arXiv:2301.06632v1 [math.OC] 16 Jan 2023

非光滑随机逼近中的渐近正态性与最优性

Damek Davis* Dmitriy Drusvyatskiy† Liwei Jiang‡

摘要

在他们的开创性工作中，Polyak 和 Juditsky 证明，用于求解光滑方程的随机逼近算法满足中心极限定理。此外，后续研究指出，从 H´ajek 和 Le Cam 的局部极小极大意义来看，该方法的渐近协方差在所有估计程序中是最优的。这一研究方向长期悬而未决的问题是，对于随机非线性规划或随机变分不等式等重要的非光滑问题，是否也能得到类似的保证。在这项工作中，我们证明确实如此。

1 引言

Polyak 和 Juditsky [30] 著名地证明，对于用于最小化光滑强凸函数的随机梯度法，中心极限定理成立：迭代点运行平均与极小值点之差除以迭代计数器的平方根后，收敛到一个正态随机向量。此外，极限分布的协方差矩阵在严格意义上是所有估计程序中“最优”的。一个长期悬而未决的问题是，非光滑优化乃至更一般的均衡问题是否也具有类似的保证——渐近正态性和最优性。在这项工作中，我们在温和条件下得到了此类保证；这些条件既适用于具体情形（例如非线性规划），也适用于一般的线性扰动。

我们将考虑的问题最适合用随机变分不等式来建模。为此，考虑求解包含关系 $x^\star$ 的任务

$$0 \in \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[A(x,z)] + N_{\mathcal{X}}(x). \tag{1.1}$$

这里，$\mathcal{P}$ 是只能通过抽样获得的概率分布，$A(\cdot, z)$ 对几乎处处的 $z \sim \mathcal{P}$ 而言是一个光滑映射，而 $N_{\mathcal{X}}(x)$ 表示闭集 $\mathcal{X}$ 的法锥。随机

*美国纽约州伊萨卡市14850，康奈尔大学运筹学与信息工程学院；people.orie.cornell.edu/dsd95/。Davis 的研究得到 Alfred P. Sloan 研究奖学金和美国国家科学基金会 DMS 项目资助（编号 2047637）的支持。

†华盛顿大学数学系，美国华盛顿州西雅图市98195；www.math.washington.edu/~ddrusv。Drusvyatskiy 的研究得到美国国家科学基金会 DMS-1651851 和 CCF-2023166 项目资助的支持。

‡康奈尔大学运筹学与信息工程学院。美国纽约州伊萨卡市14850, USA; orie.cornell.edu/research/grad-students/liwei-jiang

变分不等式 (1.1) 在当代优化中无处不在。例如，约束优化问题的最优性条件

$$\min_{x} \quad \mathop{\mathbb{E}}_{z\sim\mathcal{P}} f(x,z) \qquad \text{subject to } x \in \mathcal{X},$$

通过在 (1.1) 中令 $A(x,\ z) = \nabla f(x,\ z)$，可将其纳入框架 (1.1)。更一般地，随机博弈的纳什均衡 $x^\star = (x_1^\star,\ \ldots\ldots,\ x_m^\star)$ 是以下系统的解

$$x_j^\star \in \mathop{\operatorname{argmin}}_{x_j\in\mathcal{X}_j} \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[f_j(x,z)] \qquad \text{for all } j = 1,\ldots,m,$$

其中，f_j 和 $\mathcal{X}_j$ 分别是参与者 j 的损失函数和策略集。这些 k 耦合包含关系的一阶最优性条件可通过令 $[A(x,\ z)]_j := \nabla_{x_j} f_j(x,\ z)$ 且 $\mathcal{X} := \mathcal{X}_1,\ \ldots\ldots,\ \mathcal{X}_m$，建模为 (1.1)。

求解 (1.1) 有两种标准方法：样本均值近似（SAA）和随机前向-后向算法（SFB）。前一种方法通过抽取一批样本 $z_1,\ z_2,\ \ldots\ldots,\ z_k$ 独立同分布$\sim \mathcal{P}$，并求解经验近似的一个解 x_k

$$0 \in \frac{1}{k}\sum_{i=1}^{k}[A(x,z_i)] + N_{\mathcal{X}}(x). \tag{1.2}$$

相比之下，随机前向-后向（SFB）算法以在线方式运行，在每次迭代 k 中抽取一个样本 $z_k \sim \mathcal{P}$，并将下一个迭代点 x_{k+1} 定义为

$$x_{k+1} \in P_{\mathcal{X}}(x_k - \alpha_k \cdot A(x_k, z_k)). \tag{1.3}$$

这里，$P_{\mathcal{X}}(\cdot)$表示到 $\mathcal{X}$上的最近点投影。在约束优化的情形下，$A(x,\ z) = \nabla f(x,\ z)$是某个损失函数 $f(x,\ z)$的梯度，此时过程 (1.3) 简化为随机投影梯度算法。与样本平均近似（SAA）相比，SFB 这类在线算法通常更可取，因为每次迭代的开销较小，而且可以在线执行；而 SAA 则需要求解辅助优化问题 (1.2)。尽管样本均值近似估计量的渐近分布目前已得到充分理解 $[16, 18, 33],$，但对于非光滑和受约束情形下 FSB 迭代点的渐近表现，我们的认识仍然有限。本文旨在填补这一空白。我们工作的主要结果如下。

> 在合理的假设下，随机前向-后向迭代的运行平均与样本平均近似具有相同的渐近分布。此外，从 H´ajek 和 Le Cam $[19,35]$ 的局部极小极大意义上看，样本平均近似和随机前向-后向迭代都是渐近最优的。

接下来，我们将详细介绍我们的结果及其推论。具体而言，经典理论已知（例如 $[16,18,33]$），样本平均近似 (1.2) 的渐近表现很大程度上受到解 $x^\star$ 对 (1.1) 左端扰动的敏感性的影响。为了单独考察这一影响，令 $S(v)$ 表示扰动系统的解 x

$$v \in \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[A(x,z)] + N_{\mathcal{X}}(x).$$

在全文中，我们假设解 $S(v)$ 在 $x^\star$ 附近平滑变化。更确切地说，我们假设 S 的图在 $(0, x^\star)$ 附近局部与某个光滑映射 $\sigma(\cdot)$ 的图重合。用变分分析 [11], 的术语来说，映射 $\sigma(\cdot)$ 称为 S 在 $(0, x^\star)$ 附近的光滑局部化。已知在各种具体情形下，以及对半代数问题进行一般线性扰动时，这一假设均成立 [12]。

接下来，我们将介绍相关背景并陈述研究结果。众所周知，根据 [18,33]，在温和的假设下，SAA (1.2) 的解 x_k 渐近服从正态分布：

$$\sqrt{k}(x_k - x^\star) \xrightarrow{D} \mathsf{N}\big(0, \nabla\sigma(0) \cdot \mathrm{Cov}(A(x^\star, z)) \cdot \nabla\sigma(0)^\top\big). \tag{1.4}$$

因此，解映射的雅可比矩阵 $\nabla\sigma(0)$ 会出现在样本均值近似估计量的渐近协方差中。事实上，我们将论证这是不可避免的。本工作的首要贡献是证明，SAA 的渐近表现相对于所有估计程序而言都是局部极小极大最优的——这里的含义是 H´ajek 和 Le Cam [19, 35]所说的局部极小极大。粗略地说，这意味着，对于任何基于 k 个样本输出 $\hat{x}_k$ 的估计程序，都存在一列扰动 $\mathcal{P}_k$，满足 $\frac{d\mathcal{P}_k}{d\mathcal{P}} = 1 + O(k^{-1/2})$，使得 $\hat{x}_k$ 在这列扰动后的问题上的表现渐近而言不优于 SAA 在目标问题上的表现。我们指出，随机非线性规划的类似下界已在 [15], 中得到，而我们的论证受到其中所用技巧的启发。除了这个下界之外，本工作的主要结果是证明，在合理的假设下，随机前向-后向迭代的运行平均具有与 (1.4) 相同的渐近性质，因此是渐近最优的。

该保证 我们提出的结果对于随机非线性规划也颇具意义 :

$$\min_x \; f(x) = \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[f(x,z)] \qquad \text{subject to} \qquad g_i(x) \le 0 \qquad \forall i = 1, \ldots, m. \tag{1.5}$$

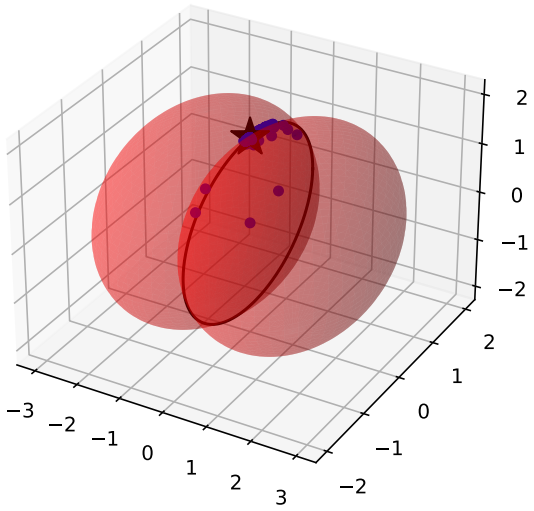
这里，每个 g_i 都是光滑函数，且映射 $x \mapsto f(x, z)$ 对几乎处处的 $z \sim \mathcal{P}$ 都是光滑的。该问题的最优性条件可通过变分不等式 (1.1) 建模，其中对应关系为 $A(x, z) = \nabla f(x, z)$，以及 $\mathcal{X} = \{x : g_i(x) \le 0 \; \forall i = 1, \ldots, m\}$。此时，随机前向-后向算法便成为随机投影梯度法。我们的结果表明，在活跃梯度线性无关、严格互补性和强二阶充分条件这三个标准条件下，随机前向-后向迭代 $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$ 的运行平均渐近正态且达到最优：

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{D} N\left(0, \nabla\sigma(0) \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot \nabla\sigma(0)\right).$$

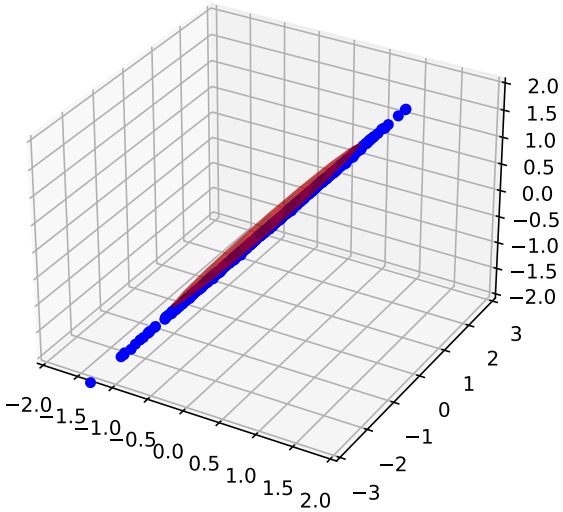
此外，正如 人们早已熟知，雅可比矩阵 $\nabla\sigma(0)$ 有一个显式描述 a s

$$\nabla\sigma(0) = (P_{\mathcal{T}} \nabla^2_{xx} \mathcal{L}(x^\star, y^\star) P_{\mathcal{T}})^\dagger, \tag{1.6}$$

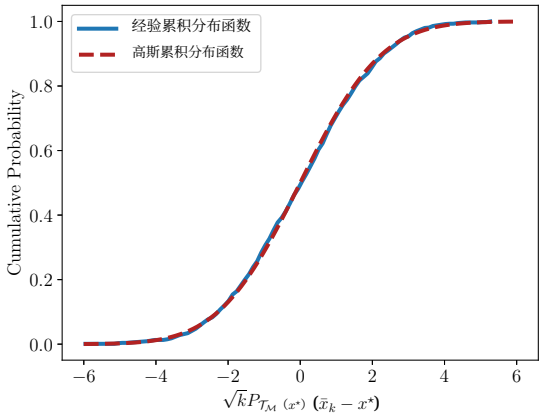
其中，$\nabla^2_{xx}\mathcal{L}(x^\star, y^\star)$ 是拉格朗日函数的海森矩阵，符号 $\dagger$ 表示 Moore-Penrose 伪逆，而 $P_{\mathcal{T}}$ 是到线性子空间 $\{\nabla g_i(x^\star)\}_{i\in\mathcal{I}}^\perp$ 和 $\mathcal{I} = \{i$ 的投影：$g_i(x^\star) = 0\}$ 是活跃指标集合。图 1展示了这一结果的一个示例，其中绘出了投影随机梯度法在两个球的交集上最小化线性函数时的表现。鉴于现有文献，这一结果或许令人惊讶。具体而言，Duchi 和 Ruan [15] 揭示了一个惊人的



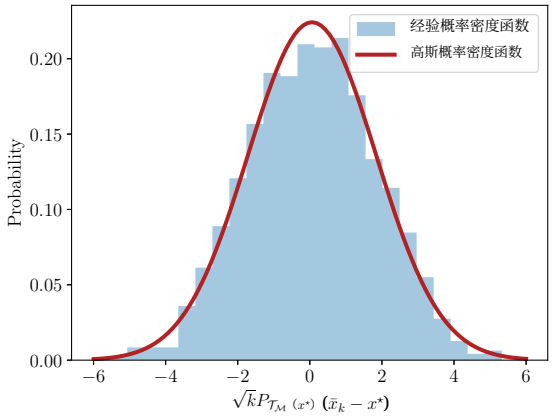
(a) 可行域与迭代点 x_k



（b）偏差 $\sqrt{k}(\bar{x}_k - \bar{x^\star})$ 和 95% 置信区域。



(c) 经验累积分布函数与高斯累积分布函数



(d) 直方图与高斯密度。

图1：随机投影梯度法在两个球的交集上最小化 $\mathbb{E}_g[-x_1 + \langle g, x\rangle]$ 的性能，这两个球的半径均为 2，球心分别为 $(-1,\ 0,\ 0)$ 和 $(1,\ 0,\ 0)$。期望是对高斯分布 $g \sim N(0,\ I)$ 取的。最优解 $(0,\ 0,\ \sqrt{3})$（以星号标记）位于活跃流形 $\mathcal{M}$ 上，该流形是图中以黑色描绘的圆。左上图展示了从原点初始化、步长为 $\eta_k = k^{-3/4}$ 、运行 1000 次迭代的一次过程所生成的迭代点。右上图展示了在 $K = 10^6$ 条件下进行 100 次运行得到的重缩放偏差 $\sqrt{k}(\bar{x}_K - x^\star)$。这两幅图清楚地表明，迭代点迅速趋近活跃流形，并且渐近地，偏差 $\sqrt{k}(\bar{x}_k - x^\star)$ 仅在 $x^\star$ 处的 $\mathcal{M}$ 切空间方向上有非零分量。第二行的两幅图分别展示了切向分量 $\sqrt{k}P_{T_{\mathcal{M}}(x^\star)}(\bar{x}_k - x^\star)$ 的直方图和经验累积分布函数，并叠加了高斯分布对应的函数。

SAA 与至少一种称为对偶平均法的标准在线方法在估计质量之间存在差距 [27, 37], ，该方法用于随机非线性优化。事实上，即使对于在球上最小化线性函数期望的问题，对偶平均法也表现出次优的渐近协方差 [15, 第5节。2][1] 相比之下，我们看到随机投影梯度法是渐近最优的。

现在让我们回到一般问题（1.1）和随机前向-后向算法（1.3）。为了推导所声称的 SFB 渐近保证，我们将施加一些额外假设。首先，除了假设 $\sigma(\cdot)$ 在原点附近是光滑的之外，我们还假设存在原点的一个邻域 U，使得 $\sigma(U)$ 是一个光滑流形。这个假设并不苛刻，因为例如当矩阵 $\nabla\sigma(\cdot)$ 在原点的某个邻域内具有常秩时，该假设会自动成立。用 [14],的术语来说，集合 $\mathcal{M} = \sigma(U)$ 被称为围绕 $\bar{x}$的一个活跃流形。回到随机非线性规划的情形，活跃流形就是活动不等式的零集合

$$\mathcal{M} = \{x : g_i(x) = 0 \ \ \forall i \in \mathcal{I}\}.$$

参见图1。活跃流形的变体在非线性规划中得到了广泛研究，其名称包括可识别曲面 [36], 部分光滑集 [23],$\mathcal{UV}$-结构 [22, 24]、 $g \circ F$可分解函数 [34], 和最小可识别集 [14]。

我们论证的核心思想，是将 SFB 的非光滑动态与 $\mathcal{M}$上的光滑随机逼近算法联系起来。更具体地说，我们将证明，在温和条件下，流形 $\mathcal{M}$ 上的影子序列 $y_k := P_{\mathcal{M}}(x_k)$除了一个小误差外，表现得是光滑的

$$y_{k+1} = y_k - \alpha_k P_{T_{\mathcal{M}}(y_k)}(A(y_k, z_k) + \nu_k) + o(\alpha_k), \tag{1.7}$$

其中 $T_{\mathcal{M}}(y_k)$表示 $\mathcal{M}$ 在 y_k 处的切空间，而 ν_k 是零均值噪声。因此，我们可以借助 Polyak 和 Juditsky [30]的技术，得到影子序列 y_k的渐近性质，进而推断原始迭代点 x_k的信息。我们注意到，在约束优化情形下，迭代（1.7）变为在 $\mathcal{M}$上对 f 的限制所实施的非精确黎曼梯度法。

(1.7) 的有效性依赖于两个额外条件，这些条件在 [8], 中引入，用来关联 $\mathcal{X}$ 与 $\mathcal{M}$的“一阶”行为。具体而言，对于接近 $x^\star$的 $x \in \mathcal{X}$ 和 $y \in \mathcal{M}$，我们假设:

- $\langle N_{\mathcal{X}}(x) \cap \mathbb{S}^{d-1}, x - y\rangle = o(\|x - y\|)$ [(b)-regularity]
- $N_{\mathcal{X}}(x) \cap \mathbb{S}^{d-1} \subset N_{\mathcal{M}}(y) + O(\|x - y\|)\mathbb{B}$ [strong (a)-regularity]

（b）正则性条件仅断言，连接 $x \in \mathcal{X}$ 和 $y \in \mathcal{M}$的割线在 x 和 y 趋近于 $x^\star$附近的同一点时，会与 $\mathcal{X}$ 相切。相比之下，强（a）正则性断言，法锥 $N_{\mathcal{X}}$（x）包含于 $N_{\mathcal{M}}$（y）中，误差至多为线性阶O（$\|x - y\|$）——这是一种 Lipschitz 条件。这两种正则性条件在 [8], 中被提出并得到充分发展，其中还给出了大量示例和微积分规则。特别是，对于随机非线性规划，这两个条件都会自动成立。

[1]相比之下，在 $\mathcal{X}$是多面体凸函数这一特殊情形下，对偶平均法是最优的 [15]。

概要。本文其余部分安排如下。第 2 节介绍本文将用到的基本符号和构造。存在光滑局部化σ（$\cdot$）是我们工作的一个核心假设。第 3 节为样本平均近似（SAA）建立渐近收敛保证和局部极小极大下界。第4节介绍我们将考虑的算法类别。第5节陈述迭代方法渐近正态性的主要结果。第 6 节讨论支撑我们工作主要结果的两大支柱。

2 记号与基本构造

本节记录了本文中将贯穿使用的基本记号。为此，符号 $\mathbf{R}^d$ 表示一个带有内积 $\langle\cdot,\cdot\rangle$ 和诱导范数$\|x\| = \sqrt{\langle x, x\rangle}$的欧几里得空间。符号 $\mathbf{B}$ 表示$\mathbf{R}^d$中的闭单位球，而 $B_r(x)$表示以点 x为中心、半径为 r 的闭球。对于任意函数 $f\colon \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$，其定义域、图和上图集 分别定义为

$$\begin{aligned}\operatorname{dom} f &:= \{x \in \mathbf{R}^d : f(x) < \infty\},\\ \operatorname{gph} f &:= \{(x, f(x)) \in \mathbf{R}^d \times \mathbf{R} : x \in \operatorname{dom} f\},\\ \operatorname{epi} f &:= \{(x, r) \in \mathbf{R}^d \times \mathbf{R} : r \geq f(x)\},\end{aligned}$$

分别如此。若 $\operatorname{epi} f$ 是闭集，则称 f 为闭的；等价地，f 是下半连续的。f 的近端映射（参数为 $\alpha > 0$）定义为

$$\operatorname{prox}_{\alpha f}(x) := \operatorname*{argmin}_y \left\{ f(y) + \frac{1}{2\alpha}\|y - x\|^2 \right\}.$$

点 $x \in \mathbf{R}^d$ 到集合 $Q \subset \mathbf{R}^d$ a 的距离和投影 re

$$d(x, Q) := \inf_{y \in Q} \|y - x\| \qquad \text{and} \qquad P_Q(x) := \operatorname*{argmin}_{y \in Q} \|y - x\|,$$

分别如此。集合 Q的示性函数记为 $\delta_Q(\cdot)$，其在 Q 上取值为零，在其外取值为$+\infty$。符号 $o(h)$表示任意满足 $o(h)/h \to 0$（当 $h \searrow 0$时）的函数 $o(\cdot)$。

2.1 光滑流形

接下来，我们回顾光滑流形的一些定义；详细内容请参阅 [3, 20]。在全文中，所有光滑流形 $\mathcal{M}$ 均假设嵌入于 $\mathbf{R}^d$ 中，并且我们将 $\mathcal{M}$ 的切空间和法空间视为 $\mathbf{R}^d$的子空间。因此，如果对于任意点 $x \in \mathcal{M}$，都存在一个开邻域 $U \subset \mathbf{R}^d$ 以及一个 C^p-光滑映射 F，其从 U 映射到某个欧几里得空间 $\mathbf{R}^n$，使得雅可比矩阵 $\nabla F(x)$ 是满射，并且等式 $\mathcal{M} \cap U = F^{-1}(0)$ 成立，那么集合$\mathcal{M} \subset \mathbf{R}^d$ 就是一个 C^p 流形（具有 $p \geq 1$）。此时，$F = 0$称为 $\mathcal{M}$的局部定义方程，而 $\mathcal{M}$ 在x 处的切空间和法空间分别定义为 $T_{\mathcal{M}}(x) := \operatorname{Null}(\nabla F(x))$ 和 $N_{\mathcal{M}}(x) := (T_{\mathcal{M}}(x))^{\perp}$。注意，对于具有 $p \geq 1$的 C^p 流形 $\mathcal{M}$，投影 $P_{\mathcal{M}}$ 在 $\mathcal{M}$ 中每个点 x 的某个邻域上是 C^{p-1}-光滑的，并且在切空间中原点的某个邻域上是 C^p-光滑的

$T_{\mathcal{M}}(x)$ [25]。此外，对于所有接近 $\mathcal{M}$ 的 x，包含关系 $\text{range}(\nabla P_{\mathcal{M}}(x)) \subseteq T_{\mathcal{M}}(P_{\mathcal{M}}(x))$ 成立，并且对于所有 $x \in \mathcal{M}$，等式 $\nabla P_{\mathcal{M}}(x) = P_{T_{\mathcal{M}}(x)}$ 成立。

设 $\mathcal{M} \subset \mathbf{R}^d$ 是某个 C^p-流形，其中 $p \geq 1$。若存在一个定义在 x 的开邻域 $U \subset \mathbf{R}^d$ 上的 C^p 函数 $f\colon U \to \mathbf{R}$，且该函数在 $U \cap \mathcal{M}$ 上与 f 一致，则称函数 $f\colon \mathcal{M} \to \mathbf{R}$ 在点 $x \in \mathcal{M}$ 附近是 C^p-光滑的。于是，f 在 x 处的协变梯度是向量 $\nabla_{\mathcal{M}} f(x) := P_{T_{\mathcal{M}}(x)}(\nabla f(x))$。当 f 和 $\mathcal{M}$ 是 C^2-光滑的时，f 在 x 处的协变海森矩阵定义为满足下式的唯一自伴双线性型 $\nabla^2_{\mathcal{M}} f(x)\colon T_{\mathcal{M}}(x) \times T_{\mathcal{M}}(x) \to \mathbf{R}$：

$$\frac{d^2}{dt^2} f(P_{\mathcal{M}}(x + tu)) \mid_{t=0} \;=\; \langle \nabla^2_{\mathcal{M}} f(x) u, u \rangle \qquad \text{for all } u \in T_{\mathcal{M}}(x).$$

若 $\mathcal{M}$ 是 C^3-光滑的，则我们可以将 $\nabla^2_{\mathcal{M}} f(x)$ 简写为

$$\nabla^2_{\mathcal{M}} f(x) = P_{T_{\mathcal{M}}(x)} \nabla^2 \hat{f}(x) P_{T_{\mathcal{M}}(x)},$$

同时将右侧视为 $T_{\mathcal{M}}(x)$ 上的线性算子。

若存在一个映射 $F\colon U \to \mathbf{R}^d$，定义在 x 的某个邻域 $U \subset \mathbf{R}^d$ 上，且在 x 附近与 F 于 $\mathcal{M}$ 上一致，则称映射 $F\colon \mathcal{M} \to \mathbf{R}^m$ 在点 x 附近是 C^p 光滑的。在这种情况下，我们将协变雅可比矩阵 $\nabla F(x)\colon T_{\mathcal{M}}(x) \to \mathbf{R}^m$ 定义为如下表达式：对所有 $u \in T_{\mathcal{M}}(x)$，有 $\nabla_{\mathcal{M}} F(x) u = \nabla F(x) u$。简单计算可知，在特例 $F(x) = \nabla_{\mathcal{M}} f(x)$ 中，对于一个 C^3-光滑函数 $f\colon \mathcal{M} \to \mathbf{R}$，由 $\nabla_{\mathcal{M}} F(x)$ 在 $T_{\mathcal{M}}(x)$ 上定义的二次型与 $\nabla^2_{\mathcal{M}} f(x)$ 一致。更准确地说，等式成立：

$$P_{T_{\mathcal{M}}(x)} \cdot \nabla_{\mathcal{M}} F(x) \cdot P_{T_{\mathcal{M}}(x)} = P_{T_{\mathcal{M}}(x)} \cdot \nabla^2_{\mathcal{M}} f(x) \cdot P_{T_{\mathcal{M}}(x)}.$$

2.2 法锥与次微分

接下来，我们需要介绍非光滑分析与变分分析中的一些基本构造。我们的讨论主要遵循 Rockafellar-Wets [32]。关于这一主题的其他重要论述还包括 [2,5,26,28]。集合 $Q \subset \mathbf{R}^d$ 在点 $x \in \mathbf{R}^d$ 处的 ***Fréchet*** 法锥，记作 $N_Q(x)$，由所有满足下述条件的向量 $v \in \mathbf{R}^d$ 组成：

$$\langle v, y - x \rangle \leq o(\|y - x\|) \quad \text{as} \quad y \to x \text{ in } Q. \tag{2.1}$$

因此，若在一阶意义下，v 与所有从 x 指向 Q 内部的方向都成钝角，则 v 属于 $N_Q(x)$。一般而言，Fréchet 法锥关于基点 x 的扰动高度不连续。因此，我们如下扩充 Fréchet 法锥。Q 在 $x \in Q$ 处的**极限法锥**，记作 $N_Q(x)$，由所有向量 $v \in \mathbf{R}^d$ 组成；对此，存在序列 $x_i \in Q$ 和 $v_i \in N_Q(x_i)$，满足 $(x_i, v_i) \to (x, v)$。

函数对应的法锥类似物是次微分，也就是广义导数的集合。具体而言，考虑函数 $f\colon \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$ 以及一点 $x \in \operatorname{dom} f$。$f$ 在 x 处的 ***Fréchet*** 次微分，记作 $\partial f(x)$，由所有满足以下逼近性质的向量 $v \in \mathbf{R}^d$ 组成：

$$f(y) \geq f(x) + \langle v, y - x \rangle + o(\|y - x\|) \quad \text{as} \quad y \to x.$$

的极限次微分 f 在 x 处，记为 $\partial f(x)$，由所有满足以下条件的向量 $v \in \mathbf{R}^d$ 组成：存在序列 $x_i \in \mathbf{R}^d$ 以及 Fréchet 次梯度 $v_i \in \partial f(x_i)$，使得当 $i \to \infty$ 时，$(x_i, f(x_i)\ v_i) \to (x, f(x)\ v)$。满足 $0 \in \partial f(x)$ 的点 x 称为 f 的**临界点**。非光滑优化算法的主要目标是寻找临界点。

2.3 活跃流形

典型非光滑函数的临界点位于某个特定流形上，该流形以如下意义刻画问题的活跃性：对函数进行微小的线性扰动后，临界点不会离开该流形。此类活跃流形已通过多种方式建模，包括可识别曲面 [36]，部分光滑性 [23]、$\mathcal{UV}$-结构 [22, 24]、$g \circ F$可分解函数 [34]，以及最小可识别集 [14]。在这项工作中，我们采用以下活跃性的形式化模型，该模型已在 [14] 中明确使用。

定义 2.1（活跃流形）。考虑函数 $f\colon \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$，并固定一个集合$\mathcal{M} \subseteq \operatorname{dom} f$，其中包含满足 $0 \in \partial f(\bar{x})$ 的点 $\bar{x}$。若存在满足下述条件的常数 $\epsilon > 0$，则称 $\mathcal{M}$ 是 $\bar{x}$ 附近的活跃 C^p-流形 。

- **（光滑性）**集合 $\mathcal{M}$ 在 $\bar{x}$ 附近是一个流形，且 f 在 $\bar{x}$ 附近限制于 $\mathcal{M}$ 后是C^p-光滑的。

- **（尖锐性）**下界成立：

$$\inf\{\|v\| : v \in \partial f(x),\ x \in U \setminus \mathcal{M}\} > 0,$$

  where we set $U = \{x \in B_\epsilon(\bar{x}) : |f(x) - f(\bar{x})| < \epsilon\}$.

更一般地，如果 $\mathcal{M}$ 是倾斜函数 $f_v(x) = f(x) - \langle v, x\rangle$ 在 $\bar{x}$ 处的活跃流形，我们称 $\mathcal{M}$ 是 f 在 $\bar{x}$ 对 $\bar{v} \in \partial f(\bar{x})$ 的活跃流形。

尖锐性条件简单来说，就是在距离 $\bar{x}$ 足够近且函数值足够接近的流形外的点处，f 的次梯度始终远离零。比如，如果 f 是弱凸的，或者 f 在其定义域上连续，那么可以省略对函数值的局部化；详见 [14] 。

直观而言，活跃流形的显著特点是：函数沿流形平滑变化，并在法向方向上线性增长；图 2 展示了一个例子。以下定理概述了这一点，摘自 [7, 定理 D.2]。

命题 2.2 (识别蕴含尖锐性)。假设闭函数 $f\colon \mathbf{E} \to \mathbf{R} \cup \{\infty\}$ 在点 $\bar{x}$ 处满足 $0 \in \partial f(\bar{x})$，并且存在一个活跃流形 $\mathcal{M}$ 。那么存在常数 c、$\epsilon > 0$，使得

$$f(x) - f(P_{\mathcal{M}}(x)) \geq c \cdot \operatorname{dist}(x, \mathcal{M}), \qquad \forall x \in B_\epsilon(\bar{x}). \tag{2.2}$$

ˆ 注意，命题 2.2 中存在一个不可忽略的假设 $0 \in \partial f(\bar{x})$。事实上，在较弱的包含关系 $0 \in \partial f(\bar{x})$ 下，增长条件 (2.2) 很容易失效，单变量示例 $f(x) = -|x|$ 就说明了这一点。值得注意的是，在假设 ˆ$0 \in \partial f(\bar{x})$ 下，活跃流形在 $\bar{x}$ [14, 附近局部唯一（命题 8.2]）。

为了取得进展，我们需要活跃流形上满足另外两个条件，使 f 在流形上和流形外的次梯度紧密耦合。尽管这两个条件在 [8]，中提出，看起来可能很难满足，但实际上它们非常宽松。

第一个正则性条件的动机源于对泰勒定理的弱化，这种弱化适用于非光滑函数。具体而言，回顾一下，任何 C^1-光滑函数 f 都满足一阶逼近性质

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + o(\|y - x\|) \qquad \text{as } x, y \to \bar{x}. \tag{2.3}$$

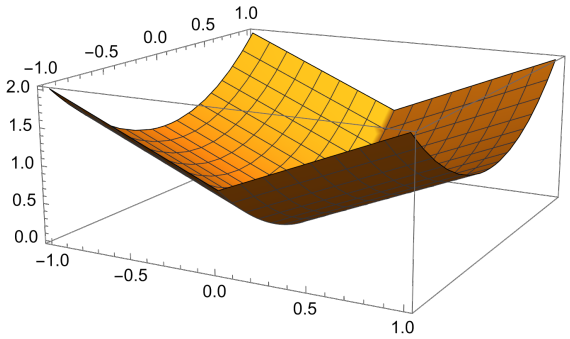


图 2： $f(x_1,\ x_2) = |x_1| + x_2^2$

这一估计对优化理论与实践至关重要，因为它量化了由梯度给出的 f 线性模型的逼近质量。当 f 不光滑时，用次梯度代替梯度的 (2.3) 式类似结果不可能对所有点 x 以及 $\bar{x}$附近的 y 一致成立，因为这将意味着可微性。更合理的要求是，仅要求当点 y 位于活跃流形 $\mathcal{M}$上时，(2.3) 成立。事实上，就我们的目的而言，只需将等式替换为不等式即可。

定义 2.3 (($b_{\leq}$)-正则性)。 考虑一个函数 $f\colon \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$，它在其定义域上局部利普希茨连续。固定一个集合 $\mathcal{M} \subset \operatorname{dom} f$，它在 $\bar{x}$ 附近是一个 C^1 流形，并且函数 f 在 $\mathcal{M}$ 上的限制在 $\bar{x}$ 附近是 C^1-光滑的。若存在 $\epsilon > 0$，使得以下估计成立，我们称 f 在 $\bar{x}$ 处沿 $\mathcal{M}$ 是 ($b_{\leq}$)-正则的：

$$f(y) \geq f(x) + \langle v, y - x\rangle + (1 + \|v\|) \cdot o(\|y - x\|), \tag{2.4}$$

对所有 $x \in \operatorname{dom} f \cap B_\epsilon(\bar{x})$、$y \in \mathcal{M} \cap B_\epsilon(\bar{x})$ 和 $v \in \partial f(x)$，均成立。

重要的是，条件 ($b_{\leq}$) 沿活跃流形 $\mathcal{M}$ 直接意味着，在 $\bar{x}$ 附近的点 x 处取得的 f 的所有负次梯度都指向活跃流形 $\mathcal{M}$。事实上，这直接源自命题 2.2，并总结于以下推论中。因此，次梯度型方法总是朝活跃流形的方向移动——这显然是一种理想的性质。

推论 2.4 （近端瞄准）。 设闭函数 $f\colon \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$ 在满足 $0 \in \hat{\partial} f(\bar{x})$的点 $\bar{x}$ 处具有一个活跃 C^1-流形 $\mathcal{M}$。假设 f 在其定义域上局部利普希茨连续，并且 f 在 $\bar{x}$附近沿 $\mathcal{M}$ 是($b_{\leq}$)-正则的。那么，存在一个常数 $\mu > 0$，使得估计式

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) + (1 + \|v\|) \cdot o(\operatorname{dist}(x, \mathcal{M})), \tag{2.5}$$

对所有靠近 $\bar{x}$ 的 $x \in$定义域 f，以及所有 $v \in \partial f(x)$，均成立。

第二个正则性条件的性质有所不同，它源于对梯度利普希茨连续性的弱化。也就是说，非光滑性本身意味着，偏差 $\partial f(x) - \partial f(y)$ 无法由参数 $\|x - y\|$ 之间的距离很好地控制。另一方面，事实证明，如果我们取 $y \in \mathcal{M}$，并且只考察切向方向上的次梯度偏差，那么误差 $P_{T_{\mathcal{M}}(y)}[\partial f(x) - \partial f(y)]$ 通常可由 $\|x - y\|$ 线性界定。此外，在典型情况下，$P_{T_{\mathcal{M}}(y)}[\partial f(y)]$ 仅由一个向量构成，即协变梯度 $\nabla_{\mathcal{M}} f(y)$。

定义 2.5（强（a）正则性）。考虑一个函数 $f\colon \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$，它在其定义域上局部利普希茨连续。固定一个集合 $\mathcal{M} \subset \operatorname{dom} f$，该集合在 $\bar{x}$ 附近是一个 C^1 流形，且函数 f 在 $\mathcal{M}$ 上的限制在 $\bar{x}$ 附近是 C^1-光滑的。若存在常数 $C, \epsilon > 0$ 满足以下条件，我们就称 f 沿 $\mathcal{M}$ 在 $\bar{x}$ 附近强（a）正则：

$$\|P_{T_{\mathcal{M}}(y)}(v - \nabla_{\mathcal{M}} f(y))\| \leq C(1 + \|v\|)\|x - y\|, \tag{2.6}$$

对所有 $x \in \operatorname{dom} f \cap B_\epsilon(\bar{x})$、$y \in \mathcal{M} \cap B_\epsilon(\bar{x})$ 和 $v \in \partial f(x)$。

这两个正则性条件很容易通过示性函数推广到集合。具体而言，若集合 $Q \subset \mathbf{R}^d$ 的示性函数 δ_Q 沿着流形 $\mathcal{M}$ 在 $\bar{x}$处是 $(b_{\leq})$-正则的（分别是强 (a)-正则的），则称集合 $Q \subset \mathbf{R}^d$ 沿着流形 C^1 在 $\bar{x} \in \mathcal{M}$ 处是 $(b_{\leq})$-正则的（分别是强 (a)-正则的）。

论文 [8] 介绍了大量具有活跃流形的函数，在这些流形上，条件 $(b_{\leq})$ 和强条件 (a) 均成立。这里，我们详细讨论一个非线性规划的例子，更多例子请参见 [8]。

例 2.1（非线性规划）。考虑以下非线性规划问题

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t. } & g_i(x) \leq 0 \qquad \text{for } i = 1, \ldots, m \\ & g_i(x) = 0 \qquad \text{for } i = m+1, \ldots, n, \end{aligned} \tag{2.7}$$

其中，f 和 g_i 是定义在 $\mathbf{R}^d$ 上的 C^p-光滑函数。令 $\mathcal{X}$ 表示该问题所有可行点组成的集合。现在考虑一个关于函数 $f + \delta_{\mathcal{X}}$ 的临界点 $\bar{x} \in \mathcal{X}$，并定义活跃指标集

$$\mathcal{I} = \{i : g_i(\bar{x}) = 0\}.$$

假设以下条件成立：

- **(LICQ)** the gradients $\{\nabla g_i(\bar{x})\}_{i \in \mathcal{I}}$ are linearly independent.

那么集合

$$\mathcal{M} = \{x : g_i(x) = 0 \ \forall i \in \mathcal{I}\}$$

在 $\bar{x}$ 局部是一个 C^p 光滑流形。此外，三个函数 f、$\delta_{\mathcal{X}}$ 和 $f + \delta_{\mathcal{X}}$沿着 $\mathcal{M}$ 在 $\bar{x}$ 附近均为（$b_{\leq}$）正则且强（a）正则。为确保 $\mathcal{M}$ 是 $f + \delta_{\mathcal{X}}$的一个活跃流形，还需要一个额外条件。定义拉格朗日函数

$$\mathcal{L}(x, y) := f(x) + \sum_{i=1}^{n+m} y_i g_i(x).$$

$\bar{x}$ 的临界性和线性无关约束规范保证存在一个满足以下条件的（唯一）拉格朗日乘子向量 $\bar{y} \in \mathbf{R}^m_+ \times \mathbf{R}^n$：$\nabla_x \mathcal{L}(\bar{x}, \bar{y}) = 0$，且对所有 $i \notin \mathcal{I}$，$\bar{y}_i = 0$。假设以下标准条件成立：

- **(严格互补性)** $\bar{y}_i > 0$ 对所有 $i \in \mathcal{I} \cap \{1, \ldots, m\}$。

那么，$\mathcal{M}$ 确实是 $\bar{x}$处 $f + \delta_{\mathcal{X}}$ 的一个活动 C^p 流形。

2.4 光滑可逆映射与活动流形

统计估计程序的性能在很大程度上取决于问题对扰动的敏感性。各种估计问题又可以建模为求解包含关系 $0 \in F(x)$的任务，其中 F是某个集值映射，而我们只能通过抽样来近似其取值。接下来，我们将回顾基于逆映射／隐函数定理范式的基本扰动理论，并紧密遵循专著 [11]。

一个集值映射 $F\colon \mathbf{R}^d \rightrightarrows \mathbf{R}^m$ 是指将点 $x \in \mathbf{R}^d$ 映射到集合$F(x) \subset \mathbf{R}^m$的对应关系。集值映射总是存在集值逆映射：

$$F^{-1}(y) = \{x : y \in F(x)\}.$$

F 的定义域和图分别定义为

$$\operatorname{dom} F := \{x : F(x) \neq \emptyset\} \qquad \text{and} \qquad \operatorname{gph} F := \{(x, y) : y \in F(x)\}.$$

我们将关注系统 $v \in F(x)$ 的解对左端项扰动 v的敏感性，或者等价地，映射$v \mapsto F^{-1}(v)$的变分行为。特别地，我们将关注这样一种情形：在基点 (v, x) 附近， F^{-1} 的图局部与某个光滑映射的图重合。以下定义阐述了这一点。

定义 2.6（光滑可逆性）。 考虑一个集值映射 $F\colon \mathbf{R}^d \rightrightarrows \mathbf{R}^m$ 以及一对 $(\bar{x}, \bar{v}) \in \operatorname{gph} F$。若存在一个单值的 C^p-光滑映射 $\sigma(\cdot)$ 和 $(\bar{v}, \bar{x})$ 的一个邻域 U，满足下列条件，则称 F在 $(\bar{x}, \bar{v})$ 附近可逆，且逆映射为 $\sigma(\cdot)$：

$$U \cap \operatorname{gph} F^{-1} = U \cap \operatorname{gph} \sigma.$$

这个定义乍看之下可能有些奇怪： F并没有任何“光滑”之处，但我们却要求 F^{-1} 的图在$(\bar{v}, \bar{x})$附近与某个光滑函数的图重合。恰恰相反，我们将看到，在各种情形下，这一假设确实成立。特别是，在非线性规划中，光滑可逆性是典型情形。

例 2.2（非线性规划）。 回到例（2.1），其中有 $p \geq 2$，定义集值映射

$$F(x) = \nabla f(x) + N_{\mathcal{X}}(x).$$

那么，众所周知 经典地已知， F 在$(\bar{x},0)$处 C^{p-1} 可逆，当且仅当矩阵 x

$$\Sigma := P_{T_{\mathcal{M}}(\bar{x})} \nabla^2_{xx} \mathcal{L}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$$

在 $T_{\mathcal{M}}(\bar{x})$上非奇异。在这种情况下，逆映射的雅可比矩阵为 $\nabla \sigma(0) = \Sigma^{\dagger}$，其中 $\dagger$ 表示 Moore-Penrose 伪逆。值得注意的是， Σ 也可以等价地写成$P_{T_{\mathcal{M}}(\bar{x})} \nabla^2_{\mathcal{M}} f(\bar{x}) P_{T_{\mathcal{M}}(\bar{x})}$.

光滑可逆性与活跃流形密切相关，而例 2.2 是一个简单的推论。事实上，以下更一般的结论成立。该结果由一个结合活跃流形与隐函数定理的标准论证得出。证明见附录 A.1。我们对所考虑的函数只需作一个温和的假设。

也就是说，根据 [29, 定义 2.1]，如果对于任意收敛到某个对 $(\bar{x}, \bar{v}) \in \operatorname{gph} \partial f$ 的序列 $(x_i, v_i) \in \operatorname{gph} \partial f$，函数值 $f(x_i)$ 都收敛到 $f(\bar{x})$，则称函数 f 在点 $\bar{x}$ 处次微分连续。特别地，在其定义域上连续的函数以及闭凸函数都是次微分连续的。

定理 2.7 （光滑可逆性与活动流形）。考虑集合值映射 p

$$F(x) := A(x) + \partial f(x),$$

其中 $A: \mathbb{R}^d \to \mathbb{R}^d$ 是 C^p-光滑的，且 $f: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ 在点 $\bar{x}$附近次微分连续。假设 f 在某点 $\bar{x}$ 处对于$-A(\bar{x}) \in \partial f(\bar{x})$具有一个 C^{p+1} 活跃流形 $\mathcal{M}$。令 $G(x) = 0$ 是 $\mathcal{M}$ 在 $\bar{x}$ 附近的任意一组 C^{p+1}-光滑局部定义方程，并令 $\hat{f}$ 是任意一个 C^{p+1}-光滑函数，它在 $\mathcal{M}$中 $\bar{x}$ 的某个邻域上与 f 一致。定义映射

$$\mathcal{H}(x, y) := A(x) + \nabla \hat{f}(x) + \nabla G(x)^\top y.$$

那么存在唯一的乘子向量 $\bar{y}$ 满足条件 $0 = \mathcal{H}(\bar{x}, \bar{y})$。此外，$F$ 在$(0, \bar{x})$ 附近可逆，当且仅当矩阵

$$\Sigma := P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$$

在 $T_{\mathcal{M}}(\bar{x})$ 上非奇异，此时等式 $\nabla \sigma(0) = \Sigma^\dagger$ 成立。

3 样本均值近似的渐近正态性与最优性。

在分析随机逼近算法的渐近性能之前，不妨先回顾样本均值近似（样本平均近似，SAA）的保证；与随机逼近算法相比，其假设、结论和论证都更容易表述。本节将介绍这一内容：我们推导非光滑问题中样本均值近似估计量的渐近分布，并证明它按照 H´ajek 和 Le Cam [19, 35]的意义渐近局部极小极大最优。整个章节中，我们聚焦于求解满足以下变分包含关系的点 $x^\star$:

$$0 \in A(x) + H(x) \qquad \text{where} \qquad A(x) = \mathop{\mathbb{E}}_{z \sim \mathcal{P}} A(x, z). \tag{3.1}$$

这里，$H\colon \mathbf{R}^d \rightrightarrows \mathbf{R}^d$ 是任意的集值映射，$\mathcal{P}$ 是测度空间 $(\mathcal{Z}, \mathcal{F})$ 上的固定概率分布，而 $A\colon \mathbf{R}^d \times \mathcal{Z} \to \mathbf{R}^d$ 是可测映射。在本节余下部分，我们将始终采用以下假设。

假设 A. 映射 $F := A + H$ 在 $(0, \bar{x})$ 处是 C^1-光滑可逆的，其逆映射为 $\sigma(\cdot)$。

求解 (3.1) 的样本平均近似方法如下。令 $S = \{z_1, \ldots, z_k\}$ 为从 $\mathcal{P}$ 中抽取的独立同分布样本，并令 x_k 为该问题的一个解

$$0 \in A_S(x) + H(x) \qquad \text{where} \qquad A_S(x) := \frac{1}{k} \sum_{i=1}^{k} A(x, z_i), \tag{3.2}$$

（假设解存在）。我们首先证明，样本均值近似的解渐近服从正态分布，其协方差为 $\nabla\sigma(0)\cdot\mathrm{Cov}(A(\bar{x},\ z))\cdot\nabla\sigma(0)^{\top}$。尽管这一结果的变体广为人知 [16,18,33],，我们仍在附录 B.1 中给出一个简短证明，着重说明解映射 $\sigma(\cdot)$ 的应用。为此，我们提出以下标准假设。

假设 B (可积性与光滑性)。 假设存在一个邻域U ，位于 $\bar{x}$ 周围，并满足以下条件。

1. 对几乎所有的 z，映射 $A(\cdot,z)$在每个 $x\in U$处都可微。
2. 二阶矩界成立：

$$\sup_{x\in U}\ \mathop{\mathbb{E}}_{z\sim\mathcal{P}}\|A(x,z)\|^2<\infty\qquad\text{and}\qquad\sup_{x\in U}\ \mathop{\mathbb{E}}_{z\sim\mathcal{P}}\left[\|\nabla A(x,z)\|_{\mathrm{op}}^2\right]<\infty.$$

以下定理表明，只要 x_k 最终停留在 $\bar{x}$的一个充分小的邻域内，误差 $\sqrt{k}(x_k-\bar{x})$就渐近服从正态分布，其协方差为 $\nabla\sigma(0)\cdot\mathrm{Cov}(M(\bar{x},z))\cdot\nabla\sigma(0)^{\top}$。验证问题 (3.2) 是否存在充分接近 $\bar{x}$ 的解 x_k，是一个独立且已得到广泛研究的课题，此处不作讨论。

定理 3.1 (样本均值近似)。 假设 *A* 和 *B* 成立。特别地，存在 ϵ_1、 $\epsilon_2>0$ 以及一个 C^1-光滑映射 σ： $B_{\epsilon_1}(0)\to B_{\epsilon_2}(\bar{x})$，满足

$$\mathrm{gph}\,\sigma=(B_{\epsilon_1}(0)\times B_{\epsilon_2}(\bar{x}))\cap\mathrm{gph}\,F^{-1}.$$

*Supp*此外，假设存在一个平方可积函数 $L(z)$，满足

$$\|\nabla A(x_1,z)-\nabla A(x_2,z)\|\le L(z)\|x_1-x_2\|\qquad\forall x_1,x_2\in B_{\epsilon_2}(\bar{x}).\tag{3.3}$$

如有必要，缩小 ϵ_2，以确保 $\epsilon_2\le\min\left\{\frac{\mathrm{lip}(\sigma)^{-1}}{2\mathbb{E}L},\sqrt{\frac{\epsilon_1}{2\mathbb{E}L}}\right\}$。设 $S=\{z_1,\dots,z_k\}$ 是从 $\mathcal{P}$ 抽取的独立同分布样本，并设 x_k 是方程 (3.2) 解的一个可测选择，使得 $\Pr[x_k\in B_{\epsilon_2}(\bar{x})]\to1$ 在 k 趋于无穷时成立。则渐近正态性成立：

$$\sqrt{k}(x_k-\bar{x})\xrightarrow{D}\mathsf{N}\big(0,\nabla\sigma(0)\cdot\mathrm{Cov}(A(\bar{x},z))\cdot\nabla\sigma(0)^{\top}\big).$$

接下来我们将证明，在局部极小极大意义下，(3.1) 中的渐近协方差是所有关于 $\bar{x}$ 的估计量中可能达到的最佳值，这是 H´ajek 和 Le Cam [19,35] 的结果。具体而言，我们将通过对抗性选取的小扰动序列 $\mathcal{P}'$ 作用于 $\mathcal{P}$，给出任何 估计程序在求解由此诱导的 (3.1) 时的性能下界。我们将使用 ϕ-散度来衡量扰动的大小

$$\Delta_\phi\big(\mathcal{P}'\parallel\mathcal{P}\big)=\int_{\mathcal{Z}}\phi\left(\frac{d\mathcal{P}'}{d\mathcal{P}}\right)d\mathcal{P},$$

由任意 C^3-光滑凸函数 ϕ: $(0,\ \infty)\to\mathbb{R}$ 满足 $\phi(1)=0$所诱导。定义 B_ε 的一个可容许邻域 为所有满足 $\Delta_\phi(\mathcal{P}'\parallel\mathcal{P})\le\varepsilon$ 且存在扰动变分方程 $0\in\mathbb{E}_{z\sim\mathcal{P}'}[A(x,\ z)]+H(x)$的解的概率分布的集合。 以下是本节的主要结果。在该定理中，我们用 $\mathbb{E}_{P_k'}$ 表示相对于 k 独立同分布观测 $z_i\sim\mathcal{P}'$的期望。

定理 3.2（渐近最优性）。假设假设 *A* 和 *B* 成立。令$\mathcal{L}\colon\ \mathbb{R}^d \to [0,\infty)$ 为任意一个对称、拟凸且下半连续函数，并令 $\widehat{x}_k\colon\ \mathcal{Z}^k \to \mathbb{R}^d$ 为一个估计量序列。那么不等式

$$\lim_{c\to\infty} \liminf_{k\to\infty} \sup_{\mathcal{P}'\in B_{c/k}} \mathbb{E}_{P'_k}\big[\mathcal{L}\big(\sqrt{k}(\widehat{x}_k - \bar{x}_{\mathcal{P}'})\big)\big] \geq \mathbb{E}\big[\mathcal{L}(Z)\big]$$

成立，其中 $Z \sim \mathrm{N}\big(0,\ \nabla\sigma(0)\cdot \mathrm{Cov}(A(\bar{x},\ z))\cdot \nabla\sigma(0)^\top\big)$。

定理 3.2 的证明见附录 B.2，其灵感来自 Duchi-Ruan 关于随机非线性规划的类似定理 [15]。

4 随机逼近：假设与示例

接下来我们转向随机逼近算法，并在本节中介绍我们将要考虑的算法及相关假设。我们将给出的具体示例都旨在求解变分包含，但这类问题的具体细节多少有些令人分心。因此，我们只提炼出使我们的结果成立所需的基本要素。为此，我们的目标是找到一个满足以下包含关系的点 x

$$0 \in F(x), \tag{4.1}$$

其中 $F\colon\ \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ 是一个集值映射。在全文中，我们固定一个这样的解 $\bar{x}$，使其满足 (4.1)。我们将假设，某种意义上，问题 (4.1) 是“变分光滑”的。也就是说，存在一个特定的流形 $\mathcal{M}$——在具体示例中即活跃流形——包含 $\bar{x}$，并且映射 $x \mapsto P_{T_{\mathcal{M}}(x)}F(x)$ 是单值的，且在 $\mathcal{M}$ 的 $\bar{x}$ 邻域内为 C^1-光滑。以下假设将对此作出精确定义。

假设 C（光滑化简）。假设存在一个 C^p 流形 $\mathcal{M} \subset \mathbf{R}^d$，使得以下性质成立。

(C1) 映射 $F_{\mathcal{M}}\colon \mathcal{M} \to \mathbf{R}^d$ 定义如下：

$$F_{\mathcal{M}}(x) := P_{T_{\mathcal{M}}(x)}F(x)$$

在 $\mathcal{M}$ 中的某个邻域内，映射是单值的。

(C2) 存在 $(\bar{x},0)$ 的一个邻域 U，使得

$$U \cap \operatorname{gph} F = U \cap \operatorname{gph}(F_{\mathcal{M}} + N_{\mathcal{M}}).$$

我们注意到，可以用协变雅可比矩阵 $\nabla_{\mathcal{M}}F_{\mathcal{M}}(\bar{x})$ 轻松刻画 F 的光滑可逆性。以下引理给出了这一结论。

引理 4.1（解映射的雅可比矩阵）。映射 F 在 C^p-光滑意义下可逆于（$\bar{x}$，0）附近，且局部化为 $\sigma(\cdot)$，当且仅当线性映射 $P_{T_{\mathcal{M}}(\bar{x})}\nabla F_{\mathcal{M}}(\bar{x})P_{T_{\mathcal{M}}(\bar{x})}$ 在 $T_{\mathcal{M}}(\bar{x})$ 上非奇异。在此情形下，局部化的雅可比矩阵为

$$\nabla\sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})}\nabla_{\mathcal{M}}F_{\mathcal{M}}(\bar{x})P_{T_{\mathcal{M}}(\bar{x})})^{\dagger}.$$

证明。令 Φ 为 F 在 $\bar{x}$ 的邻域 $V \subset \mathbf{R}^d$ 上的光滑延拓。根据假设（C2）， F 和 $\Phi + N_{\mathcal{M}}$ 的图在 $(\bar{x}, 0)$ 附近重合，因此我们可以专注于 $(\Phi + N_{\mathcal{M}})^{-1}$ 的光滑局部化是否存在。应用引理 A.2，取 $\bar{y} = 0$，可知 $\Phi + N_{\mathcal{M}}$ 在 $(\bar{x}, 0)$ 附近 C^p-光滑可逆，当且仅当线性映射 $P_{T_{\mathcal{M}}}(\bar{x})\nabla\Phi(\bar{x})P_{T_{\mathcal{M}}}(\bar{x})$ 在 $T_{\mathcal{M}}(\bar{x})$ 上非奇异。在此情形下，局部化的雅可比矩阵为 $\nabla\sigma(0) = (P_{T_{\mathcal{M}}}(\bar{x})\nabla\Phi(\bar{x})P_{T_{\mathcal{M}}}(\bar{x}))^{\dagger}$。注意到等式$\nabla F_{\mathcal{M}}(\bar{x})P_{T_{\mathcal{M}}}(\bar{x}) = \nabla_{\mathcal{M}}\Phi(\bar{x})P_{T_{\mathcal{M}}}(\bar{x})$，证明得证。 □

我们考虑的随机逼近算法假设可以访问广义梯度映射：

$$G\colon \mathbb{R}_{++} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d.$$

给定 $x_0 \in \mathbb{R}^d$，算法迭代执行以下更新

$$x_{k+1} = x_k - \alpha_k G_{\alpha_k}(x_k, \nu_k), \tag{4.2}$$

其中， $\alpha_k > 0$ 是控制序列， ν_k 是随机噪声。我们将在第 6 节稍后对噪声 ν_k 提出相关假设。

我们对 G 作出两个假设。第一个假设（假设 D）类似于经典的利普希茨假设，并确保步长相对于 $\|\nu\|$至多呈线性缩放。

假设 D （步长）。 我们假设存在一个常数 $C > 0$ 和 $\bar{x}$ 的一个邻域 $\mathcal{U}$，使得估计式

$$\sup_{x \in \mathcal{U}_F} \|G_\alpha(x, \nu)\| \leq C(1 + \|\nu\|),$$

对所有 $\nu \in \mathbb{R}^d$和 $\alpha > 0$均成立，其中我们令 $\mathcal{U}_F := \mathcal{U} \cap \mathrm{dom}\ F$。

第二个假设精确地说明了映射 G 与 $F_{\mathcal{M}}$之间的关系。

假设 E （强（a）与瞄准）。 我们假设存在常数 $C, \mu > 0$ 和 $\bar{x}$ 的一个邻域 $\mathcal{U}$，使得对于所有 $\nu \in \mathbb{R}^d$ 和 $\alpha > 0$，以下条件均成立，其中我们令 $\mathcal{U}_F := \mathcal{U} \cap \mathrm{dom}\ F$。

(E1) （切向比较）对于所有 $x \in \mathcal{U}_F$，有

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\alpha(x, \nu) - F(P_{\mathcal{M}}(x)) - \nu)\| \leq C(1 + \|\nu\|)^2(\mathrm{dist}(x, \mathcal{M}) + \alpha).$$

(E2) （近端瞄准）对于 $x \in \mathcal{U}_F$，有

$$\langle G_\alpha(x, \nu) - \nu, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \mathrm{dist}(x, \mathcal{M}) - (1 + \|\nu\|)^2(o(\mathrm{dist}(x, \mathcal{M})) + C\alpha).$$

这里需要说明几点。假设 (E1) 确保运动方向$G_{\alpha_k}(x_k, \nu_k)$在流形 $\mathcal{M}$ 的切向方向上能够很好地近似 $F_{\mathcal{M}}(P_{\mathcal{M}}(x))$。假设 (E2) 确保从 $G_{\alpha_k}(x_k, \nu_k)$ 中减去噪声后，更新方向 $x_k - x_{k+1}$ 在局部指向流形 $\mathcal{M}$。我们稍后将证明，这确保迭代点 x_k 以受控的速率趋近流形 $\mathcal{M}$ 。

4.1 变分包含的随机逼近示例

本节其余部分将介绍满足假设 D 和 E 的算法示例。在所有情况下，我们都将考虑求解变分包含

$$0 \in A(x) + \partial g(x) + \partial f(x). \tag{4.3}$$

其中，$A\colon \mathbf{R}^d \to \mathbf{R}^d$ 是任意单值连续映射，$f\colon \mathbf{R}^d \to \mathbf{R}$ 是闭函数，而 $g\colon \mathbf{R}^d \to \mathbf{R}$ 是有下界的闭函数。[2] 如引言中所述，变分包含涵盖了多种问题，其中最值得注意的是非线性规划的一阶最优性条件和博弈中的纳什均衡。为了将 (4.3) 与 (4.1) 对应起来，我们将集值映射 F 定义为

$$F(x) := A(x) + \partial g(x) + \partial f(x).$$

在全文中，我们固定一个满足包含关系 (4.3) 的点 $x^\star$。

问题 (4.3) 的一种经典算法是随机前向-后向迭代，其步骤是在 $A + \partial g$ 上执行“前向步”，并在 f 上执行近端步。具体而言，给定当前迭代点 x_t，算法执行以下更新

$$\left\{\begin{array}{l}\text{Choose } w_t \in \partial g(x_t)\\ \text{Choose } x_{t+1} \in \operatorname{prox}_{\alpha_t f}(x_t - \alpha_t(A(x_t) + w_t + \nu_t))\end{array}\right\}, \tag{4.4}$$

其中，ν_t 是一个噪声序列。与该算法对应的算子 $G_\alpha(x, \nu)$ 为

$$G_\alpha(x,\nu) := \frac{x - s_f(x - \alpha(A(x) + s_g(x) + \nu))}{\alpha},$$

其中，$s_g(x)$ 是次微分 $\partial g(x)$ 的任意一个选取，而 $s_f(x)$ 是近端映射 $\operatorname{prox}_{\alpha f}(x)$ 的任意一个选取。本节的目标是在对 A、g 和 f 作出若干合理假设的条件下，验证该算子满足假设 E。

特别是，文献中广泛采用针对 G 的局部有界条件 D，并且已有多种充分条件。下面的引理描述了其中若干条件，我们将在后文中用到。证明见附录 C.1。

引理 4.2 (局部有界性)。假设 $A(\cdot)$ 和 $s_g(\cdot)$ 在$\bar{x}$附近局部有界。那么，在以下任一情形下，假设 *D* 均成立。

1. f 是闭集 $\mathcal{X}$ 的示性函数。
2. f 是凸的，并且函数 $x \mapsto$ 距离$(0, \partial f(x))$ 在定义域 f 附近有界 $\bar{x}$.
3. f 在定义域 $g \cap$ 定义域 f 上利普希茨连续。

接下来，我们将在若干合理的情形下验证假设 E；所有证明均见附录。特别地，注意到 $F_{\mathcal{M}}$ 的以下表达式将很有用。本节中我们将反复使用这一引理，而不再明确引用。

[2] 特别地，对于所有 $x \in \mathbf{R}^d$ 和所有 $\alpha > 0$，$\operatorname{prox}_{\alpha f}(x)$ 非空。

引理 4.3 (局部切空间约化)。 假设 f 和 g 在各自的定义域上均为利普希茨连续， A 为 C^p- 光滑， $f+g$ 在 $x^\star$ 处对于 $-A(x^\star)$ 存在一个活跃 C^{p+1} 流形，并且 f 和 g 沿 $\mathcal{M}$ 在 $x^\star$ 附近均为 C^{p+1}-光滑且强(a) 正则。则假设 **C** 成立，且 $F_{\mathcal{M}}$ 可表示为简单形式

$$F_{\mathcal{M}}(x) = P_{T_{\mathcal{M}}(x)}(A(x)) + \nabla_{\mathcal{M}} g(x) + \nabla_{\mathcal{M}} f(x), \tag{4.5}$$

对所有 $x \in \mathcal{M}$ ，在 $x^\star$ 附近。

4.1.1 随机前向算法 $(f = 0)$

我们从 (4.3) 的最简单情形开始，其中 $f = 0$。在这种情况下，迭代 (4.2) 简化为纯随机前向算法，且映射 G 具有如下简单形式

$$G_\alpha(x, \nu) := A(x) + s_g(x) + \nu,$$

它与 α 无关。下面对问题数据引入以下假设。

假设 F （前向算法的假设）. 假设 $f = 0$，且 $g(\cdot)$ 和 $A(\cdot)$ 均在 $\bar{x}$ 附近利普希茨连续。假设对于 g， $\mathcal{M} \subseteq \mathcal{X}$ 在 $\bar{x}$ 处是一个 C^p-光滑流形。

(F1) (强 (a)) 函数 g 沿着 $\mathcal{M}$ 在处强 (a)-正则 $\bar{x}$.

(F2) **(近端瞄准)** 存在 $\mu > 0$ 使得该不等式成立 :

$$\langle A(\bar{x}) + v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) \qquad \text{for all } x \text{ near } \bar{x} \text{ and } v \in \partial g(x). \tag{4.6}$$

注意，推论 2.4 表明，只要包含关系 $-A(\bar{x}) \in \partial g(\bar{x})$ 成立， $\mathcal{M}$ 是 g 在 $\bar{x}$ 处针对 $v = -A(\bar{x})$ 的一个活跃流形，并且 g 沿 $\mathcal{M}$ 在 $\bar{x}$处是 $(b_{\leq})$-正则的，瞄准条件 (F2) 就成立。以下命题表明，假设 F 足以确保假设 E 成立。证明见附录。

命题 4.4 (前向方法)。 假设 ***F*** 蕴含假设 ***E***。

以下结论可立即得到。

推论 4.5 (活跃流形)。 假设 $f = 0$ ，并且 $g(\cdot)$ 和 $A(\cdot)$ 在 $\bar{x}$ 附近均为 ***Lipschitz*** 连续。进一步假设包含关系 $-A(\bar{x}) \in \partial g(\bar{x})$ 成立， g在 $\bar{x}$ 附近针对 $\bar{v} = -A(\bar{x})$具有一个 C^2 活跃流形，并且 g 沿 $\mathcal{M}$ 在 $\bar{x}$处既是(b)≤-正则的，也是强(a)-正则的。那么假设 ***E*** 成立。

4.1.2 随机投影前向算法 $(f = \delta_{\mathcal{X}})$

接下来，我们聚焦于 (4.3) 的一个特例，其中 f 是闭集 $\mathcal{X}$ 的示性函数。在这种情况下，迭代 (4.2) 化简为随机投影前向算法，而映射 G 的形式为

$$G_\alpha(x, \nu) := \frac{x - s_{\mathcal{X}}(x - \alpha(A(x) + s_g(x) + \nu))}{\alpha},$$

其中， $s_{\mathcal{X}}(x)$ 是投影映射 $P_{\mathcal{X}}(x)$ 的一个选取。为确保随机投影前向方法满足假设 E，我们引入以下假设。

假设 G（关于投影梯度映射的假设）。 假设 f 是闭集 $\mathcal{X}$ 的示性函数，且 $g(\cdot)$ 和 $A(\cdot)$ 在 $\bar{x}$ 附近均为利普希茨连续。令 $\mathcal{M} \subseteq \mathcal{X}$ 为包含 $\bar{x}$ 的一个 C^2 流形，并假设 f 在 $\bar{x}$ 附近的 $\mathcal{M}$ 上是 C^2。

(G1) (强 (a)) 函数 g 和集合 $\mathcal{X}$ 在 $\bar{x}$处沿 $\mathcal{M}$ 强 (a)-正则。

(G2) (近端瞄准) 存在 $\mu > 0$ 使得该不等式成立

$$\langle A(\bar{x}) + v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) \qquad \forall\, x \in \mathcal{X} \text{ near } \bar{x} \text{ and } v \in \partial g(x). \tag{4.7}$$

(G3) (条件 (b)) 集合 $\mathcal{X}$ 在 $\bar{x}$处沿 $\mathcal{M}$ 是 $(b_{\leq})$-正则的。

注意，与推论 2.4 类似的论证表明，只要包含关系 $-A(\bar{x}) \in \partial(g + f)(\bar{x})$成立， $\mathcal{M}$ 是 $v = -A(\bar{x})$ 的活跃流形，且 g 在 $\bar{x}$ 处沿 $\mathcal{M}$ 是 $(b_{\leq})$- 正则的，瞄准条件 (G2) 就成立。

以下命题表明，假设 G 足以确保假设 E 成立。

命题 4.6 (投影前向法)。 若假设 *D* 和假设 *G* 成立，则假设 *E* 也成立。

以下结论显然成立。

推论 *4.7*（活跃流形）。设 f 是闭集 $\mathcal{X}$的示性函数，且 $g(\cdot)$ 和 $A(\cdot)$ 在 $\bar{x}$附近均为利普希茨连续。此外，假设包含关系 $-A(\bar{x}) \in \hat{\partial}(g + f)(\bar{x})$成立， $g + f$ 在 $\bar{x}$ 附近对向量 $\bar{v} = -A(\bar{x})$具有一个 C^2 活跃流形，并且 g 和 f 在 $\bar{x}$处沿 $\mathcal{M}$ 均为 $(b_{\leq})$-正则且强 (a)-正则。则假设 *E* 成立。

4.1.3 随机前向-后向法 $(g = 0)$

最后，我们关注 (4.3) 的一个特殊情形，其中 $g = 0$。在这种情况下，迭代 (4.2) 化为随机前向-后向算法，映射 G 变为

$$G_\alpha(x, \nu) := \frac{x - s_f(x - \alpha(A(x) + \nu))}{\alpha},$$

为了确保随机近端梯度方法满足假设 E，我们引入以下假设。

假设 H (前向-后向方法的假设)。 假设 $g = 0$ 和 $f(\cdot)$ 以及 $A(\cdot)$ 在 $\bar{x}$附近的 $\operatorname{dom} f$ 上利普希茨连续。此外，假设存在一个包含 $\bar{x}$ 的C^2 流形 $\mathcal{M} \subset \mathcal{X}$，使得 f 在 $\mathcal{M}$ 附近的 C^2-光滑。

(H1)（强 (a)） 函数 f 在 $\mathcal{M}$ 处沿着.....是强 (a)-正则的 $\bar{x}.$

(H2)（近端瞄准） 存在 $\mu > 0$ 使得不等式 y

$$\langle A(\bar{x}) + v, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \operatorname{dist}(x, \mathcal{M}) - (1 + \|v\|)o(\operatorname{dist}(x, \mathcal{M})) \tag{4.8}$$

ˆ 对所有 $x \in \operatorname{dom} f$ near $\bar{x}$ and $v \in \partial f(x)$ 都成立。

注意，推论 2.4 表明，只要包含关系 $-A(\bar{x}) \in \partial f(\bar{x})$ 成立，$\mathcal{M}$ 是 f 在 $\bar{x}$ 处关于 $v = -A(\bar{x})$ 的活跃流形，并且 f 在 $\bar{x}$ 处沿 $\mathcal{M}$ 是 $(b_{\leq})$-正则的，瞄准条件 (H2) 就成立。

以下命题表明，假设 H 足以确保假设 E 成立.

命题 4.8 (前向-后向方法)。若假设 D 和 H 成立，则假设 E 也成立。

由此可立即得出以下结论。

推论 4.9 (活跃流形)。假设 $g = 0$，且 f 和 $A(\cdot)$ 在 f 的定义域附近关于 $\bar{x}$ 均为 *Lipschitz* 连续。进一步假设包含关系 $-A(\bar{x}) \in \partial f(\bar{x})$ 成立。假设 f 在 $\bar{v} = -A(\bar{x})$ 意义下，在 $\bar{x}$ 附近存在一个 C^2 活跃流形，并且 f 沿 $\mathcal{M}$ 在 $\bar{x}$ 处既满足 $(b)_{\leq}$-正则性，也满足强 (a)-正则性。那么假设 E 成立。

5 渐近正态性

接下来，我们对步长 α_k 和噪声序列 ν_k提出两个假设。第一个是假设是标准的，下面将其概述。

假设 I（基本假设）。假设如下。

(J1) 映射 G 是可测的。

(J2) 存在常数 $c_1, c_2 > 0$ 和 $\gamma \in (1/2, 1]$ 使得

$$\frac{c_1}{k^\gamma} \leq \alpha_k \leq \frac{c_2}{k^\gamma}.$$

(J3) $\{\nu_k\}$ 是关于递增的 σ-域序列的鞅差序列

$$\mathcal{F}_k = \sigma(x_j : j \leq k \text{ and } \nu_j : j < k),$$

并且存在函数 $q\colon \mathbb{R}^d \to \mathbb{R}_+$，该函数在有界集合上有界，且满足

$$\mathbb{E}[\nu_k \mid \mathcal{F}_k] = 0 \qquad \text{and} \qquad \mathbb{E}[\|\nu_k\|^4 \mid \mathcal{F}_k] < q(x_k).$$

我们用 $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$ 表示条件期望。

(J4) 对所有 $k \geq 1$，均有包含关系 $x_k \in \operatorname{dom} F$。

假设 I 中的所有条件都是随机逼近方法文献中的标准条件，并且与 [9,假设 C] 中的条件相呼应。唯一的例外是关于 $\|\nu_k\|$ 的四阶矩界，该界规定 ν_k 的尾部稍轻。对于我们考虑的情形，这个界似乎是必要的。

为了证明渐近正态性结果，我们对噪声序列 ν_k 施加一个额外假设，这一假设也见于 [15, 假设 D]。在陈述该假设之前，作为动机，考虑以下随机变分不等式 (4.3)：

$$0 \in A(x) + \partial f(x) + \partial g(x) \qquad \text{where} \qquad A(x) = \mathop{\mathbb{E}}_{z \sim \mathcal{P}} A(x, z).$$

那么，算法 (4.4) 中的噪声 ν_k 具有如下形式

$$\nu_k = A(x_k; z_k) - A(x_k).$$

等价地，我们可以将右端分解为

$$\nu_k = \underbrace{A(\bar{x}; z_k) - A(\bar{x})}_{=:\nu_k^{(1)}} + \underbrace{(A(x_k; z_k) - A(\bar{x}; z_k)) + (A(\bar{x}) - A(x_k))}_{=:\nu_k^{(2)}(x_k)},$$

两个分量 $\nu_k^{(1)}$ 和 $\nu_k^{(2)}(x_k)$ 在以下意义上有本质区别。一方面，只要协方差 $\mathrm{Cov}(A(\bar{x}, z))$ 存在，和 $\frac{1}{\sqrt{k}}\sum_{i=1}^k \nu_i^{(1)}$ 显然会收敛到均值为零的正态向量。另一方面，$\nu_k^{(2)}(x_k)$ 在以下意义上很小：$\mathbb{E}_k\|\nu_k^{(2)}(x_k)\|^2 \le 2 \cdot \mathbb{E}_z[L(z)^2] \cdot \|x_k - \bar{x}\|^2$, 其中 $L(z)$ 是 $A(\cdot, z)$ 的 Lipschitz 常数。基于这个例子，我们对噪声序列引入如下假设。

假设 J。固定一点 $\bar{x} \in \mathrm{dom}\, F$，使得假设 C 成立，并令 U 为一个矩阵，其列向量构成 $T_{\mathcal{M}}(\bar{x})$ 的正交基。我们假设噪声序列具有可分解结构 $\nu_k = \nu_k^{(1)} + \nu_k^{(2)}(x_k)$，其中 $\nu_k^{(2)} : \mathrm{dom}\, F \to \mathbb{R}^d$ 是一个满足以下条件的随机函数

$$\mathbb{E}_k[\|U^\top \nu_k^{(2)}(x)\|^2] \le C\|x - \bar{x}\|^2 \qquad \text{for all } x \in \mathrm{dom}\, F \text{ near } \bar{x},$$

以及某个 $C > 0$。此外，我们假设对所有 $x \in \mathrm{dom}\ F$，都有 $\mathbb{E}_k[\nu_k^{(1)}] = \mathbb{E}_k[\nu_k^{(2)}(x)] = 0$，并且满足以下极限：

$$\frac{1}{\sqrt{k}} \sum_{i=1}^k U^\top \nu_i^{(1)} \xrightarrow{D} N(0, \Sigma).$$

其中某个对称半正定矩阵为 Σ。

现在我们可以陈述本文的主要结果——随机逼近算法的渐近正态性。

定理 5.1 (渐近正态性)。假设假设 *C*、*D*、*E*、*I* 和 *J* 成立。假设 $\gamma \in (\frac{1}{2}, 1)$，且由过程 (4.2) 生成的序列 $\bar{x}_k$ 几乎必然收敛到 $\bar{x}$。假设存在一个常数 $\mu > 0$，满足

$$\langle \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x})v, v\rangle \ge \mu\|v\|^2 \qquad \textit{for all} \qquad v \in T_{\mathcal{M}}(\bar{x}). \tag{5.1}$$

那么 F 在 $(\bar{x},0)$ 附近是 C^p-光滑可逆的，其逆映射为 $\sigma(\cdot)$，并且平均迭代点 $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$ 满足

$$\sqrt{k}(\bar{x}_k - \bar{x}) \xrightarrow{D} N\left(0, \nabla\sigma(0) \cdot \Sigma \cdot \nabla\sigma(0)^\top\right).$$

此外，$\nabla\sigma(0)$ 也可等价地写成 $\nabla\sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})}\nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) P_{T_{\mathcal{M}}(\bar{x})})^\dagger$。

这个定理的结论令人惊讶：尽管序列 x_k 从未到达流形，但 $\sqrt{k}(\bar{x}_k - \bar{x})$ 的极限分布却支撑在切空间 $T_{\mathcal{M}}(\bar{x})$ 上。因此，从渐近角度看，与 $\mathcal{M}$ 垂直的“非光滑方向”会很快被“平均掉”。当对所有 k，$\|G_{\alpha_k}(x_k, \nu_k)\|$ 都与 0 保持正距离时，这意味着 x_k 必须在流形两侧来回振荡，而不是从某一侧趋近流形。

5.1 非线性规划中的渐近正态性

作为定理 5.1 的一个简单示例，我们现在详细说明它对随机非线性规划中的随机投影梯度法所带来的结果；这一方法已在例 2.1 中讨论过。具体而言，考虑问题 (2.7)，并令 $\bar{x}$ 为一个局部极小点。假设 g_i 在 $\bar{x}$ 附近是 C^3-光滑的，且 f 具有如下形式：$f(x) = \mathbb{E}_{z\sim\mathcal{P}} f(x, z)$，其中 $\mathcal{P}$ 是某个概率分布，并且每个函数 $f(\cdot, z)$ 在 $\mathcal{X}$ 附近都是 C^3-光滑的。考虑以下随机投影梯度法：

$$\begin{aligned} &\text{Sample: } z_k \sim P \\ &\text{Update: } x_{k+1} \in P_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k; z_k)). \end{aligned} \tag{5.2}$$

为了理解算法的渐近性质，如例 2.1 所示，令 $\bar{y}$ 为拉格朗日乘子向量，并假设满足线性无关约束规范和严格互补性。此外，假设满足二阶充分条件：存在 $\mu > 0$ 使得

$$w^\top \left[\nabla^2_{xx} \mathcal{L}(\bar{x}, \bar{y})\right] w \geq \mu \|w\|^2 \qquad \text{for all } w \in T_{\mathcal{M}}(\bar{x}).$$

注意，如例 2.2 所述，这一条件仅要求 $f := f_0 + \delta_{\mathcal{X}}$ 的协变海森矩阵

$$\nabla^2_{\mathcal{M}} f(\bar{x}) = P_{T_{\mathcal{M}}(\bar{x})} \nabla^2_{xx} \mathcal{L}(x^\star, y^\star) P_{T_{\mathcal{M}}(\bar{x})}$$

在 $T_{\mathcal{M}}(\bar{x})$ 上正定。最后，为了确保我们的噪声序列

$$\begin{aligned} \nu_k &= \nabla f(x_k; z_k) - \nabla f(x_k) \\ &= \underbrace{\nabla f(\bar{x}; z_k) - \nabla f_0(\bar{x})}_{=:\nu_k^{(1)}} + \underbrace{(\nabla f(x_k; z_k) - \nabla f(\bar{x}; z_k) + \nabla f(\bar{x}) - \nabla f(x_k))}_{=:\nu_k^{(2)}(x_k)}, \end{aligned}$$

满足假设 I 和 J，我们假设随机性具有足够良好的性质：

(G5)（随机梯度）作为 x 的函数，四阶矩

$$x \in \mathcal{X} \mapsto \mathbb{E}_{z\sim\mathcal{P}}[\|\nabla f(x; z) - \nabla f(x)\|^4]$$

在有界集合上有界。此外，存在 $C > 0$ 使得

$$\mathbb{E}_{z\sim\mathcal{P}}[\|\nabla f(x; z) - \nabla f(\bar{x}; z)\|^2] \leq C\|x - \bar{x}\|^2 \qquad \text{for all } x \in \mathcal{X}.$$

最后，梯度 $P_{T_{\mathcal{M}}(\bar{x})} \nabla f(\bar{x}; z)$ 的协方差 $\Sigma = \mathsf{Cov}(P_{T_{\mathcal{M}}(\bar{x})} \nabla f(\bar{x}; z))$ 有限。

基于这些假设，我们得到如下非线性规划的渐近正态性结果——这是定理 5.1 的直接推论。

推论 5.2 (非线性规划中的渐近正态性**)。** 设 $\gamma \in (\frac{1}{2}, 1)$，并考虑由随机投影梯度法 (5.2) 生成的迭代点 x_k。那么，若 x_k 以概率 *1* 收敛到 $\bar{x}$，平均迭代点 $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$ 满足

$$\sqrt{k}(\bar{x}_k - \bar{x}) \xrightarrow{d} N\left(0, \nabla\sigma(0) \cdot \mathrm{Cov}(\nabla f(\bar{x}; z)) \cdot \nabla\sigma(0)^\top\right),$$

其中 $\nabla\sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})} \nabla^2_{xx} \mathcal{L}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})})^\dagger$。

正如引言中所述，这似乎是针对一般非线性规划问题中含有 C^3 数据的标准随机投影梯度法的首个渐近正态性保证，即使在凸情形下也是如此。最后，我们指出，即便对于简单的优化问题，对偶平均过程也可能达到次优收敛 [15]。这令人意外，因为此类方法能够识别活跃流形 [21]（另见 [15, 第 4 节。1]），而随机投影梯度法则不能。

6 定理 5.1 证明的两大支柱

主要结果定理 5.1 的证明见附录。本节概述证明的主要要点。具体而言，点 $\bar{x}$ 处的假设 E 保证了两种有用的行为，前提是算法（4.2）的迭代点 $\{x_k\}$ 始终处于 $\bar{x}$周围的一个小球内。首先，x_k 必须以受控的速率趋近于包含 $\bar{x}$ 的流形 $\mathcal{M}$，这是近端瞄准条件的推论。其次，沿流形的迭代点影子 $y_k = P_{\mathcal{M}}(x_k)$ 构成一个采用隐式回缩的近似黎曼随机梯度序列。此外，该序列的近似误差随距离 $\mathrm{dist}(x_k, \mathcal{M})$ 和 α_k 衰减，而这些量会迅速趋于零。

概述这两种行为模式的正式陈述需要局部论证。因此，我们将频繁用到以下停时：给定指标 $k \geq 1$ 和常数 $\delta > 0$，定义

$$\tau_{k,\delta} := \inf\{j \geq k \colon x_j \notin B_\delta(\bar{x})\}. \tag{6.1}$$

请注意，停时隐式依赖于 $\bar{x}$，该点满足假设 E。[8, 命题 5.1]中证明的以下命题表明，序列 x_k会迅速趋近流形。我们注意到，[8, 命题 5.1]专门针对优化问题而非求集值映射的零点而陈述；然而，在这个更一般的情形下，论证完全相同。

命题 6.1 (支柱I：朝流形前进)。假设假设 *C*、*D*、*E*、*I* 成立。令 $\gamma \in (1/2, 1]$ 并假设当 $\gamma = 1$ 时 $c_1 \geq 32/\mu$。那么对于所有 $k_0 \geq 1$ 以及充分小的 $\delta > 0$，存在一个常数 C，使得以下结论对于停时 $\tau_{k_0,\,\delta}$（定义见 (6.1)）成立：

1. 存在一个随机变量$V_{k_0,\,\delta}$ 使得

   (a) 极限成立：

   $$\frac{k^{2\gamma-1}}{\log(k+1)^2}\mathrm{dist}^2(x_k, \mathcal{M})1_{\tau_{k_0,\delta}>k} \xrightarrow{a.s.} V_{k_0,\delta}.$$

   (b) 该和几乎处处有限：

   $$\sum_{k=1}^{\infty} \frac{k^{\gamma-1}}{\log(k+1)^2}\mathrm{dist}(x_k, \mathcal{M})1_{\tau_{k_0,\delta}>k} < +\infty.$$

2. 我们有

(*a*) 期望平方距离满足：

$$\mathbb{E}[\mathrm{dist}^2(x_k,\mathcal{M})1_{\tau_{k_0,\delta}>k}] \le C\alpha_k \qquad for\ all\ k \ge 1.$$

(b) 尾和有界：

$$\mathbb{E}\left[\sum_{i=k}^{\infty}\alpha_i \mathrm{dist}(x_i,\mathcal{M})1_{\tau_{k_0,\delta}>i}\right] \le C\sum_{i=k}^{\infty}\alpha_i^2 \qquad for\ all\ k \ge 1.$$

接下来，我们研究影子 $y_k = P_{\mathcal{M}}(x_k)$ 沿流形的演化，并证明当 x_k 趋近流形时，y_k 局部上是一个误差渐近衰减的不精确黎曼随机梯度序列。因此，我们可以利用命题 6.1 控制误差。以下命题已在 [8, 命题 5.2] 中证明。需要注意的是，[8, 命题 5.2] 专门针对优化问题而非寻找集值映射的零点；不过，在这一更一般情形下，论证是相同的。

命题 6.2 (支柱 II：影子迭代)。假设条件 *C*、*D*、*E*、*I* 成立。那么，对于所有 $k_0 \ge 1$ 以及充分小的 $\delta > 0$，存在一个常数 C，使得在停时 $\tau_{k_0,\delta}$（定义见 (6.1)）下，以下结论成立：存在一个由$\mathcal{F}_{k+1}$-可测随机向量组成的序列 $E_k \in \mathbb{R}^d$，使得

1. 影子序列

$$y_k = \begin{cases} P_{\mathcal{M}}(x_k) & if\ x_k \in B_{2\delta}(\bar{x}) \\ \bar{x} & otherwise. \end{cases}$$

满足 $y_k \in B_{4\delta}(\bar{x}) \cap \mathcal{M}$ 对所有 k 均成立，并且递推关系如下：

$$\boxed{y_{k+1} = y_k - \alpha_k F_{\mathcal{M}}(y_k) - \alpha_k P_{T_{\mathcal{M}}(y_k)}(\nu_k) + \alpha_k E_k \qquad for\ all\ k \ge 1.} \tag{6.2}$$

此外，对于这样的 k，我们有 $\mathbb{E}_k[P_{T_{\mathcal{M}}(y_k)}(\nu_k)] = 0$。

2. 令 $\gamma \in (1/2, 1]$，并假设若 $\gamma = 1$，则 $c_1 \ge 32/\mu$。

(a) 对于 $k_0 \le k \le \tau_{k_0,\delta} - 1$，有以下界：

i. 最大值

$\{\mathbb{E}_k[\|E_k\|1_{\tau_{k_0,\delta}>k}, \mathbb{E}_k[\|E_k\|^2 1_{\tau_{k_0,\delta}>k}]\} \le C$。 *ii.*

$\mathbb{E}[\|E_k\|^2 1_{\tau_{k_0,\delta}>k}] \le C\alpha_k$。 *iii.* 该和有限：

$$\sum_{k=1}^{\infty}\frac{k^{\gamma-1}}{\log(k+1)^2}\max\{\|E_k\|1_{\tau_{k_0},\delta>k}, \mathbb{E}_k[\|E_k\|]1_{\tau_{k_0},\delta>k}\} < +\infty.$$

(b) 尾和有界

$$\mathbb{E}\left[1_{\tau_{k_0,\delta}=\infty}\sum_{i=k}^{\infty}\alpha_i\|E_k\|\right] \le C\sum_{i=k}^{\infty}\alpha_i^2 \qquad for\ all\ k \ge 1.$$

借助这两个支柱，我们将对序列 x_k 的研究分为两个正交分量：在切向／光滑方向上，我们研究序列 y_k；该序列源自一种误差快速衰减的非精确梯度法，适用于光滑优化的技术。在法向／非光滑方向上，我们逐步逼近流形，从而可以根据 y_k 的相应性质推断出 x_k 的强性质。

参考文献

[1] Hilal Asi 和 John C Duchi. 随机（近似）近端点方法：收敛性、最优性与自适应性。*SIAM* 优化杂志, 29(3):2257–2290, 2019年。[2] Jonathan Borwein 和 Adrian S Lewis. 凸分析与非线性优化：理论与实例。Springer Science & Business Media, 2010。[3] Nicolas Boumal. 《光滑流形上的优化导论》。在线提供，*8*月, 2020年。[4] Louis H. Y. Chen. 关于条件 Borel-Cantelli 引理的简短注记。概率年刊, 6(4):699–700, 1978年。[5] Francis H Clarke、Yuri S Ledyaev、Ronald J Stern 和 Peter R Wolenski. 非光滑分析与控制理论, 第178卷。Springer Science & Business Media, 2008。[6] Joshua Cutler、Mateo D´ıaz 和 Dmitriy Drusvyatskiy. 决策依赖分布下的随机逼近：渐近正态性与最优性。*arXiv* 预印本 *arXiv:2207.04173*, 2022年。[7] Damek Davis、Dmitriy Drusvyatskiy 和 Vasileios Charisopoulos. 几何步长衰减的随机算法在尖锐函数上呈线性收敛。*arXiv* 预印本 *arXiv:1907.09547*, 2019年。[8] Damek Davis、Dmitriy Drusvyatskiy 和 Liwei Jiang. 活跃流形、分层结构及非光滑优化中向局部极小值的收敛。*arXiv* 预印本 *arXiv:2108.11832v2*, 2021。[9] Damek Davis、Dmitriy Drusvyatskiy、Sham Kakade 和 Jason D Lee. 随机次梯度方法在温和函数上收敛。计算数学基础, 20(1):119–154, 2020年。[10] A Dembo. 概率论讲义：斯坦福统计学310课程。访问日期：*10*月, 1:2016, 2016年。[11] Asen L Dontchev 和 R Tyrrell Rockafellar. 隐函数与解映射, 第543卷。施普林格, 2009年。[12] D. Drusvyatskiy、A. D. Ioffe 和 A. S. Lewis. 半代数优化中的一般极小化行为。*SIAM* 优化杂志, 26(1):513–534, 2016年。

[13] Dmitriy Drusvyatskiy 和 Adrian S Lewis. 最优性、可识别性与敏感性。*arXiv:1207.6628v1*。[14] Dmitriy Drusvyatskiy 和 Adrian S Lewis. 最优性、可识别性与敏感性。《数学规划》, 147(1):467–498, 2014。[15] John C Duchi 和 Feng Ruan. 随机优化中的渐近最优性。《统计年鉴》, 49(1):21–48, 2021。[16] Jitka Dupacov´a 和 Roger Wets. 统计估计量及随机优化问题最优解的渐近行为。《统计年鉴》, 16(4):1517–1549, 1988。[17] Rick Durrett.《概率论：理论与例子》, 第49卷。剑桥大学出版社, 2019年。[18] Alan J King 和 R Tyrrell Rockafellar. 统计估计与随机规划中解的渐近理论。《运筹学数学》, 18(1):148–162, 1993。[19] Lucien Le Cam、Lucien Marie LeCam 和 Grace Lo Yang.《统计渐近理论：若干基本概念》。Springer Science & Business Media, 2000。[20] John M Lee. 光滑流形。载于《光滑流形导论》, 第1–31页。施普林格, 2013。[21] Sangkyun Lee 和 Stephen J. Wright. 正则化随机在线学习中对偶平均的流形识别。《机器学习研究杂志》, 13(55):1705–1744, 2012。[22] Claude Lemar´echal、Fran¸cois Oustry 和 Claudia Sagastiz´abal. 凸函数的 $\mathcal{U}$-拉格朗日函数。《美国数学会汇刊》, 352(2):711–729, 2000。[23] Adrian S Lewis. 活跃集、非光滑性与敏感性。*SIAM* 优化杂志, 13(3):702–725, 2002。[24] Robert Mifflin 和 Claudia Sagastiz´abal. 凸最小化的 VU-算法。《数学规划》, 104(2):583–608, 2005。[25] Scott A Miller 和 J´erˆome Malick. 非光滑凸最小化的牛顿方法： VU-拉格朗日方法、黎曼牛顿方法与序列二次规划方法之间的联系。《数学规划》, 104(2):609–633, 2005。[26] Boris S Mordukhovich.《变分分析与广义微分 I：基础理论》, 第330卷。Springer Science & Business Media, 2006。[27] Yurii Nesterov. 凸问题的原始-对偶次梯度方法。《数学规划》, 120(1):221–259, 2009。

[28] Jean-Paul Penot。《无导数微积分》，第266卷。Springer Science & Business Media，2012年。[29] Ren´e Poliquin 和 R Rockafellar。变分分析中的近端正则函数。美国数学会学报，348(5):1805–1838，1996年。[30] Boris T Polyak 和 Anatoli B Juditsky。通过平均加速随机逼近。*SIAM* 控制与优化杂志，30(4):838–855，1992年。[31] Herbert Robbins 和 David Siegmund。关于非负几乎超鞅的一个收敛定理及其若干应用。载于统计学中的优化方法，第233–257页。爱思唯尔，1971年。[32] R Tyrrell Rockafellar 和 Roger J-B Wets。《变分分析》，第317卷。Springer Science & Business Media，2009年。[33] Alexander Shapiro。随机规划中统计估计量的渐近性质。《统计年鉴》，17(2):841–858，1989年。[34] Alexander Shapiro。一类非光滑复合函数。《运筹学数学》，28(4):677–692，2003年。[35] Aad W Van der Vaart。《渐近统计学》，第3卷。剑桥大学出版社，2000年。[36] Stephen J Wright。约束优化中的可识别曲面。*SIAM Journal onControl and Optimization*，31(4):1063–1079，1993年。[37] Lin Xiao。用于正则化随机学习和在线优化的对偶平均法。神经信息处理系统进展，22，2009年。

A 第 2 节的证明

A.1 定理 2.7 的证明

定理 2.7 的证明由两个引理得出。第一个引理使得可以将包含关系 $v \in A(x) + \partial f(x)$ 的灵敏度分析化简为完全光滑的情形。更准确地说，[13, 命题 10.12]，中证明的以下基本结果表明，只要 f 存在活跃流形，次微分 ∂f 的图就有光滑描述。

引理 A.1 (光滑化简)。设 f 是一个次微分连续函数，它在点 $\bar{x}$ 处关于向量 $\bar{w} \in \hat{}\partial f(\bar{x})$ 存在一个 C^2 活跃流形 $\mathcal{M}$。则等式成立：

$$\operatorname{gph} \partial f = \operatorname{gph} \partial(f + \delta_{\mathcal{M}}) \qquad \textit{locally around} \qquad (\bar{x}, \bar{w}).$$

注意，令 $\hat{} f$ 为一个 C^2 光滑函数，该函数在 $\bar{x}$ 附近的 $\mathcal{M}$ 上与 f 一致，我们可以写成

$$\partial(f + \delta_{\mathcal{M}})(x) = \partial(\hat{f} + \delta_{\mathcal{M}})(x) = \nabla \hat{f}(x) + N_{\mathcal{M}}(x).$$

因此，在与引理 A.1 相同的假设下，等式：

$$\operatorname{gph} \partial f = \operatorname{gph} (\nabla \hat{f} + N_{\mathcal{M}}) \qquad \text{holds locally around} \qquad (\bar{x}, \bar{w}).$$

由该引理可知，我们现在可以将重点放在如下形式的变分包含上$v \in \Phi(x) + N_{\mathcal{M}}(x)$，其中 Φ 和 $\mathcal{M}$ 均为光滑的。这类包含关系的扰动理论完全是经典的，并总结于以下引理中。

引理 A.2 (光滑变分不等式)。 考虑一个集值映射

$$F(x) = \Phi(x) + N_{\mathcal{M}}(x) \tag{A.1}$$

并设 $\bar{x}$ 是满足 $0 \in F(\bar{x})$的点。假设 $\Phi: \mathbf{R}^d \to \mathbf{R}^d$ 是一个 C^p-光滑映射，且 $\mathcal{M} \subset \mathbf{R}^d$ 在 $\bar{x}$附近是一个 C^{p+1}-光滑流形。令 $G(x) = 0$ 为 $\mathcal{M}$ 的任意一组 C^{p+1}-光滑局部定义方程，并定义映射

$$\mathcal{H}(x, y) = \Phi(x) + \nabla G(x)^\top y.$$

那么存在唯一向量 $\bar{y}$ 满足条件 $0 = \mathcal{H}(\bar{x}, \bar{y})$。此外， F 在 $(0, \bar{x})$ 附近可逆，当且仅当 $P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$ 在 $T_{\mathcal{M}}(\bar{x})$处非奇异。在这种情况下，等式成立：

$$\nabla \sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})})^\dagger.$$

证明。 $\bar{y}$ 的存在性源于表达式 $N_{\mathcal{M}}(\bar{x}) = \operatorname{range}(\nabla G(\bar{x})^\top)$，而唯一性则源于 $\nabla G(\bar{x})$的满射性。

我们首先证明逆向蕴含，并推导出所述的雅可比矩阵表达式。假设 $P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$ 在 $T_{\mathcal{M}}(\bar{x})$ 上确实是非奇异的。那么存在 $\epsilon > 0$，使得对于任意 $v \in \epsilon \mathbb{B}$ 和 $x \in B_\epsilon(\bar{x})$，包含关系 $v \in \Phi(x) + N_{\mathcal{M}}(x)$ 成立，当且仅当存在满足以下条件的 y

$$\left.\begin{aligned} v &= \Phi(x) + \nabla G(x)^\top y \\ 0 &= G(x) \end{aligned}\right\}. \tag{A.2}$$

将 t 右端视为 (x, y) 的映射，其在 $(\bar{x}, \bar{y})$ 处的雅可比矩阵为 y

$$\begin{bmatrix} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) & \nabla G(\bar{x})^\top \\ \nabla G(\bar{x}) & 0 \end{bmatrix}. \tag{A.3}$$

简单计算可知，该矩阵是可逆的，因为 $P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$ 在 $T_{\mathcal{M}}(\bar{x})$ 上是非奇异的。因此，反函数定理保证，对于所有足够小的 v，系统 (A.3) 在 $\bar{x}$ 附近存在唯一解 $\sigma(v)$，且该解关于 v C^p 光滑。对 (A.3) 求逆可得雅可比矩阵 $\nabla \sigma(0) = (P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})})^\dagger$ 的表达式。

反之，假设 F 在 $(0, \bar{x})$ 附近是 C^p 光滑可逆的，其逆映射为 $\sigma(\cdot)$。固定一个向量 $v \in \mathbf{R}^d$。那么，对所有充分小的 $t > 0$，都存在唯一向量 $y(t) \in \mathbf{R}^d$ 满足

$$tv = \Phi(\sigma(tv)) + \nabla G(\sigma(tv))^\top y(t).$$

将方程 $0 = \Phi(\bar{x}) + \nabla G(\bar{x})^\top \bar{y}$ 两边相减，并投影到 $P_{T_{\mathcal{M}}(\bar{x})}$，得到

$$P_{T_{\mathcal{M}}(\bar{x})} v = P_{T_{\mathcal{M}}(\bar{x})} \left[\frac{\Phi(\sigma(tv)) - \Phi(\bar{x})}{t} \right] + P_{T_{\mathcal{M}}(\bar{x})} \left[\frac{\nabla G(\sigma(tv))^\top}{t} y(t) \right].$$

显然， $y(t)$是连续的，因此右侧在$P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) \nabla \sigma(0) v$ 处于 t 趋于零时收敛。综上，由于 v 是任意的，我们已证明矩阵恒等式

$$P_{T_{\mathcal{M}}(\bar{x})} = P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) \nabla \sigma(0).$$

考虑到 $\nabla\sigma(0)$的值域包含于 $T_{\mathcal{M}}(\bar{x})$，可知$P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$ 的值域必须等于 $T_{\mathcal{M}}(\bar{x})$。因此，$P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$ 在 $T_{\mathcal{M}}(\bar{x})$上必为非奇异，正如所述。

□

现在我们可以完成定理 2.7 的证明了。具体而言，引理 A.1与 $A(\cdot)$的连续性共同直接推出，在$(\bar{x},0)$附近， F 的图与下述映射的图重合

$$x \mapsto \Phi(x) + N_{\mathcal{M}}(x),$$

ˆ，其中我们令 $\Phi(x) = A(x) + \nabla f(x)$。引理 A.2 直接推出， F 在 $(0, \bar{x})$ 附近是 C^p-可逆的，且逆映射为 $\sigma(\cdot)$，当且仅当 $\Sigma = P_{T_{\mathcal{M}}(\bar{x})} \nabla_x \mathcal{H}(\bar{x}, \bar{y}) P_{T_{\mathcal{M}}(\bar{x})}$ 在$T_{\mathcal{M}}(\bar{x})$上非奇异。在这种情况下，等式 $\nabla\sigma(0) = \Sigma^\dagger$ 成立。

B 第 3 节中的证明

B.1 定理 3.1 的证明

定义随机向量 $w_k := A(x_k) - A_S(x_k)$，并定义事件

$$\mathcal{E}_k := \{x_k \in B_{\epsilon_2}(\bar{x})\} \qquad \text{and} \qquad \mathcal{Z}_k = \{w_k \in B_{\epsilon_1}(0)\}.$$

注意，根据我们的假设， $1_{\mathcal{E}_k} \xrightarrow{p} 1$ 。根据 x_k 的定义可知，包含关系$w_k \in (A + H)(x_k)$成立。因此，等式成立：

$$\sqrt{k}(x_k - \bar{x}) 1_{\mathcal{E}_k \cap \mathcal{Z}_k} = \sqrt{k}[\sigma(w_k 1_{\mathcal{E}_k \cap \mathcal{Z}_k}) - \sigma(0)]. \tag{B.1}$$

应用德尔塔方法（引理 E.2(7)）可知，只要 $\sqrt{k} w_k 1_{\mathcal{E}_k \cap \mathcal{Z}_k} \xrightarrow{D} \mathrm{N}(0, \Sigma)$ 对某个矩阵 Σ 成立，则 (B.1) 右端依分布收敛到 $\mathrm{N}(0, \nabla\sigma(0)\Sigma\nabla\sigma(0)^\top)$。因此，我们的任务归结为计算 $\sqrt{k} w_k 1_{\mathcal{E}_k \cap \mathcal{Z}_k}$ 的极限。

我们首先 展示$1_{\mathcal{Z}_k} \xrightarrow{p} 1$。围绕 $\bar{x}$ 对 $A(\cdot)$ 和 $A(\cdot, z_i)$ 作一阶展开，得到 ds

$$\begin{aligned} &\|w_k - \underbrace{(A(\bar{x}) - A_S(\bar{x}))}_{O_P(1/\sqrt{k})} + \underbrace{(\nabla A(\bar{x}) - \nabla A_S(\bar{x}))(x_k - \bar{x})}_{O_P(1/\sqrt{k})} \| 1_{\mathcal{E}_k} \\ &\le \tfrac{1}{2} \Big(\mathbb{E}L + \underbrace{\frac{1}{k} \sum_{i=1}^{k} L(z_i)}_{\mathbb{E}L + O_P(1/\sqrt{k})} \Big) \|x_k - \bar{x}\|^2 1_{\mathcal{E}_k}, \end{aligned} \tag{B.2}$$

其中，方括号中的结论由中心极限定理得出。重新整理后，得到

$$\|w_k 1_{\mathcal{E}_k}\| \le \epsilon_2^2 \mathbb{E}L + O_P(1/\sqrt{k}) \le \frac{\epsilon_1}{2} + O_P(1/\sqrt{k}).$$

由此得证 $1_{\mathcal{Z}_k \cap \mathcal{E}_k} \xrightarrow{p} 1$。考虑到 $1_{\mathcal{E}_k} \xrightarrow{p} 1$，我们推出 $1_{\mathcal{Z}_k} \xrightarrow{p} 1$，如所述。

接下来，我们证明 $\|x_k - \bar{x}\| = O_P(1/\sqrt{k})$。注意，(B.1) 直接推出 $\|x_k - \bar{x}\| 1_{\mathcal{E}_k \cap \mathcal{Z}_k} \le \mathrm{lip}(\sigma) \|w_k 1_{\mathcal{E}_k \cap \mathcal{Z}_k}\|$. 将此与 (B.2) 结合，并将等式两边乘以 $1_{\mathcal{Z}_k}$，我们得到

$$\mathrm{lip}(\sigma)^{-1} \|x_k - \bar{x}\| 1_{\mathcal{E}_k \cap \mathcal{Z}_k} = O_P(1/\sqrt{k}) + (\mathbb{E}L + O_P(1/\sqrt{k})) \|x_k - \bar{x}\|^2 1_{\mathcal{E}_k \cap \mathcal{Z}_k}.$$

整理后可得

$$(\mathrm{lip}(\sigma)^{-1} - \mathbb{E}L \|x_k - \bar{x}\|) \cdot \|x_k - \bar{x}\| 1_{\mathcal{E}_k \cap \mathcal{Z}_k} = O_P(1/\sqrt{k}).$$

注意到在 $1/2$ 事件$\mathcal{E}_k$中，左侧的系数有下界 $\|x_k - \bar{x}\| 1_{\mathcal{E}_k \cap \mathcal{Z}_k} = O_P$，我们推出 $1/\sqrt{k}$($1_{\mathcal{E}_k \cap \mathcal{Z}_k} \xrightarrow{p} 1$)。考虑到 $\|x_k - \bar{x}\| = O_P$，我们推出$1/\sqrt{k}$()，如所述。

最后，将 (B.2) 两边乘以 $\sqrt{k}$ ，得到

$$\sqrt{k} w_k 1_{\mathcal{E}_k} = \underbrace{\sqrt{k}(A(\bar{x}) - A_S(\bar{x})) 1_{\mathcal{E}_k}}_{\xrightarrow{D} \mathsf{N}(0, \mathrm{Cov}(A(\bar{x}, z))} + \underbrace{(\nabla A(\bar{x}) - \nabla A_S(\bar{x})) \sqrt{k}(x_k - \bar{x}) 1_{\mathcal{E}_k}}_{\xrightarrow{p} 0}$$
$$+ \underbrace{\frac{1}{2} \left(\mathbb{E}L(z) + \frac{1}{k} \sum_{i=1}^{k} L(z_i) \right) O(\sqrt{k} \|x_k - \bar{x}\|^2) 1_{\mathcal{E}_k}}_{\xrightarrow{p} 0},$$

其中，所声称的极限由中心极限定理以及 $\sqrt{k}\|x_k - \bar{x}\|$ 依概率有界这一事实推出。两边乘以 $1_{\mathcal{Z}_k}$，我们便得到 $\sqrt{k} w_k 1_{\mathcal{E}_k \cap \mathcal{Z}_k} \xrightarrow{D} \mathsf{N}\big(0, \mathrm{Cov}(A(\bar{x}, z))\big)$。因此，回到 (B.1) 并应用德尔塔方法，可知 $\sqrt{k}(x_k - \bar{x}) 1_{\mathcal{E}_k \cap \mathcal{Z}_k} \xrightarrow{D} \mathsf{N}\big(0, \nabla\sigma(0) \mathrm{Cov}(A(\bar{x}, z)) \nabla\sigma(0)^\top\big)$。最后需要指出，$\sqrt{k}(x_k - \bar{x}) 1_{\mathcal{E}_k \cap \mathcal{Z}_k}$ 与 $\sqrt{k}(x_k - \bar{x})$ 具有相同的极限分布，因为 $1_{\mathcal{E}_k \cap \mathcal{Z}_k} \xrightarrow{p} 1$。

B.2 定理 3.2 的证明

本节中，我们始终假设假设 A 和假设 B 成立。现在我们证明，式 (3.1) 中的渐近协方差在所有用于估计 $\bar{x}$ 的估计量中是最优的。具体而言，我们将对任何用于寻找均衡点的估计程序的性能给出下界，其中目标问题受到一系列对抗性选取的小扰动。为定义这一序列，先定义集合

$$\mathcal{G} := \{ g \colon \mathcal{Z} \to \mathbf{R}^d : \mathop{\mathbb{E}}_{z \sim \mathcal{P}} [g(z)] = 0, \ \mathop{\mathbb{E}}_{z \sim \mathcal{P}} \|g(z)\|^2 < \infty \}.$$

现在固定一个函数 $g \in \mathcal{G}$ 和一个任意的 C^3-光滑函数 $h\colon \mathbb{R} \to [-1,1]$，使其前三阶导数有界，且对所有 $t \in [-1/2, 1/2]$，$h(t) = t$。现在对每个 $u \in \mathbb{R}^d$，定义一个新的概率分布 $\mathcal{D}^u$，其密度由下式给出

$$d\mathcal{P}_u(z) := \frac{1 + h\big(u^\top g(z)\big)}{C(u)}\, d\mathcal{P}(z), \tag{B.3}$$

其中 $C(u)$ 是归一化常数 $C(u) := 1 + \int h\big(u^\top g(z)\big)\, d\mathcal{P}(z)$。因此，每个向量 $u \in \mathbf{R}^d$ 都会诱导出扰动问题

$$0 \in L(x,u) + H(x) \qquad \text{where} \qquad L(x,u) = \mathop{\mathbb{E}}_{z\sim\mathcal{P}_u} A(x,z). \tag{B.4}$$

令人放心的是，下面的引理表明，映射 $(x,u) \mapsto L(x,u)$ 在 $(\bar{x},0)$ 附近 C^1。

引理 B.1. 映射 $(x,u) \mapsto L(x,u)$ 在 $(\bar{x},0)$ 附近 C^1，其偏导数为

$$\nabla_x L(\bar{x},0) = \nabla A(\bar{x}) \qquad \textit{and} \qquad \nabla_u L(x,0) = \mathop{\mathbb{E}}_{z\sim\mathcal{P}} A(\bar{x},z) g(z)^\top.$$

证明。我们先考虑归一化常数 $C(u) = 1 + \int h\big(u^\top g(z)\big)\, d\mathcal{P}(z)$。控制收敛定理[3] 直接推出，$C(\cdot)$ 二阶可微，且 $\nabla C(u) = \int h'\big(u^\top g(z)\big) g(z)^\top\, d\mathcal{P}(z)$，以及 $\nabla^2 C(u) = \int h''\big(u^\top g(z)\big) g(z) g(z)^\top\, d\mathcal{P}(z)$。此外，控制收敛定理[4] 推出，$A(x)$ 是 C^1-光滑的，且 $\nabla A(x) = \mathbb{E}_{z\sim\mathcal{P}} \nabla A(x,z)$。因此，现在只需证明 $\hat{L}(x,u) := \int h(u^\top g(z)) A(x,z)\, d\mathcal{P}(z)$ 是 C^1-光滑的。对 u 中应用控制收敛定理，直接推出 $L(x,u)$ 关于 u 可微，且 $\nabla_u \hat{L}(x,u) = \int h'(u^\top g(z)) A(x,z) g(z)^\top\, d\mathcal{P}(z)$；此外，$\nabla_u \hat{L}(x,u)$ 关于 (x,u) 连续。类似地，控制收敛定理[5] 推出，$\hat{L}(x,u)$ 关于 x 可微，且 $\nabla_x \hat{L}(x,u) = \int h(u^\top g(z)) \nabla A(x,z)\, d\mathcal{P}(z)$，并且 $\nabla_x \hat{L}(x,u)$ 关于 (x,u) 连续。因此，$L(\cdot,\cdot)$ 在 $(\bar{x},u)$ 附近是 C^1-光滑的。注意，表达式 $\nabla_x L(\bar{x},0) = \nabla A(\bar{x})$ 显然成立，因为对所有 x，$L(x,0) \equiv A(x)$。为得到 $\nabla_u L(x,0)$ 的表达式，注意 $\nabla C(0) = 0$ 且 $\nabla^2 C(0) = 0$，因此 $C(u) = 1 + o(\|u\|^2)$。由此立即可得，对所有 x，$\nabla_u \hat{L}(x,0) = \nabla_u L(x,0)$，证明完毕。

□

如果这些 B.4 问题的解在 u 中变化剧烈，那么这一系列问题就没有多大用处。相反，下面的引理表明，对所有足够小的 u，每个 B.4 问题在 U 中都有唯一解，而且该解在 u 中平滑变化。我们将使用以下标准记号。如果集值映射 $F(\cdot)$ 在点对 $(\bar{u}, \bar{v})$ 附近的局部化为映射 $\sigma(\cdot)$，且二者的图像 $\operatorname{gph}\sigma$ 和 $\operatorname{gph} F$ 在 $(\bar{u}, \bar{v})$ 附近局部重合，则称 $\sigma(\cdot)$ 是集值映射 $F(\cdot)$ 在点对 $(\bar{u}, \bar{v})$ 附近的局部化。

引理 B.2（解映射的导数）。解映射

$$S(u) = \{x : 0 \in L(x,u) + H(x)\}.$$

存在一个单值局部化 $s(\cdot)$，其在$(0,\bar{x})$ 附近，并且在0 处可微，雅可比矩阵为

$$\nabla s(0) = -\nabla\sigma(0) \cdot \mathop{\mathbb{E}}_{z\sim\mathcal{P}} [A(\bar{x},z) g(z)^\top].$$

[3]利用 h' 和 h'' 有界，且 $\mathbb{E}_{z\sim\mathcal{P}}\|g(z)\|^2 < \infty$[4]利用存在 $\bar{x}$ 的一个邻域 U，使得 $\sup_{x\in U}\|\nabla A(x,\ z)\|^2$ 可积。[5] 利用 h 有界，且存在 $\bar{x}$ 的一个邻域 U，使得 $\sup_{x\in U}\|\nabla A(x,\ z)\|^2$ 可积。

证明。假设 A 保证映射 $A+H$ 在 $(\bar{x}, 0)$ 附近可逆，且其逆映射为 $\sigma(\cdot)$。现在定义 A 在 $\bar{x}$ 处的线性化 $\Psi(x) := A(\bar{x}) + \nabla A(\bar{x})(x-\bar{x})$。由 [11, 定理 2B.10], 可知，映射 $\Psi + H$ 在 $(0, \bar{x})$ 附近也可逆，其逆映射为 $\hat{}\sigma$ ，并满足 $\nabla\hat{}\sigma(0) = \nabla\sigma(0)$。注意，根据引理 B.1，我们也可以将 Ψ 等价地写为 $\Psi(x) := L(\bar{x}, 0) + \nabla_x L(\bar{x}, 0)(x-\bar{x})$。应用 [11, 定理 2D.6],可知，映射 $S(u)$ 存在一个单值局部化 $s(\cdot)$，其在 $(0, \bar{x})$ 附近，并且在 0 处可微，且满足 $\nabla s(0) = -\nabla\hat{}\sigma(0) \circ \nabla_u L(\bar{x}, 0)$。应用引理 B.1 即完成证明。

□

根据引理 B.2，对于所有足够小的 u，我们定义解 $\bar{x}_u := s(u)$。本节其余部分将用于证明以下定理。我们用 $\mathbb{E}_{P_u^k}$ 表示关于 k 独立同分布观测 $z_i \sim \mathcal{P}_u$的期望。

定理 B.3 (局部极小极大)。 设 $\mathcal{L}\colon \mathbb{R}^d \to [0,\infty)$ 为对称、拟凸且下半连续的，设 $\widehat{x}_k\colon \mathcal{Z}^k \to U$ 为估计量序列，并令 $g(x,z) := A(\bar{x},z) - A(x)$。则不等式

$$\lim_{c\to\infty} \liminf_{k\to\infty} \sup_{\|u\|\le\sqrt{c/k}} \mathbb{E}_{P_{k,u}}\big[\mathcal{L}\big(\sqrt{k}(\widehat{x}_k - \bar{x}_u)\big)\big] \ge \mathbb{E}\big[\mathcal{L}(Z)\big] \tag{B.5}$$

成立，其中 $Z \sim \mathsf{N}\big(0, \nabla\sigma(0)\cdot \mathsf{Cov}(A(\bar{x},z))\cdot \nabla\sigma(0)^\top\big)$。

定理 B.3 直接蕴含定理 3.2。事实上，正如 [15, 第 3 节2], 中所指出的，对于任意给定的 $C>0$，存在 $C>0$，使得条件 $\|u\| \le \sqrt{C/k}$ 对所有充分大的 k均蕴含 $\mathcal{P}_u \in B_{c/k}$。

B.2.1 定理 B.3 的证明

定理 B.3 的证明将基于 H´ajek 和 Le Cam 的局部极小极大定理，例如见 [19, 定理 6.6.2]。该定理依赖于两个概念：（1）分布序列的局部渐近正态性；（2）映射序列的正则性。由于稍后将说明的原因，第一个概念的确切定义对我们并不重要。然而，第二个性质将发挥关键作用。

定义 B.4 （正则映射序列）。 设 $W \subset \mathbb{R}^n$ 是零点的一个邻域。若映射序列 $\{\Gamma_k\colon W \to \mathbb{R}^d\}$ 满足

$$\lim_{k\to\infty} \sqrt{k}\big(\Gamma_k(u) - \Gamma_k(0)\big) = Du \qquad \textit{for all } u \in W$$

现在我们可以陈述局部极小极大定理。

定理 B.5。 设 $\{\Omega_k, \mathcal{F}_k, Q_{k,u}\}_{u\in\mathbb{R}^d}$ 是精度为 $Q \succeq 0$ 的局部渐近正态族，且 $\{\Gamma_k\colon \mathbb{R}^d \to \mathbb{R}^d\}$ 是导数为 D 的正则序列。设 $\mathcal{L}\colon \mathbb{R}^d \to [0,\infty)$ 是对称、拟凸且下半连续的。那么，对于任意估计量序列$T_k\colon \Omega_k \to \mathbb{R}^d$，我们有

$$\sup_{U_0\subset\mathbb{R}^d, |U_0|<\infty} \liminf_{k\to\infty} \max_{u\in U_0} \mathbb{E}_{Q_{k,u}}\big[\mathcal{L}\big(\sqrt{k}\big(T_k - \Gamma_k(u)\big)\big)\big] \ge \mathbb{E}\big[\mathcal{L}(Z)\big], \tag{B.6}$$

其中，当 $Z \sim \mathsf{N}\big(0, DQ^{-1}D^\top\big)$ 可逆时， Q 成立；如果 Q 是奇异的，则对所有 $\lambda > 0$, (B.6) 成立，其中 $Z \sim \mathsf{N}\big(0, D(Q+\lambda I)^{-1}D^\top\big)$。

我们旨在如下应用定理 B.5。对于每个 u，我们将 $(\Omega_k, \mathcal{F}_k, Q_{k,u})$ 取为概率空间 $(\Omega\ \mathcal{F}, \mathcal{P}_{u/\sqrt{k}})$ 的 k 重乘积，并令 $\Gamma_k(u) = s(u/\sqrt{k})$。引理 8.3] 已证明，序列 $\{\Omega_k, \mathcal{F}_k, Q_{k,u}\}_{u\in\mathbf{R}^d}$ 是局部渐近正态的，精度为 $D = \mathop{\mathbb{E}}_{z\sim\mathcal{P}}\big[g(z)g(z)^\top\big]$. 。此外，以下引理确立了序列 Γ_k 的正则性。

引理 B.6. 序列 $\Gamma_k\colon \mathbb{R}^d \to \mathbb{R}^d$ 是正则的，其导数为 $-\nabla\sigma(0)\cdot\mathbb{E}_{z\sim\mathcal{P}}[g(z)A(\bar{x}, z)^\top]$。

证明。 利用引理 B.2，对 $s(\cdot)$ 在 $\bar{x}$ 处作一阶展开，得到

$$\sqrt{k}(\Gamma_k(u) - \Gamma_k(0)) = \sqrt{k}(s(u/\sqrt{k}) - \bar{x}) = -\nabla\sigma(0) \cdot \mathop{\mathbb{E}}_{z\sim\mathcal{P}}[g(z)A(\bar{x}, z)^\top]u + \frac{o(k^{-1/2})}{k^{-1/2}}.$$

令 k 趋于无穷，即完成证明。 □

现在应用定理 B.5。设 $\mathcal{L}\colon \mathbb{R}^d \to [0,\infty)$ 对称、拟凸且下半连续， $\widehat{x}_k\colon \mathcal{Z}^k \to \mathbb{R}^d$ 为估计量序列，并且 $c > 0$。令

$$\begin{aligned} g(z) &:= A(\bar{x}, z) - A(\bar{x}), \\ \Sigma &:= \mathop{\mathbb{E}}_{z\sim\mathcal{P}}\big[g(z)g(z)^\top\big], \\ W &:= \nabla\sigma(0). \end{aligned}$$

为将不等式 (B.6) 与 (B.5) 联系起来，注意到对于任意有限子集 $U_0 \subset \mathbb{R}^d$，我们有

$$\liminf_{k\to\infty} \sup_{\|u\|\le\sqrt{c/k}} \mathbb{E}_{P_{k,u}}\big[\mathcal{L}\big(\sqrt{k}(\widehat{x}_k - \bar{x}_u)\big)\big] \ge \liminf_{k\to\infty} \max_{u\in U_0} \mathbb{E}_{Q_{k,u}}\big[\mathcal{L}\big(\sqrt{k}\big(\widehat{x}_k - \Gamma_k(u)\big)\big)\big].$$

其中 $c = \max_{u\in U_0}\|u\|^2$。对所有有限的 $U_0 \subset \mathbb{R}^d$ 取上确界，并应用定理 B.5，得到

$$\lim_{c\to\infty} \liminf_{k\to\infty} \sup_{\|u\|\le\sqrt{c/k}} \mathbb{E}_{P_{k,u}}\big[\mathcal{L}\big(\sqrt{k}(\widehat{x}_k - \bar{x}_u)\big)\big] \ge \mathbb{E}\big[\mathcal{L}(Z_\lambda)\big] \tag{B.7}$$

其中，对任意 $\lambda > 0$， $Z_\lambda \sim \mathsf{N}\big(0, W\Sigma(\Sigma + \lambda I)^{-1}\Sigma^\top W^\top\big)$。由基本线性代数可知

$$\lim_{\lambda\downarrow 0} \Sigma\big(\Sigma + \lambda I\big)^{-1}\Sigma = \Sigma.$$

因此，基于单调收敛定理的直接论证（参见例如 [6, Section 5.1.2]）可推出，当 $\lambda \downarrow 0$ 时，(B.7) 右侧趋于 $\mathbb{E}\big[\mathcal{L}(Z)\big]$，其中 $Z \sim \mathsf{N}(0, W\Sigma W^\top)$。证明完毕。

C 第4节中的证明

C.1 引理 4.2 的证明

在整个证明过程中，设 $\epsilon > 0$ 和 L 满足

$$\max\{\|s_g(x)\|, \|A(x)\|\} \le L \qquad \forall x \in B_\epsilon(\bar{x}).$$

为证明断言 (1)，对于所有足够接近 $\bar{x}$ 的 x，我们计算

$$\begin{aligned}\alpha\|G_\alpha(x,\nu)\| &= \|x - s_f(x-\alpha(A(x)+s_g(x)+\nu))\| \\ &\le \operatorname{dist}_{\mathcal{X}}(x-\alpha(A(x)+s_g(x)+\nu)) + \alpha\|A(x)+s_g(x)+\nu\| \\ &\le 2\alpha\|A(x)+s_g(x)+\nu\| \\ &= 2\alpha(2L+\|\nu\|).\end{aligned}$$

在证明的其余部分，我们设 $x_+ := x - \alpha G_\alpha(x,\ \nu)$。现在验证命题 2。为此，假设 f 是凸的，并且通过增大 L，我们可以确保对于所有 $x \in B_\epsilon(\bar{x}) \cap \operatorname{dom} f.$，都有 $\operatorname{dist}(0,\ \partial f(x))\} \le L$。取一个长度最小的向量 $z \in \partial f(x)$。代数运算表明 $x = \operatorname{prox}_{\alpha f}(x+\alpha z)$。由于 $\operatorname{prox}_{\alpha f}$ 是非扩张的，对于所有 $x \in B_\epsilon(\bar{x})$，我们可推出

$$\begin{aligned}\alpha\|G_\alpha(x,\nu)\| &= \|x - \operatorname{prox}_{\alpha f}(x-\alpha(A(x)+s_g(x)+\nu))\| \\ &\le \alpha\|A(x)+s_g(x)+z+\nu\| \\ &\le \alpha(3L+\|\nu\|),\end{aligned}$$

如所断言。

接下来验证命题 3。为此，假设 f 在 $\operatorname{dom} g \cap \operatorname{dom} f$ 上是 L-利普希茨连续的。设 $w := A(x) + s_g(x) + \nu$，并注意到 x_+ 的定义直接保证

$$\begin{aligned}\frac{1}{2\alpha}\|x_+ - x\|^2 &\le f(x) - f(x_+) - \langle w, x_+ - x\rangle \\ &\le L\|x_+ - x\| + \|w\|\|x_+ - x\|.\end{aligned}$$

因此，对于所有 $x \in B_\epsilon(\bar{x})$，我们有 $\alpha^{-1}\|x_+ - x\| \le 2(L + \|w\|) \le 2(3C + \|\nu\|)$，正如所需。

C.2 引理 4.3 的证明

强 (a)-正则性蕴含以下等式：$P_{T_{\mathcal{M}}(x)}\partial f(x) = \{\nabla_{\mathcal{M}} f(x)\}$ 以及 $P_{T_{\mathcal{M}}(x)}\partial g(x) = \{\nabla_{\mathcal{M}} g(x)\}$，对所有接近 $x^\star$ 的 $x \in \mathcal{M}$ 均成立。所述表达式 (4.5) 立即成立。接下来，引理 A.1 表明，$\operatorname{gph} \partial(f+g)$ 与 $\operatorname{gph}\ [\nabla_{\mathcal{M}}(f+g) + N_{\mathcal{M}}]$ 在 $(x^\star, -A(x^\star))$ 附近重合。因此，在 $(x^\star, 0)$ 附近局部有以下等式成立：

$$\operatorname{gph}[A + \partial(f+g)] = \operatorname{gph}[A + \nabla_{\mathcal{M}} f + \nabla_{\mathcal{M}} g + N_{\mathcal{M}}] = \operatorname{gph}(F_{\mathcal{M}} + N_{\mathcal{M}}),$$

如所述。

C.3 命题 4.4 的证明

证明。 固定 $x \in \mathcal{X}$、$\alpha > 0$ 和 $\nu \in \mathbf{R}^d$。假设 D 显然成立，例如可由引理 4.2(1)推出。为了验证假设 (E1)，我们计算

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}\ (G_\alpha(x,\nu) - F(P_{\mathcal{M}}(x)) - \nu)\| = \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(A(x) + s_g(x) - F(P_{\mathcal{M}}(x)))\|.$$

因此，当 x 充分接近 $\bar{x}$ 时，我们可以将右侧上界估计为

$$\begin{aligned}&\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}[s_g(x)-\nabla_{\mathcal{M}}g(P_{\mathcal{M}}(x))]+P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}[A(x)-A(P_{\mathcal{M}}(x))]\|\\&\leq C\cdot\operatorname{dist}(x,\mathcal{M}),\end{aligned}$$

其中，不等式由 (F1) 以及 $A(\cdot)$ 的局部利普希茨连续性得出。因此，假设 (E1) 成立。最后，为了验证假设 (E2)，我们计算

$$\begin{aligned}\langle G_\alpha(x,\nu)-\nu,x-P_{\mathcal{M}}(x)\rangle&=\langle A(\bar{x})+s_g(x),x-P_{\mathcal{M}}(x)\rangle+\langle A(x)-A(\bar{x}),x-P_{\mathcal{M}}(x)\rangle\\&\geq\mu\cdot\operatorname{dist}(x,\mathcal{M})-\|A(x)-A(\bar{x})\|\cdot\operatorname{dist}(x,\mathcal{M})\\&\geq\tfrac{\mu}{2}\cdot\operatorname{dist}(x,\mathcal{M}),\end{aligned}$$

对所有充分接近 $\bar{x}$ 的 x 都成立。证明完成。 □

C.4 命题 4.6 的证明

取足够小的 $\epsilon>0$，使得对所有 $x\in B_\epsilon(\bar{x})\cap\mathcal{X}$，以下条件均成立。首先，(4.7) 成立。特别地，由于 $A(\cdot)$ 在 $\bar{x}$ 附近局部利普希茨连续，我们可以确定

$$\langle A(x)+v,x-P_{\mathcal{M}}(x)\rangle\geq\frac{\mu}{2}\cdot\operatorname{dist}(x,\mathcal{M}),\tag{C.1}$$

对所有 $v\in\partial g(x)$ 均成立。其次，我们要求存在某个 $L>0$，使得

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(s_g(x)-\nabla_{\mathcal{M}}g(P_{\mathcal{M}}(x))\|\leq L\cdot\operatorname{dist}(x,\mathcal{M}),\tag{C.2}$$

$$\|P_{T_{\mathcal{M}}(z)}(u)\|\leq L\|x-z\|,\tag{C.3}$$

对所有单位范数的 $u\in N_{\mathcal{X}}(x)$ 以及所有 $z\in B_\epsilon(\bar{x})\cap\mathcal{M}$ 均成立，这是 (G1) 的推论。第三，我们可以将 $\epsilon>0$ 取足够小，使得

$$\langle z,x-x'\rangle\geq o(\|x-x'\|)\tag{C.4}$$

对于所有 $z\in N_{\mathcal{X}}(x)$ 的单位范数，以及 $x'\in\mathcal{M}\cap B_\epsilon(\bar{x})$——这是 (G3) 的一个推论。我们将固定 $x\in B_{\epsilon/2}(\bar{x})\cap\mathcal{X}$ 和任意的 $\alpha>0$ 及 $\nu\in\mathbb{R}^d$，并选择任意的 $y\in P_{\mathcal{M}}(x)$。定义

$$w=G_\alpha(x,\nu)-\nu-A(x)-s_g(x)\qquad\text{and}\qquad x_+=s_{\mathcal{X}}(x-\alpha(A(x)+s_g(x)+\nu)).$$

注意包含关系 $w\in N_{\mathcal{X}}(x^+)$。此外，缩小 $\epsilon>0$ 假设 D 可直接推出

$$\max\{\|w\|,\|G_\alpha(x,\nu)\|\}\leq C(1+\|\nu\|)\qquad\text{and}\qquad\|x_+-x\|\leq C(1+\|\nu\|)\alpha,$$

对于某个常数 $C>0$。我们将在证明中经常使用这些估计。最后，我们令 C 是一个与 x,α 和 ν 无关的常数，并且其值随行变化。假设 (E1)：首先假设 $x_+\in B_\epsilon(\bar{x})$。利用 (C.3)，我们计算得

$$\begin{aligned}\|P_{T_{\mathcal{M}}(y)}w\|&\leq L\|w\|\|x_+-y\|\\&\leq L\|w\|(\|x_+-x\|+\operatorname{dist}(x,\mathcal{M}))\\&\leq C(1+\|\nu\|)^2\alpha+C(1+\|\nu\|)\operatorname{dist}(x,\mathcal{M}).\end{aligned}\tag{C.5}$$

另一方面，若 $x_+ \notin B_\epsilon(\bar{x})$，则我们计算得

$$\|P_{T_{\mathcal{M}}(y)} w\| \le \|w\| \le \frac{2}{\epsilon}\|w\|\|x_+ - x\| \le C(1+\|\nu\|)^2\alpha. \tag{C.6}$$

无论哪种情况，假设 (E1) 现在都成立，因为由 (C.2) 可得

$$\begin{aligned}\|P_{T_{\mathcal{M}}(y)}(A(x)+s_g(x)-F(y)\| &\le \|P_{T_{\mathcal{M}}(y)}(\nabla_{\mathcal{M}} g(y) - s_g(x))\| \\ &\quad + \|P_{T_{\mathcal{M}}(y)}[A(x)-A(y)]\| \\ &\le C\mathrm{dist}(x,\mathcal{M}),\end{aligned}$$

这正是我们要证明的。

假设 (E2)：我们将其分解写为

$$\langle G_\alpha(x,\nu)-\nu, x-y\rangle = \underbrace{\langle A(x)+s_g(x), x-y\rangle}_{R_1} + \underbrace{\langle w, x_+ - y\rangle}_{R_2} + \underbrace{\langle w, x-x_+\rangle}_{R_3}. \tag{C.7}$$

估计式 (C.1) 给出

$$R_1 \ge \tfrac{\mu}{2}\cdot \mathrm{dist}(x,\mathcal{M}). \tag{C.8}$$

接下来分两种情况讨论。首先，设 $x_+ \in B_\epsilon(\bar{x})$。利用包含关系 $w \in N_{\mathcal{X}}(x_+)$ 和 (C.4)，我们计算得

$$\begin{aligned}R_2 &\ge \|w\|\cdot o(\|x_+ - y\|) \ge \|w\|\cdot(o(\|y-x\|) - \|x-x_+\|) \\ &\ge -C(1+\|\nu\|)^2(o(\mathrm{dist}(x,\mathcal{M})) + \alpha).\end{aligned} \tag{C.9}$$

接下来，由柯西–施瓦茨不等式可得

$$|R_3| \le \|w\|\|x-x_+\| \le C(\alpha(1+\|\nu\|)^2). \tag{C.10}$$

结合 (C.7)–(C.10) 可得所述界 (E2)。

现在反设 $x_+ \notin B_\epsilon(\bar{x})$，因此 $\|x-y\| \le \|x-x_+\|$。由此我们推得 $R_2 + R_3 = \langle w, x-y\rangle \ge -\|w\|\|x-y\| \ge -C\alpha(1+\|\nu\|)^2$ 成立。将此估计与 (C.7) 和 (C.8) 结合，即可验证断言 (E2)。

C.5 命题 4.8 的证明

取足够小的 $\epsilon \in (0,\ 1)$，使得对所有 $x \in B_\epsilon(\bar{x}) \cap \mathrm{dom}\ f$，下述条件均成立。首先，(4.8) 成立，因此考虑到 $A(\cdot)$ 的利普希茨连续性，我们可以等价地写为

$$\langle A(x)+v, x - P_{\mathcal{M}}(x)\rangle \ge \frac{\mu}{2}\cdot\mathrm{dist}(x,\mathcal{M}) - (1+\|v\|)o(\mathrm{dist}(x,\mathcal{M})), \tag{C.11}$$

对所有 $v \in \partial f(x)$。其次，我们要求对于某个 $L>0$，有

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(u - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x))\| \le L\sqrt{1+\|u\|^2}\cdot\mathrm{dist}(x,\mathcal{M}) \tag{C.12}$$

对所有 $u \in \partial f(x)$，这是强 (a) 正则性的结果。第三，我们假设 $\nabla_{\mathcal{M}} f$ 在 $B_\epsilon(\bar{x}) \cap \mathcal{M}$上是$L$-利普希茨连续的。第四，我们假设 $A(\cdot)$ 是 L-利普希茨连续的。缩小 ϵ 后，我们还可以假设 $\epsilon \leq \frac{\mu}{8L}$。最后，我们也可以假设映射 $x \mapsto P_{\mathcal{M}}(x)$ 和$x \mapsto P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}$ 在 $B_\epsilon(\bar{x})$ 上是利普希茨连续的。

固定 $x \in B_{\epsilon/2}(\bar{x}) \cap \mathrm{dom}\, f$ and $\nu \in \mathbb{R}^d$ and，设 $y := P_{\mathcal{M}}(x)$。我们定义这些向量

$$w = G_\alpha(x,\nu) - A(x) - \nu \qquad \text{and} \qquad x_+ = s_f(x - \alpha(A(x)+\nu)).$$

ˆ 简单的代数运算表明包含关系 $w \in \partial f(x_+)$。此外，假设 D 直接推出

$$\max\{\|w\|, \|G_\alpha(x,\nu)\|\} \leq C(1+\|\nu\|) \qquad \text{and} \qquad \|x_+ - x\| \leq C(1+\|\nu\|)\alpha.$$

我们将在证明中经常使用这些估计。最后，令 C 为一个与 x, α and ν 无关的常数，该常数会随推导步骤而改变。

假设 (E1)：首先，假设 $x_+ \in B_\epsilon(\bar{x})$。利用三角不等式，我们写出

$$\begin{aligned}
&\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\alpha(x,\nu) - F(P_{\mathcal{M}}(x)) - \nu)\| \\
&= \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(w + A(x) - A(P_{\mathcal{M}}(x)) - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x)))\| \\
&\leq \underbrace{\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(w - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x_+)))\|}_{R_1} + \underbrace{\|A(x) - A(P_{\mathcal{M}}(x))\|}_{R_2} + \underbrace{\|\nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x))) - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x_+)))\|}_{R_3}.
\end{aligned}$$

利用三角不等式和估计式 (C.12)，我们得到

$$\begin{aligned}
R_1 &\leq \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x_+))}(w - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x_+)))\| + \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x_+))} - P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}\|_{\mathrm{op}} \cdot \|w - \nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x_+))\| \\
&\leq C(1+\|w\|) \cdot \mathrm{dist}(x_+, \mathcal{M}) + C\|x - x_+\| \cdot (1+\|w\|) \\
&\leq C(1+\|\nu\|)\mathrm{dist}(x_+, \mathcal{M}) + C(1+\|\nu\|)^2\alpha \\
&\leq C(1+\|\nu\|)(\mathrm{dist}(x, \mathcal{M}) + C\|x - x_+\|) + C(1+\|\nu\|)^2\alpha \\
&\leq C(1+\|\nu\|)\mathrm{dist}(x, \mathcal{M}) + C(1+\|\nu\|)^2\alpha.
\end{aligned}$$

此外，显然我们有 $R_2 \leq C\mathrm{dist}(x, \mathcal{M})$ 和 $R_3 \leq C\|x - x_+\| \leq (1+\|\nu\|)\alpha.$。条件 (E1) 立即成立。

现在假设 $x_+ \notin B_\epsilon(\bar{x})$，因此 $\|x_+ - x\| \geq \epsilon/2$。那么，我们可以写成

$$\begin{aligned}
\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\alpha(x,\nu) - \nu - \nabla f_{\mathcal{M}}(P_{\mathcal{M}}(x)))\| &\leq \|G_\alpha(x,\nu)\| + \|\nu\| + \|\nabla f_{\mathcal{M}}(P_{\mathcal{M}}(x))\| \\
&\leq \frac{2}{\epsilon}(\|G_\alpha(x,\nu)\| + \|\nu\| + \|\nabla f_{\mathcal{M}}(P_{\mathcal{M}}(x))\|)\|x - x_+\| \\
&\leq C(1+\|\nu\|)^2\alpha,
\end{aligned}$$

这正是所需的结果。

假设 (E2)：我们从以下分解开始

$$\begin{aligned}
\langle G_\alpha(x,\nu) - \nu, x - y\rangle = &\underbrace{\langle A(x_+) + w, x_+ - P_{\mathcal{M}}(x_+)\rangle}_{R_1} \\
&+ \underbrace{\langle A(x) - A(x_+), x - y\rangle}_{R_2} + \underbrace{\langle A(x_+) + w, (x - P_{\mathcal{M}}(x)) - (x_+ - P_{\mathcal{M}}(x_+))\rangle}_{R_3}.
\end{aligned}$$

现在我们对 $x_+ \in B_\epsilon(\bar{x})$ 情形下右侧的两项进行界定。利用 (C.11)，我们估计

$$\begin{aligned} R_1 &\geq \tfrac{\mu}{2} \cdot \operatorname{dist}(x_+, \mathcal{M}) - (1 + \|v\|) o(\operatorname{dist}(x_+, \mathcal{M})) \\ &\geq \tfrac{\mu}{2} \cdot (\operatorname{dist}(x, \mathcal{M}) - \|x - x_+\|) - (1 + \|v\|)(o(\operatorname{dist}(x, \mathcal{M})) + \|x - x_+\|) \\ &\geq \tfrac{\mu}{2} \cdot \operatorname{dist}(x, \mathcal{M}) - (1 + \|\nu\|)^2 (o(\operatorname{dist}(x, \mathcal{M})) + C\alpha). \end{aligned}$$

接下来，我们计算

$$|R_2| \leq \|A(x) - A(x_+)\| \cdot \operatorname{dist}(x, \mathcal{M}) \leq 2L\epsilon \cdot \operatorname{dist}(x, \mathcal{M}) \leq \frac{\mu}{4} \operatorname{dist}(x, \mathcal{M}).$$

接下来，利用映射 $I - P_{\mathcal{M}}$ 在 $B_\epsilon(\bar{x})$ 上的利普希茨连续性，我们推得

$$|R_3| \leq C\|A(x_+) + w\| \cdot \|x - x_+\| \leq C(1 + \|\nu\|)^2 \alpha.$$

将 μ 替换为 $\mu/4$，即可立即得到所声称的近端瞄准条件。

现在来看 $x_+ \notin B_\epsilon(\bar{x})$ 的情形，因此距离 $\operatorname{dist}(x, \mathcal{M}) \leq \frac{\epsilon}{2} \leq \|x - x^+\|$。于是我们计算

$$\begin{aligned} \langle G_\alpha(x, \nu) - \nu, x - P_{\mathcal{M}}(x) \rangle &\geq -\operatorname{dist}(x, \mathcal{M}) \cdot \|G_\alpha(x, \nu) - \nu\| \\ &= \operatorname{dist}(x, \mathcal{M}) - \operatorname{dist}(x, \mathcal{M})(1 + \|G_\alpha(x, \nu) - \nu\|) \\ &\geq \operatorname{dist}(x, \mathcal{M}) - C\|x - x_+\|(1 + C(1 + \|\nu\|)) \\ &\geq \operatorname{dist}(x, \mathcal{M}) - C(1 + \|\nu\|)^2 \alpha, \end{aligned}$$

正如所需。证明完毕。

D 定理 5.1 的证明

现在概述证明中使用的常见记号和约定。令 $\mathcal{U}$ 表示邻域 $\bar{x}$，在该邻域内所有基本假设均成立。缩小 $\mathcal{U}$ 后，不妨假设投影映射 $P_{\mathcal{M}}$ 在 $\mathcal{U}$ 中为 C^2，特别地，$P_{\mathcal{M}}$ 是具有 Lipschitz雅可比矩阵的 Lipschitz 映射。在整个证明过程中，我们会在需要时多次缩小 $\mathcal{U}$。

现在，将停时 (6.1) 记为 $\tau := \tau_{k_0,\ \delta}$ 并将噪声界记为 $Q := \sup_{x \in B_\delta(\bar{x})} q(x)$。注意，根据命题 6.2，影子序列 y_k 满足 $y_k \in B_{4\delta}(x_k) \cap \mathcal{M} \subseteq B_\epsilon(\bar{x}) \cap \mathcal{M}$，并且递推关系 (6.2) 成立。此外，令 C 表示一个依赖于 k_0 和 δ 的常数，该常数可能在不同的推导步骤中发生变化。

D.1 强局部极小值点附近的改进速率

首先，我们在增长条件 (5.1) 下得到改进的收敛速率。为此，我们首先需要以下引理，以确保 $F_{\mathcal{M}}$ 在$B_{2\delta}(\bar{x})$中具有足够的曲率。

引理D.1（曲率）。估计式

$$\langle F_{\mathcal{M}}(y), y - \bar{x} \rangle \geq \frac{\mu}{2} \|y - \bar{x}\|^2,$$

对所有充分接近 $\bar{x}$ 的 $x \in \mathcal{M}$ 都成立。

证明。令 Φ 为 $F_{\mathcal{M}}$ 在 $U \subset$的邻域内的一个光滑延拓$\mathbf{R}^d$。考虑任意一个收敛到 $\bar{x}$ 的序列 $x_i \in \mathcal{M}$。取一个子序列后，不妨设单位向量 $\frac{x_i-\bar{x}}{\|x_i-\bar{x}\|}$ 收敛到某个单位向量 $w \in T_{\mathcal{M}}($$\bar{x}$)。令 $H_i := \int_0^1 \nabla\Phi(\overline{\bar{x}+\tau(}x_i-\bar{x}))\, d\tau$ 表示 $\bar{x}$ 与 x_i之间的平均雅可比矩阵。注意，当 i 趋于无穷时， H_i 显然趋于$\nabla\Phi($$\bar{x}$)。由微积分基本定理可得

$$\frac{\langle \Phi(x_i)-\Phi(\bar{x}), x_i-\bar{x}\rangle}{\|x_i-\bar{x}\|^2} = \left\langle H_i\left(\frac{x_i-\bar{x}}{\|x_i-\bar{x}\|}\right), \frac{x_i-\bar{x}}{\|x_i-\bar{x}\|}\right\rangle \underset{i\to\infty}{\to} \langle \nabla\Phi(\bar{x})w, w\rangle \geq \mu.$$

由于 $x_i \in \mathcal{M}$ 是任意一个收敛到 $\bar{x}$ 的序列，因此结论成立 s. □

接下来，我们给出一个熟知的单步改进保证。

引理 D.2 (单步改进)。 对于所有充分小的 δ，存在一个常数C，使得对于任意 $k \geq k_0$，我们有

$$\mathbb{E}[\|y_{k+1}-\bar{x}\|^2 1_{\tau>k}] \leq \left(1-\frac{\alpha_k\mu}{2}\right)\mathbb{E}[\|y_k-\bar{x}\|^2 1_{\tau>k}] + C\alpha_k^2. \tag{D.1}$$

证明。 展开 $\|y_{k+1}-\bar{x}\|^2$，可得

$$\begin{aligned}
&\|y_{k+1}-\bar{x}\|^2 1_{\tau>k}\\
&= \left\|y_k-\alpha_k F_{\mathcal{M}}(y_k)-\alpha_k P_{T_{\mathcal{M}}(y_k)}(\nu_k)+\alpha_k E_k-\bar{x}\right\|^2 1_{\tau>k}\\
&= \|y_k-\alpha_k F_{\mathcal{M}}(y_k)+\alpha_k E_k-\bar{x}\|^2 1_{\tau>k} + \alpha_k^2\left\|P_{T_{\mathcal{M}}(y_k)}(\nu_k)\right\|^2 1_{\tau>k}\\
&- 2\alpha_k\left\langle y_k-\alpha_k F_{\mathcal{M}}(y_k)+\alpha_k E_k-\bar{x}, P_{T_{\mathcal{M}}(y_k)}(\nu_k)\right\rangle 1_{\tau>k}\\
&= \underbrace{\|y_k-\alpha_k F_{\mathcal{M}}(y_k)-\bar{x}\|^2 1_{\tau>k}}_{P_1} + \alpha_k^2\|E_k\|^2 1_{\tau>k} + 2\alpha_k\underbrace{\langle y_k-\alpha_k F_{\mathcal{M}}(y_k)-\bar{x}, E_k\rangle 1_{\tau>k}}_{P_2}\\
&+ \alpha_k^2\left\|P_{T_{\mathcal{M}}(y_k)}(\nu_k)\right\|^2 1_{\tau>k} - 2\alpha_k\underbrace{\left\langle y_k-\alpha_k F_{\mathcal{M}}(y_k)+\alpha_k E_k-\bar{x}, P_{T_{\mathcal{M}}(y_k)}(\nu_k)\right\rangle 1_{\tau>k}}_{P_3}.
\end{aligned} \tag{D.2}$$

利用引理D.1，我们可以将 P_1 上界为

$$\begin{aligned}
P_1 &= \left(\|y_k-\bar{x}\|^2 - 2\alpha_k\langle F_{\mathcal{M}}(y_k), y_k-\bar{x}\rangle + \alpha_k^2\|\nabla f_{\mathcal{M}}(y_k)\|^2\right)1_{\tau>k}\\
&\leq \left((1-\alpha_k\mu)\|y_k-\bar{x}\|^2 + C\alpha_k^2\right)1_{\tau>k}
\end{aligned}$$

接下来，利用命题 6.2 (2(a)i) 和假设 I，我们看到 $\mathbb{E}_k[\|E_k\|^2 1_{\tau>k}]$ 和$\mathbb{E}_k[\|P_{T_{\mathcal{M}}(y_k)}($$\nu_k$)$\|^2 1_{\tau>k}]$ 均被一个数值常数所界。还需对 P_2 和 P_3 作界。先考虑前者，利用杨氏不等式，我们计算得

$$P_2 \leq \frac{\mu\|y_k-\alpha_k F_{\mathcal{M}}(y_k)-\bar{x}\|^2 1_{\tau>k}}{8} + \frac{2\|E_k\|^2 1_{\tau>k}}{\mu} = \frac{\mu P_1}{8} + \frac{2\|E_k\|^2 1_{\tau>k}}{\mu}.$$

接下来，再次利用 杨氏不等式，我们对 P_3 的条件期望进行如下估计 :

$$\mathbb{E}_k[P_3] = \alpha_k\mathbb{E}_k[\left\langle E_k, P_{T_{\mathcal{M}}(y_k)}(\nu_k)\right\rangle 1_{\tau>k}] \leq \frac{\alpha_k\mathbb{E}_k\|E_k\|^2 1_{\tau>k}}{2} + C\alpha_k.$$

因此，回到引理 D.2 并利用命题 6.2(2(a)ii)，我们得到如下估计:

$$\mathbb{E}[\|y_{k+1}-\bar{x}\|^2 1_{\tau>k}] \leq (1-\alpha_k\mu/2)\mathbb{E}[\|y_k-\bar{x}\|^2 1_{\tau>k}] + C\alpha_k^2.$$

证明至此完成。 □

接下来，我们可以对递推关系进行迭代，以确保 $\|y_k - \bar{x}\|$ 快速收敛。由此我们还可以得到误差 E_k 大小的估计。为简化记号，我们记作 $\tau_{k_0} := \tau_{k_0,\,\delta}$，因为我们将考虑 k_0 的多个取值。按照这些约定，我们有如下命题，它将有助于确保某些序列的可求和性。

引理 D.3. 存在 $C > 0$ 使得

1. $\mathbb{E}[\|y_k - \bar{x}\|^2 1_{\tau_{k_0}>k}] \le C/k^\gamma$ 对所有 $k \ge 1$均成立。2. $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|y_k - \bar{x}\|^2 < \infty$ 几乎处处成立。3. $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \|y_k - \bar{x}\|^2 \to 0$ 几乎处处成立。4. $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|E_k\| < +\infty$ 几乎处处成立。5. $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \|E_k\| < +\infty$ 几乎处处成立。

证明。 将 $s_k = \mathbb{E}[\|y_k - \bar{x}\|^2 1_{\tau_{k_0}>k}]$ 代入引理 D.2 和 E.9，即可得出第 1 部分。现在证明第 2 部分。根据第 1 部分，我们有

$$\mathbb{E}\left[\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|y_k - \bar{x}\|^2 1_{\tau_{k_0}>k}\right] \le \sum_{k=1}^{\infty} \frac{C}{k^{\gamma+\frac{1}{2}}} < \infty.$$

因此，$\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|y_k - \bar{x}\|^2 1_{\tau_{k_0}>k}$ 几乎处处有限。考虑到 $x_k \to \bar{x}$几乎处处成立，和式 $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|y_k - \bar{x}\|^2$ 必定几乎处处有限。第 3 部分现在可由克罗内克引理 E.6 立即得出。

接下来，我们证明第 4 部分。由命题 6.2(2(a)iii)，我们知道误差序列 E_k几乎处处满足

$$\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|E_k\| 1_{\tau_{k_0}>k} < +\infty.$$

由于 $x_k \to \bar{x}$ 几乎处处成立，我们推知几乎处处有 $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \|E_k\| < +\infty$，正如所需。第 5 部分由克罗内克引理 E.6 得出。 □

D.2 定理 5.1 证明的完成

现在我们转向 定理 5.1 的证明。为此，我们引入一个额外的序列

$$z_k := P_{\bar{x}+T_{\mathcal{M}}(\bar{x})}(y_k). \tag{D.3}$$

显然，对于所有足够小的 δ， z_k 都能很好地近似 y_k。事实上，由于 $\mathcal{M}$ 的光滑性，存在 $C > 0$ 使得

$$\|y_k - z_k\| 1_{\tau_{k_0}>k} \le C \|y_k - \bar{x}\|^2 1_{\tau_{k_0}>k}. \tag{D.4}$$

接下来 结果表明，只需研究 $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (z_k - \bar{x})$ 的分布。

引理 D.4 (化归为一个辅助序列)。 如果 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(z_k-\bar{x})$ 依分布收敛到某个 $\mathcal{D}$，那么 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(x_k-\bar{x})$ 依分布收敛到 $\mathcal{D}$。

证明。 注意到

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(x_k-\bar{x})=\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(z_k-\bar{x})+\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(x_k-y_k)+\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(y_k-z_k).$$

根据引理 E.2 (4)，只要证明

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(x_k-y_k)\to 0 \qquad \text{and} \qquad \frac{1}{\sqrt{n}}\sum_{k=1}^{n}(y_k-z_k)\to 0,$$

几乎处处成立，结论便得证。为此，我们回顾命题 6.1(1b)保证几乎处处有

$$\sum_{k=1}^{\infty}\frac{1}{\sqrt{k}}\left\|x_k-y_k\right\|1_{\tau_{k_0}>k}<+\infty$$

由于 $x_k\to\bar{x}$ 几乎处处成立，对于几乎每条样本路径，我们都可以找到一个 k_0 使得 $\tau_{k_0}=\infty$。因此，几乎处处有 $\sum_{k=1}^{\infty}\frac{\|x_k-y_k\|}{\sqrt{k}}<\infty$. 应用克罗内克引理 E.6，几乎处处有

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\left\|x_k-y_k\right\|\to 0,$$

这意味着 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(x_k-y_k)\to 0$。另一方面，根据引理 D.3 和不等式 (D.4)，我们有

$$\sum_{k=1}^{\infty}\frac{1}{\sqrt{k}}\left\|y_k-z_k\right\|1_{\tau_{k_0}>k}\le\sum_{k=1}^{\infty}\frac{C}{\sqrt{k}}\left\|y_k-\bar{x}\right\|^2 1_{\tau_{k_0}>k}<+\infty$$

同样，由于对于几乎每条样本路径，我们都可以找到 k_0 使得 $\tau_{k_0}=\infty$，因此有 $\sum_{k=1}^{\infty}\frac{1}{\sqrt{k}}\left\|y_k-z_k\right\|<+\infty$，如所需。 □

根据引理 D.4，现在只需计算 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(z_k-\bar{x})$ 的极限分布。这正是以下引理的内容。请注意，在这些引理中，我们以一种与定理 5.1 中所述形式等价、但更便于计算的形式给出了渐近协方差矩阵。

引理 D.5. 序列 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(z_k-\bar{x})$ 依分布收敛到

$$N\left(0,U(U^{\top}\nabla F_{\mathcal{M}}(\bar{x})U)^{-1}\cdot\Sigma\cdot(U^{\top}\nabla F_{\mathcal{M}}(\bar{x})U)^{-1}U^{\top}\right),$$

其中 U 是一个矩阵，其列向量构成 $T_{\mathcal{M}}(\bar{x}$的标准正交基).

证明。回想一下，$UU^\top$是到 $T_{\mathcal{M}}(\bar{x})$上的正交投影。因此，我们可以写作$z_k = \bar{x} + UU^\top($ $y_k - \bar{x})$。此外，从（6.2）两边减去 $\bar{x}$并乘以$U^\top$，可得

$$\begin{aligned} U^\top(y_{k+1} - \bar{x}) &= U^\top(y_k - \bar{x}) - \alpha_k U^\top F_{\mathcal{M}}(y_k) - \alpha_k U^\top P_{T_{\mathcal{M}}(y_k)}(\nu_k) + \alpha_k U^\top E_k \\ &= U^\top(y_k - \bar{x}) - \alpha_k U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) UU^\top(y_k - \bar{x}) \\ &\quad - \alpha_k (U^\top F_{\mathcal{M}}(y_k) - U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) UU^\top(y_k - \bar{x})) \\ &\quad - \alpha_k U^\top P_{T_{\mathcal{M}}(\bar{x})}(\nu_k) - \alpha_k (U^\top P_{T_{\mathcal{M}}(y_k)}(\nu_k) - U^\top P_{T_{\mathcal{M}}(\bar{x})}(\nu_k)) + \alpha_k U^\top E_k. \end{aligned}$$

定义 $\Delta_k = U^\top(y_k - \bar{x})$、 $H = U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) U$, $\zeta_k = U^\top P_{T_{\mathcal{M}}(y_k)}(\nu_k) - U^\top P_{T_{\mathcal{M}}(\bar{x})}(\nu_k)$，以及

$$R(y) = U^\top F_{\mathcal{M}}(y) - U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) UU^\top(y - \bar{x}).$$

根据我们的假设，对于每个向量 z，矩阵 H 满足

$$\langle Hz, z\rangle = \langle \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) Uz, Uz\rangle \geq \sigma \|Uz\|^2 = \sigma \|z\|^2.$$

因此， H 是一个强单调矩阵。此外，注意等式 $U^\top P_{T_{\mathcal{M}}(\bar{x})}(\nu_k) = U^\top UU^\top \nu_k = U^\top \nu_k$。因此，我们可以将 Δ_k 的更新重写为

$$\Delta_{k+1} = \Delta_k - \alpha_k H \Delta_k - \alpha_k U^\top \nu_k - \alpha_k \left(R(y_k) + \zeta_k - U^\top E_k \right).$$

在证明的其余部分，我们证明 $\frac{1}{\sqrt{n}} \sum_{k=1}^n \Delta_n$ 依分布收敛到$N\left(0, (U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) U)^{-1} \Sigma (U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x}) U)^{-1}\right)$。根据定义，$\frac{1}{\sqrt{n}} \sum_{k=1}^N (z_k - \bar{x}) = \frac{1}{\sqrt{n}} \sum_{k=1}^n U \Delta_k$。这就得出了该结论。我们注意到，我们的证明与 [1, 定理 2]的证明非常相似。

为证明这一断言，定义矩阵

$$B_k^n = \alpha_k \sum_{i=k}^n \prod_{j=k+1}^i (I - \alpha_j H) \qquad \text{and} \qquad A_k^n = B_k^n - H^{-1}.$$

Polyak 和 Juditsky [30, 引理 2] 表明， $\bar{\Delta}_n = \frac{1}{n} \sum_{k=1}^n \Delta_k$ 满足等式

$$\begin{aligned} \sqrt{n}\bar{\Delta}_n &= \frac{1}{\sqrt{n}} \sum_{k=1}^n H^{-1} U^\top \nu_k \\ &\quad + \frac{1}{\sqrt{n}} \sum_{k=1}^n A_k^n U^\top \nu_k + \frac{1}{\sqrt{n}} \sum_{k=1}^n B_k^n [R(y_k) + \zeta_k - U^\top E_k] + O\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

其中 $\sup_{k,n} \max\{\|B_k^n\|, \|A_k^n\|\} < +\infty$且 $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \|A_k^n\| = 0$。等价地，展开 $\nu_k = \nu_k^{(1)} + \nu_k^{(2)}(x_k)$ 后，我们得到

$$\begin{aligned} \sqrt{n}\bar{\Delta}_n &= \frac{1}{\sqrt{n}} \sum_{k=1}^n H^{-1} U^\top \nu_k^{(1)} \\ &\quad + \frac{1}{\sqrt{n}} \sum_{k=1}^n A_k^n U^\top \nu_k^{(1)} + \frac{1}{\sqrt{n}} \sum_{k=1}^n B_k^n [R(y_k) + \zeta_k - U^\top E_k + \nu_k^{(2)}(x_k)] + O\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

假设 J 确保总和 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} H^{-1}U^\top \nu_k^{(1)}$ 依分布收敛至

$$N\left(0, (U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x})U)^{-1}\Sigma(U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(\bar{x})U)^{-1}\right).$$

因此，只要能证明表达式 $\sqrt{n}\Delta_n$ 中的其他各项和几乎处处收敛于 0，断言便成立。我们将通过以下一系列断言来证明这一点。

断言 1。我们有

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k^{(1)} \xrightarrow{\text{a.s.}} 0.$$

证明。利用 $\sup_{k,n}\|A_k^n\| < \infty$ 以及 $\mathbb{E}\|\nu_k\|^2 1_{\tau_{k_0}>k}$ 有界这一事实，我们推出

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k^{(1)}\right\|^2 1_{\tau_{k_0}>k}\right] &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[\left\|A_k^n U^\top \nu_k^{(1)} 1_{\tau_{k_0}>k}\right\|^2\right] \\ &\le \frac{C}{n}\sum_{k=1}^{n}\|A_k^n\| \\ &\to 0.\end{aligned}$$

因此，$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k 1_{\tau_{k_0}>k}$ 是一个 L^2-有界鞅。由 [17, 定理 4.4.6], 可知 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k 1_{\tau_{k_0}>k} \xrightarrow{L^2} 0$。另一方面，根据 [17, 定理 4.2.11], $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k 1_{\tau_{k_0}>k}$ 几乎必然收敛。因此，由于对几乎每条样本路径，都存在 k_0 使得 $\tau_{k_0} = \infty$，我们有 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n U^\top \nu_k \xrightarrow{\text{几乎处处}} 0$，正如所需。 □

断言 2. 我们有

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} B_k^n U^\top R(y_k) \xrightarrow{\text{a.s.}} 0.$$

证明。令 Φ 为 $F_{\mathcal{M}}$ 在 $\bar{x}$ 的邻域 $U \subset \mathbf{R}^d$ 上的光滑延拓。于是我们可推出

$$\begin{aligned}R(y) &= U^\top(\Phi(y) - \Phi(\bar{x}) - \nabla\Phi(\bar{x})UU^\top(y-\bar{x})) \\ &= U^\top\nabla\Phi(\bar{x})(I - UU^\top)(y-\bar{x}) + O(\|y-\bar{x}\|^2) \\ &= U^\top\nabla\Phi(\bar{x})P_{N_{\mathcal{M}}(\bar{x})}(y-\bar{x}) + O(\|y-\bar{x}\|^2)\end{aligned}$$

由于 $\mathcal{M}$ 是 C^2-光滑的，立即可得，当 $y \in \mathcal{M}$ 趋于 $\bar{x}$ 时，$\|P_{N_{\mathcal{M}}(\bar{x})}(y-\bar{x})\| \le O(\|y-\bar{x}\|^2)$。因此，我们有 $\|R(y)\| = O(\|y-\bar{x}\|^2)$。此外，根据假设 $x_k \xrightarrow{\text{几乎处处}} \bar{x}$，我们有 $y_k \xrightarrow{\text{几乎处处}} \bar{x}$。因此，存在一个依赖于样本路径的常数 C，使得 $\|R(y_k)\| \le C\|y_k - \bar{x}\|^2$ 几乎处处成立。B_k^n 的一致有界性与引理 D.3 因此意味着 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} B_k^n U^\top R(y_k) \xrightarrow{\text{几乎处处}} 0$。 □

断言 3. 我们有

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} B_k^n \zeta_k \xrightarrow{\text{a.s.}} 0.$$

证明。对于 $k \geq 1$，定义截断变量 $\zeta^{(k_0)}{}_k = \zeta_k 1_{\tau_{k_0}>k}$。注意，只需证明

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} B_k^n \zeta_k^{(k_0)} \xrightarrow{\text{a.s.}} 0,$$

因为对于每条样本路径，都存在一个 k_0 使得 $\tau_{k_0} = \infty$。因此，我们将在整个过程中使用这些截断变量。

转向证明，我们首先证明 $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \zeta^{(k_0)}{}_k \xrightarrow{P} 0$ 和 $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} A_k^n \zeta^{(k_0)}{}_k \xrightarrow{P} 0$。回想一下，$\zeta_k = U^\top P_{T_{\mathcal{M}}(y_k)}(\nu_k) - U^\top P_{T_{\mathcal{M}}(\bar{x})}(\nu_k)$，因此我们有

$$\mathbb{E}\left[\zeta_k^{(k_0)} \mid \mathcal{F}_k\right] = \mathbb{E}\left[\zeta_k \mid \mathcal{F}_k\right] 1_{\tau_{k_0}>k} = 0.$$

由于 $x \mapsto P_{T_{\mathcal{M}}(x)}$ 在 $\mathcal{M}$ 中的 $\bar{x}$ 的某个邻域上局部 Lipschitz 连续，因此对于某个 $C > 0$ 以及所有充分小的 δ，有如下界：

$$\left\|\zeta_k^{(k_0)}\right\|^2 \leq C \left\|y_k - \bar{x}\right\|^2 1_{\tau_{k_0}>k},$$

特别地，有

$$\mathbb{E}\left[\left\|\zeta_k^{(k_0)}\right\|^2 \mid \mathcal{F}_k\right] \leq C^2 \left\|y_k - \bar{x}\right\|^2 1_{\tau_{k_0}>k}.$$

结合引理 D.3(1)，我们知道 $\zeta_k^{(k_0)}$ 是一个鞅差序列，并且几乎处处，

$$\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}\left[\left\|\zeta_k^{(k_0)}\right\|^2 \mid \mathcal{F}_k\right] \leq C^2 \sum_{k=1}^{\infty} \frac{1}{k} \left\|y_k - \bar{x}\right\|^2 < \infty.$$

因此，由引理 E.5 可得

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \zeta_k^{(k_0)} \xrightarrow{\text{a.s.}} 0.$$

特别地，有 $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \zeta^{(k_0)}{}_k \xrightarrow{P} 0.$

接下来，我们证明对于任意 $k_0 < \infty$，都有 $n^{-1/2} \sum_{k=1}^{n} A_k^n \zeta^{(k_0)}{}_k \xrightarrow{P} 0$。为此，注意到根据引理 D.3，存在 $C' > 0$ 使得

$$\mathbb{E}\left[\left\|\zeta_k^{(k_0)}\right\|^2\right] \leq C\mathbb{E}\left[\left\|y_k - \bar{x}\right\|^2 1_{\tau_{k_0}>k}\right] \leq C'\alpha_k. \tag{D.5}$$

因此，以下极限成立

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}} \sum_{k=1}^{n} A_k^n \zeta_k^{(k_0)}\right\|^2\right] = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[\left\|A_k^n \zeta_k^{(k_0)}\right\|^2\right] \leq \frac{C'\alpha_k}{n} \sum_{k=1}^{n} \left\|A_k^n\right\|^2 \leq \frac{C'\alpha_k \sup_{k,n} \left\|A_k^n\right\|}{n} \sum_{k=1}^{\infty} \left\|A_k^n\right\| \to 0,$$

其中，第一个等式由鞅差性质得出，第二个不等式则由 $\zeta_k^{(k_0)}$ 的矩有界性得出。因此，我们已经证明

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n \zeta_k^{(k_0)} \xrightarrow{L^2} 0,$$

这意味着 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k^n \zeta^{(k_0)}{}_k \xrightarrow{P} 0$。

因此，我们已经证明 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} B_k^n \zeta^{(k_0)}{}_k \xrightarrow{P} 0.$。现在证明 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} B_k^n \zeta_k^{(k_0)}$ 几乎必然收敛。由于几乎处处极限和依概率极限在二者均存在时相同，这就完成了证明。

为此，定义序列

$$Z_{n,k_0} = \sum_{k=1}^{n} B_k^n \zeta_k^{(k_0)}.$$

如果我们能证明，对于任意有限的 k_0，序列$n^{-1/2} Z_{n,k_0}$ 几乎必然收敛，则结论成立。为此，注意到 $B_k^{n+1} - B_k^n = \alpha_k \prod_{i=k+1}^{n}(I - \alpha_i H)$。因此，定义

$$W_k^n = \prod_{i=k}^{n}(I - \alpha_i H), \qquad V_{n,k_0} = \sum_{k=1}^{n} \alpha_k W_{k+1}^{n+1} \zeta_k^{(k_0)},$$

we 推得 V_{n,k_0} 是 $\mathcal{F}_{n+1}$ 可测的，并且 Z_{n,k_0} 具有如下分解 :

$$Z_{n,k_0} = Z_{n-1,k_0} + V_{n-1,k_0} + \alpha_n \zeta_n^{(k_0)} = \sum_{k=1}^{n-1} V_{k,k_0} + \sum_{k=1}^{n} \alpha_k \zeta_k^{(k_0)}.$$

注意，和式 $\sum_{k=1}^{n} \alpha_k \zeta_k^{(k_0)}$ 是一个增量平方可求和的平方可积鞅，因此根据定理 4.2.11], 它几乎必然收敛 [17, 。由此，我们得到如下极限 $n^{-1/2}\sum_{k=1}^{n} \alpha_k \zeta_k^{(k_0)} \xrightarrow{\text{a.s.}} 0$。因此，只需证明 $n^{-1/2}\sum_{k=1}^{n-1} V_{k,k_0}$几乎必然收敛。

为此，令 λ 表示 H的最小特征值。于是我们有

$$\mathbb{E}\left[\|V_{n,k_0}\|^2\right] = \sum_{k=1}^{n} \alpha_k^2 \left\|W_{k+1}^{n+1}\right\|^2 \mathbb{E}\left[\left\|\zeta_k^{(k_0)}\right\|^2\right] \le C' \sum_{k=1}^{n} \alpha_k^3 \left\|W_{k+1}^{n+1}\right\|^2, \tag{D.6}$$

其中，不等式由界 $\mathbb{E}\left[\left\|\zeta^{(k_0)}{}_k\right\|^2\right] \le C'\alpha_k$ 得出（见公式（D.5））。引理 1（第 3 部分） [, 表明，存在常数 $\beta > 0$ 和 $K < \infty$，使得对于所有 k 和 $t \ge k$，以下估计成立:

$$\left\|W_{k+1}^{n+1}\right\|^2 \le K \exp\left(-\beta \sum_{i=k+1}^{n} \alpha_i\right).$$

将此估计代入（D.6），完全按照 [1, 引理 A.7] with$\rho = 3$ 的证明可知，存在某个常数 C，使得

$$\mathbb{E}\left[\|V_{n,k_0}\|^2\right] \le \frac{C \log n}{n^{2\gamma}}.$$

因此，对于任意 $\epsilon > 0$，我们都可以找到某个 C，使得

$$\mathbb{E}\left[\|V_{n,k_0}\|^2\right] \leq \frac{C}{n^{2\gamma-\epsilon}}.$$

现在定义 $T_{n,\ k_0} = \frac{1}{\sqrt{n}}\sum_{k=1}^{n} V_{k,\ k_0}$. 我们断言，$T_{n,\ k_0}$ 几乎处处具有有限长度。事实上，对于任意 $\epsilon > 0$，存在 C，$C' > 0$ 使得

$$\begin{aligned}
\mathbb{E}\left[\|T_{n,k_0} - T_{n+1,k_0}\|\right] &\leq \left|\frac{1}{\sqrt{n+1}} - \frac{1}{\sqrt{n}}\right| \sum_{k=1}^{n} \mathbb{E}\left[\|V_{k,k_0}\|\right] + \frac{1}{\sqrt{n+1}}\mathbb{E}\left[\|V_{n+1,k_0}\|\right] \\
&\leq \frac{C}{n^{\frac{3}{2}}} \sum_{k=1}^{n} \frac{1}{k^{\gamma-\epsilon}} + \frac{1}{\sqrt{n}} \frac{1}{n^{\gamma-\epsilon}} \\
&\leq \frac{C'}{n^{\gamma+1/2-\epsilon}}.
\end{aligned}$$

由于 $\gamma \in (\frac{1}{2}, 1)$，因此我们有 $\sum_n \mathbb{E}\left[\|T_{n,k_0} - T_{n+1,k_0}\|\right] < \infty$。于是，该和几乎处处有限：$\sum_n \|T_{n,\ k_0} - T_{n+1,\ k_0}\| < +\infty$。这意味着 $T_{n,\ k_0} = n^{-1/2}\sum_{k=1}^{n} V_{k,\ k_0}$几乎必然收敛。回顾 $V_{k,\ k_0}$的定义，我们发现 $n^{-1/2}Z_{n,\ k_0}$ 几乎处处收敛，由此完成证明。

□

断言 4。我们有

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} B_k^n \nu_k^{(2)}(x_k) \xrightarrow{\text{a.s.}} 0.$$

证明。这可以通过与命题 2 类似的论证来证明。事实上，注意到序列 $\xi_k = \nu_k^{(2)}(x_k)1_{\tau > k_0}$ 是一个鞅差序列，这些界对某个$C > 0$

$$\mathbb{E}_k[\|\xi_k\|^2] \leq C\|x_k - \bar{x}\|^2 1_{\tau > k_0} \qquad \text{and} \qquad \mathbb{E}[\|\xi_k\|^2] \leq C\alpha_k,$$

以及 $\sum_{k=\mathfrak{T}}^{\infty} \frac{1}{k}\mathbb{E}_k[\|\xi_k\|^2] \leq \sum_{k=\mathfrak{T}}^{\infty} \frac{1}{k}\|x_k - \bar{x}\|^2 1_{\tau > k_0} < +\infty$成立。证明命题 2 时只用到了关于 $\zeta_k^{(k_0)}$ 的这些事实。 □

断言 5。我们有

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} B_k^n U^\top E_k \xrightarrow{\text{a.s.}} 0.$$

证明。这直接由引理 D.3(5 以及 $\sup_{k,\ n}\|B_k^n\| < \infty$ 这一事实得出。

□

考虑到这些断言，证明即告完成。 □

E 关于随机变量序列的补充事实。

定义 E.1. 设 $\{X_k\}_{k\geq 1}$ 和 X 是定义在概率空间 $(\Omega\ \mathcal{F},\ P)$ 上的随机向量，取值于$\mathbf{R}^d$。

1. X_k 几乎必然收敛到 X，记作 $X_k \xrightarrow{a.s.} X$，如果对于几乎所有 $\omega \in \Omega$，向量 $X_k(\omega)$ 收敛到 $X(\omega)$。

2. X_k 依概率收敛到 X，记作 $X_k \xrightarrow{p} X$，若对每个 $\epsilon > 0$，都有 $\lim_{k\to\infty} P(\|X_k - X\| \le \epsilon) \to 1$。

3. X_k 依分布收敛到 X，记作 $X_k \xrightarrow{D} X$，若对每个有界连续函数 $G: \mathbf{R}^d \to \mathbf{R}$，都有 $\lim_{k\to\infty} \mathbb{E}G(x_k) = \mathbb{E}G(X)$。

4. X_k 是依概率有界的，记作 $X_k = O_p(1)$，若对每个 $\epsilon > 0$，存在 M_ϵ使得对所有充分大的指标 k，都有 $P(\|X_n\| > M_\epsilon) < \epsilon$。

引理 E.2. 设 $\{X_k\}_{k\ge1}$、$\{Y_k\}_{k\ge1}$、X 和 Y 是某个欧几里得空间中的随机向量，且设 a、b 是确定性的。以下陈述成立。

1. 推论如下：

$$X_k \xrightarrow{a.s.} X \quad \Longrightarrow \quad X_k \xrightarrow{p} X \quad \Longrightarrow \quad X_k \xrightarrow{D} X \quad \Longrightarrow \quad X_k = O_p(1).$$

2. 若 $X_k \xrightarrow{p} X$ 且 $Y_k \xrightarrow{p} Y$，则 $aX_k + bY_k \xrightarrow{p} aX + bY$ 且 $X_kY_k \xrightarrow{p} XY$。几乎处处收敛也有类似的结论。

3. 若 $X_n = O_p$（1）且 $Y_k \xrightarrow{p} 0$，则 $X_kY_k \xrightarrow{p} 0$。

4.（斯卢茨基定理 *I*）若 $X_k \xrightarrow{D} X$ 且 $Y_k \xrightarrow{p} a$，则 $X_k + Y_k \xrightarrow{D} X + a$ 且 $X_kY_K \xrightarrow{D} aX$。

5.（斯卢茨基定理 *II*）若 $X_k \xrightarrow{p} X$ 且 $Y_k \xrightarrow{p} Y$，则 $X_k + Y_k \xrightarrow{p} X + Y$ 且 $X_kY_K \xrightarrow{p} XY$。

6. 若 $X_kY_k \xrightarrow{D} X$ 且 $Y_k \xrightarrow{p} c$，则 $X_k \xrightarrow{D} X/c$，只要 $c \neq 0$ 成立。

7.（德尔塔方法）假设对于某个 $\mu \in \boldsymbol{R}^d$ 中的 Σ 和某个矩阵 $\Sigma \in \boldsymbol{R}^{d\times d}$，有 $\sqrt{k}(X_k - \mu) \xrightarrow{d} \boldsymbol{N}(0$，则对于任意在 μ 处可微的映射 $g: \boldsymbol{R}^d \to \boldsymbol{R}^m$，我们有 $\sqrt{k}(g(X_k) - g(\mu)) \xrightarrow{d} \boldsymbol{N}(0, \nabla g(\mu)\Sigma\nabla g(\mu)^\top)$。

引理 E.3 (Robbins-Siegmund [31])。设 $A_k, B_k, C_k, D_k \ge 0$ 为适应于滤过 $\{\mathcal{F}_k\}$ 的非负随机变量，并满足

$$\mathbb{E}[A_{k+1} \mid \mathcal{F}_k] \le (1 + B_k)A_k + C_k - D_k.$$

于是，在事件 $\{\sum_k B_k < \infty, \sum_k C_k < \infty\}$ 上，存在一个随机变量 $A_\infty < \infty$，使得 $A_k \xrightarrow{a.s.} A_\infty$ 且 $\sum_k D_k < \infty$ 几乎处处成立。

引理 E.4 (条件 Borel-Cantelli 引理 [4])。设 $\{X_n\colon n \ge 1\}$ 是在概率空间 $(\Omega\ \mathcal{F}, \mathbb{P})$ 上定义的一列非负随机变量，且 $\{\mathcal{F}_n\colon n \ge 0\}$ 是一列 σ 的子代数。设 $M_n = \mathbb{E}[X_n \mid \mathcal{F}_{n-1}]$ 对 $n \ge 1$成立。若 $\{\mathcal{F}_n\colon n \ge 0\}$ 单调不减，即它是一个滤过，则 $\sum_{n=1}^\infty X_n < \infty$ 在 $\{\sum_{n=1}^\infty M_n < \infty\}$上几乎处处成立。

引理 E.5 ([10,练习 5.3。35])。 设 M_k 是一个适应于滤过$\{\mathcal{F}_k\}$ 的 L^2 鞅，且 $b_k \uparrow \infty$ 是一个正确定性序列。那么，若

$$\sum_{k\geq 1} b_k^{-2}\mathbb{E}\left[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}\right] < +\infty,$$

我们有 $b_n^{-1}M_n \xrightarrow{a.s.} 0$。

引理 E.6 （克罗内克引理）。 设 $\{x_k\}_k$ 是一个实数无限序列，使得级数 $\sum_{k=1}^{\infty} x_k$ 存在且有限。则对于任意发散正非减序列 $\{b_k\}$，我们有

$$\lim_{K\to\infty} \frac{1}{b_K}\sum_{k=1}^{K} b_k x_k = 0.$$

T以下三个引理的证明可在 [8, 第 A.7.2 节中找到].

引理 E.7。 固定 $k_0 \in \mathbb{N}$、$c > 0$ 和 $\gamma \in$ （1/2、 1]。设 $\{X_k\}$、$\{Y_k\}$ 和 $\{Z_k\}$ 是适应于滤过 $\{\mathcal{F}_k\}$ 的非负随机变量。假设满足以下关系：

$$\mathbb{E}[X_{k+1} \mid \mathcal{F}_k] \leq (1 - ck^{-\gamma})X_k - Y_k + Z_k \qquad for\ all\ k \geq k_0.$$

进一步假设，当 $\gamma = 1$ 时有 $c \geq 6$ 。定义常数 $a_k := {}^{k^{2\gamma-1}}\log^2(k+1)$ 。则存在一个随机变量 $V < \infty$ ，使得在事件 $\{\sum_{k=1}^{\infty} a_{k+1}Z_k < +\infty\}$ 上，以下结论成立：

1. 极限成立
$$a_k X_k \xrightarrow{a.s.} V.$$

2. 该和有限
$$\sum_{k=1}^{\infty} a_{k+1}Y_k < +\infty.$$

引理 E.8. 固定 $k_0 \in \mathbb{N}$、 c、 $C > 0$ 和 $\gamma \in (1/2、\ 1]$。假设 $\{s_k\}_k$ 是一个非负序列，满足

$$s_k \leq \frac{c}{12\gamma} \qquad and \qquad s_{k+1}^2 \leq s_k^2 - ck^{-\gamma}s_k + Ck^{-2\gamma}, \qquad for\ all\ k \geq k_0,$$

Then，存在一个常数 C_{ub} ，仅取决于 c、 C、 γ 和 k_0 ，使得

$$s_k \leq C_{ub}k^{-\gamma}, \qquad \forall k \geq 1.$$

引理 E.9. 固定 $k_0 \in \mathbb{N}$、 c、 $C > 0$ 和 $\gamma \in (1/2、\ 1]$。假设 $\{s_k\}_k$ 是一个非负序列，满足

$$s_{k+1} \leq (1 - ck^{-\gamma})s_k + Ck^{-2\gamma}, \qquad for\ all\ k \geq k_0,$$

进一步假设，当 $\gamma = 1$ 时 $c \geq 16$ 。那么，存在一个常数 C_{ub} ，仅取决于 c、 C、 γ 和 k_0 ，使得

$$s_k \leq C_{ub}k^{-\gamma}, \qquad \forall k \geq 1.$$