高斯-赛德尔方案等，[16], 中所考察的主化-最小化过程，包括序列二次规划方案和移动球方法，以及一些 [38] 中的束方法——都会在 X 中生成具有某种下降性质的序列。具体而言，对于某个常数 $\delta > 0$，连续的

迭代点 x、 x_+ 始终满足

$$f(x) - f(x_+) \;\geq\; \delta\big(|\overline{\nabla} f|(x_+)\big)^2. \tag{3.1}$$

欧几里得空间或希尔伯特空间之外的算法也可能保证相同的条件：正如 [30], 所述，一个基本的例子是近端点方法 [10,28]。

定理 3.2（函数值收敛）设在度量空间 X上，考虑任意下半¯连续真函数 f: $X \to \mathbf{R}$，该函数下有界，并满足用于最小化的 KL 性质；再考虑任意序列 (x_k) 属于 X，且满足下降性质（*3.1*），初始值 $f(x_0)$ 接近 下确界 f。那么函数值 $f(x_k)$ 单调递减并收敛到 下确界 f。此外，若相应的去奇异函数适度增长，则间隙满足

$$f(x_k) - \inf f \;=\; o\Big(\frac{1}{k}\Big).$$

证明：考虑第 k 次迭代时的间隙 $\tau_k = g(x_k) \geq 0$ 。根据假设，我们有

$$\tau_{k-1} - \tau_k \;\geq\; \delta\big(|\overline{\nabla} f|(x_k)\big)^2,$$

因此 τ_k 单调不增。若 $\tau_k = 0$ 在某次迭代时为零 k，则无需证明，因此我们可以假设对于所有 k， τ_k 都很小且非零。

如果结果不成立，那么 τ_k 递减至某个极限 $\bar{\tau} > 0$。利用链式法则，KL 性质确保

$$\phi'(\tau_k) \cdot |\overline{\nabla} f|(x_k) \;\geq\; 1$$

对于所有 k，因此

$$\tau_{k-1} - \tau_k \;\geq\; \frac{\delta}{\big(\phi'(\tau_k)\big)^2}. \tag{3.3}$$

但根据 $k \to \infty$，左侧收敛到零，而右侧收敛到严格为正的极限 $\delta/\big(\phi'(\bar{\tau})\big)^2$，这就产生了矛盾。

现在假设去奇异函数 ϕ 具有适度增长。给定任意常数$\mu > 0$，由于 $0 < \tau_k \downarrow 0$，我们知道 $\tau_k < \mu$ 和 $\tau_k\phi'(\tau_k) < \sqrt{\mu\delta}$ 对所有充分大的 k都成立。特别地，存在一个整数 m，使得只要 $k \geq m$,

$$\tau_{k-1} - \tau_k \;>\; \frac{1}{\mu}\tau_k^2,$$

因此

$$\frac{1}{\tau_{k-1}} \;<\; \frac{1}{\tau_k} - \frac{1}{\mu + \tau_k} \;<\; \frac{1}{\tau_k} - \frac{1}{2\mu}.$$

通过归纳法可得

$$\frac{1}{\tau_k} \;\geq\; \frac{1}{\tau_m} + \frac{k-m}{2\mu}$$

so

$$k\tau_k \;\le\; \frac{2\mu\tau_m k}{2\mu+\tau_m(k-m)}.$$

由此可推得

$$\limsup_{k\to\infty} k\tau_k \;\le\; 2\mu.$$

由于常数 $\mu>0$ 是任意的，结论得证。 □

该 $o(\frac{1}{k})$ 速率令人联想到凸优化中近端点方法的表现 [25, 定理 3。1]。在采用标准去奇异函数$\phi(\tau)=\kappa\tau^{1-\theta}$且 $\kappa>0$并且 $0\le\theta<1$时，我们可以对这一结果稍作改进。关键不等式 (3.3) 可写为

$$\tau_{k-1}-\tau_k \;\ge\; \mu\tau_k^{2\theta}$$

对某个常数 $\mu>0$。通过初等归纳可得，间隙收敛至零的速率有四种不同情形（参见 [4]）：有限收敛（$\theta=0$）、超线性（$0<\theta<\frac{1}{2}$）、线性（$\theta=\frac{1}{2}$）；最后，当 $\frac{1}{2}<\theta<1$,

$$f(x_k)-\inf f \;=\; O\left(\left(\frac{1}{k}\right)^{\frac{1}{2\theta-1}}\right).$$

第 一种情形显然成立，而第二和第三种情形可由不等式推出

$$\frac{\tau_{k-1}}{\tau_k} \;\ge\; 1+\mu\tau_k^{2\theta-1}$$

对所有充分大的 k 均成立。在第二种情形下，右端随 k 无限增大；在第三种情形下，它等于 $1+\mu$。假设最后一种情形不成立。对于任意常数 $M\ge 1$，假设

$$\tau_k \;>\; M\left(\frac{1}{k}\right)^{\frac{1}{2\theta-1}}$$

对于某个 $k\ge 2$，然而将 k 替换为 $k-1$ 时，同一个不等式却不成立。我们得到

$$M\left(\frac{1}{k}\right)^{\frac{1}{2\theta-1}} + \mu M^{2\theta}\left(\frac{1}{k}\right)^{\frac{2\theta}{2\theta-1}} < \tau_k+\mu\tau_k^{2\theta} \;\le\; \tau_{k-1} \;\le\; M\left(\frac{1}{k-1}\right)^{\frac{1}{2\theta-1}}$$

然后，凸性意味着当 M 足够大时会产生矛盾：

$$1+\frac{\mu M^{2\theta-1}}{k} \;<\; \left(1-\frac{1}{k}\right)^{-\frac{1}{2\theta-1}} \;\le\; \left(1-\frac{2}{k}\right)+\frac{2}{k}\left(\frac{1}{2}\right)^{-\frac{1}{2\theta-1}} \;=\; 1+\frac{2(2^{\frac{1}{2\theta-1}}-1)}{k}.$$

归纳过程由此完成。

4 基于斜率的 KL 性质证明

从斜率的角度来看，非光滑半代数函数满足 Kurdyka–Łojasiewicz性质这一事实，是 Lojasiewicz 提出的原始光滑结果的一个简单推论 [32] ，Kurdyka 也有相关结果 [29]。像原始的非光滑论证那样诉诸次梯度[12, 13], ，反而会掩盖这一简洁性。在这里给出的简短证明中，我们也 [41], 完全聚焦于斜率。

事实上，我们的论述适用于比半代数几何广泛得多的情形，而是采用一种称为“在 o-极小结构中可定义性”的性质。不熟悉可定义集和函数的读者可以放心地将注意力集中在半代数这一特殊情形上。参考文献 [19] 对可定义性作了简明的介绍。

我们固定一个极小结构 $\mathcal{O}$，本文所考虑的各种函数和集合相对于该结构都是可定义的。我们从一个光滑结果开始（参见 [13, 定理 11]），推广了 [29,定理 1]。

定理 4.1 在欧几里得空间 $\mathbb{R}^n$ 中，考虑任意有界的 $\mathcal{C}^{(1)}$-光滑可定义子流形 $\mathcal{M}$，以及任意可微可定义函数 $f\colon \mathcal{M} \to \mathbb{R}$。则存在一个凹可定义去奇异函数 ϕ 和一个常数 $\rho > 0$，使得

$$|\nabla_{\mathcal{M}}(\phi \circ f)(x)| \geq 1 \quad whenever\ 0 < f(x) < \rho, \tag{4.2}$$

其中，$\nabla_{\mathcal{M}}$ 表示 $\mathcal{M}$上的协变梯度。

¯证明：考虑可定义函数 $\psi : (0, \infty) \to \mathbb{R}$，定义如下

$$\psi(t) = \inf\{|\nabla_{\mathcal{M}} f(x)| : f(x) = t\}.$$

（通常，空集上的下确界为 $+\infty$。）不妨假设 $\liminf_{t \searrow 0} \psi(t) = 0$，否则，对于所有充分小的 $\rho > 0$，由 $\phi(t) = \frac{1}{\rho} t$ for $t \geq 0$ 定义的去奇异函数具有所需性质。可定义函数的一个标准性质（[19, 定理 2.1]）随即推出 $\lim_{t \searrow 0} \psi(t) = 0$。

考虑可定义集

$$S = \{(x, t) \in \mathcal{M} \times (0, +\infty) : f(x) = t,\ |\nabla_{\mathcal{M}} f(x)| < \psi(t) + t^2\}.$$

由于流形 $\mathcal{M}$ 有界，通过一个简单的论证可以证明，存在一个点$\bar{x} \in \mathbb{R}^n$ 满足 $(\bar{x}, 0) \in \operatorname{cl} S$。进一步的标准性质（参见[19, 定理 2.1 和 3.2]）现在保证，点 $(\bar{x}, 0)$ 可以通过一条光滑曲线在S中到达。更准确地说，存在一条 $\mathcal{C}^{(1)}$-光滑曲线 $\gamma\colon (0, 1) \to \mathcal{M}$，使得复合函数 $g = f \circ \gamma$满足 $\big(\gamma(s)\ g(s)\big) \in S$，其中 $0 < s < 1$，并且当 $s \searrow 0$时，我们有 $\big(\gamma(s)\ g(s)\big) \to (\bar{x}, 0)$，且 $\big(\gamma'(s)\ g'(s)\big)$ 收敛于某个极限 $(v, r) \in \mathbb{R}^n \times \mathbb{R}_+$。此外，我们可以假设导数 g' 要么为常数，要么严格单调。

每当 $0 < s < 1$ 我们有

$$g'(s) \;=\; \langle \nabla_{\mathcal{M}} f\big(\gamma(s)\big), \gamma'(s)\rangle \;\leq\; |\nabla_{\mathcal{M}} f\big(\gamma(s)\big)| \cdot |\gamma'(s)|.$$

此外，我们观察到

$$\begin{aligned} r \;&=\; \lim_{s\searrow 0} \langle \nabla_{\mathcal{M}} f\big(\gamma(s)\big), \gamma'(s)\rangle \\ &\leq\; \lim_{s\searrow 0} \Big(|\gamma'(s)| \big(\psi\big(g(s)\big) + g(s)^2\big)\Big) \;=\; 0, \end{aligned}$$

因此，事实上 $r = 0$。因此，导数 g' 必须严格递增，故 g 也严格递增，并且对所有 $s \in (0,1)$ 都满足 $g(s) < sg'(s)$。任取区间 $\big(0$ 中的常数 ρ，$\lim_{s\nearrow 1} g(s)\big)$，并令 $\xi\colon (0,\rho] \to (0,1)$ 表示反函数 g^{-1} 的限制，该函数也严格递增。

现在考虑任意满足 $0 < f(x) < \rho$ 的点 $x \in \mathcal{M}$。令 $t = f(x)$ 且 $s = \xi(t)$，则当 ρ 足够小时，我们有

$$\begin{aligned} |\nabla_{\mathcal{M}}(\xi \circ f)(x)| \;&=\; \xi'\big(f(x)\big) \cdot |\nabla_{\mathcal{M}} f(x)| \;\geq\; \xi'(t)\cdot\psi(t) \;=\; \frac{\psi\big(g(s)\big)}{g'(s)} \\ &>\; \frac{|\nabla_{\mathcal{M}} f\big(\gamma(s)\big)| - g(s)^2}{g'(s)} \;>\; \frac{1}{|\gamma'(s)|} - sg(s) \;>\; \frac{1}{1+|v|}. \end{aligned}$$

注意，导数 $\xi' = \frac{1}{g'\circ\xi} > 0$ 严格递减，且 $\lim_{t\searrow 0} \xi(t) = 0$。因此，由 $\phi : \mathbf{R}_+ \to \mathbf{R}_+$ 定义的函数

$$\phi(t) \;=\; \begin{cases} 0 & (t=0) \\ (1+|v|)\cdot\xi(t) & (0<t<\rho) \\ (1+|v|)\cdot\xi(\rho) & (t\geq\rho) \end{cases}$$

具有所需的性质。 □

通过一个简单的分层论证，我们得到以下适用于非光滑可定义函数的 Kurdyka-Lojasiewicz 型结果。

定理 4.3 在欧几里得空间 $\mathbf{R}^n$ 中，任取有界可定义集 U 和可定义函数 $f\colon U \to \mathbf{R}$。则存在一个凹的可定义去奇异化函数 ϕ 和一个常数 $\rho > 0$，使得

$$|\nabla(\phi\circ f)|(x) \geq 1 \quad \textit{whenever } 0 < f(x) < \rho. \tag{4.4}$$

证明 一个标准的分层 [42] 将 U 划分为有限个有界的 $\mathcal{C}^{(1)}$-光滑可定义子流形 $\mathcal{M}_1, \ldots, \mathcal{M}_k$，它们属于 $\mathbf{R}^n$，且每个限制 $f|_{\mathcal{M}_i}$

都是可微可定义函数。对每个限制应用定理 4.1，可得到一个凹可定义去奇异函数 ϕ_i 和一个常数 $\rho > 0$，使得

$$|\nabla_{\mathcal{M}_i}(\phi_i \circ f|_{\mathcal{M}_i})(x)| \geq 1 \quad \text{whenever } x \in \mathcal{M}_i \text{ with } 0 < f(x) < \rho.$$

将 [19, 定理 2。1] 应用于函数族 $\phi_i - \phi_j$，其中 $i \neq j$，必要时通过缩小 ρ，我们可以推得存在一个特定指标 j 使得$\phi_j' \geq \phi_i'$ 在区间 $(0, \rho)$ 上对每个指标 i成立。取 $\phi = \phi_j$。现在考虑任意一点$x \in U$，满足 $0 < f(x) < \rho$。对于某个 i，我们有 $x \in \mathcal{M}_i$，于是

$$\begin{aligned} |\nabla(\phi \circ f)|(x) &\geq |\nabla(\phi_i \circ f)|(x) \geq |\nabla(\phi_i \circ f|_{\mathcal{M}_i})|(x) \\ &= |\nabla_{\mathcal{M}_i}(\phi_i \circ f|_{\mathcal{M}_i})(x)| \geq 1 \end{aligned}$$

其中第一个不等式由不等式 $\phi' \geq \phi_i'$ 和链式法则推出，第二个不等式则由斜率的定义推出。□

注意，当 U 和 f 为次解析时，根据 Puiseux 展开，我们知道去奇异函数可以采用 $\phi(\tau) = \kappa\tau^{1-\theta}$ 的形式，其中 $\theta \in [0, 1$ 为某个指数，$\kappa > 0$ 为常数（见 [29]）。

我们在此给出的证明比[13], 中依赖次梯度几何的原始推导更加简洁。不过，这种简洁是有代价的。正如我们前面所指出的，KL 不等式（4.4）也可以等价地用极限斜率表示，因此可写成如下性质：

函数 $\phi \circ f$在点 x处的次梯度范数均至少为 1。

相比之下，[13] 证明了一个更强的结果，即同一性质也适用于*Clarke*次梯度。特别是，当函数 f 为 Lipschitz 函数时，这保证上述性质对次梯度的所有凸组合均成立。依赖较弱的、基于极限斜率的 KL 性质，可能会缩小收敛分析的适用范围：迭代点3.1处的基本下降条件（x）要求在下一个迭代点 x_+ 处存在一个范数不大于$\frac{1}{\sqrt{\delta}}\sqrt{f(x) - f(x_+)}$ 的次梯度；而采用更强的 KL 性质则更为灵活，因为它允许使用 Clarke 次梯度。另一方面，主化-最小化算法等具体方法确实会强制满足条件（3.1），因此这种区别并无影响。对于本文给出的收敛结果，我们尚未发现更强的 KL 性质在 [13]中会带来任何实际影响：本文证明的较简单的基于斜率的版本似乎已经足够。

5 收敛到极小点

正如我们所见，函数值收敛可由基本的下降条件（3.1）推出。实践中，这一条件通常源自以下两个斜率下降条件：

存在常数 $\alpha > 0$ 和 $\beta > 0$，使得相邻的迭代点 x、x_+ 始终满足

$$f(x) - f(x_+) \geq \alpha d^2(x, x_+), \tag{5.1}$$

$$d(x, x_+) \geq \beta |\overline{\nabla} f|(x_+). \tag{5.2}$$

我们首先注意到，这些条件源自 [33], 的开创性分析，并且始终可以满足。现在我们假设底层度量空间 X是完备的。

命题 5.3（斜率下降序列的存在性）在完备度量空间 X中，考虑任意一个下半连续函数 $f\colon\ X \to \mathbf{R}$，该函数下有界。则对于任意点 $x \in \operatorname{dom} f$以及满足 α, $\beta > 0$的常数$\alpha\beta < \frac{1}{2}$，存在一个点 $x_+ \in X$ 满足斜率下降条件（*5.1*）和（*5.2*）。

证明 事实上，我们证明一个更强的性质，即用斜率代替极限斜率。函数 $h = f + \alpha d_x^2$ 也下半连续且下有界。我们声称，在所有点 $y \in X$ 处，其斜率满足

$$|\nabla h|(y) \geq |\nabla f|(y) - 2\alpha d(x, y).$$

为证明这一点，我们可以只考虑一个满足 $|\nabla h|(y)$ 有限的点 $y \in \operatorname{dom} h$。考虑任意常数 $\epsilon' > \epsilon > |\nabla h|(y)$。由于 y 是函数$h + \epsilon d_y$的局部极小点，我们可以推得，所有充分接近 y 的点 $z \in X$ 都满足

$$\begin{aligned} f(y) &\leq f(z) + \alpha\big(d(x, z)^2 - d(x, y)^2\big) + \epsilon d(y, z) \\ &\leq f(z) + \alpha\big(2d(x, y)d(y, z) + d(y, z)^2\big) + \epsilon d(y, z) \\ &\leq f(z) + \big(2\alpha d(x, y) + \epsilon'\big)d(y, z), \end{aligned}$$

因此， y 也局部极小化函数

$$f + \big(2\alpha d(x, y) + \epsilon'\big)d_y$$

并由此可得 $|\nabla f|(y) \leq 2\alpha d(x, y) + \epsilon'$。令 $\epsilon' \downarrow |\nabla h|(y)$，即可证明该结论。

若 x 使函数 h取得极小值，则 $|\nabla h|(x) = 0$，从而 $|\nabla f|(x) = 0$，因此我们可以选取 $x_+ = x$。否则，定义 $t = \frac{1}{2}(h(x) + \inf h)$，并注意到从 x 到水平集 $\{y \in X : h(y) \leq t\}$ 的距离 δ 严格为正。Ekeland 原理 [24] 蕴含存在一个点 $x_+ \in X$，同时满足不等式 $h(x_+) \leq t$和 $|\nabla h|(x_+) \leq (\frac{1}{\beta} - 2\alpha)\delta$。我们可得 $d(x, x_+) \geq \delta$，并且

$$|\nabla f|(x_+) \leq |\nabla h|(x_+) + 2\alpha d(x, x_+) \leq \frac{1}{\beta} d(x, x_+),$$

s不等式（5.2）得证。不等式（5.1由 h（x_+）$\leq t < h$（x）得证。 □

上述论证尤其表明，任何近端点 x_+ （即使函数 $f+\alpha d_x^2$取最小值的点）都必须满足斜率下降条件（5.1）和（5.2），假设 $\alpha\beta \le \frac{1}{2}$。下半连续凸函数 f 的近端点在希尔伯特空间 X 中由 Minty 定理[35], 保证存在（且唯一）；在更一般的度量空间 [10, 28]中也有类似结论。 X 然而，一般而言，除非度量空间是 proper（即所有闭球都是紧的），否则近端点可能不存在。

以下结果及其证明概括了 [6] 中的论证。我们假设去奇异函数是凹的，并利用目标函数的下半连续性，但不依赖其他连续性性质。

定理 5.4（迭代点收敛）在完备度量空间 X上，考虑任意下半连续函数 f: $X \to \mathbf{R}$ ，它下有界，且在最小化问题中满足 KL 性质，并具有凹去奇异函数 ϕ; 再考虑任意序列(x_k) ，其取值于 X 并满足斜率下降条件(5.1) 和(5.2) ，且初始值 $f(x_0)$接近inf f。那么， x_k收敛到 f的一个极小点。

证明 我们可以假设没有任何迭代点使 f 取得极小值，因为否则，第一个不等式将保证此后序列保持不变。目标函数值$f(x_k)$单调递减，因此均为有限值；KL 不等式则保证间隙 g 满足

$$1 \;\le\; |\overline{\nabla}(\phi\circ g)|(x_k) \;=\; \phi'\big(g(x_k)\big)\cdot \overline{|\nabla g|}(x_k).$$

我们推得，每个极限斜率 $|\overline{\nabla} g|(x_k) \;=\; |\overline{\nabla} f|(x_k)$ 都非零，因此对于所有 $k>0$，距离 $d_k = d(x_{k-1}, x_k)$ 都非零。

利用去奇异函数 ϕ 的凹性，我们观察到，对于所有 $k>0$,

$$\begin{aligned}\phi\big(g(x_k)\big)-\phi\big(g(x_{k+1})\big) &\ge \phi'\big(g(x_k)\big)\big(g(x_k)-g(x_{k+1})\big)\\ &\ge \frac{f(x_k)-f(x_{k+1})}{|\overline{\nabla} f|(x_k)} \;\ge\; \frac{\alpha d_{k+1}^2}{\beta d_k},\end{aligned}$$

so

$$\frac{\beta}{\alpha}\Big(\phi\big(g(x_k)\big)-\phi\big(g(x_{k+1})\big)\Big) \;\ge\; \frac{2d_kd_{k+1}-d_k^2}{d_k} \;\ge\; 2d_{k+1}-d_k.$$

就这些数而言

$$\lambda_k \;=\; \frac{\beta}{\alpha}\phi\big(g(x_k)\big)+d_k \;>\; 0 \qquad \text{for } k>0,$$

我们已经证明

$$d(x_k, x_{k+1}) \;\le\; \lambda_k - \lambda_{k+1}.$$

由此可知，序列（x_k）是柯西序列，因此收敛于某个极限 x^*。但是，定理 3.2表明 f（x_k）递减至下确界 f，因此下半连续性现在确保 x^* 使 f达到最小值。 □

一般而言，我们对误差 $d(x_k, x^*)$ 趋于零的速度一无所知，即使在近端点方法这一典型情形下也是如此。然而，借助对去奇异函数 ϕ的一些了解，证明可以得到更多结论。根据 [4, 6], 我们有

$$d(x_k, x^*) \;\le\; \sum_{j=k}^{m-1} d(x_j, x_{j+1}) \;+\; d(x_m, x^*) \;\le\; \lambda_k - \lambda_m + d(x_m, x^*).$$

令 $m \to \infty$ 可见】【。

$$d(x_k, x^*) \;\le\; \lambda_k \;=\; \frac{\beta}{\alpha}\phi\big(g(x_k)\big) + d_k \;\le\; \frac{\beta}{\alpha}\phi\big(g(x_k)\big) + \sqrt{\frac{g(x_{k-1})}{\alpha}},$$

并由此得到复杂度估计。按照前一节的讨论，对于标准去奇异函数 $\phi(\tau) = \kappa\tau^{1-\theta}$ for $\kappa > 0$and $0 \le \theta < 1$，根据收敛速率，我们得到四种不同的情形：有限（$\theta = 0$）、超线性（$0 < \theta < \frac{1}{2}$）、线性（$\theta = \frac{1}{2}$），以及在 $\frac{1}{2} < \theta < 1$ 情形下的情况

$$d(x_k, x^*) \;=\; O\left(\left(\frac{1}{k}\right)^{\frac{1-\theta}{2\theta-1}}\right).$$

6 收敛到临界点

先前的收敛结果依赖于目标函数f 的初始值接近其下确界。在没有这一假设的情况下，若对 f 不作任何全局凸性类假设，我们则转而寻求一个临界点 $x^* \in X$，即极限斜率 $|\overline{\nabla} f|(x^*)$ 为零。例如，对于流形上的光滑函数，这一思路与经典概念一致。显然，任何局部极小点都是临界点。

在本节中，我们给出保证全局收敛性质的条件：任何满足斜率下降条件（5.1）和（5.2）的序列都必须收敛到一个临界点。该方法基本遵循 [6], 尤其首先依赖于以下性质。

¯定义 6.1 设 f 是一个下半连续真函数：$X \to \mathbb{R}$。若对于所有取有限值的点 $\bar{x} \in X$，都存在一个凹去奇异函数 ϕ 和一个常数 $\rho > 0$，使得

$$(6.2) \qquad d(x, \bar{x}) < \rho \quad \text{and} \quad f(x) - \bar{f} \in (0, \rho) \quad \Rightarrow \quad |\nabla\big(\phi \circ (f - \bar{f})\big)|(x) \;\ge\; 1,$$

¯其中 $\bar{f} = f(\bar{x})$。

关于这个定义，有几点说明。首先，在非临界点 $\bar{x}$ 处，所要求的性质会自动成立，这很容易验证，因此这一概念实际上

<escape>涉及临界点。其次，我们注意到，所有下半连续的半代数（或全局次解析）函数都是KL函数 [12, 13, 41]。最后，与之前一样，我们可以在定义中用极限斜率等价地替换斜率。</escape>

由第 4 节，我们推导出以下结果。

推论6.3 所有下半连续半代数（或更一般地，次解析）函数都是KL函数。

任何 KL 函数的一个重要特征是 f 可以在任意紧集上选定一个 统一的去奇异函数 $X \subset X$，在该集合上 f 为常数。更准确地说，存在一个去奇异函数 ϕ 和一个常数 $\rho > 0$，使得 KL 性质（6.2）对所有点 $\bar{x} \in X$都成立（就斜率和极限斜率而言）。证明是一个常规的紧致性论证：利用 $\phi_{\bar{x}}$ 以及常数 $\rho_{\bar{x}}$，对于每个点 $\bar{x}$，相应的邻域覆盖了 X，因此我们可以选出一个对应于有限个点 $\{x_i\}$的子覆盖，然后令常数 $\rho =$ 为最小值$_i\, \rho_{x_i}$，并令去奇异函数 $\phi = \sum_i \phi_{x_i}$。由这一统一性性质可知，任何具有紧的极小点集合的 KL 函数都具有最小化的 Kurdyka–Łojasiewicz 性质，定义见 2.3。

断言实践中的所有优化目标都是 KL函数，也并不算太夸张。我们假设的第二个条件并非那么普遍，但仍很常见（参见 [3, 定理 2.3。1] 和 [22]）。

定义 6.4 若真函数 $f\colon X \to \mathrm{R}$ 满足：对于所有点 $\bar{x} \in \mathrm{dom}\ f$ 以及序列 $x_k \to \bar{x}$，只要函数值 $f(x_k)$ 和斜率$|\nabla f|(x_k)$ 一致有界，就有 $f(x_k) \to f(\bar{x})$，则称其在斜率有界集上连续。

显然，若 f 连续，或者更一般地，在其定义域上连续，则这一性质成立；但欧几里得空间上的下半连续函数若为凸函数，或者更一般地，若“次微分连续” [40]，这一性质也成立。

定理 6.5（全局收敛）在完备度量空间 X上，考虑任意 KL函数 $f\colon X \to \mathbf{R}$，它下有界且在斜率有界集上连续。那么，任何满足斜率下降条件(x_k) 且属于 X 的相对紧序列 (5.1) 和 (5.2) 必收敛到 f的一个临界点。

证明 该值 $\mu =$下确界$_k\, f(x_k)$ 是有限的。注意 $f(x_k) - f(x_{k+1}) \to 0$，因为这些值 $f(x_k$ 单调递减至 μ。斜率下降条件 (5.1) 和 (5.2) 现在意味着

$$d(x_k, x_{k+1}) \to 0 \qquad \text{and} \qquad |\overline{\nabla} f|(x_k) \to 0. \tag{6.6}$$

此外，如果 $f(x_k) = \mu$曾经成立过，那么 $|\overline{\nabla} f|(x_k) = 0$且对所有 $j \geq k$，$x_j = x_k$，因此无需证明任何内容。于是，我们可以假设对于所有 k，$f(x_k) > \mu$。

¯序列 (x_k) 的极限点集合 X 是闭集，并且包含于¯该序列的闭包中，而根据假设，该闭包是紧集。因此， X 非空且紧。一个初等论证表明，$\operatorname{dist}_{\bar{X}}(x_k) \to 0$。

考虑任意一个子序列 $K \subset \mathbb{N}$，使得序列 $(x_k)_{k\in K}$ 收敛。记其极限为 $\bar{x}$。由于 $k \to \infty$ 属于 K，我们知道 $f(x_k) \to \mu$ 且 $|\overline{\nabla} f|(x_k) \to 0$，因此根据极限斜率的定义，存在相应的点 x'_k，对于 $k \in K$，使得 $x'_k \to \bar{x}$， $f(x'_k) \to \mu$，并且当 $k \to \infty$ 属于 K 时， $|\nabla f|(x'_k) \to 0$。因此 $\bar{x}$ 是临界点，而斜率有界集上的连续性意味着 $f(\bar{x}) = \mu$。因此 f 在¯集合 X 上为常数；此外，该集合完全由临界点组成。

根据定义 6.1之后关于一致性的讨论，由于集合 X 是紧的，存在一个凹去奇算函数 ϕ 和一个常数 $\rho > 0$使得¯ ¯性质（6.2）对所有点 $f = \mu$ 均成立，其中 $\bar{x} \in X$。对于所有充分大的 k，我们知道

$$\operatorname{dist}_{\bar{X}}(x_k) < \rho \qquad \text{and} \qquad f(x_k) - \mu \in (0, \rho),$$

so

$$|\overline{\nabla}(\phi \circ (f - \mu))|(x_k) \;\geq\; 1.$$

完全按照定理5.4证明中的相同论证，我们可以推得¯整个序列（x_k）收敛，并且由于极限属于 X，因此该极限是临界点。 □

7 主化-最小化方法

为了阐释本文的理念，我们提出一种完全基于度量的方法，用于对主化-最小化（MM）算法进行收敛分析。这些算法在概念上是简单的迭代技术，用于在集合 X 上最小化目标函数 $f\colon\ X \to \mathbb{R}$。在每个点 $x \in X$，我们将一个包含 x 的模型子集 $D(x)$ 与 X 相关联，并定义一个主化模型函数 h_x：也就是说， $h_x \geq f$，且在点 x 处取等号。最简单的例子是 $D(x)$ 恒等于 X。随后，我们考虑相应的包络函数 $F\colon\ X \to \mathbb{R}$，以及集合值迭代映射 $p\colon\ X \rightrightarrows X$，它在每个点 $x \in X$ 处定义为

$$F(x) = \inf_{D(x)} h_x \qquad \text{and} \qquad p(x) \;=\; \operatorname*{argmin}_{D(x)} h_x$$

以及MM 序列，其中相继的迭代点 x、 x_+ 始终满足 $x_+ \in p(x)$。主化-最小化算法在实践中已得到广泛应用，最初见于统计学文献[21]。

一个基本的初始问题是，MM 序列 (x_k) 是否是目标函数 f 的极小化序列；这里我们仅指 $f(x_k) \to \inf f$。我们首先整理一些初步观察。

命题 7.1 设在集合 X 上针对目标函数 f 使用 MM 方法。那么 f 是相应包络函数 F 的上界，且 $\inf f = \inf F$。MM 序列是 f 的极小化序列，当且仅当它是 F 的极小化序列。此外，迭代映射 p 满足

$$p(\operatorname{argmin} F) \subset \operatorname{argmin} f \subset \operatorname{argmin} F,$$

若 p 处处取值非空，且模型对所有不同的点 $x \in X$ 和 $y \in D(x)$ 均满足 $h_x(y) > f(y)$，则等号成立。

证明 所有 $x \in X$ 均满足 $F(x) \le h_x(x) = f(x)$，因此可得出主化结论。我们还有

$$\inf_{y\in X} F(y) \le \inf_{y\in X} f(y) \le \inf_{y\in D(x)} h_x(y) = \inf_x F(x),$$

因此， f 和 F 具有相同的下确界，且 $\operatorname{argmin} f \subset \operatorname{argmin} F$。为推导关于 MM 序列 (x_k) 的结论，注意到 $f(x_k) \ge F(x_k) \ge f(x_{k+1})$，因为 $F(x_k) = h_{x_k}(x_{k+1})$。最后，考虑 F 的任意极小点 x。如果存在一个点 $y \in p(x)$，那么我们有

$$\inf F = F(x) = h_x(y) \ge f(y) \ge \inf f = \inf F.$$

因此，等号处处成立，故 y 是 f 的极小化解。如果我们还具有额外的严格支配条件，那么上述关系意味着 $y = x$。□

如果迭代映射可能取空值，最后一条结论可能不成立。例如，目标函数 $f(x) = e^x$ for $x \in X = \mathbb{R}$，模型集合 $D(x) = \mathbb{R}$ for all x，以及模型函数 $h_x(y) = (1 + |x - y|)e^y$ for $y \in \mathbb{R}$。包络函数 F 恒为零，而迭代映射 p 始终取空值。

上述命题启发了我们后续的分析，这种方法的灵感来自[16]，但再次避开了次梯度或法锥等变分分析工具。我们不直接将第3 节中的收敛结果应用于目标函数 f，而是借助命题7.1 转而研究包络函数 F。

我们回到这样的情形： X 是一个度量空间，并考虑双变量函数 $h\colon X^2 \to \mathbb{R}$，这些函数是局部利普希茨的；等价地说，它们关于每个变量分别都是局部利普希茨的。我们的分析的新意在于，对关于第二个变量的利普希茨性质作如下强化；这显然是一个新概念。

定义 7.2 给定度量空间 X，近似器是一个局部利普希茨函数 $h\colon X^2 \to \mathbb{R}$，使得对每个点 $\bar{x} \in X$，都存在一个常数 $\gamma \ge 0$，使所有靠近 $\bar{x}$ 的点 x、y、$z \in X$ 都满足该不等式

$$|h(x,y) - h(z,y)| \le \gamma d(x,z)\big(d(x,y) + d(z,y)\big). \tag{7.3}$$

下面给出一些简单的例子。

例 7.4（平方距离）由 $h(x,y) = d^2(x,y)$ 定义的函数，对于$x, y \in X$，是一个近似器，因为它是局部利普希茨的，且满足

$$\begin{aligned} |h(x,y) - h(z,y)| &= |d^2(x,y) - d^2(z,y)| = |d(x,y) - d(z,y)|(d(x,y) + d(z,y)| \\ &\le d(x,z)\big((d(x,y) + d(z,y)\big) \end{aligned}$$

对所有 $z \in X$。

例 7.5（阈值）对于任意半径 $\rho > 0$，由以下表达式定义的函数

$$h(x,y) = \max\{d(x,y) - \rho, 0\} = \operatorname{dist}_{B_\rho(x)}(y) \qquad (x, y \in X)$$

是一个近似器。它是 Lipschitz 函数的最大值，因此也是 Lipschitz 的。此外，对于任意点 $z \in X$，若 $d(x,y) \ge \rho$ 且 $d(z,y) \ge \rho$，则

$$h(x,y) - h(z,y) = d(x,y) - d(z,y) \le d(x,z) \le \frac{1}{2\rho} d(x,z)\big(d(x,y) + d(z,y)\big),$$

而若 $d(x,y) \ge \rho \ge d(z,y)$，则

$$\begin{aligned} h(x,y) - h(z,y) &= d(x,y) - \rho \le d(x,y) - d(z,y) \le d(x,z) \\ &\le \frac{1}{\rho} d(x,z)\big(d(x,y) + d(z,y)\big). \end{aligned}$$

由此可得

$$h(x,y) - h(z,y) \le \frac{1}{\rho} d(x,z)\big(d(x,y) + d(z,y)\big),$$

该情形 $d(x,y) \le \rho$ 是平凡的。所需性质现可由对称性得出。

下面的例子尤为有用。

例 7.6（线性近似）在欧几里得空间 $X = \mathbf{R}^n$ 中，考虑一个函数 $g\colon \mathbf{R}^n \to \mathbf{R}$，其梯度 ∇g 是 L-Lipschitz 的。由线性近似定义的函数为

$$h(x,y) = g(x) + \langle \nabla g(x), y - x \rangle \qquad (x, y \in X)$$

是一个近似器。

下面的观察很容易验证。

命题 7.7 近似器的和与数乘也是逼近子。

我们的论述依赖于对底层可行域 X 的一些强结构假设。特别是假设作为度量空间， X 是适当的，也就是说，所有闭球都是紧的。因此， X 是完备且局部紧的。

命题 7.8（一致逼近器）考虑一个正度量空间 X和一个函数 h: $X^2 \to \mathbf{R}$。那么 h 是一个近似器，当且仅当对于所有有界子集 $X' \subset X$，函数 h 在 $(X')^2$ 上是 Lipschitz 的，并且存在一个常数$\gamma > 0$ ，使得逼近器不等式（7.3）对所有点 x， y， $z \in X'$均一致成立。

证明 给定的性质显然意味着 h是一个近似器，因此我们重点讨论其逆命题。假设 h 是一个近似器，那么有界集合上的一致 Lipschitz 性质是一个标准的紧致性论证，而下面对不等式（7.3）的证明与之类似。如果该性质不成立，则存在有界集合 X' 中的序列（x_k）、（y_k）和（z_k），以及常数 $\gamma_k \to +\infty$ ，使得对于所有$k = 1$，2，3，.....，不等式

$$|h(x_k, y_k) - h(z_k, y_k)| \;>\; \gamma_k d(x_k, z_k)\big(d(x_k, y_k) + d(z_k, y_k)\big). \tag{7.9}$$

由于 X 是适当的，我们可以取子序列，并假设 $x_k \to \bar{x}$、 $y_k \to \bar{y}$ 和 $z_k \to \bar{z}$。记 λ 为 h 在 $X' \times X'$ 上的一个 Lipschitz 常数，我们可以推得

$$\lambda \;>\; \gamma_k\big(d(x_k, y_k) + d(z_k, y_k)\big),$$

因此 $d(x_k, y_k)$ 和 $d(z_k, y_k)$ 都收敛于零。因此我们有 $\bar{x} = \bar{y} = \bar{z}$，所以不等式（7.9）与近似器的定义相矛盾。□

引理 7.10（距离函数作为近似器）考虑一个真度量空间 X 和一个集值映射 D: $X \rightrightarrows X$ ，使得对每个点 $x \in X$，像 $D(x)$ 都是闭的，并且包含 x。那么定义的函数为

$$h(x, y) \;=\; \operatorname{dist}_{D(x)}(y)$$

当且仅当 h 是近似器，并且对于每个点 $\bar{x} \in X$，存在一个常数 $\gamma \geq 0$ ，使得所有接近 $\bar{x}$ 且满足 $y \in D(z)$ 的点 x、 y、 $z \in X$ 都满足该不等式

$$\operatorname{dist}_{D(x)}(y) \;\leq\; \gamma d(x, z)\big(d(x, y) + d(z, y)\big).$$

证明 直接推出显而易见，因此我们着重讨论逆命题。对于任意点 $x \in X$，函数 $H(x, \cdot)$ 作为距离函数，是 1-Lipschitz。转而

考虑第二个变量和逼近器不等式，首先注意，对于所有点x、 y、 $z \in X$，任取 $u \in \mathrm{Proj}^{D(z)}(y)$ 和 $w \in \mathrm{Proj}^{D(x)}(u)$，我们有

$$\mathrm{dist}_{D(x)}(y) - \mathrm{dist}_{D(z)}(y) \;=\; \mathrm{dist}_{D(x)}(y) - d(u,y) \;\le\; d(w,y) - d(u,y) \;\le\; d(w,u)$$

so

$$\mathrm{dist}_{D(x)}(y) - \mathrm{dist}_{D(z)}(y) \;\le\; \mathrm{dist}_{D(x)}(u). \tag{7.11}$$

由于 $z \in D(z)$，我们推得 $d(u,y) \le d(z,y)$。因此，当 x、 y、 z都接近 $\bar{x}$时， u也接近 $\bar{x}$，而我们的假设意味着

$$\begin{aligned}
\mathrm{dist}_{D(x)}(y) - \mathrm{dist}_{D(z)}(y) \;&\le\; \mathrm{dist}_{D(x)}(u) \;\le\; \gamma d(x,z)\big(d(x,u) + d(z,u)\big) \\
&\le\; \gamma d(x,z)\big(d(x,y) + d(u,y) + d(z,y) + d(u,y)\big) \\
&\le\; \gamma d(x,z)\big(d(x,y) + 3d(z,y)\big) \\
&\le\; 3\gamma d(x,z)\big(d(x,y) + d(z,y)\big).
\end{aligned}$$

Th逼近器性质 (7.3)由对称性得出，其中常数为 3γ . □

例 7.5 恰好展示了这种近似器。

注 7.12 引理 h 中定义的函数的 Lipschitz 连续性 7.10等同于映射 的伪利普希茨连续性 D: 见 [39]。

在接下来的论述中，除了假设可行域X 是一个正度量空间外，我们还假设它是某个可能¯ ¯更大的度量空间 X的闭子集，并赋有继承度量。引入 X 纯粹是为了运用一些凸性论证。我们使用（单位速度）测地线的概念，它是从 R 的一个紧区间到 X的等距映射；其端点和中点分别是该区间端点和中点的像。在¯的讨论中，我们常将测地线与其像视为一体。如果每一对点都是某条测地线的端点，我们就称 X 为测地空间。若子集 D 包含所有端点位于 D中的测地线，则称其为测地凸。在这种情况下，函数g: $D \to$ R 是测地凸的，当且仅当它与 D 中每条测地线的复合函数都是凸函数；若它与 D 中每条测地线的复合函数都是 μ-强凸的，则称其为 μ-强测地凸（对于 $\mu > 0$）。

¯ ¯当对于所有点 $x \in X$ 而言，函数 $\frac{1}{2}d^2(x,\cdot)$ 都是测地 1-强凸函数时，测地空间 X 称为 CAT(0)空间。这一性质对于¯近端点迭代 [10, 28]至关重要。特别是，它意味着 X 中的球都是测地凸的。根据霍普夫–里诺定理 [36, 定理 7.2]以及 [18, 命题 I.3.7]的适当版本，完备、局部紧的测地空间都是适当空间。同时属于 CAT(0)空间的此类空间包括所有完备、单连通且具有非正截面曲率的有限维黎曼流形。

为方便起见，我们在下文汇总各项假设。

假设 7.13 可行域 X 是一个测地空间的真度量子空间，目标函数 $f\colon X \to \mathbb{R}$ 下有界，且优化问题 $\inf_X f$ 具备以下主化-最小化工具。

- 一个近似映射 $D\colon X \rightrightarrows X$，它为每个可行点$x \in X$ 指定一个包含 x 的闭且测地凸子集 $D(x)$。
- 一个常数 $\mu \geq 0$ 和一个建模函数 $h\colon X^2 \to \mathbb{R}$，为每个$x \in X$ 指定一个测地 μ-强凸函数 $h_x\colon D(x) \to \mathbb{R}$，满足

$$h(x,y) = h_x(y) \geq f(y) \quad \textit{for } y \in D(x), \quad \textit{with equality if } y = x.$$

此外，h 以及由下式定义的函数

$$(x,y) \mapsto \operatorname{dist}_{D(x)}(y) \qquad (x,y \in X)$$

都是定义 7.2意义下的逼近子。

在继续之前，我们先作一些观察。假设中隐含着（作为逼近子性质的推论），可行域 X 是闭的，目标函数 f 是局部利普希茨的。不过，我们强调，两者都不必是测地凸的。此外，该假设还保证近似映射 D 的图是闭的；事实上，它还是连续的，因为函数 $(x,y) \mapsto \operatorname{dist}_{D(x)}(y)$ 是连续的。我们还看到，每个近似集合 $D(x)$ 都是 X 的闭二阶近似，而每个模型 h_x 都是 f 的连续二阶近似：对于靠近 x 的点 $y \in X$，有

$$\operatorname{dist}_{D(x)}(y) = O\big(d^2(x,y)\big) \qquad \text{and} \qquad (h_x - f)(y) \;=\; O\big(d^2(x,y)\big).$$

例 7.14（近端点方法）考虑一个适当的 CAT(0)空间 X，以及一个下有界的测地凸、局部利普希茨函数 $f\colon X \to \mathbb{R}$。设对所有点 $x \in X$都有 $D(x) = X$，并且对于任意常数 $\mu > 0$，定义相应的二次惩罚建模函数

$$h(x,y) \;=\; f(y) + \frac{\mu}{2} d^2(x,y). \tag{7.15}$$

则假设 7.13成立。相应的包络函数是 f 的 Moreau-Yosida 近似，迭代映射是近端点迭代的度量版本 [10], ，而 MM 序列就是近端点序列。

在任意适当的 CAT(0)空间 X中，对于任意半径 $\rho > 0$，可定义一个非平凡的近似映射 $D(x) = B_\rho(x)$，如例 7.5所示。

MM 序列就是与迭代映射p相关联的不动点迭代。我们首先研究不动点性质 $x \in p(x)$。显然，目标函数 f 的每个极小点都是 p的不动点。此外，我们还有以下逆命题。

命题 7.16（不动点与零斜率）若假设*7.13* 成立，则在迭代映射的任意不动点处，目标函数的斜率为零。

证明 考虑任意不动点 x_*。如果目标函数 f 的斜率严格为正，那么存在一个常数 $\sigma > 0$ 和一个点列 $x_k \neq x$，该点列在 X 中收敛到 x，且所有点都满足

$$\frac{f(x) - f(x_k)}{d(x, x_k)} > \sigma.$$

根据我们的假设，存在点 $y_k \in D(x)$ 满足 $d(y_k, x_k) = o\big(d(x_k, x)\big)$。不动点性质保证 $f(x) \leq h_x(y_k)$，因此对于所有足够大的 k，我们有

$$\frac{h_x(x_k) - f(x_k)}{d(x, x_k)} = \frac{h_x(y_k) - f(x_k)}{d(x, x_k)} > \sigma,$$

利用模型 h_x 的利普希茨性质。但这与我们的假设所推出的 $h_x(x_k) - f(x_k) = o\big(d(x_k, x)\big)$ 矛盾。□

命题 7.17 在正则测地空间上，如果一个测地强凸函数是连续且下有界的，那么它具有唯一的极小点。

证明 将给定的函数记为 g，将强凸性常数记为 μ。固定任意点 $\bar{x} \in X$。对于任意满足 $g(x) \leq g(\bar{x})$ 的点 $x \in X$，考虑一条端点为 x和 $\bar{x}$的测地线。中点 x' 满足

$$\inf g \leq g(x') \leq \frac{1}{2}g(x) + \frac{1}{2}g(\bar{x}) - \frac{\mu}{8}d^2(x, \bar{x}) \leq g(\bar{x}) - \frac{\mu}{8}d^2(x, \bar{x}),$$

因此 x位于由 $d^2(x, \bar{x}) \leq \frac{8}{\mu}(g(\bar{x}) - \inf g)$ 定义的闭球中。根据连续性，g在这个紧集上有一个极小点 $\hat{x}$，而 $\hat{x}$必定是 g的全局极小点。前述不等式表明该极小点唯一。□

如果空间是 CAT(0)，则这一结果在更一般的情况下也成立 [34, 引理 1。3] .

命题 7.18 假设 *7.13* 成立。那么，迭代映射是单值且连续的。此外，对于任意有界的相关 MM 序列，存在一个常数 $\beta \geq 0$，使得包络函数 F 在所有连续的三个迭代点 x_-、x、x_+处满足下降条件

$$F(x_-) - F(x) \geq \frac{\mu}{2}d^2(x, x_+), \tag{7.19}$$

$$|\nabla F|(x) \leq \beta d(x, x_+). \tag{7.20}$$

证明 前一个结果表明，迭代映射 p 是单值的。接下来我们证明二次增长条件

$$h_x(y) - h_x\big(p(x)\big) \;\geq\; \frac{\mu}{2} d^2\big(y, p(x)\big) \qquad \text{whenever } y \in D(x). \tag{7.21}$$

为此，对于测地凸集 $D(x)$ 中的任意点 $y \neq p(x)$，考虑距离 $\delta = d(p(x),\, y) > 0$ 以及一条测地线 $\gamma\colon [0,\delta] \to D(x)$，满足 $\gamma(0) = p(x)$ 且 $\gamma(\delta) = y$。根据 μ-强凸性，对于 $\tau \in [0,\delta]$，我们有 $\gamma(t) \in D(x)$，并且

$$h_x\big(\gamma(\tau)\big) \;-\; \frac{\mu}{2}\tau^2 \;\leq\; \Big(1 - \frac{\tau}{\delta}\Big) h_x\big(p(x)\big) \;+\; \frac{\tau}{\delta}\Big(h_x(y) - \frac{\mu}{2}\delta^2\Big).$$

由于 $p(x)$ 在 $D(x)$ 上使 h_x 取得最小值，对于充分小的 $\tau > 0$，我们可推出

$$0 \;\leq\; h_x\big(\gamma(\tau)\big) \;-\; h_x\big(p(x)\big) \;\leq\; \frac{\tau}{\delta}\Big(h_x(y) - h_x\big(p(x)\big) - \frac{\mu}{2}\delta^2\Big) \;+\; \frac{\mu}{2}\tau^2.$$

两　除以 τ 并令 $\tau \downarrow 0$ 趋于极限，即可证明增长条件（7.21）。

考虑序列（x_k）中的任意元素，该序列在 X 中收敛到 $\bar{x}$。由于 $p(\bar{x}) \in D(\bar{x})$，由 D 的连续性可知，存在点 $y_k \in D(x_k)$，满足 $y_k \to p(\bar{x})$。增长条件（7.21）推出，对所有 k，不等式

$$h_{x_k}(y_k) - h_{x_k}\big(p(x_k)\big) \;\geq\; \frac{\mu}{2} d^2\big(y_k, p(x_k)\big).$$

对两边取上极限可得 $d\big(y_k, p(x_k)\big) \to 0$，因此 $p(x_k) \to p(\bar{x})$。由此可知 p 是连续的。

在任意点 $x \in X$，不等式 (7.21) 表明

$$\begin{aligned} F(x) - F\big(p(x)\big) \;&=\; h_x\big(p(x)\big) - h_{p(x)}\Big(p\big(p(x)\big)\Big) \\ &\geq\; h_{p(x)}\big(p(x)\big) - h_{p(x)}\Big(p\big(p(x)\big)\Big) \;\geq\; \frac{\mu}{2} d^2\Big(p(x), p\big(p(x)\big)\Big). \end{aligned}$$

这证明了下降条件 (7.19)。

为证明最后的下降条件 (7.20)，我们可以假设 $|\nabla F|(x) > 0$，因为否则无需证明，因此存在一个点列 $x_k \to x$ 属于X 且满足

$$|\nabla F|(x) \;=\; \lim_k \frac{F(x) - F(x_k)}{d(x, x_k)} \;=\; \lim_k \frac{h\big(x, p(x)\big) - h\big(x_k, p(x_k)\big)}{d(x, x_k)}.$$

考虑序列

$$\alpha_k \;=\; d\big(x, p(x_k)\big) + d\big(x_k, p(x_k)\big).$$

由于 $p(x_k) \in D(x_k)$，且函数 $(x, y) \mapsto \operatorname{dist}^{D(x)}(y)$ 是一个近似器，因此我们知道，对于某个常数 $\gamma \geq 0$，不等式

$$\operatorname{dist}_{D(x)}\big(p(x_k)\big) \;\leq\; \gamma\alpha_k d(x, x_k)$$

对所有 k，因此存在一个有界点列 $y_k \in D(x)$ 满足

$$d\big(y_k, p(x_k)\big) \;\leq\; \gamma\alpha_k d(x, x_k).$$

如有必要，增大 γ，我们还可以确保对所有 k，以下不等式成立

$$\begin{aligned} h\big(x, p(x_k)\big) - h\big(x_k, p(x_k)\big) &\leq \gamma\alpha_k d(x, x_k) \\ h(x, y_k) - h\big(x, p(x_k)\big) &\leq \gamma d\big(y_k, p(x_k)\big), \end{aligned}$$

由于 h 也是一个近似器，利用我们的假设以及 p 的连续性，可推出

$$\begin{aligned} |\nabla F|(x) &\leq \liminf_k \frac{h\big(x, y_k\big) - h\big(x_k, p(x_k)\big)}{d(x, x_k)} \\ &\leq \liminf_k \frac{h\big(x, p(x_k)\big) + \gamma d\big(y_k, p(x_k)\big) - h\big(x_k, p(x_k)\big)}{d(x, x_k)} \\ &\leq \gamma(\gamma+1) \liminf_k \alpha_k \;=\; 2\gamma(\gamma+1) d\big(x, p(x)\big). \end{aligned}$$

证明至此完成。 □

遵循 [23], 所倡导的通用框架，这一结果为 MM 算法的自然“小步长”停止准则提供了依据，因为如果两个连续的迭代点 x, x_+ 足够接近，那么包络函数在 x 处的斜率必定很小。

定理 7.22（MM 序列收敛）假设 *7.13* 成立。若相应的包络函数限制在闭球上时，对于最小化始终满足 KL 性质，且具有凹去奇异函数，那么任何有界MM序列 (x_k) 初始值为 $f(x_0)$ 且接近 $\inf f$ 的序列，都会收敛到 f 的一个极小点。

证明 证明采用了与以下内容相似的技巧 [16, 38]，模仿定理 5.4 的证明。根据前面的结果，连续的迭代点三元组 $\ldots x_-, x, x_+ \ldots$ 满足

$$F(x_-) - F(x) \;\geq\; \frac{\mu}{2} d^2(x, x_+) \qquad \text{and} \qquad d(x, x_+) \;\geq\; \beta|\nabla F|(x). \tag{7.23}$$

我们可以假设没有任何迭代点能够使 F 达到最小值，否则，第一个不等式将保证此后该序列保持不变。定义间隙函数$g = F - \inf F$。包络函数值 $F(x_k)$ 单调递减，而 KL 不等式保证间隙 g 满足

$$1 \;\leq\; |\nabla(\phi \circ g)|(x_k) \;=\; \phi'\big(g(x_k)\big) \cdot |\nabla g|(x_k).$$

我们推导出，每个斜率 $|\nabla g|(x_k) = |\nabla F|(x_k)$ 都不为零，因此对于所有 $k > 0$，距离$d_k = d(x_{k-1}, x_k)$ 都不为零。

利用去奇异函数 ϕ 的凹性，我们可以观察到，对所有 $k > 0$,

$$\phi\big(g(x_{k-1})\big) - \phi\big(g(x_k)\big) \geq \phi'\big(g(x_{k-1})\big)\big(g(x_{k-1}) - g(x_k)\big) \geq \frac{F(x_{k-1}) - F(x_k)}{|\nabla F|(x_{k-1})} \geq \frac{\mu d_{k+1}^2}{2\beta d_k},$$

so

$$\frac{2\beta}{\mu}\Big(\phi\big(g(x_{k-1})\big) - \phi\big(g(x_k)\big)\Big) \geq \frac{2d_k d_{k+1} - d_k^2}{d_k} \geq 2d_{k+1} - d_k.$$

我们推得正数 $\lambda_k = \frac{2\beta}{\mu}\phi\big(g(x_{k-1})\big) + d_k$ 满足关键不等式 $d(x_k, x_{k+1}) \leq \lambda_k - \lambda_{k+1}$，因此序列 (x_k) 是柯西序列，收敛到某个极限 x^*。定理 3.2 表明 (x_k) 是 F 的极小化序列，因此也是 f 的极小化序列。于是，由连续性可知 x^*是 f 的极小点，证毕。 □

注 7.24 这一结果意味着，在广泛的一般情形下，半代数优化中的 MM 序列会收敛。更确切地说，假设可行域是欧几里得空间的一个半代数子集，并且近似映射和建模函数也都是半代数的。在这种情况下，包络函数F 也是半代数的，因此满足该条件。

定理 7.25（MM 序列的全局收敛）假设 *7.13* 成立。如果相应的包络函数是一个具有凹去奇异函数的 KL 函数（特别是，当可行域、近似映射和建模函数都为半代数时，便满足这一条件），那么任何有界的 MM 序列都会收敛到目标函数斜率为零的点。

证明 我们沿用定理 6.5的证明，但改用不等式（7.23），如前一个证明中所示。如果目标函数 f 具有有界水平集，那么任何 MM 序列都必定有界，因为相应的目标函数值单调递减。该证明表明，由迭代 $x \leftarrow p(x)$ 生成的序列收敛到某个极限 x^*。由于迭代映射 p 连续， x^* 必定是不动点，由此我们根据命题 7.16 推得 $|\nabla f|(x^*) = 0$ 。□

请注意，如果目标函数具有有界水平集，MM 序列总是有界的。最后，我们列举两个来自文献的示例。

凸复合优化 [31]

给定欧几里得空间 X和 W、一个 $\mathcal{C}^{(2)}$-光滑映射 $G\colon X \to W$、一个凸函数$q\colon W \to \mathbb{R}$，以及任意常数 $\mu > 0$，我们考虑复合目标函数 $f\colon X \to \mathbb{R}$,

一个函数 $H\colon X^2 \to W$，用于描述对 G的线性近似；以及“近端线性化”函数 $h\colon X^2 \to \mathbb{R}$，对于 X中的点 x, y，定义为

$$\begin{aligned} f(x) &= q\big(G(x)\big), \\ H(x,y) &= G(x) + DG(x)[y-x], \\ h(x,y) &= q\big(H(x,y)\big) + \frac{\mu}{2}|x-y|^2, \end{aligned}$$

其中，线性映射 $DG(x)\colon X \to W$ 是 G 在 x处的导数。通过定义平凡的近似映射 $D(x) = X$ 对所有点 $x \in X$均成立，我们证明在合理条件下，假设 7.13 成立。

我们假设 f 下有界。我们还假设 q 是 L-Lipschitz，且 G 的二阶导数，即双线性映射 $D^2G(x)\colon X^2 \to W$，其范数对所有 x 都有一个常数上界 M。（为简化表述，这些都是全局假设；更精细的推导允许常数依赖于给定的有界定义域。）对于任意点 $x \in X$ 和单位向量 $w \in W$，考虑由下式定义的函数 $e\colon X \to \mathbb{R}$】【。

$$e(y) = \langle w, H(x,y) - G(y)\rangle \qquad (y \in X).$$

注意，对于所有向量 $u \in X$，我们都有

$$De(y)[u] = \langle w, \big(DG(x) - DG(y)\big)[u]\rangle.$$

根据中值定理，存在一点 $z \in [x,y]$，使得

$$e(y) = e(x) + De(z)[y-x] = \langle w, \big(DG(x) - DG(z)\big)[y-x]\rangle,$$

并且还有一个点 $u \in [x,z]$，使得

$$\big(DG(x) - DG(z)\big)[y-x] = D^2G(u)[y-x, x-z].$$

因此

$$\big|G(x) + DG(x)[y-x] - G(y)\big| \le M|y-x|\,|x-z|,$$

so

$$\Big|q\big(G(y)\big) - q\big(G(x) + DG(x)[y-x]\big)\Big| \le LM|x-y|^2.$$

给定 $\mu > 2LM$，我们得出主化性质。

转而考虑其余假设，固定一个点 $y \in X$。我们首先证明所有点 x、$z \in X$ 都满足

$$|H(x,y) - H(z,y)| \le M|x-z|\big(|x-y| + |z-y|\big). \tag{7.26}$$

任取单位向量 $w \in W$，并定义函数 $c\colon X \to \mathbb{R}$，使得 $c(x) = \langle w, H(x,y)\rangle$。于是，根据中值定理，存在一点 $u \in [x,z]$，使得

$$c(x) - c(z) = Dc(u)[x-z] = \langle w, D^2G(u)[y-u, x-z]\rangle.$$

由于 w 是任意的，我们的结论现在成立，因为 $|y-u| \le |x-y| + |z-y|$。利用不等式（7.26），我们推得

$$\begin{aligned} |h(x,y) - h(z,y)| &\le LM|x-z|\big(|x-y| + |z-y|\big) + \frac{\mu}{2}\big||x-y|^2 - |z-y|^2\big| \\ &= LM|x-z|\big(|x-y| + |z-y|\big) + \mu|\langle x-z, x+z-2y\rangle| \\ &\le \Big(LM + \frac{\mu}{2}\Big)|x-z|\big(|x-y| + |z-y|\big). \end{aligned}$$

假设7.13成立。

另外，假设函数 q and G 是半代数的。那么，对于所有充分大的常数 $\mu > 0$，由近端线性迭代生成的任何有界序列

$$x \leftarrow \operatorname*{argmin}_y \Big\{ q\big(G(x) + DG(x)[y-x]\big) + \frac{\mu}{2}|x-y|^2 \Big\}$$

收敛到一个使复合函数 $q \circ G$ 的斜率为零的点。

移动球方法 [7, 16]

对于具有 Lipschitz 梯度的可微函数 f、f_1、f_2、……、f_m: $\mathrm{R}^n \to \mathrm{R}$，我们考虑在可行域上最小化目标函数 f 的问题。

$$X = \{x \in \mathbf{R}^n : f_i(x) \le 0 \text{ for each } i\}.$$

我们假设这个经典的非线性规划下有界，并且满足 Mangasarian-Fromovitz约束资格条件（MFCQ）：换言之，对于所有点 $x \in X$，存在一个向量 $d \in \mathrm{R}^n$ 满足

$$\langle \nabla f_i(x), d\rangle < 0 \qquad \text{for all } i \in I(x) = \{i\colon f_i(x) = 0\}.$$

为了构造一种能够保持可行性的优化方法，我们的主化-最小化工具由一个近似映射 D 和一个建模函数 h 构成，定义如下:

$$\begin{aligned} h(x,y) &= f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2, \\ h_i(x,y) &= f_i(x) + \langle \nabla f_i(x), y-x\rangle + \frac{L_i}{2}\|y-x\|^2 \quad \text{for each } i, \\ D(x) &= \bigcap_i D_i(x) \quad \text{with} \quad D_i(x) = \{y \in \mathbf{R}^n : h_i(x,y) \le 0\} \quad \text{for each } i, \end{aligned}$$

其中，L、L_1、……、L_m 分别是 ∇f、∇f_1、……、∇f_m 的 Lipschitz 模量。对于每个点 $x \in X$，集合 $D_i(x)$ 是一个以点 $x - \frac{1}{L_i}\nabla f_i(x)$ 为中心、半径为

$$\rho_i(x) = \sqrt{\frac{1}{L_i^2}\|\nabla f_i(x)\|^2 - \frac{2}{L_i} f_i(x)}.$$

因此得名。我们的目标是证明假设 7.13 成立，从而确保该方法在合理条件下收敛。

只需验证函数 h 以及 $(x,y) \mapsto \operatorname{dist}^{D(x)}(y)$（对于点$x, y \in \mathbf{R}^n$）都是近似器，因为其他条件很容易由 L-光滑函数 $g\colon \mathbf{R}^n \to \mathbf{R}$ 的标准二次上界推出：

$$g(y) \;\le\; g(x) + \langle \nabla g(x), y - x\rangle + \frac{L}{2}\|y-x\|^2.$$

事实是 h 是一个近似器，这一点可由例 7.4 和 7.6以及命题 7.7得出。事实上，直接验证以下不等式很容易看出，所有点 x、y、$z \in \mathbf{R}^n$ 都满足不等式

$$|h(x,y) - h(z,y)| \;\le\; \frac{3L}{2}|x-z|\big(|x-y| + |z-y|\big) \tag{7.27}$$

以及函数 h_i 所对应的不等式。

为了证明函数（x, y） $\mapsto \operatorname{dist}^{D(x)}(y)$ 也是一个近似器，我们需要以下初等事实 [37]：对于包含球 S_1、S_2, ..., $S_m \subset \mathbf{R}^n$的闭凸集合，每个集合都包含球 $B_\delta(0)$，所有点 $x \in \mathbf{R}^n$ 都满足

$$\operatorname{dist}_{\cap_i S_i}(x) \;\le\; \frac{\|x\|}{\delta}\max_i \big\{\operatorname{dist}_{S_i}(x)\big\}. \tag{7.28}$$

固定一点 $\bar{x} \in X$。根据 MFCQ 条件，存在一个单位向量 $d \in \mathbf{R}^n$，满足

$$\langle \nabla f_i(\bar{x}), d\rangle < 0 \quad \text{for all } i \in I(\bar{x}).$$

对于任意充分小的常数 $\epsilon \in (0, \min_i 2L_i)$，所有点 $x \in B_\epsilon(\bar{x})$ 均满足不等式

$$f_i(x) \le -\epsilon \quad \text{for all } i \notin I(\bar{x}) \qquad \text{and} \qquad \nabla f_i(x)^\top d \le -\epsilon \quad \text{for all } i \in I(\bar{x}). \tag{7.29}$$

并且，对于充分大的 $M \ge \max_i L_i$，不等式 $\|\nabla f_i(x)\| \le M$ 对每个$i = 1$、2、.....、m 均成立。

定义常数 $t = \min\{1, \frac{\epsilon}{4M}\}$ 和 $\delta = \min\{1, \frac{\epsilon}{2M}\}t$。对于任意点 $x \in B_\epsilon(\bar{x}) \cap X$ 和任意向量 $e \in B_\delta(0)$，考虑以下量

$$h_i(x, x+td+e) \;=\; f_i(x) + \langle \nabla f_i(x), td+e\rangle + \frac{L_i}{2}\|td+e\|^2.$$

对于所有 $i \notin I(\bar{x})$，左端均不大于 $-\epsilon + 4Mt$；对于所有 $i \in I(\bar{x})$，左端均不大于 $-\frac{\epsilon}{2}t + 2Mt^2$，因此在这两种情况下，左端均为非正数。由此我们推得

$$B_\delta(x+td) \;\subset\; \bigcap_i D_i(x) \;=\; D(x).$$

Pro性质（7.28）因此意味着所有点 $y \in \mathbf{R}^n$ 都满足不等式

$$\operatorname{dist}_{D(x)}(y) \;\le\; \frac{\|y-x-td\|}{\delta}\max_i\big\{\operatorname{dist}_{D_i(x)}(y)\big\}. \tag{7.30}$$

考虑点 $x \in B_\epsilon(\bar{x}) \cap X$, $z \in X$ 和 $y \in D(z)$。若 $y \notin D_i(x)$，则

$$\begin{aligned}
\operatorname{dist}_{D_i(x)}(y) &= \big\|y-x+\frac{1}{L_i}\nabla f_i(x)\big\| \;-\; \rho_i(x)\\
&= \frac{2}{L_i}\cdot\frac{h_i(x,y)}{\big\|y-x+\frac{1}{L_i}\nabla f_i(x)\big\| \;+\; \rho_i(x)}\\
&\le \frac{2}{L_i}\cdot\frac{h_i(x,y)-h_i(z,y)}{\big\|y-x+\frac{1}{L_i}\nabla f_i(x)\big\| \;+\; \rho_i(x)}\\
&\le \frac{h_i(x,y)-h_i(z,y)}{\sqrt{\|\nabla f_i(x)\|^2-2L_i f_i(x)}} \;\le\; \frac{1}{\epsilon}\big(h_i(x,y)-h_i(z,y)\big)\\
&\le \frac{3L_i}{2\epsilon}|x-z|\big(|x-y|+|z-y|\big),
\end{aligned}$$

由不等式 (7.29) 得出倒数第二个不等式。性质 (7.30) 现在意味着

$$\operatorname{dist}_{D(x)}(y) \;\le\; \frac{3M\|y-x-td\|}{2\epsilon\delta}|x-z|\big(|x-y|+|z-y|\big).$$

由于上述不等式对所有 $y \in D(z)$ 都成立，因此结合 (7.11) 可得函数 $(x,y) \mapsto \operatorname{dist}^{D(x)}(y)$ 局部具有利普希茨连续性。因此，根据引理 7.10，函数 $(x,y) \mapsto \operatorname{dist}_{D(x)}(y)$ 是一个近似器，从而完成了对假设 7.13 的验证。

参考文献

[1] P.-A. Absil、R. Mahony 和 B. Andrews。解析代价函数下降法迭代序列的收敛性。SIAM 优化期刊，6(2):531–547，2005 年。[2] P.-A. Absil、R. Mahony 和 R. Sepulchre。矩阵流形上的优化算法。普林斯顿大学出版社，新泽西州普林斯顿，2008 年。[3] L. Ambrosio、N. Gigli 和 G. Savar´e。度量空间与概率测度空间中的梯度流。Birkh¨auser-Verlag，巴塞尔，第 2 版，2008 年。[4] H. Attouch 和 J. Bolte。关于具有解析特征的非光滑函数的近端算法的收敛性。数学规划，116:5–16，2009 年。

[5] H. Attouch、J. Bolte、P. Redont 和 A. Soubeyran。用于非凸问题的近端交替最小化和投影方法：一种基于 Kurdyka-Lojasiewicz 不等式的方法。《运筹学数学》，35：438–457，2010年。[6] H. Attouch、J. Bolte 和 B.F. Svaiter。半代数问题和温和问题的下降方法的收敛性：近端算法、前向-后向分裂和正则化高斯-赛德尔方法。数学规划，137(1-2，A辑)：91–129，2013年。[7] A. Auslender、R. Shefi 和 M. Teboulle。一种用于一类光滑约束最小化问题的移动球近似方法。《SIAM优化杂志》，20(6)：3232–3259，2010年10月。[8] D. Az´e 和 J.-N. Corvellec。度量空间上下半连续函数的误差界刻画。ESAIM：COCV，10：409–425，2004年。[9] D. Az´e 和 J.-N. Corvellec。通过函数变换得到的非线性误差界。《优化理论与应用杂志》，172：9–32，2017年。[10] M. Ba˘c´ak。度量空间中的近端点算法。《以色列数学杂志》，194：689–701，2013年。[11] A. Blanchet 和 J. Bolte。一个函数不等式族：Lojasiewicz不等式和位移凸函数。《泛函分析杂志》，275：1650–1673，2018年。[12] J. Bolte、A. Daniilidis 和 A.S. Lewis。非光滑次解析函数的 Lojasiewicz 不等式及其在次梯度动力系统中的应用。SIAM 优化期刊，17：1205–1223，2006年。[13] J. Bolte、A. Daniilidis、A.S. Lewis 和 M. Shiota。可分层函数的 Clarke 次梯度。SIAM 优化期刊，18(2)：556–572，2007年。[14] J. Bolte、A. Daniilidis、O. Ley 和 L. Mazet。Lojasiewicz不等式的刻画：次梯度流、谷底曲线、凸性。《美国数学会汇刊》，362(6)：3319–3363，2010年。[15] J. Bolte、T.P. Nguyen、J. Peypouquet 和 B. Suter。从凸函数的误差界到一阶下降方法的复杂度。数学规划，165：471–507，2017年。[16] J. Bolte 和 E. Pauwels。半代数和温和规划中的主化-最小化过程及序列二次规划方法的收敛性。《运筹学数学》，41：442–465，2016年。

[17] N. Boumal。《光滑流形上的优化导论》。剑桥大学出版社，剑桥，2022。[18]M.R. Bridson 和 A. Haefliger。《非正曲率度量空间》。施普林格出版社，柏林，1999。[19] M. Coste。《o-极小几何导论》。RAAG 笔记，81页，雷恩数学研究所，1999年11月。[20] E. De Giorgi、A. Marino 和 M. Tosques。度量空间中的演化问题与最大下降曲线。意大利林琴国家科学院自然科学、物理、数学与自然科学类汇刊（8），68：180–187，1980。[21] A.P. Dempster、N.M. Laird 和 D.B. Rubin。通过 EM 算法从不完整数据中进行最大似然估计。英国皇家统计学会杂志B辑，39：1–38，1977。[22] D. Drusvyatskiy、A.D. Ioffe 和 A.S. Lewis。下降曲线。SIAM 控制与优化杂志，53(1)：114–138，2015。[23] D. Drusvyatskiy、A.D. Ioffe 和 A.S. Lewis。使用类泰勒模型进行非光滑优化：误差界、收敛与终止准则。数学规划，185：357–383，2021。[24] I. Ekeland。论变分原理。数学分析与应用杂志，47：324–353，1974。[25] O. G¨uler。凸函数极小化中近端点算法的收敛性。SIAM 控制与优化杂志，29：403–419，1991。[26] D. Hauer 和 J.M. Maz´on。度量空间中梯度流的 Kurdyka–Łojasiewicz–Simon 不等式。美国数学会汇刊，372：4917–4976，2019。[27] A.D Ioffe。《正则映射的变分分析》。施普林格国际出版社，2017。[28] J. Jost。凸泛函与映入非正曲率空间的广义调和映射。海尔维提数学评论，70：659–673，1995。[29] K. Kurdyka。o-极小结构中可定义函数的梯度。傅里叶研究所年刊（格勒诺布尔），48(3)：769–783，1998。[30] A.S. Lewis 和 Tonghua Tian。度量空间中的可识别性、KL 性质与次梯度曲线。 arXiv:2205.02868，2022。

[31] A.S. Lewis 和 S.J. Wright。复合最小化的近端方法。数学规划，页码1–46，2015。[32] S. Lojasiewicz。半解析集合。预印本，IHES，1965。[33] Z.-Q. Luo 和 P. Tseng。可行下降方法的误差界与收敛分析：一种通用方法。运筹学年刊，46/47(1-4)：157–178，1993。优化问题中的退化。[34] U.F. Mayer。非正曲率度量空间上的梯度流与调和映射。分析与几何通讯，6：199–253，1998。[35] G.J. Minty。希尔伯特空间中的单调（非线性）算子。杜克数学杂志，29：341–346，1962。[36] S.B. Myers。度量空间中的弧与测地线。美国数学会汇刊，57：217–227，1945。[37] A. Nedi´c。用于凸集交问题的随机投影算法。载于第49届 IEEE 决策与控制会议（CDC），页码7655–7660，2010。[38] D. Noll。利用 Kurdyka–Łojasiewicz 不等式实现非光滑下降方法的收敛。《优化理论与应用杂志》，160：553–572，2014。[39] R.T. Rockafellar。多值函数的 Lipschitz 性质。非线性分析，9：867–885，1985。[40] R.T. Rockafellar 和 R.J-B. Wets。《变分分析》。数学科学基础丛书，第317卷，施普林格出版社，柏林，1998。[41] A Valette。奇异点处的 Lojasiewicz 不等式。美国数学会会刊，147：1109–1117，2019。[42] L. van den Dries。驯服拓扑与 o-极小结构，伦敦数学学会讲义系列第248卷。剑桥大学出版社，剑桥，1998。[43] Hongyi Zhang 和 S. Sra。测地凸优化的一阶方法。机器学习研究杂志，49：1–22，2016。