通过近端技术研究EM算法的收敛性

DOMINIKUS NOLL

摘要。我们将期望最大化算法表示为一种广义近端方法，并据此研究其收敛性。在诸如不完整数据对数似然函数具有可定义性（即满足o-极小结构理论意义下的可定义性）等自然假设下，我们研究迭代点的收敛性，而不仅仅是函数值的收敛性。

关键词： EM算法 · 可定义集 · o-极小结构 · 近端方法 ·Kullback-Leibler散度 ·信息几何

MSC 2020: 65K05 · 49J52 · 62D10 · 62B11

arXiv:2601.02252v1 [math.ST] 5 Jan 2026

1. 引言

EM算法保证不完全数据负对数似然单调下降[30], ，但迭代序列的收敛性可能以各种方式失效 [78]。在缺乏强制性的情况下，迭代点可能逃逸至无穷远，而函数值却仍然收敛。即使迭代点保持有界，它们仍可能不收敛、发生循环 [76], ，或生成一个聚点连续统。当迭代点趋近自然参数域的边界时，收敛分析会进一步复杂化，因为此处的似然通常表现不佳。

即使能够收敛，EM迭代点也可能不趋近于不完全数据负对数似然的全局最小值，而是趋向局部最小值、鞍点，甚至局部最大值 [53]。这是非凸优化中的一种众所周知的现象；在实践中，通常能够找到较好的局部极值就已经可以接受。

在本文中，我们允许最大似然问题包含参数约束。这为曲指数族建模提供了一种便捷方式；但约束也可以传达关于未知参数的先验知识，允许实现受限最大似然 [46, p. 191],[14, 70, 63, 71], 处理截断族 [47], ，使迭代点远离自然参数域的边界，或仅仅是为了强制其保持有界，参见例如 [48, 36, 56, 41]。尽管具有实践意义，约束也进一步复杂化了 EM 算法分析。

EM算法与近端点方法（PPM）之间存在相似之处，相关研究已注意到这一点 [75, 21, 22, 23, 25]。PPM中的二次惩罚项被基于Kullback-Leibler信息距离的正则项所取代。鉴于已有研究探讨了非凸情况下PPM的收敛性 [4, 5, 64, 40, 72, 68], ，因此这些技术或许也适用于Kullback-Leibler正则项。研究这种联系的另一个有力理由是：当EM算法包含约束时，PPM可以被视为EM算法的一个特例。

目前，我们重新审视这一研究方向，并引入一个新要素：要求不完全数据对数似然在 o-极小结构理论意义下具有可定义性 [32, 31, 33, 77], ——这一假设在实践中总是成立。我们的研究发现，Kullback-Leibler 距离仅具有部分正则化作用，且这种作用集中在线性子空间上；该子空间的维数取决于给定观测数据时，缺失数据的条件Fisher信息矩阵的秩。因此，即使满足可定义性，也只能推出迭代点在该子空间上的投影收敛。要确保整个 EM序列收敛，还需要其他条件

*图卢兹大学数学研究所，法国。1

。积极的一面是，这为我们解释某些 EM 实例为何无法收敛提供了线索。

从 [59], 将EM算法解释为在数据集与模型集之间进行交替Bregman投影的方法这一视角出发，可以进一步深入了解完整EM序列 $\theta^{(k)}$ 的收敛性。M步生成的EM迭代点 $\theta^{(k)}$ 与E步生成的迭代点 $\vartheta^{(k)}$ 同步产生，算法在二者之间交替进行。由于在相当一般的假设下（其中还包括可定义性），这些 $\vartheta^{(k)}$ 会收敛，因此这为推导 $\theta^{(k)}$的收敛性提供了新的契机。

尽管通常预期EM算法以线性速率收敛，但也有关于次线性速率的报告，见 [53, 第34]页。作为我们分析的一部分，对于某些 $0 < \rho < \infty$，我们获得了次线性最坏情况速率 $\|\theta^{(k)} - \theta^*\| = O(k^{-\rho})$，并且在相当自然的假设下可以对其进行验证。线性速率出现的情形也得到了精确界定。

近期关于EM算法收敛性的研究方法包括： [44, 42], 作者利用镜像下降来应用非线性优化的结果； [19], Polyak-Łojasiewicz不等式使作者能够在非参数设定下推导复杂度结果。由于Polyak-Łojasiewicz不等式是Kurdyka-Łojasiewicz不等式的一个特例，因此它也纳入了我们当前的分析，并对应线性收敛的情形。

本文结构如下。第 2 节回顾优化与可定义性理论中的相关事实。第 3 节回顾EM算法的设定，包括曲指数族的情形。第 4 节探讨Kullback-Leibler信息、其作为正则项的作用，以及它与缺失数据的Fisher信息之间的联系。第 5 节对条件族进行参数降维，揭示Kullback-Leibler正则项的部分性质。第 6 节证明部分正则化下的收敛性，并在第 6.2 节给出最坏情况速率。第 7 节将此结果应用于EM算法，并讨论将部分收敛提升为整个序列收敛的情形$\theta^{(k)}$。第 8 节将介绍交替Bregman投影。第 9 节讨论指数族之外的推广。第 10 节给出示例。

记号

对于函数 $\Psi(x, y)$，我们用 $\nabla_x \Psi$ 表示 $\nabla_1 \Psi$ 关于 x的导数，用 $\nabla_y \Psi$表示 $\nabla_2 \Psi$ 关于 y的导数。关于 x的二阶导数记为 $\nabla^2_{11} \Psi(x, y)$ 或 $\nabla^2_{xx} \psi(x, y)$；类似地，混合二阶导数记为 $\nabla^2_{22} \Psi(x, y)$ 或 $\nabla^2_{yy} \Psi(x, y)$，以及 $\nabla^2_{xy} \Psi(x, y)$。次微分按 [55, 69] 的意义理解，记为 $\partial \psi(x)$。对于函数 $\Psi(x, y)$，有 $\partial_1 \Psi(x, y) = \partial \Psi(\cdot, y)(x)$，以及 $\partial_2 \Psi(x, y) = \partial \Psi(x, \cdot)(y)$。$\mathbb{R}^n$ 上的欧几里得标量积和范数分别记为 $x \cdot y$和 $\|x\|$。欧几里得球记为 $B(x, \delta)$，集合 K的欧几里得 δ-邻域记为 $N(K, \delta)$。

2. 准备工作

本节回顾优化和可定义性理论中的一些事实。我们遵循如下约定：一般优化算法中的迭代点记为 x_k；而对于 EM 算法，迭代点是待估计的参数，因此称为 $\theta^{(k)}$。

2.1. 近端点算法。 对于适当的下半连续函数 $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$，经典的近端点方法通过以下方式生成迭代点 x_k

$$x_{k+1} \in \operatorname*{argmin}_{x \in \mathbb{R}^n} f(x) + \frac{1}{2\lambda_k} \|x - x_k\|^2, \tag{1}$$

其中 $\lambda_k > 0$ 且其中 $\|\cdot\|$ 是欧几里得范数 [51, 52]。对于凸 f，已知 [67, 35]当 f存在最小值时，序列 x_k 收敛到 $x^* \in \operatorname{argmin} f$ 当且仅当 $\sum_{j=1}^k \lambda_j \to \infty (k \to \infty)$，并且当 $\lambda_k \to \infty$、 [69, 35]、 [49, 时，收敛速度甚至是超线性的（定理 2.1]）。在非凸情形下，收敛更难实现，但已有一些部分结果，参见 [4, 5, 64, 40, 72, 68]。

2.2. **Bregman近端法**。有人提出用 Bregman正则项 $\lambda_k^{-1}D(x,\ x_k)$ 替代二次惩罚项 $\frac{1}{2\lambda_k}\|x-x_k\|^2$，从而得到如下迭代方案

$$x_{k+1} \in \operatorname*{argmin}_{x\in\mathbb{R}^n} f(x) + \lambda_k^{-1} D(x, x_k), \tag{2}$$

参见 [20, 10]。这里，给定一个 Legendre型函数 ψ，与 ψ 关联的 Bregman距离为

$$D(x,y) = \begin{cases} \psi(x) - \psi(y) - \nabla\psi(y)\cdot(x-y) & \text{if } y \in \operatorname{int}(\operatorname{dom}\psi) =: G \\ +\infty & \text{otherwise} \end{cases} \tag{3}$$

其中近端点方法对应于特殊情形 $\psi(x) = \frac{1}{2}\|x\|^2$。

在指数族的背景下，ψ 是累积量生成函数，也称为对数归一化函数。ψ 的 Legendre性称为陡峭性 [18], ，而 $\nabla^2\psi \succ 0$对应于族的极小性。下文始终假设 ψ 属于 C^2类。

2.3. **更一般的正则项**。在（1）和（2）的基础上，我们接下来考虑如下形式的方案

$$x_{k+1} \in \operatorname*{argmin}_{x\in\mathbb{R}^n} f(x) + \lambda_k^{-1} \Psi(x, x_k), \tag{4}$$

其中采用了更一般的正则项 $\Psi(x^+,\ x)$，并允许将 $D(x^+,\ x)$ 作为特殊情形。与 $D(x^+,\ x)$ 一样，$\Psi(x^+,\ x)$ 也只在部分范围内有定义。以 Bregman正则项为典范，我们提出以下框架：

(i) 存在一个开集 G，使得 $G\times G \subset \operatorname{dom}\nabla_{11}^2\Psi = \operatorname{dom}\nabla_1\Psi \subset \operatorname{dom}\Psi \subset G\times\overline{G}$，且 Ψ，$\nabla_1\Psi$，$\nabla_{11}^2\Psi$ 在其定义域上联合连续。(ii) $\Psi \geq 0$，并且 $\Psi(x,\ x) = 0$ 对 $x \in G$成立。

如果 Ψ 满足以下更强的性质，我们就称其为分离的

(ii') $\Psi \geq 0$ 且 $\Psi(x^+,\ x) = 0$ 当且仅当 $x^+ = x$。

我们称 为 Ψ 在 $x \in G$ 处具有范数下界，如果存在 $\delta > 0$ 和 $m > 0$ 使得 t

(iii) $\Psi(y,z) \geq m\|y-z\|^2$ for all $y, z \in B(x,\delta)$.

如果 (iii) 对每个 $x \in K$都成立，我们就称 Ψ在集合 $K \subset G$ 上具有逐点下范数界。尽管 δ，m 取决于 x，但标准的紧致性论证（见引理 7）表明，当 $K \subset G$ 为紧集时，我们可以对所有 $x \in K$ 得到相同的 δ，m。最后，根据 (i)，不失一般性，我们可以假设 $\operatorname{dom} f \subset G$。

2.4. **部分正则化项**。假设 $\Psi_V(v^+,\ v)$ 仅对线性子空间 V 中的元素 $v^+,\ v$ 满足 (i)、(ii)。那么，$\Psi(x^+,\ x) = \Psi_V(Px^+,\ Px)$，其中 P 是到 V 的正交投影，在 $\mathbb{R}^n$ 上给出了一个正则项。我们称这样的 Ψ 为部分正则化项，因为其作用仅限于迭代点 x 的 V-分量 Px，而不会影响 $V^\perp$-分量。在 x-空间中，部分正则化项满足 (i) 和 (ii)，而 (ii') 或 (iii) 至多可在 V 中得到满足。

2.5. **Kurdyka-Łojasiewicz 不等式**。以下定义对我们的方法至关重要。

定义 1。（Kurdyka-Łojasiewicz 不等式）。下半连续函数 $f\colon \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ 在 $\bar{x} \in \operatorname{dom}(\partial f)$ 处具有 KŁ性质，当且仅当存在 $\gamma,\ \eta > 0$、U 的一个邻域 $\bar{x}$，以及一个连续凹函数 $\phi\colon [0,\eta) \to \mathbb{R}_+$（称为去奇异函数），使得

i. $\phi(0) = 0$，ii. ϕ 在 $(0,\eta)$ 上属于 C^1 类，iii. $\phi'(s) > 0$ 适用于 $s \in (0,\eta)$，iv. 对所有 $x \in U \cap \{x\colon f(\bar{x}) < f(x) < f(\bar{x}) + \eta\}$ 满足 KŁ 不等式

$$\phi'(f(x) - f(\bar{x}))\operatorname{dist}(0, \partial f(x)) \geq \gamma \tag{5}$$

成立，其中 ∂f 是 [69]的次微分。

注 1. 当去奇异函数为 $\phi(s) = s^{-\theta}$且满足某个$\theta \in [\frac{1}{2}, 1)$时，我们称 f 满足Łojasiewicz不等式，这意味着 $\phi(s) = \frac{s^{1-\theta}}{1-\theta}$。有时会将 $\theta = \frac{1}{2}$这一情形单独称为Polyak-Łojasiewicz不等式。 –

注 2. 当 K 是一个紧集，且 f 在该集合上对所有 $\bar{x} \in K$ 都取常值 $f(\bar{x}) = f^*$ 时，式 (5) 在 K 的某个邻域 U 上一致成立。

注 3. 关于通过 KŁ 性质得到收敛，参见例如 [1, 5, 6, 7, 16, 57, 58]。众所周知，在 o-极小结构 [31, 32, 33]，中的可定义性（简称 可定义性）蕴含 KŁ 不等式。参见 [45]，；对于非光滑情形，参见 f, [17, 的定理 11]，，其中证明了 ϕ 可以取为凹函数。我们使用全局次解析集的 o- 极小结构 $\mathbb{R}_{an}$，也使用更大的结构 $\mathbb{R}_{an,\,\exp}$，它允许指数函数和对数函数，参见 [31, 15, 27, 74, 77, 54]。

3. EM算法

我们考虑一族概率测度 $\mathbb{P}_\theta$，其相对于完整数据空间 X 上的 σ-有限基准测度 μ 的密度为 $p(x, \theta)$，$d\mathbb{P}_\theta(x) = p(x, \theta)d\mu(x)$，其中 $\theta \in \Theta$ 是我们希望通过最大似然估计的未知参数。然而，被观测到的并非 $x \in X$，而是不完整数据空间 Y 中的随机变量 $y = h(x)$，其中 $h : X \to Y$ 可测且通常不可逆。因此，观测数据 y 相对于边缘测度 $\nu = \mu \circ h^{-1}$ 的密度为

$$q(y, \theta) = \int_{h^{-1}(y)} p(x, \theta) d\mu_y(x), \tag{6}$$

其中，族 $(\mu_y)_{y \in Y}$ 是测度 μ 相对于 Y 上的边缘测度 $\nu = \mu \circ h^{-1}$ 的测度分解，且每个 μ_y 集中于 $h^{-1}(y) \subset X$。这里，测度分解是指

$$\int_X f(x) d\mu(x) = \int_Y \left[\int_{h^{-1}(y)} f(x) d\mu_y(x) \right] d\nu(y) \tag{7}$$

对于 μ-可积的 f。代入 $f = \chi_{h^{-1}(B)} p(\cdot, \theta)$，得到关系式

$$(\mathbb{P}_\theta \circ h^{-1})(B) = \mathbb{P}_\theta(h^{-1}(B)) = \int_B \left[\int_{h^{-1}(y)} p(x, \theta) d\mu_y(x) \right] d\nu(y)$$

j为（6）提供依据。这使我们能够为每个 $y \in Y$ 定义条件密度 y

$$k(x|y, \theta) = \frac{p(x, \theta)}{q(y, \theta)},\ x \in h^{-1}(y), \tag{8}$$

相对于测度 μ_y。关于建立解体存在性所需的假设，例如参见 [37]。

给定可用样本 y，不完整数据空间 Y 中的最大似然问题是以下优化程序

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in M} -\log q(y, \theta), \tag{9}$$

其中 $M \subset \Theta$ 是允许进行优化的模型参数集合。在此阶段，EM算法的基本思路是假设最小化问题（9）较为繁琐，而且从算法角度而言，最好在完整数据空间中进行最大似然估计

$$\tilde{\theta} \in \operatorname*{argmin}_{\theta \in M} -\log p(x, \theta). \tag{10}$$

由于没有可用的样本 x，无法直接执行 (10)，因此采用以下迭代过程。给定未知参数 $\theta^{(k)} \in M$ 的当前估计值，对每个 $\theta \in M$，计算在给定 y 和 $\theta^{(k)}$ 的条件下，$\log\, p(x, \theta)$ 的条件期望：

$$Q(\theta, \theta^{(k)}) := \mathbb{E}_{\theta^{(k)}}(\log p(x, \theta)|y),$$

这会（对于固定的 y, $\theta^{(k)}$）得到一个关于 θ 的函数。根据以下公式，这称为 E步

$$\mathbb{E}_\theta(\phi(x,\theta')|y) = \int_{h^{-1}(y)} \phi(x,\theta')k(x|y,\theta)d\mu_y(x),$$

其中 $k(x|y,\ \theta)$ 由（8）给出。得到该结果后，执行 M 步

$$\theta^{(k+1)} \in \operatorname*{argmin}_{\theta\in M} -Q(\theta,\theta^{(k)}),$$

从而得到新的模型参数估计 $\theta^{(k+1)} \in M$。众所周知，EM算法在每一步都会降低负对数似然 $-\log q(y,\ \theta^{(k)})$，优化 $\theta \in M$ 并不会改变这一点，即使只求得一个局部最小值也是如此。然而，如前所述，函数值单调下降并不能保证迭代点 $\theta^{(k)}$ 收敛，而我们目前要考察的正是迭代点的收敛性。

算法EM算法

- **步骤 1** (E步)。给定当前模型参数估计 $\theta^{(k)} \in M$，构造函数$\theta \mapsto Q(\theta,\ \theta^{(k)}) = \mathbb{E}_{\theta^{(k)}}(\log\ p(x,\ \theta)|y)$，使其在 M上可用于优化。
- **步骤 2** (M步)。计算 $\theta^{(k+1)} \in \operatorname{argmin}_{\theta\in M} -Q(\theta,\ \theta^{(k)})$。返回步骤 1。

备注4。约束的一种特殊情况是曲线族 $M = \{\theta \in \Theta :\ \theta = \theta(u)\ u \in U\}$，其中 $\theta(u)$是 θ的重新参数化，但我们的方法允许更一般的集合。例3给出了 M 的一个典型实例。

3.1. **指数族的性质**。若密度 $p(x,\theta)$在 X 上相对于基准测度 μ 具有如下形式，则它们构成一个 n 维指数族：

$$p(x,\theta) = e^{\theta\cdot T(x)-\psi(\theta)}, \tag{11}$$

其中，$T(x)$是充分统计量，$\theta \in \Theta = \{\theta \in \mathbb{R}^n :\ p(\cdot,\theta) \in L^1(X,d\mu)\}$ 是自然参数，而 $\psi(\theta)$是定义在 Θ上的对数归一化函数，并满足

$$\psi(\theta) = \text{对数} \int_X e^{\theta\cdot T(x)}d\mu(x)\text{。} \tag{12}$$

自然参数空间也可写作 $\Theta = \{\theta \in \mathbb{R}^n :\ \int_X e^{\theta\cdot T(x)}d\mu(x) < \infty\}$，且它是 $\mathbb{R}^n$的一个凸子集。假设 Θ 是满维的n并不失一般性；否则，可以进行参数约化，得到一个维数更低的等价表示（11） $m < n$：

引理*1.* 假设自然参数空间 Θ 包含在一个维数为 $m < n$ 的仿射子空间中。那么，存在 $\mathbb{P}_\theta$ 的一个等价表示，使其成为一个 m 维指数族 $d\mathbb{P}_\theta(x) = p'(x,\theta')d\mu'(x) = e^{\theta'\cdot T'(x)-\psi'(\theta')}d\mu'(x)$，$\theta' \in \Theta' \subset \mathbb{R}^m$，其中现在 $\dim(\Theta') = m$。如果在式（11）中，统计量 $T(x)$ 在 X 上仿射无关，那么 $T'(x)$ 也仿射无关。

证明：不失一般性，设参数为 $\theta = (\theta_1,\theta_2)$，其中 $\theta_2 = A\theta_1 + a$。$x = e^{\theta_1\cdot T_1(x)+\theta_2\cdot T_2(x)-\psi(\theta)}$ $x = e^{\theta_1\cdot T_1(x)+A^T\theta_2\cdot T_2(x)-\psi(\theta_1,A\theta_1+a)}e^{a\cdot T_2(x)}$ $x = e^{\theta_1\cdot T'(x)-\psi'(\theta_1)}$ x，矩阵 A 的大小为 $(\ )$，秩为 。那么 $d\mathbb{P}_\theta(\ )(\ ,\theta)d\mu(\ )(\ )(\ )(\ )d\mu(\ )(\ (\ )(\ ))(\ ,\ )(\ )d\mu(\ )(\ )(\ )d\mu'(\ )$，其中 $T'(x) = T_1(x) + A^TT_2(x)$，$\psi'(\theta_1) = \psi(\theta_1,A\theta_1+a)$，且 $d\mu'(x) = e^{a\cdot T_2}(x)d\mu(x)$；参数空间 $\Theta' = \{\theta_1 : (\theta_1, A\theta_1+a) \in \Theta\}$ 现在是满维的，维数为 m。由于 m 是最小可能值，因此 $\theta_1 \in \Theta'$ 之间不再存在仿射依赖。还要注意，$\mu \ll \mu'$ 和 $\mu' \ll \mu$ 表明这两种表示等价。

最后，假设 $a\cdot T(x)$ 几乎处处为常数蕴含 $a = 0$。那么，如果对于几乎所有 $x \in X$，$a'\cdot T'(x) = c$，则有 $(a', A^Ta')\cdot(T_1(x)\ T_2(x)) = c$，因此 $(a', A^Ta') = (0,0)$，由此得到 $a' = 0$，所以 T' 也是仿射无关的。□

因此，假设 Θ在 (11) 中已具有满维 n，我们将 Θ的内部记为 G。函数 ψ 也称为累积量生成函数，因为它满足

$$\mathbb{E}_\theta[T(x)] = \nabla\psi(\theta), \quad \mathbb{V}_\theta[T(x)] = \nabla^2\psi(\theta),$$

对于高阶矩也有类似关系；参见 [18]。

定义 2。 若函数 $T_i(\cdot)$ 仿射无关，则称族 (11) 为极小，即不存在 $a \neq 0$ 使得对于 μ-几乎处处的 $x \in X$， $a \cdot T(x) = const$ 。

引理 2。 在极小性条件下，对于每个 $\theta \in G = \mathrm{int}(\Theta)$，都有 $\nabla^2\psi(\theta) \succ 0$ 。此外，映射 $\theta \mapsto p(\cdot, \theta)$ 在 G 上是单射的。

证明： 1) 由 $\nabla^2\psi(\theta)a = 0$ 可得 $0 = a \cdot \nabla^2\psi(\theta)a = a \cdot \mathbb{V}_\theta[T(x)]a = \mathbb{V}_\theta[a \cdot T(x)] = \mathbb{E}_\theta|a \cdot T(x) - \mathbb{E}_\theta(a \cdot T(x))|^2$，因此 $a \cdot T(x) = \mathbb{E}_\theta[a \cdot T(x)] = a \cdot \nabla\psi(\theta)$ 对 $= const$ μ-几乎处处成立，从而得到$a = 0$。

2) 令 $p(\cdot, \theta) = p(\cdot, \theta')$ μ-几乎处处成立，则 $\theta \cdot T(x) - \psi(\theta) = \theta' \cdot T(x) - \psi(\theta')$ 几乎处处成立，因此 $(\theta - \theta') \cdot T(x) + \psi(\theta') - \psi(\theta) = 0$ 几乎处处成立，于是向量 $a = \theta - \theta'$ 使 $a \cdot T(x) = \psi(\theta) - \psi(\theta')$ 几乎处处为常数，从而得到 $\theta = \theta'$。□

如果对数归一化函数 ψ 是 Legendre型 [69, 9, 18]，则称指数族为陡峭。如果 Θ 是一个开集，即 $\Theta = G$，则称该族为正则。在这种情况下，该族自动是陡峭的，但陡峭类更大 [18]。

3.2. 指数族的 EM算法。有时，EM算法被描述为在 E步中补全数据与在 M步中在完整数据空间中进行最大似然估计之间来回交替。这一般并不成立，参见 [34], ；但当$p(x, \theta)$ 属于指数族 (11) 时则成立。也就是说，在这种情况下，$\log\ p(x, \theta) = \theta \cdot T(x) - \psi(\theta)$，因此

$$\mathbb{E}_{\theta^{(k)}}[\log p(x, \theta)|y] = \theta \cdot \mathbb{E}_{\theta^{(k)}}[T(x)|y] - \psi(\theta),$$

右侧的第一项选取一个完全数据统计量 $t^{k+1} = T(x_{k+1})$，这一事实可表述为：E步在于补全数据。M步保持不变，但由于代入了 $T(x_{k+1})$，得到 $Q(\theta,\ \theta^{(k)}) = \log\ p(x_{k+1},\ \theta) = \theta \cdot T(x_{k+1}) - \psi(\theta)$，因此可恰当地称为完整数据空间中的最大似然。总而言之，对于指数族，EM算法的形式如下

算法指数族的EM算法

- **步骤 1** （E步）。 给定当前的模型参数估计 $\theta^{(k)} \in M$，通过计算 $T(x_{k+1}) = \mathbb{E}_{\theta^{(k)}}[T(x)|y]$ 来补全数据。
- **步骤 2** （M步）。 计算 $\theta^{(k+1)} \in \mathrm{argmin}_{\theta \in M}\, \psi(\theta) - \theta \cdot T(x_{k+1})$。返回步骤 1。

注 5. 当充分统计量是仿射的，即 $T(x) = Ax + b$ 时，E步甚至可以基于计算 $x_{k+1} = \mathbb{E}_{\theta^{(k)}}(x|y)$，因为这样 $\mathbb{E}_{\theta^{(k)}}[T(x)|y] = \mathbb{E}_{\theta^{(k)}}(Ax + b|y) = A\mathbb{E}_{\theta^{(k)}}(x|y) + b$，使得“补全数据 ”这一表达更具启发性。

最后我们指出，当完全数据如式（11）所示，来自样本空间 $p(x,\ \theta)$ 上的指数族时，条件密度 $k(x|y,\ \theta)$ 对于给定的$y \in Y$，也构成样本空间 $h^{-1}(y)$ 上的指数族，其基准测度 μ_y 来自对 μ 的测度分解。这一点可由下式看出：

$$k(x|y, \theta) = \frac{p(x, \theta)}{q(y, \theta)} = \frac{e^{\theta \cdot T(x) - \psi(\theta)}}{\int_{h^{-1}(y)} e^{\theta \cdot T(x') - \psi(\theta)} d\mu_y(x')} =: e^{\theta \cdot T(x) - \psi_y(\theta)}, \tag{13}$$

将.....代入后得到

$$\psi_y(\theta) = \log \int_{h^{-1}(y)} e^{\theta \cdot T(x)} d\mu_y(x), \tag{14}$$

它与（12）相对应，并且可以写成 $d\mathbb{P}_\theta^{x|y} = k(x|y,\ \theta)d\mu_y(x)$。对于指数族，测度 $k(\cdot|y,\ \theta)d\mu_y$ 彼此等价，即 $k(\cdot|y,\ \theta)d\mu_y \ll k(\cdot|y,\ \theta')d\mu_y$ 对所有 $\theta,\ \theta' \in \Theta$ 均成立。在一般情形下，出于实际考虑，也可以将此作为一项假设（参见 [21, 第 III B 节]）。

注 6. 即使完全数据 $p(x,\ \theta)$ 来自指数族，不完整数据也未必如此 $q(y,\ \theta)$。数据缺失本身就可能导致指数结构丧失。但也有一些情形，特定的非指数型分布族 $q(y,\ \theta)$ 则并非如此，而是有意构造为来自指数族的缺失数据。这一术语源自 Sundberg [73], ，他列举了多种示例 $q(y,\ \theta)$，包括指数族的有限混合、删失数据、卷积、折叠分布、负二项分布等。目前我们将参数约束 $\theta \in M$ 也纳入其中，因此，完全数据 $p(x,\ \theta)$（由其导出 $q(y,\ \theta)$）例如可以是曲线型的。

关于EM序列的适定性，我们必须牢记， M 必须是闭集，因为目标函数 f 必须是下半连续的。这可能与 Θ 通常不是闭集这一事实相冲突。因此，我们只能假设 $M \subset \Theta$，这样 $M \cap \partial\Theta$ 可能非空，而该集合中的点会带来麻烦。

4. Kullback-Leibler information measure

X 上的Kullback-Leibler信息距离定义为

$$\mathcal{K}(q||p) = \mathbb{E}_q\left(\log \frac{q}{p}\right) = \int_X q(x) \log \frac{q(x)}{p(x)} d\mu(x),$$

在依赖参数的情形下，我们使用如下记号

$$K(\theta||\theta^+) = \mathcal{K}(p(\cdot,\theta)||p(\cdot,\theta^+)).$$

将KL距离限制在 $h^{-1}(y)$ 上时，其形式为

$$K_y(\theta||\theta^+) = \mathbb{E}_\theta\left(\log \frac{k(\cdot|y,\theta)}{k(\cdot|y,\theta^+)}\middle| y\right) = \int_{h^{-1}(y)} k(x|y,\theta) \log \frac{k(x|y,\theta)}{k(x|y,\theta^+)} d\mu_y(x), \tag{15}$$

其中， μ_y 由 μ 的测度分解得到；根据假设 $k(\cdot|y,\ \theta)d\mu_y \ll k(\cdot|y,\ \theta')d\mu_y$ 对所有 $\theta,\ \theta' \in \Theta$ 成立，该值是有限的。接下来是一个关键的观测。

命题 1。 (参见 [21, 命题 1])。期望最大化算法是一般方案(4)的一种实现，其中 $f(\theta) = -$对数 $q(y,\theta) + i_M(\theta)$、 $\lambda_k = 1$ 以及由(15)给出的 $\Psi(\theta,\theta^{(k)}) = K_y(\theta^{(k)}||\theta)$。

证明： 重新排列（8）式并积分，得到

$$\log q(y,\theta) = \mathbb{E}_{\theta^{(k)}}(\log p(x,\theta)|y) - \mathbb{E}_{\theta^{(k)}}(\log k(x|y,\theta)|y)$$

对于任意 k。因此，EM算法中的M步为

$$\theta^{(k+1)} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^n} -\log q(y,\theta) + i_M(\theta) - \mathbb{E}_{\theta^{(k)}}\left(\log k(x|y,\theta)\middle|y\right).$$

在目标函数中加上常数项 $\mathbb{E}_{\theta^{(k)}}\left(\log\ k(x|y,\theta^{(k)})\middle|y\right)$ 不会改变优化程序，因此我们有

$$\theta^{(k+1)} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^n} -\log q(y,\theta) + i_M(\theta) - \mathbb{E}_{\theta^{(k)}}\left(\log \frac{k(x|y,\theta)}{k(x|y,\theta^{(k)})}\middle| y\right), \tag{16}$$

并且最后一项等于 $K_y(\theta^{(k)}||\theta)$。 □

注7. 1）这解释了为什么EM算法会降低负对数似然$-\log\ q(y,\ \theta)$，因为这是方案（4）的一般性质，另见定理1。

2）Kullback-Leibler散度在函数空间中具有分离性，即 $\mathcal{K}(q||p)\ =\ 0$蕴含$q=p$几乎处处成立。因此 $K_y(\theta||\theta^+)\ =0$蕴含 $k(\cdot|y,\ \theta)\ =k(\cdot|y,\ \theta^+)$ μ_y-几乎处处成立。然而，后者并不总能推出 $\theta=\theta^+$，因为族 $k(\cdot|y,\ \theta)$在 $h^{-1}(y)$上不一定是极小的。

4.1. 指数族的解释。 现在我们将 Kullback-Leibler散度具体化到指数族的情形。

命题 2。 属于同一指数族的两个分布 $p(x,\theta)$ 和 $p(x,\theta')$ 之间的 ***Kullback-Leibler***散度为 $K(\theta||\theta^+)\ =\ D(\theta^+,\theta)$，其中 D 是由对数归一化函数 ψ诱导的 ***Bregman***散度。

证明： 由（11），且由于$\int_X p(x,\ \theta)d\mu(x)\ =1$，我们有

$$\psi(\theta)=\log\int_X e^{\theta\cdot T(x)}d\mu(x).$$

对 θ 求导可得

$$\nabla\psi(\theta)=\int_X T(x)e^{\theta\cdot T(x)}d\mu(x)\Big/\int_X e^{\theta\cdot T(x)}d\mu(x).$$

现在 $e^{\psi(\theta)}=\int_X e^{\theta\cdot T(x)}d\mu(x)$，因此 $\nabla\psi(\theta)\ =\int_X T(x)e^{\theta\cdot T(x)-\psi(\theta)}d\mu(x)\ =\int_X T(x)p(x,\ \theta)d\mu(x)$ $=E_\theta[T(x)]$, 随机变量 $T(x)$ 关于分布 $d\mathbb{P}_\theta=p(\cdot,\ \theta)d\mu$ (另见 [53, (1.57)]) 的期望。于是

$$\begin{aligned}
K(\theta||\theta^+)&=\int_X p(x,\theta)\log\frac{p(x,\theta)}{p(x,\theta^+)}\,d\mu(x)\\
&=\int_X p(x,\theta)\left[\psi(\theta^+)-\psi(\theta)+(\theta-\theta^+)\cdot T(x)\right]d\mu(x)\\
&=\int_X p(x,\theta)\left[D(\theta^+,\theta)+(\theta^+-\theta)\cdot\nabla\psi(\theta)+(\theta-\theta^+)\cdot T(x)\right]d\mu(x)\qquad(17)\\
&=D(\theta^+,\theta)+\int_X p(x,\theta)\left[(\theta^+-\theta)\cdot(\nabla\psi(\theta)-T(x))\right]d\mu(x)\\
&=D(\theta^+,\theta)+(\theta^+-\theta)\cdot(\nabla\psi(\theta)-E_\theta[T(x)])\\
&=D(\theta^+,\theta).
\end{aligned}$$

这就证明了该结论。 □

Bregman距离或散度通常针对 Legendre型函数 ψ [9, 69] 来定义。正如前面提到的，对于对数归一化函数，这称为陡峭性 [18, 定义3.2]。实践中的大多数指数族都是正则的，即 Θ 是开集，在这种情况下，陡峭性自动成立。

引理3. 假设指数族 $p(x,\theta)$ 是极小的。那么，由对数归一化函数 ψ 诱导的 ***Bregman***散度是分离的，即 $D(\theta^+,\theta)\ =0$ 蕴含 $\theta^+=\theta$。

证明：由 $D(\theta^+,\theta)\ =0$可得 $\psi(\theta^+)\ -\psi(\theta)\ -\nabla\psi(\theta)\ \cdot(\theta^+-\theta)\ =0$。在$\theta$处进行泰勒展开，得到 $\psi(\theta^+)\ =\psi(\theta)\ +\nabla\psi(\theta)\ \cdot(\theta^+-\theta)+\frac{1}{2}(\theta^+-\theta)\ \cdot\nabla^2\psi(\bar{\theta})(\theta^+-\theta)$，其中某个 $\bar{\theta}$位于连接 θ^+与 θ的开线段上，并且取决于 θ,θ^+。因此$(\theta^+-\theta)\ \cdot\nabla^2\psi(\bar{\theta})(\theta^+-\theta)\ =0$。但由引理 2，极小性给出 $\nabla^2\psi(\bar{\theta})\succ 0$，故 $\theta^+=\theta$。□

引理 4。 设指数族 $p(x,\theta)$ 是极小的。则 $\nabla\psi$ 在 G上是单射。对于 $\eta=\nabla\psi(\theta)\ \in G^*=\mathrm{int}(\mathrm{dom}\,\psi^*)$，有 $\nabla\psi^*(\eta)\ =\theta$。特别地，如果该族是陡峭的，则 $\nabla\psi^*$是 $\nabla\psi$的逆函数，其中 $G=\mathrm{int}(\mathrm{dom}\,\psi)$被 1-1 映射到 $G^*=\mathrm{int}(\mathrm{dom}\ \psi^*)$ 上。

证明：令 $\nabla\psi(\theta) = \nabla\psi(\theta^+)$。对于测试向量 h，将 $\theta \mapsto h\cdot\nabla\psi(\theta)$ 在 θ^+处进行泰勒展开，得到 $h\cdot\nabla\psi(\theta) = h\cdot\nabla\psi(\theta^+) + h\cdot\nabla^2\psi(\bar{\theta})(\theta-\theta^+)$，其中 $\bar{\theta} = \bar{\theta}(\theta,\theta^+,h)$ 位于连接 θ和 θ^+ 的开线段上，并且取决于 θ,θ^+,h。将其与上面的第一个方程相比较可得$h\cdot\nabla^2\psi(\bar{\theta})(\theta-\theta^+) = 0$。取 $h = \theta-\theta^+$ 作为测试向量，得到 $(\theta-\theta^+)\cdot\nabla^2\psi(\bar{\theta})(\theta-\theta^+) = 0$，其中 $\bar{\theta} = \bar{\theta}(\theta,\theta^+,\theta-\theta^+)$，并且由于 $\nabla^2\psi(\bar{\theta}(\theta,\theta^+,\theta-\theta^+)) \succ 0$，我们有 $\theta = \theta^+$。

对于第二部分，回顾一下，$\partial\psi$和 $\partial\psi^*$ 互为逆函数，即 $\eta\in\partial\psi(\theta)$当且仅当 $\theta\in\partial\psi^*(\eta)$；参见 [69, 推论 23.5.1]。由 ψ 的严格凸性可知，其共轭函数 ψ^* 在 $G^* = \operatorname{int}\operatorname{dom}\psi^*$上可微。因此，若 $\eta = \nabla\psi(\theta)\in G^*$，则 $\nabla\psi^*(\eta) = \theta$。由于我们可能有$\nabla\psi(G)\not\subset G^*$，关于 $\eta = \nabla\psi(\theta)\in\partial G^*$ 我们所能确定的只有 $\theta\in\partial\psi^*(\eta)$。当 $p(x,\theta)$是陡峭的，则$\partial\psi(\theta) = \emptyset$ 对 $\theta\in\partial\Theta$成立，进而 $\nabla\psi$ 将 G 1-1映射到 G^*，其逆函数为$(\nabla\psi)^{-1} = \nabla\psi^*$。□

注8. 一般而言，$G = \operatorname{int}(\operatorname{dom}\psi) \subset \operatorname{dom}(\nabla\psi) \subset \operatorname{dom}\psi = \Theta$，且 $G^* = \operatorname{int}(\operatorname{dom}\psi^*) \subset \operatorname{dom}(\nabla\psi^*) \subset \operatorname{dom}\psi^*$；参见 [66, 定理 23.4]。

4.2. 缺失数据的 Fisher 信息。 式 (16) 中的正则项具有统计学解释。众所周知，$K_y(\theta||\theta) = 0$，且 $K_y \geq 0$，因此对于固定的 θ，$\theta^+ \mapsto K_y(\theta||\theta^+)$ 的全局最小值 0 尤其在 $\theta^+ = \theta$ 处取得。显然，根据必要最优性条件，$\nabla_2 K_y(\theta||\theta) = 0$ 且 $\nabla^2_{22}K_y(\theta||\theta) \succeq 0$。我们考察是否可以期待更强的充分最优性条件 $\nabla^2_{22}K_y(\theta||\theta) \succ 0$。回到定义，我们有

$$K_y(\theta||\theta^+) = \int_{h^{-1}(y)} k(x|y,\theta)\log\frac{k(x|y,\theta)}{k(x|y,\theta^+)}d\mu_y(x) = \mathbb{E}_\theta\left[\log\frac{k(\cdot|y,\theta)}{k(\cdot|y,\theta^+)}\middle|y\right].$$

对 θ^+ 求两次导数（参见 [8, 定理 5.8，第 7.1] 节）可得

$$\nabla^2_{22}K_y(\theta||\theta^+) = \mathbb{E}_\theta[-\nabla^2_{\theta^+\theta^+}\log k(\cdot|y,\theta^+)|y],$$

因此我们得到

$$\nabla^2_{22}K_y(\theta||\theta^+)\big|_{\theta^+=\theta} = \mathbb{E}_\theta[-\nabla^2_{\theta\theta}\log k(\cdot|y,\theta)|y] =: \mathcal{I}_m(\theta,y),$$

这可识别为给定 y 时缺失数据的条件期望费希尔信息矩阵；参见 [53, 3.52]。根据缺失信息原理 [61], $\mathcal{I}_m$（θ，y）给出了完全数据与不完整数据之间的信息损失。对恒等式 $\mathbb{E}_\theta[\nabla_\theta \log k(\cdot|y,\theta)|y] = 0$ 关于 θ 求导（再次参见 [8, 定理 5.8，第 7.1] 节），可得另一个公式

$$\mathcal{I}_m(\theta,y) = \mathbb{E}_\theta[\nabla_\theta k(\cdot|y,\theta)\nabla_\theta k(\cdot|y,\theta)^T|y] \succeq 0. \tag{18}$$

对于指数族，还可以进一步说明。

引理 5. </style}假设 （$x|y,\theta$）</style} 是一个指数族。那么 （$\theta||\theta$）</style}（θ）</style}。此外，若该族关于样本空间 （y）</style} 是极小的，则 （θ）*</style}* $\succ 0$。

证明： 由 (17) 可得，对于由 ψ_y 诱导的 Bregman距离 D_y，有 $K_y(\theta||\theta^+) = D_y(\theta^+,\theta)$，进而 $\nabla^2_{\theta^+\theta^+}K_y(\theta||\theta^+) = \nabla^2\psi_y(\theta^+)$。最后一部分由引理 2 得出。□

5. 条件族的降维

我们考虑来自一个 n 维指数族的不完整数据的约束最大似然，即 $\operatorname{im}(T) \subset \mathbb{R}^n$，其中 $\dim(\Theta) = n$。鉴于引理1，我们还假设完全数据族 $p(x,\theta)$ 是极小的。然而，这并不意味着条件 n 维指数族 $k(x|y,\theta)$ 在 $h^{-1}(y)$ 上也同样是极小的。事实上，缺失数据情形（例 1）表明，我们反而应该预期相反的情况。因此需要进行降维论证。

命题3。 n 维条件指数族 $d\mathbb{P}_\theta^{x|y} = k(\cdot|y,\theta)d\mu_y$ 具有等价的极小表示，即 m 维指数族 $d\mathbb{P}_{\theta'}^{x|y} = \bar{k}(x|y,\theta')d\mu_y$，其中 $\theta' = P\theta$ 是从 $\mathbb{R}^n$ 到 m 维子空间 V 的正交投影 P。条件分布满足 $\mathbb{P}_\theta^{x|y} = \mathbb{P}_{\theta'}^{x|y}$，此时充分统计量 $T' = P\circ T$ 是极小的。经过这一约简后，*Kullback-Leibler* 散度之间的关系为 $K_y(\theta||\theta^+) = \bar{K}_y(\theta'||\theta'^+)$。

证明： 我们有 $T\colon X\to\mathbb{R}^n$，且 $T(h^{-1}(y)) \subset t_0 + V$属于 $\mathbb{R}^n$ 中的一个维数极小的线性子空间 V，其中不失一般性，我们可以假设 $t_0\in V^\perp$。令 P为到 V上的正交投影。记 $T(x) = t_0 + v(x)$ 为 $v(x)\in V$。我们有

$$\begin{aligned}\psi_y(\theta) &= \log\int_{h^{-1}(y)} e^{\theta\cdot t_0}e^{\theta\cdot v(x)}d\mu_y(x)\\ &= \theta\cdot t_0 + \log\int_{h^{-1}(y)} e^{P\theta\cdot v(x)}d\mu_y(x) =: \theta\cdot t_0 + \bar{\psi}_y(P\theta),\end{aligned}\tag{19}$$

这里我们利用了对于 $v\in V$有 $\theta\cdot v = P\theta\cdot v$ 这一事实，因此最右侧的 $\bar{}$项 $\psi_y(P\theta)$ 仅依赖于 $P\theta$。现在由 $Pt_0 = 0$可得 $PT(x) = v(x)$，因此

$$\begin{aligned}\log k(x|y,\theta) &= \theta\cdot T(x) - \psi_y(\theta)\\ &= \theta\cdot t_0 + \theta\cdot v(x) - \psi_y(\theta)\\ &= \theta\cdot t_0 + \theta\cdot v(x) - \bar{\psi}_y(P\theta) - \theta\cdot t_0 \qquad \text{(using (19))}\\ &= P\theta\cdot v(x) - \bar{\psi}_y(P\theta) =: \log\bar{k}(x|y,P\theta),\end{aligned}\tag{20}$$

再次利用对于 $v\in V$有 $\theta\cdot v = P\theta\cdot v$。由此得到如下表示

$$d\mathbb{P}_\theta^{x|y} = k(\cdot|y,\theta)d\mu_y = \bar{k}(\cdot|y,\theta')d\mu_y = d\mathbb{P}_{\theta'}^{x|y},\tag{21}$$

其中新的充分统计量为 $v = P\circ T$，新参数为 $\theta' = P\theta$。注意，$v = P\circ T$在 $h^{-1}(y)$ 上仿射无关，因为根据 V的选择， $T(h^{-1}(y)) - t_0\subset V$在 V中满维m，并且我们有 $T(h^{-1}(y)) - t_0 = P[T(h^{-1}(y)) - t_0] = v(h^{-1}(y))$，因此后者在 V中也满维。

接下来需要比较这两种表示所产生的 Kullback–Leibler 散度。

$$\begin{aligned}\bar{K}_y(\theta'||\theta'^+) &= \int_{h^{-1}(y)}\log\frac{\bar{k}(x|y,\theta')}{\bar{k}(x|y,\theta'^+)}\bar{k}(x|y,\theta')d\mu_y(x)\\ &= \int_{h^{-1}(y)}\log\frac{\bar{k}(x|y,\theta')}{\bar{k}(x|y,\theta'^+)}k(x|y,\theta)d\mu_y(x) \qquad \text{(using (21))}\\ &= \int_{h^{-1}(y)}\left[(P\theta - P\theta^+)\cdot v(x) - \bar{\psi}_y(P\theta) + \bar{\psi}_y(P\theta^+)\right]k(x|y,\theta)d\mu_y(x)\\ &= \int_{h^{-1}(y)}\left[(\theta-\theta^+)\cdot v(x) - \psi_y(\theta) + \theta\cdot t_0 + \psi_y(\theta^+) - \theta^+\cdot t_0\right]k(x|y,\theta)d\mu_y(x)\\ &= \int_{h^{-1}(y)}\left[(\theta-\theta^+)\cdot(v(x)+t_0) - \psi_y(\theta) + \psi_y(\theta^+)\right]k(x|y,\theta)d\mu_y(x)\\ &= \int_{h^{-1}(y)}\left[(\theta-\theta^+)\cdot T(x) - \psi_y(\theta) + \psi_y(\theta^+)\right]k(x|y,\theta)d\mu_y(x)\\ &= \int_{h^{-1}(y)}\log\frac{k(x|y,\theta)}{k(x|y,\theta^+)}k(x|y,\theta)d\mu_y(x) = K_y(\theta||\theta^+).\end{aligned}$$

最后，为了符号简洁性，我们将 v 写作 T'。 □

$\nabla^2_{\theta\theta}\psi_y(\theta)$与 $\nabla^2_{\theta'\theta'}\bar{\psi}_y(\theta')$之间的关系如下。

推论1. 存在一个 $n \times n$ 正交矩阵 Q ，使得

$$Q\nabla^2_{\theta\theta}\psi_y(\theta)Q^T = \left[\begin{array}{c|c} \nabla^2_{\theta'\theta'}\bar{\psi}_y(\theta') & 0 \\ \hline 0 & 0 \end{array}\right], \quad Q\theta = \left[\begin{array}{c} \theta' \\ \theta'' \end{array}\right], \quad \nabla^2_{\theta'\theta'}\bar{\psi}_y(\theta') \succ 0, \tag{22}$$

并且特别地， $m = \dim(\theta')$ 是 $\nabla^2_{\theta\theta}\psi_y(\theta)$的秩。*KL*散度 $K_y(\theta||\theta^+)$ 仅取决于坐标 θ' ，并且在子空间 $V = \mathrm{im}(P)$上具有分离性，其中 P 是正交投影 $\theta' = P\theta$。此外，$\nabla^2_{22}\bar{K}_y(\theta'||\theta') \succ 0$。

证明： 命题3的证明中使用的投影 $P\colon \mathbb{R}^n \to V$和 $I - P\colon \mathbb{R}^n \to V^\perp$ 由正交 $n \times n$ 矩阵 Q 表示，该矩阵给出了式(22)中的变量变换，其中 $\theta' \in V, \theta'' \in V^\perp$。于是， $\bar{k}$ 和对数归一化函数 $\bar{\psi}_y$ 仅取决于 θ'。

由于族 $k(x|y, \theta)$ 是极小的，因此确有 $\bar{k}(\cdot|y, \theta')$ 是一一对应的， $\bar{K}_y(\theta'||\theta'^+) = 0$ 首先给出$\bar{k}(\cdot|y, \theta') = \bar{k}(\cdot|y, \theta'^+)$ μ_y-几乎处处成立，进而得到 $\theta' = \theta'^+$。最后一个结论由引理 5 得出。□

注 9。秩 m 取决于 y。由于 $\dim(\theta') = m$，我们称 θ' 为精确的参数，它在正交坐标变换下唯一。其含义是，族 $k(x|y, \theta)$ 关于 $n - m$ 冗余的参数 θ'' 是过参数化的，并且具有一个统计量，其分量 $T_j(x)$ $n - m$ 过多；而 θ' 仅保留所需参数的精确数量 m 。

注 10. 尽管 $T(h^{-1}(y))$ 的拓扑维数通常小于 n，但 $m = n$ $T(h^{-1}(y))$ 的仿射维数却可能如此。比如， $T(h^{-1}(y))$ 可能是三维空间中的一条空间曲线，其拓扑维数为 1，但仿射维数为 3。具有讽刺意味的是，对于这类弯曲纤维 $h^{-1}(y)$，我们的收敛结果先验地比纤维为仿射子空间这一更常见的情形更好。

我们发现，(16) 中的 $\Psi(\theta^+, \theta) = K_y(\theta||\theta^+)$ 是子空间 $V = \mathrm{im}(P)$ 上的部分正则化项，符合第 2.4 节所述的意义。因此，现在需要给出部分正则化下的核心收敛结果，我们将在下一节中阐述。

6. 内点收敛

我们证明，在迭代序列有界，且该序列及其聚点集都位于 G 定义域 Ψ的内部这一假设下，广义近端方法 (4) 部分收敛。由于这些结果具有普遍性，我们改用优化领域中熟悉的记号。

我们考虑一种扩展，其中（4）以近似的方式求解，即

$$e_k = g_k + \lambda_{k-1}^{-1}\nabla_1\Psi(x_k, x_{k-1}) \text{ and } f(x_k) + \lambda_{k-1}^{-1}\Psi(x_k, x_{k-1}) \le f(x_{k-1}), \tag{23}$$

其中 $g_k \in \partial f(x_k)$，且次梯度误差 e_k 满足以下条件之一：

a. $\sum_k \lambda_{k-1}\|e_k\| < \infty$； b. $\lambda_{k-1}\|e_k\| \le M'\|\nabla_1\Psi(x_k, x_{k-1})\|$，
其中某个固定的大数 $M' > 0$； c. $\|e_k\| \le M''\|g_k\|$，其中某个
大数 $M'' > 0$。

定理1。（部分收敛）。假设 f 在 G 上满足 *KŁ* 不等式。考虑由近似近端方法生成的序列 x_k ，该序列有界，且与其聚点均位于 G 中。设 Ψ_V 具有逐点下范数界，并且在具有投影 P 的子空间 V 上是分离的。假设 $\lambda_k/\lambda_{k-1} \le r < \infty$ 和 $\lambda_k \le R < \infty$。那么序列 Px_k 收敛。当 $\sum_k \lambda_k = \infty$ 时， x_k 至少有一个聚点是临界点；若 $\lambda_k \ge \eta > 0$，则所有聚点都是临界点。

证明： 1) 在 $e_{k+1} = 0$ 的情形下由 (4)，以及由 (23)，我们有 ve

$$f(x_{k+1}) + \lambda_k^{-1}\Psi(x_{k+1}, x_k) \le f(x_k) + \lambda_k^{-1}\Psi(x_k, x_k) = f(x_k), \tag{24}$$

使用 $\Psi(x,\ x) = 0$。对于 $\Psi(x_{k+1},\ x_k) > 0$，其值严格递减。仅当 $\Psi(x_{k+1},\ x_k) = 0$ 时，才会出现 $f(x_{k+1}) = f(x_k)$ 这种情形；由 (23) 可推出 $e_{k+1} \in \partial f(x_{k+1})$。此时，算法在 $e_{k+1} = 0 \in \partial f(x_{k+1})$ 时停止，但在 $e_{k+1} \neq 0$ 时继续。因此，我们可以集中讨论算法不会在有限步内终止的情形。

2）根据假设，序列 x_k 和 $f(x_k)$ 有界，x_k 的聚点属于 G，并且由单调收敛可得 $f(x_k) \to f^* \in \mathbb{R}$。由此结合（24），利用 $\Psi \geq 0$，立即得到 $\lambda_k^{-1}\Psi(x_{k+1},\ x_k) \to 0$。由于 $\lambda_k^{-1} \geq R^{-1} > 0$，我们推得 $\Psi(x_{k+1},\ x_k) \to 0$。我们来证明这意味着 $P(x_{k+1} - x_k) \to 0$。事实上，假设存在一个无限子序列 $k \in N \subset \mathbb{N}$，使得 $\|P(x_{k+1} - x_k)\| \geq \epsilon > 0$。利用 x_k 的有界性，抽取一个子子序列 $k' \in N' \subset N$，使得 $x_{k'} \to x,\ \ x_{k'+1} \to x'$。于是 $\Psi(x_{k'+1},\ x_{k'}) \to \Psi(x',\ x) = 0$；又因为 Ψ 在 V 上是分离的，所以 $Px' = Px$，从而必有 $P(x_{k'+1} - x_{k'}) \to 0$，这就产生了矛盾。

3) 我们论证，这意味着从某个下标开始，$\Psi(x_{k+1}, x_k) \geq m\|P(x_{k+1} - x_k)\|^2$。事实上，令 K 为 x_k的聚点构成的紧集。对空间 V中的正则项 Ψ_V应用引理 7，我们可得存在 $m, \delta > 0$，使得对所有满足 $\|P(y-z)\| < \delta$的 $y, z \in N(K, \delta)$，均有 $m\|P(y-z)\|^2 \leq \Psi(y, z)$。但显然，从某个下标 k开始，$x_{k+1}, x_k \in N(K, \delta)$；又由于 $P(x_{k+1} - x_k) \to 0$，我们也有 $\|P(x_{k+1} - x_k)\| < \delta$从某个下标开始成立。由此得证。

4) 序列 x_k 有界，因此其聚点集 K 是紧集，并且 f 在 K上取常值 f^*。因此，KŁ 不等式 (5) 在 K的某个邻域 U 上成立。由于 U之外只有有限个 x_k，重新标记该序列后，我们可以假设 (5) 对整个序列都成立。

由（5）中去奇异函数 ϕ 的凹性可得

$$\begin{aligned}\phi(f(x_k) - f^*) - \phi(f(x_{k+1}) - f^*) &\geq \phi'(f(x_k) - f^*)(f(x_k) - f^* - (f(x_{k+1}) - f^*)) \\ &= \phi'(f(x_k) - f^*)(f(x_k) - f(x_{k+1})) \\ &\geq \phi'(f(x_k) - f^*)\lambda_k^{-1}\Psi(x_{k+1}, x_k) \\ &\geq \phi'(f(x_k) - f^*)\lambda_k^{-1} m\|Px_{k+1} - Px_k\|^2.\end{aligned} \tag{25}$$

这里第三行用到了（24），而最后一行用到了 V 的范数下界；如上所述，该下界基于引理 7，并应用于 Ψ_V。

5）根据 Kurdyka-Łojasiewicz 不等式（5），我们有

$$\phi'(f(x_k) - f^*)\|g_k\| \geq \gamma,$$

使用 $g_k \in \partial f(x_k)$。应用部分范数上界以及阶段 $k-1$ 的近似最优性，可得

$$\begin{aligned}\phi'(f(x_k) - f^*)^{-1} \leq \gamma^{-1}\|g_k\| &= \gamma^{-1}\left(\|e_k\| + \lambda_{k-1}^{-1}\|\nabla_1\Psi(x_k, x_{k-1})\|\right) \\ &\leq \gamma^{-1}\left(\|e_k\| + M\lambda_{k-1}^{-1}\|Px_k - Px_{k-1}\|\right),\end{aligned} \tag{26}$$

其中最后一个估计用到了 V 的范数上界。该上界使用下一节中的引理 6，并应用于 Ψ_V。结合（25）可得

$$\phi(f(x_k) - f^*) - \phi(f(x_{k+1}) - f^*) \geq \frac{\gamma\lambda_k^{-1} m\|Px_{k+1} - Px_k\|^2}{\|e_k\| + M\lambda_{k-1}^{-1}\|Px_k - Px_{k-1}\|}.$$

由于对于 $a,\ b,\ c \geq 0$，$a^2 \leq bc$ 蕴含 $a \leq \frac{1}{2}b + \frac{1}{2}c$，我们得到

$$\begin{aligned}\|Px_{k+1} - Px_k\| \leq &\tfrac{1}{2}\left(\|Px_k - Px_{k-1}\| + M^{-1}\lambda_{k-1}\|e_k\|\right) \\ &+ \tfrac{1}{2}\gamma^{-1}\lambda_k/\lambda_{k-1}M/m\left(\phi(f(x_k) - f^*) - \phi(f(x_{k+1}) - f^*)\right)\end{aligned} \tag{27}$$

并设定 $C = \frac{Mr}{\gamma m}$，同时使用 $\lambda_k/\lambda_{k-1} \leq r$，可得

$$\|Px_k - Px_{k+1}\| \leq \frac{1}{2}\|Px_k - Px_{k-1}\| + \frac{\lambda_{k-1}}{2M}\|e_k\| + \frac{C}{2}\left[\phi(f(x_k) - f^*) - \phi(f(x_{k+1}) - f^*)\right]. \tag{28}$$

将此式从 $k=1$ 求和至 $k=n$，得到

$$\sum_{k=1}^{n}\|Px_{k+1}-Px_k\| \le \frac{1}{2}\sum_{k=1}^{n}\|Px_k-Px_{k-1}\|+\frac{1}{2M}\sum_{k=1}^{n}\lambda_{k-1}\|e_k\|+\frac{C}{2}\left[\phi(f(x_1)-f^*)-\phi(f(x_{n+1})-f^*)\right].$$

因此

$$\begin{aligned}\sum_{k=1}^{n}\|Px_{k+1}-Px_k\| &\le \|Px_1-Px_0\|+C\left[\phi(f(x_1)-f^*)-\phi(f(x_{n+1})-f^*)\right]\\ &\quad-\|Px_{n+1}-Px_n\|+\frac{1}{2M}\sum_{k=1}^{n}\lambda_{k-1}\|e_k\|\\ &\le \|x_1-x_0\|+C\phi(f(x_1)-f^*)+\frac{1}{2M}\sum_{k=1}^{\infty}\lambda_{k-1}\|e_k\|.\end{aligned}$$

在条件 a. 下，右侧的级数收敛，因此 $\sum_k\|Px_{k+1}-Px_k\|<\infty$，从而 Px_k 是一个柯西序列，它收敛到某个 $Px^*\in V$，其中 x^* 是 x_k 的一个聚点。事实上，对于 x_k 的所有聚点 x^*，Px^* 都相同。

6) 接下来只需证明，在条件 a. 下，至少有一个聚点 x^* 是临界点。由于 $x_k\in G$，我们有

$$e_{k+1}\in\partial f(x_{k+1})+\lambda_k^{-1}\nabla_1\Psi(x_{k+1},x_k),$$

因此，对于 $g_{k+1}\in\partial f(x_{k+1})$ 和 $v_{k+1}=\nabla_1\Psi(x_{k+1},x_k)$，我们可以分别写作 $\lambda_k e_{k+1}=\lambda_k g_{k+1}+v_{k+1}$ 和 $v_{k+1}=\nabla_1\Psi$。现在，$\|\nabla_1\Psi(x_{k+1},x_k)\|\le M\|Px_{k+1}-Px_k\|$，并且 $\sum_k\|Px_{k+1}-Px_k\|<\infty$，因此 $v\in\ell_1$；又由于根据假设 a. 还有 $\lambda\cdot e\in\ell_1$，我们必有 $\lambda\cdot g\in\ell_1$。根据假设，$\lambda\notin\ell_1$，这意味着 g 不可能与 0 保持正距离。换言之，至少对一个子序列 k'，有 $g_{k'}\to 0$，于是由 $x_{k'}\to x^*$ 可得 $0\in\partial f(x^*)$，又有 $g_{k'}\in\partial f(x_{k'})$，其中 ∂f 是上半连续的。

7) 如果满足更强的条件 $\lambda_k\ge\eta>0$，那么每个聚点 x^* 都是临界点，因为在这种情况下，整个序列都必须满足 $g\in\ell_1\subset c_0$。

8) 接下来讨论条件 b. 和 c.。在条件 b. 下，我们可以直接从估计 (26) 中去掉项 $\|e_k\|$；条件 c. 也同样如此。这样一来，证明的其余部分便简化了，因为我们可以按照 $e_k=0$ 来处理。□

注 11. 1) 对于 $P=I$, $\Psi(x^+,x)=\frac{1}{2}\|x^+-x\|^2$, $\eta\le\lambda_k\le R$ 以及 $e_k=0$，Attouch 和 Bolte [5] 已证明了这一点。另见 [16, 定理 24]。

2) 当 $P=I$ 时，只需满足 $\sum_k\lambda_{k-1}\|e_k\|<\infty$，极限点 x^* 就是临界点。

3）约束 $x\in M$ 通过令 $f=f_0+i_M$直接纳入。$f=f_0+i_M$ 在 G 上的下紧性保证了序列 x_k 有界；如果 $M\subset G$是闭有界的，则显然满足该条件。

4）如果假设 $Px_k-Px_{k-1}\to 0$，则可以不要求 Ψ_V的可分离性。 5）如果假设当迭代点 M 趋近边界 ∂G时，范数上下界仍然以相同的 m，成立，那么在边界处也成立收敛性。然而，在期望最大化算法的语境下，这并不是一个现实的假设。

6.1. 收敛性引理。回顾一下，一般正则项 $\Psi(x^+,x)$关于第一个变量在 $G\times G$上属于 C^2类，其中对于每个 $y\in G$，$G\subset\mathrm{dom}\Psi(\cdot,y)$，并且对于每个 $x\in G$，$G\subset\mathrm{dom}\Psi(x,\cdot)$。此外，$\nabla_1\Psi(x,y)$和$\nabla^2_{11}\Psi(x,y)$联合连续。我们有$\Psi(\cdot,x)\ge 0$ 以及 $\Psi(x,x)=0$，因此根据必要最优性条件，$\nabla_1\Psi(x,x)=0$ 且 $\nabla^2_{11}\Psi(x,x)\succeq 0$。

引理 6。（范数上界）。设 $K\subset G$ 为紧凸集。则存在 $M>0$ 和 $\delta>0$，使得对所有 x、$y\in N$（K、δ），都有 $\|\nabla_1\Psi(x,y)\|\le M\|x-y\|$。

证明： 利用 $K\times K$ 的紧致性以及 $(x,y)\mapsto\nabla^2_{11}\Psi(x,y)$ 在 $G\times G$ 上的连续性，选取 $M>0$ 使得对所有 $x,y\in K$，都有 $\lambda_{\max}(\nabla^2_{11}\Psi(x,y))\le M/2$。然后找一个邻域

$N(K,\delta)$，它属于 K，使得对所有 $x,y\in N(K,\delta)$，都有 $\lambda_{\max}(\nabla_{11}^2\Psi(x,y))\le M$。固定 $\|h\|=1$，则 $x\mapsto h\cdot\nabla_1\Psi(x,y)$ 在 y 处的泰勒展开为

$$h\cdot\nabla_1\Psi(x,y)=h\cdot\nabla_1\Psi(y,y)+h\cdot\nabla_{11}^2\Psi(\bar y,y)(x-y)$$

对于某个 $\bar y\in(x,\ y)$，开线段，其中 $\bar y$取决于 $x,\ y$和 h。利用 $\nabla_1\Psi(y,\ y)=0$以及 $\lambda_{\max}(\nabla_{11}^2\Psi(\bar y,\ y))\le M$，我们得到 $h\cdot\nabla_1\Psi(x,\ y)\le\|h\|M\|x-y\|=M\|x-y\|$，由此得到所述估计，因为 $\|h\|=1$ 是任意的。□

我们称之为范数上界。当 $\nabla_{11}^2\Psi$ 严格为正时，有如下

引理 7。（范数下界）。设 K 是 G的一个紧凸子集。假设$\nabla_{11}^2\Psi(x,x)$ 在 K上。则存在 $\delta>0$ 和 $m>0$，使得对所有满足 $y\in N(K,\delta)$ 且 $\|x-y\|<\delta$的 x,y，均有 $m\|x-y\|^2\le\Psi(x,y)$。

证明： 1) 由于 K 是紧集，且 $\nabla_{11}^2\Psi(x,\ x)\succ 0$ 在 G上，因此存在 $m>0$ 使得$\nabla_{11}^2\Psi(x,\ x)\succeq m>0$对所有 $x\in K$都成立。由此可得 K的一个邻域 $U=N(K,\ \delta)$，使得 $\nabla_{11}^2\Psi(y,\ z)\succeq m/2$对所有满足 $z\in N(K,\ \delta)$ 和 $\|y-z\|<\delta$ 的 $y,\ z$ 都成立。

事实上，对于 $x\in K$，选取 $\epsilon_x>0$使得 $\nabla_{11}^2\Psi(y,z)\succeq m/2$对所有 $y,z\in B(x,\epsilon_x)$都成立。这是由于 $(y,z)\mapsto\nabla_{11}^2\Psi(y,z)$ 的连续性。现在令 $\Delta_K=\{(x,x):\ x\in K\}$ 为对角线，则 $\Delta_K\subset\bigcup_{x\in K}B(x,\epsilon_x/2)\times B(x,\epsilon_x/2)$，因此由 Δ_K的紧致性可知，存在有限多个 $x_i\in K$ 使得 $\Delta_K\subset B(x_1,\epsilon_1/2)\times B(x_1,\epsilon_1/2)\cup\cdots\cup B(x_n,\epsilon_n/2)\times B(x_n,\epsilon_n/2)$，其中 $\epsilon_i=\epsilon_{x_i}$。

现在令 $\delta:=\min_{i=1,\dots,n}\epsilon_i/4$。假设 $z\in N(K,\delta)$ 且 $\|y-z\|<\delta$。取满足$\|z-x\|<\delta$ 的 $x\in K$，则 $\|y-x\|<2\delta$。对于某个 x_i，有 $(x,x)\in B(x_i,\epsilon_i/2)\times B(x_i,\epsilon_i/2)$，因此 $\|y-x_i\|<2\delta+\epsilon_i/2<\epsilon_i$ 且 $\|y-x_i\|<\delta+\epsilon_i/2<\epsilon_i$。于是，根据 $B(x_i,\epsilon_{x_i})$ 的定义，$\nabla_{11}^2\Psi(y,z)\succeq m/2$。

2) 现在对 $\Psi(\cdot,y)$ 在 y 处作二阶泰勒展开，得到

$$\Psi(x,y)=\Psi(y,y)+\nabla_1\Psi(y,y)\cdot(x-y)+\tfrac12(x-y)\cdot\nabla_{11}^2\Psi(\bar y,y)(x-y)$$

对于某个 $\bar y\in(x,y)$，该点位于开线段上。因此，若 $\|x-y\|<\delta$ 且 $y\in N(K,\delta)$，则也有 $\|\bar y-y\|<\delta$；于是根据第 1) 部分，$\lambda_{\min}(\nabla_{11}^2\Psi(\bar y,y))\ge m/2$。接着利用 $\Psi(y,y)=0$ 且 $\nabla_1\Psi(y,y)=0$，可得 $\Psi(x,y)\ge(m/4)\|x-y\|^2$。□

利用引理 7 的证明，我们还可以得到以下结果

引理 8. 设 $K\subset G$ 为紧集，且 $\Psi(x^+,x)$ 在每个 $x\in K$处都有范数下界。那么存在 $\delta>0$ 和 m，使得对所有满足$\|y-z\|<\delta$的 $y,z\in N(K,\delta)$，均有 $m\|y-z\|^2\le\Psi(y,z)$。

6.2. **收敛速率**。

推论 2. 考虑定理 1 中的情形 $\lambda_k\ge\eta>0$ 和 $e_k=0$。进一步假设，对于 $\theta\in[\frac12,1)$，有 $\phi(s)=s^{1-\theta}/(1-\theta)$。若 $\theta\in(\frac12,1)$，则收敛速度为$\|Px_k-Px^*\|=O(k^{-\frac{1-\theta}{2\theta-1}})$。对于 $\theta=\frac12$，收敛速度为 R-线性。

证明： 在 Łojasiewicz 情形下，方程 (28) 可特化为

$$\|Px_k-Px_{k+1}\|\le\frac12\|Px_k-Px_{k-1}\|+\frac{C}{2}\left[(f(x_k)-f^*)^{1-\theta}-(f(x_{k+1})-f^*)^{1-\theta}\right].$$

将这种形式从 $k=N$ 求和到 $k=M$，得到

$$-\frac12\|Px_{N-1}-Px_N\|+\frac12\sum_{k=N}^{M-1}\|Px_k-Px_{k+1}\|+\|Px_M-Px_{M+1}\|\le\frac{C}{2}\left[(f(x_N)-f^*)^{1-\theta}-(f(x_{M+1})-f^*)^{1-\theta}\right].$$

取极限 $M \to \infty$，得到

$$-\frac{1}{2}\|Px_{N-1} - Px_N\| + \frac{1}{2}\sum_{k=N}^{\infty}\|Px_k - Px_{k+1}\| \le \frac{C}{2}(f(x_N) - f^*)^{1-\theta}.$$

代入 $S_N = \sum_{k=N}^{\infty}\|Px_k - Px_{k+1}\|$，可得

$$-\frac{1}{2}(S_{N-1} - S_N) + \frac{1}{2}S_N \le \frac{C}{2}(f(x_N) - f^*)^{1-\theta}.$$

现在由（26）式可得 $\phi'(f(x_N) - f^*)^{-1} \le \gamma^{-1}M\lambda_{k-1}^{-1}\|Px_N - Px_{N-1}\| = \gamma^{-1}M\lambda_{k-1}^{-1}(S_{N-1} - S_N) \le \gamma^{-1}M\eta^{-1}(S_{N-1} - S_N)$。由于 $\phi'(s) = s^{-\theta}$，这意味着

$$\begin{aligned}\phi(f(x_N) - f^*) &= (1-\theta)^{-1}(f(x_N) - f^*)^{1-\theta} \\ &= (1-\theta)^{-1}\left[\phi'(f(x_N) - f^*)^{-1}\right]^{\frac{1-\theta}{\theta}} \\ &\le (1-\theta)^{-1}(M\gamma^{-1}\eta^{-1})^{\frac{1-\theta}{\theta}}(S_{N-1} - S_N)^{\frac{1-\theta}{\theta}}.\end{aligned}$$

综上可得

$$\frac{1}{2}S_N \le C'(S_{N-1} - S_N)^{\frac{1-\theta}{\theta}} + \frac{1}{2}(S_{N-1} - S_N) \tag{29}$$

对于 $C' = (1-\theta)^{-1}(M\gamma^{-1}\eta^{-1})^{\frac{1-\theta}{\theta}}$。现在对于 $\theta > \frac{1}{2}$，我们有 $\frac{1-\theta}{\theta} < 1$，因此（29）式右侧的第一项支配第二项，从而得到

$$S_N^{\frac{\theta}{1-\theta}} \le C''(S_{N-1} - S_N)$$

当 N 足够大且 C'' 为另一个常数时。根据 [60, Cor. 4(24)ff], ，这将得到一个估计 $S_N \le C'''N^{-\frac{1-\theta}{2\theta-1}}$。

接下来讨论 $\theta = \frac{1}{2}$ 的情形。此时，由 (29) 可得

$$\frac{1}{2}S_N \le C'(S_{N-1} - S_N) + \frac{1}{2}(S_{N-1} - S_N),$$

因此

$$S_N \le \frac{1 + SC'}{2 + 2C'}S_{N-1}$$

这表明 S_N Q-线性收敛，因此 Px_k R-线性收敛。 □

注 12. 当 $\phi(s) = s^{1/2}$ 时，这是可能达到的最佳情形，Łojasiewicz不等式 (5) 就化为 Polyak-Łojasiewicz不等式。在无约束情形下，对于临界点$\bar{x}$，(5) 此时为 $(f(x) - f(\bar{x}))^{-1/2}\|\nabla f(x)\| \ge \gamma$，这意味着 f 在局部有二次下界，特别是在 $\bar{x}$ 处有严格局部极小值。事实上，假设 $f(\bar{x}) = 0, \bar{x} = 0$ 并令 $y(t) = f(th)$，其中 $\|h\| = 1$ 固定，则由 (5) 可得 $-y' \ge \gamma\sqrt{y}$，进而得到 $dy/\sqrt{y} \ge \gamma dt$，再得到 $\sqrt{y} \ge \frac{\gamma}{2}t$，即 $y \ge \frac{\gamma^2}{4}t^2$，这里用到了 $y(0) = 0$。当然，这好得令人难以置信，因此我们预计 Polyak-Łojasiewicz不等式只会在特殊情形下成立。

对于 Kurdyka条件的一般情形，我们仍然可以得出一些结论：

推论 3. 考虑定理 1 中的情形 $\lambda_k \ge \eta > 0$ *and* $e_k = 0$。那么，部分收敛的速度为 $\|Px_k - Px^*\| \le C\phi(f(x_k) - f^*)$，其中 ϕ 是去奇异函数。

证明：由于 $\Psi(x,x) = 0$ 且 $\nabla_1\Psi(x,x) = 0$，对 $\Psi(\cdot,x)$ 在 x 处作泰勒展开，得到$\Psi(u,x) = \Psi(x,x) + \nabla_1\Psi(x,x)\cdot(u-x) + \frac{1}{2}(u-x)\cdot\nabla_{11}^2\Psi(\bar{x},x)(u-x) = \frac{1}{2}(u-x)\cdot\nabla_{11}^2\Psi(\bar{x},x)(u-x)$，其中 $\bar{x}$ 位于线段 (x,u) 上。由于对于紧集 $K \subset G$，我们可以找到 $c > 0$，使得$\nabla_{11}^2\Psi(x',x) \preceq c^2 I$对所有 $x', x \in K$均成立，取 K 为迭代点及其聚点构成的集合的凸包，可得由 (4) 生成的序列的估计 $\Psi(x^+,x)^{1/2} \le c\|x^+ - x\|$。

由于从当前 x出发的 (4) 的解 x^+是精确的，令 $f(x^+) = r^+$, $f(x) = r$，则对所有 $u \in \{f \le r^+\}$，有 $\Psi(x^+, x) \le \Psi(u, x)$。因此， $\Psi(x^+, x)^{1/2} \le \min_{u\in\{f\le r^+\}} \Psi(u, x)^{1/2} \le c \min_{u\in\{f\le r^+\}} \|u - x\| = c\text{dist}(\{f \le r^+\}, x) \le c\max^{v\in\{f\le r\}} \text{dist}(\{f \le r^+\}, v) \le c\,\text{haus}(\{f \le r^+\}, \{f \le r\})$，这里用到了 x 是其中一个 $v \in \{f \le r\}$。

根据 [16, 定理 20(vi)，推论 4]，KŁ性质等价于次水平集算子的 Lipschitz 连续性。更准确地说，存在 $k > 0$，使得 $\text{haus}(\{f \le r^+\}, \{f \le r\}) \le k|\phi(r^+) - \phi(r)|$，其中 ϕ 是式 (5) 的去奇异函数。代入后得到$\Psi(x^+, x)^{1/2} \le c\cdot\text{haus}(\{f \le r^+\}, \{f \le r\}) \le ck|\phi(r^+) - \phi(r)|$。利用范数下界，我们推得 $\|Px_{k+1} - Px_k\| \le m^{-1/2}ck[\phi(f(x_k) - f^*) - \phi(f(x_{k+1}) - f^*)]$,，对两边从 $k = n$到 $k = r$ 求和，得到 $\sum_{k=n}^{r} \|Px_{k+1} - Px_k\| \le m^{-1/2}ck[\phi(f(x_n) - f^*) - \phi(f(x_{r+1}) - f^*)|$。令 $r \to \infty$，即可得到所述收敛速率 $\|Px_n - Px^*\| \le m^{-1/2}ck\phi(f(x_n) - f^*)$。□

对于库 尔迪卡不等式，我们也可以如下论证。由 (28) 可得估计 e

$$-\tfrac{1}{2}(S_{n-1} - S_n) + \tfrac{1}{2}S_n \le \tfrac{C}{2}\phi(f(x_n) - f^*).$$

固定 $\alpha \in (\frac{1}{2}, 1)$，并将整数分为两类 $\mathcal{N}_1 = \{n\colon\ S_n \le \alpha S_{n-1}\}$ 和 $\mathcal{N}_2 = \{n\colon\ S_n > \alpha S_{n-1}\}$。现在，对于 $n \in \mathcal{N}_2$，我们有

$$(1 - \tfrac{1}{2\alpha})S_n \le S_n - \tfrac{1}{2}S_{n-1} \le \tfrac{C}{2}\phi(f(x_n) - f^*).$$

另一方面，对于 $n \in \mathcal{N}_1$，我们有 $S_n \le \alpha S_{n-1}$，因此此处的误差以线性速率收缩。综上，我们得到一个序列 $n_1 < m_1 < n_2 < m_2 < \dots$，使得

$$S_n \le c\phi(f(x_n) - f^*) \text{ for } n = n_k, \dots, m_k - 1$$
$$S_n \le \alpha^{n-m_k} S_{m_k}, \qquad \text{for } n = m_k, \dots, n_{k+1} - 1,$$

其中 $c = \frac{C}{2}(1 - \frac{1}{2\alpha})^{-1}$，这对推论 3 作了细微改进，因为即使对于 ϕ 去奇异化程度低于 $s^{1/2}$ 的情形，它仍保留了整个序列 S_n 以速率 α 线性收敛的可能性。

7. Convergence of the EM algorithm

现在我们将其应用于 EM算法。首先，我们假设内点性。

定理2。（约束指数族的收敛）。设 $q(y, \theta)$ 是来自一个极小的 n 维指数族 $p(x, \theta)$的不完整数据。假设 $-\log q(y, \cdot)$和 M 均为可定义的。令 $\theta^{(k)} \in M$ 为由约束期望最大化算法生成的有界序列，且该序列及其聚点均处于 G中。精确参数序列 $\theta'^{(k)}$ 收敛， $\theta'^{(k)} \to \theta'^*$。序列 $\theta^{(k)}$的每个聚点 θ^* 都能求解约束不完全数据最大似然估计问题，具有相同的投影 $P\theta^* = \theta'^*$，且条件分布 $\mathbb{P}^{x|y}_{\theta^{(k)}}$ 弱收敛至 $\mathbb{P}^{x|y}_{\theta^*}$，对所有 θ^*而言极限相同。此外，序列 $T(x_k)$ 也收敛。 _

证明：由于 $-\log q(y, \cdot)$ 和 M均为可定义的，因此目标函数 $f = -\log q(y, \cdot) + i_M$（见式（16））也可定义。因此， f 具有KŁ性质（5）。根据假设，该序列及其聚点均处于 G中，满足主要收敛结果的要求。

根据命题3，对于精确参数 θ'，我们有 $\nabla^2_{\theta'\theta'}\bar\psi_y(\theta') \succ 0$，因此部分正则化项 $\Psi(\theta^+, \theta) = \bar K_y(\theta'||\theta'^+)$满足 $\nabla^2_{22}\bar K_y(\theta'||\theta') \succ 0$。因此，它在 $\theta'^{(k)}$ 包含于维数为 m的子空间 V 中的聚点紧集上具有范数下界。此外，由于族 $k(\cdot|y, \theta')$是极小的，部分正则化项在子空间 V上具有分离性。因此，我们可以应用主收敛定理（取 $\lambda_k = 1$），从而得到 $\theta'^{(k)}$的收敛。

由于 $\mathbb{P}^{x|y}_{\theta^{(k)}} = \mathbb{P}^{x|y}_{\theta'^{(k)}}$ 根据命题3，且由于右侧序列弱收敛，因此左侧序列也弱收敛。根据 F. Riesz的经典结果，还可得 $\bar k(\cdot|y, \theta'^{(k)})$收敛于 $\bar k(\cdot|y, \theta'^*)$，在 $L^1(h^{-1}(y)\ d\mu_y)$中成立。但由于（20），序列 $k(\cdot|y, \theta^{(k)})$也在 L^1中收敛，而无论 $\theta''^{(k)}$ 是否收敛。

作为主收敛定理的推论，完整序列 $\theta^{(k)}$的每个聚点 θ^* 都是临界点，并且 $P\theta^* = \theta'^*$对每个这样的 θ^*都成立。$T(x_{k+1})$的收敛性源于： $T(x_{k+1})$是计算 $\mathbb{E}_{\theta'^{(k)}}[T(x)|y]$的结果，而后者关于 $\theta'^{(k)}$连续。 □

备注 13。 在定理1中采用更一般的 λ_k 和时，也能保证部分收敛，但这需要分别计算 $-\log\, q(y,\ \theta^+)$ 和 $K_y(\theta'||\theta'^+)$。在实践中，人们更倾向于使用 Q函数。

推论 4。 在定理2的假设下，假设 $T(h^{-1}(y))$ 不包含于 $\mathbb{R}^n$的任何真仿射子空间中。那么整个序列 $\theta^{(k)}$ 收敛。

证明： 在这种情况下不会发生降维，并且有 $\theta^{(k)} = \theta'^{(k)}$。因此$\theta^{(k)}$收敛。□

7.1. **边界处的问题**。 由 (6) 可知，$\mathrm{dom}\ q(y,\ \cdot) = \Theta$；而由 (8) 可得，对每个 $\theta' \in G$，都有 $\Theta \subset \mathrm{dom}\Psi(\cdot,\ \theta')$。但这并不能排除迭代点同时位于 $\mathrm{dom}(-\log\, q(y,\ \cdot))$ 和 $\mathrm{dom}\ \Psi(\cdot,\ \theta')$ 的边界上，从而引发问题（见例 1）。我们来仔细考察一下。

定义 3. 如果 $f(\theta) = -\log\, q(y,\theta) + i_M(\theta) = \infty$对所有 $\theta \in \partial\Theta$ 都成立，我们就称不完全数据问题是正则的；如果 $\partial f(\theta) = \partial(-\log\, q(y,\cdot) + i_M)(\theta) = \emptyset$对所有 $\theta \in \partial\Theta$ 都成立，则称其陡峭。

注14. 1) 我们有 $\partial f(\theta) = \partial(-\log\, q(y,\cdot) + i_M)(\theta) = \partial(-\log\, q(y,\cdot))(\theta) + \mathcal{N}_M(\theta)$，因为$\theta \in G$ 由 [69, 8 得出。8. c] or [55], 但是和法则在边界 $\partial\Theta$ 处失效，除非额外作出正则性假设。这就是定义中使用 $\partial(-\log\, q(y,\cdot) + i_M)$ 的原因。

2) 正则蕴含陡峭。如果 $M \subset G = \mathrm{int}(\Theta)$，那么该问题自动是正则的。

3) 当 $q(y,\cdot)$ 属于指数族且满足 $M = \Theta$ 时，按照定义所说的陡峭性等价于对数归一化函数 ψ_q of q 的陡峭性。

4) 假设完全数据族是陡峭的，即只要 $\theta \in \partial\Theta$，就有 $\partial_\theta(p(x,\cdot) + i_M)(\theta) = \emptyset$ for μ_y-a.a. $x \in h^{-1}(y)$。那么，不完全数据问题的陡峭性可由包含关系 $\partial_\theta(q(y,\cdot) + i_M)(\theta) \subset \int_{h^{-1}(y)} \partial_\theta(p(x,\cdot) + i_M)(\theta) d\mu_y(x)$ 得出（见 $[26, \text{定理 } 2.7.2]$）。当 $p(x,\theta)$ 为指数型时，就会出现这种情况，因为此时 $\partial_\theta p(x,\theta) = p(x,\theta)(T(x) + \partial\psi(\theta))$，因此 $\partial\psi(\theta) = \emptyset$ for $\theta \in \partial\Theta \cap M$ 可推出对所有 x 都有 $\partial_\theta p(x,\theta) = \emptyset$。这为该定义提供了依据。

设 $\theta^{(k)}$ 为约束EM算法生成的序列，那么每个聚点 θ^*都必须具有有限值 $f(\theta^*) < \infty$，因为 $\theta^{(k)} \in \{f \le f(\theta^{(1)})\}$，且由 f 的下半连续性可知，该集合是闭集。因此，在正则情形下，没有 θ^* 会位于边界 $\partial\Theta$上。由此，对于有界的 $\theta^{(k)}$，定理2中的内点性假设自动成立。例如，当 f 下紧时，有界性便得到保证。换言之，如果不完全数据问题是正则的，那么边界 $\partial\Theta$处不会有任何问题。

现在假设 f是陡峭的，并令 $e_k = 0$。根据最优性， $-\lambda_{k-1}^{-1}\nabla_1\Psi(\theta^{(k)},\ \theta^{(k-1)}) \in \partial f(\theta^{(k)})$, 因此 $\theta^{(k)}$ 不可能位于边界 $\partial\Theta$上，且该方法定义良好。假设 $\theta^{(k)} \to \theta^* \in \partial\Theta$ 对某个子序列 $k \in N \subset \mathbb{N}$成立。那么由陡峭性可知， θ^* 不可能是约束不完全数据最大似然估计规划的解，因为这将导致 $0 \in \partial f(\theta^*) = \partial(-\log\, q(y,\ \cdot) + i_M)(\theta^*) = \emptyset$。遗憾的是，仅凭陡峭性并不能阻止迭代点趋近于 $\partial\Theta$，但至少我们知道，在这种情况下，这些迭代点会偏离正轨。

注 15. 当 $M \subset G = \mathrm{int}(\Theta)$ 有界时，序列 $\theta^{(k)}$ 及其聚点始终位于 G中，从而避免了边界 $\partial\Theta$ 处的所有问题。反过来，假设我们希望对满足内点性条件的序列 $\theta^{(k)}$ 作出论述，即该序列有界且其聚点包含在 G 中。那么，我们可以用闭子集 $M' \subset M \cap G$ 替换约束集 $M \subset \Theta$，使得该序列可以

视为在约束 $\theta^{(k)} \in M'$下生成。集合 M'可以选为有界集；若 M 可定义，则该集合也可定义；由于 G是凸集，该集合也可以选为凸集。参见 [59,第 2.2] 节。

7.2. 对M步的影响。 在建立了内点性下的部分收敛 $\theta'^{(k)} \to \theta'^*$ 之后，我们现在探讨在什么条件下可以将其提升为完整参数序列 $\theta^{(k)}$的收敛。注意，在新的坐标 $Q\theta = (\theta',\ \theta'')$ 下，M步（16）具有如下等价形式：

$$(\theta'^{(k+1)}, \theta''^{(k+1)}) \in \operatorname*{argmin}_{\theta',\theta''} -\log q(y,\theta',\theta'') + i_M(\theta',\theta'') + \lambda_k^{-1}\Psi(\theta',\theta'^{(k)}),$$

with

$$\Psi(\theta',\theta'^{(k)}) = -\mathbb{E}_{\theta'^{(k)}}\left[\log\frac{\bar{k}(x|y,\theta')}{\bar{k}(x|y,\theta'^{(k)})}\middle| y\right] = \bar{K}_y(\theta'^{(k)}||\theta'),$$

一个与变量 θ'' 无关的部分正则项。为简便起见，将 θ 与（θ'，θ''）对应，定义 $M(\theta') = \{\theta'' :\ (\theta',\ \theta'') \in M\}$，那么我们可以将M步优化程序拆分如下：

$$(P_k)\quad \theta''^{(k+1)} \in \operatorname*{argmin}_{\theta''\in M(\theta'^{(k+1)})} -\log q(y,\theta'^{(k+1)},\theta''),$$

这只是关于 θ'' 的一系列参数化优化程序，其中 $\theta'^{(k+1)}$ 是参数，并且不再有任何正则化影响 θ''。显然，极限程序为：

$$(P_\infty)\quad \theta''^* \in \operatorname*{argmin}_{\theta''\in M(\theta'^*)} -\log q(y,\theta'^*,\theta''),$$

并且由于 $\theta'^{(k)} \to \theta'^*$，EM序列 $\theta^{(k)}$ 的每个聚点（θ'^*，θ''^*）都给出（P_∞）的一个解 θ''^*，且其不完全数据最大似然估计值相同，为 $q^* = q(y,\ \theta^*)$。由此立即得到：

命题4. 如果极限程序 (P_∞) 在临界值为 q^* 的临界点中有唯一的临界点 $\theta''^* \in M(\theta'^*)$，那么期望最大化算法 序列 $\theta^{(k)}$ 收敛。

这是一个较弱的假设；例如，它不像 [78], 中那样要求完整的不完整数据最大似然估计程序具有一个取正确最大似然估计值的唯一解 θ^*。

注16. 但是请注意，在每个程序（P_k）中，如果有多个局部解，我们可以自由选择任意一个。如果存在 $\epsilon > 0$，使得对于每个（P_k），都可以选出两个相距 ϵ 的局部解，那么备用序列 $\theta''^{(k)}$ 必然不收敛。例如，[76, 第 4] 节中的反例就是如此。

第二个推论基于以下命题。

命题 5. 假设完全数据指数族 $p(x,\theta)$ 是极小的。那么，对于固定的 θ'，$-\nabla^2_{\theta''\theta''}$对数 $q(y,\theta',\theta'') \succ 0$。

证明：采用标准记号，定义 $I(\theta,y) = -\nabla^2_{\theta\theta}\log\ q(y,\theta)$，于是 $\mathcal{I}(\theta) = \mathbb{E}_\theta[I(\theta,y)]$ 是不完整数据的期望 Fisher 信息。对于完全数据，定义$I_c(\theta,x) = -\nabla^2_{\theta\theta}\log\ p(x,\theta)$，则 $\mathcal{I}_c(\theta) = \mathbb{E}_\theta[I_c(x,\theta)]$ 是完全数据的期望 Fisher 信息。同理，还令 $\mathcal{I}_c(\theta,y) = \mathbb{E}_\theta(I_c(\theta,x)|y) = \nabla^2\psi(\theta)$，表示给定 y 时完全数据的条件期望 Fisher 信息。

现在由 (8) 得到 $\log\ p(x,\theta) = \log\ q(y,\theta) + \log\ k(x|y,\theta)$，因此两次求导得到

$$I_c(\theta,x) = I(\theta,y) - \nabla^2_{\theta\theta}\log k(x|y,\theta).$$

对给定 y 的 x 取条件期望，我们得到

$$\mathcal{I}_c(\theta,y) = I(\theta,y) + \mathcal{I}_m(\theta,y), \tag{30}$$

其中 $\mathcal{I}_m(\theta,\ y) = \mathbb{E}_\theta(-\nabla^2_{\theta\theta}\log\ k(x|y,\ \theta)|y)$ 是给定 y条件下缺失数据的期望 Fisher 信息。然而，后者先前已被确定为正则项的二阶导数。

现在，令 $k(x|y,\ \theta)$ 为一个指数族，并采用式 (22) 中的坐标变换，我们知道

$$\mathcal{I}_m(\theta,y)=\left[\begin{array}{c|c}\nabla^2_{\theta'\theta'}\bar{\psi}_y(\theta') & 0\\ \hline 0 & 0\end{array}\right],\qquad \nabla^2_{\theta'\theta'}\bar{\psi}_y(\theta')\succ 0,$$

而

$$I(\theta,y)=-\nabla^2_{\theta\theta}\log q(y,\theta',\theta'')=\left[\begin{array}{c|c}-\nabla^2_{\theta'\theta'}\log q(y,\theta',\theta'') & -\nabla^2_{\theta'\theta''}\log q(y,\theta',\theta'')\\ \hline * & -\nabla^2_{\theta''\theta''}\log q(y,\theta',\theta'')\end{array}\right].$$

根据假设，$p(x,\ \theta)$是极小的，因此我们有 $I_c(\theta,\ x)=\nabla^2\psi(\theta)\succ 0$。因此，$\mathcal{I}_c(\theta,\ y)\succ 0$也成立。因此，式(30)右侧的矩阵 $\mathcal{I}_m(\theta,\ y)$和 $I(\theta,\ y)$相加后必须得到一个满秩矩阵。由于 $\mathcal{I}_m(\theta,\ y)$的结构，这意味着$-\nabla^2_{\theta''\theta''}\log\ q(y,\ \theta)\succ 0$对于 $I(\theta,\ y)$中的下对角块成立。□

将此结论应用于 θ'^*，可知式(P_∞)的目标函数 $-\log\ q(y,\ \theta'^*,\ \cdot)$是严格凸的。因此我们得到以下推论:

定理3. 在定理2的假设下，假设 $M(\theta'^*)$ 是凸的。那么，*EM*序列 $\theta^{(k)}$ 收敛。

证明： 严格凸函数在凸域上有唯一的最小值 . □

注意，若 M 是凸的，则 $M(\theta'^*)$ 显然是凸的，但 $M(\theta'^*)$ 的凸性是一个较弱的假设。特别地，我们有以下

推论 *5.* 假设约束集为 $M=\{\theta(u):\ u\in U\}=\{(\theta'(u)\ \theta''(u)):\ u\in U\}$，其中 $U\subset\mathbb{R}^m$, $\theta'(\cdot)$ 属于 C^1 类，且 $\theta''(\cdot)$ 连续，并且二者均可定义。设 $\theta'(u^*)=\theta'^*$，并假设雅可比矩阵 $\frac{d\theta'}{du}(u^*)$ 的秩为 m。则 *EM*序列 $\theta^{(k)}$ 收敛。当 $\theta''(\cdot)$ 局部利普希茨时，$\theta^{(k)}$ 的收敛速度与 $\theta'^{(k)}$ 的收敛速度相同。

证明： 根据秩假设，映射 $u\mapsto\theta'(u)$ 局部存在左逆，即存在一个 C^1 映射 $\theta'\mapsto u(\theta')$，定义在 θ'^* 的某个邻域内，使得 $u(\theta'^*)=u^*$ 且$u(\theta'(u))=u$。于是 $\theta''=\theta''(u)=\theta''(u(\theta'))$，因此 θ'' 是 θ'的函数。于是 $M(\theta'^*)$ 是单点集，故为凸集。更直接的论证是，由于 $\theta''(\cdot)$ 和 $u(\cdot)$ 的连续性，$\theta''^{(k)}=\theta''(u(\theta'^{(k)}))$ 收敛。关于收敛速度的结论成立，是因为当 $\theta''(\cdot)$ 局部利普希茨时，$\theta''(\cdot)\circ u(\cdot)$ 也局部利普希茨。□

注释 17. 1）假设每个 (P_k) 都有唯一解，并不足以保证收敛，因为得到的 $\theta''^{(k)}$ 仍可能有多个聚点。

2）在序列 $\theta''^{(k)}$ 的聚点集合是离散的这一稍显人为的假设下，只要 $\theta''^{(k)}-\theta''^{(k-1)}\to 0$，就能得到收敛。

3）当解集 $\operatorname{argmin}_{\theta''}-\log\ q(y,\ \theta',\ \cdot)+i_{M(\theta')}$ 对参数θ' 的依赖在紧集 $\{\theta'^{(k)}:\ k\in\mathbb{N}\}\cup\{\theta'^*\}$上是上利普希茨的（参见 [43]）时，即可推出收敛，因为这会使得某个 $L>0$ 满足 $\|\theta''^{(k)}-\theta''^{(k-1)}\|\le L\|\theta'^{(k)}-\theta'^{(k-1)}\|$ ；又由于$\sum_k\|\theta'^{(k)}-\theta'^{(k-1)}\|<\infty$，序列 $\theta''^{(k)}$ 也满足柯西条件。对于非线性规划约束 M，[65, 43], 中讨论了充分条件；这些条件通常是局部的，并要求满足温和的正则性假设。这里，这些条件必须在备用序列 $\theta''^{(k)}$的所有聚点 θ''^* 处成立。

定理 4。（指数族的正则化 EM）。在定理 2 的假设下，假设将 M 步正则化为 $\min_{\theta\in M}-Q(\theta,\theta^{(k)})+\lambda_k^{-1}\|\theta''-\theta''^{(k)}\|^2$。则序列 $\theta^{(k)}$ 收敛于临界点 θ^*，该临界点是不完全数据问题的最大似然估计。 不完全数据负对数似然的值仍然单调递减。

证明： 根据命题 1，我们修改了 M步，使正则项现在为 $\Psi(\theta,\ \theta^{(k)})=\bar{K}_y(\theta'^{(k)}||\theta')+\|\theta''-\theta''^{(k)}\|^2$，它不再是部分正则项，而是完整正则项。我们应用

带有 $P = I$ 的主收敛定理，由此得到 $\theta^{(k)}$ 的收敛。由于 (4) 总是使目标函数减小，因此最后一个结论成立。□

7.3. 可定义目标函数。 我们要探究目标函数 $f(\theta) = -\log q(y,\theta) + i_M(\theta)$ 在式（16）中是否以及何时能在某个 o-极小结构中可定义，因为这是我们确保 KŁ 不等式（5）成立的方式。首先考察 $-\log q(y,\theta)$，我们先逐一分析 $q(y,\theta)$ 本身属于指数族的情形，例如 $q(y,\theta) = \exp\{\langle\theta, T(y)\rangle - \psi_q(\theta) + h(y)\}$。此时，可定义性取决于相应对数归一化函数 ψ_q的可定义性。

查阅指数族列表可知，除代数表达式外，对数归一化函数 $\psi(\theta)$ 有时还包含某些分量 θ_i ofθ 的 $\log\theta_i$ 项。这些项在$\mathbb{R}_{an,exp}$中可定义。此外，如果这些 θ_i 能预先得到上界，且其下界严格大于 0，则可在 $\mathbb{R}_{an}$ 中得到可定义性。逆伽马分布和 χ^2-分布需要针对 θ 的某些分量包含 $\log\Gamma(\theta_i)$ 项，这要求伽马函数具有可定义性。近期已有研究对此进行了探讨，例如 [62], ；就我们的目的而言，只需再次保证这些 θ_i 远离 0 即可。

第二种情况是，$q(y,\ \theta)$ 是来自指数族的不完整数据，正如 [73], 中所称，但其本身并不来自指数族。此时，由于 $-\log q(y,\ \theta) = -\log p(x_k,\ \theta) + \lambda_k^{-1}\bar{K}_y(\theta'^{(k)}||\theta')$，$\log q(y,\ \cdot)$ 的可定义性可以由 $\log p(x_k,\ \cdot)$ 的可定义性与 $\bar{K}_y(\theta'^{(k)}||\cdot)$ 的可定义性共同推出。前者在 ψ 可定义时得到保证，正如上文所述。对于后者，我们使用命题 2，该命题表明，必须满足由 ψ_y 诱导的 Bregman距离具有可定义性；而这可由 ψ_y的可定义性推出。

M 的可定义性甚至更加简单，因为 M 通常汇集了如下形式的等式约束和不等式约束：$M = \{\theta \in \mathbb{R}^n\colon\ f_i(\theta) = 0,\ i \in I,\ g_j(\theta) \le 0,\ j \in J\}$，其中$I$，$J$为有限集合，$f_i$，$g_j$为可定义函数，通常是亚解析的，甚至是代数的。请注意，M甚至可能有助于将分量 θ_i 限制在一个有界区间内，从而可以用更方便的结构 $\mathbb{R}_{an}$替代$\mathbb{R}_{an,\ exp}$，此时（5）就变成了Łojasiewicz不等式。

注记 18。 当 f 在 $\mathbb{R}_{an}$中可定义时，推论 2 给出收敛速率 $\|\theta'^{(k)} - \theta'^*\| = O(k^{-\rho})$。此外，若 $\theta'' = \theta''(\theta')$是局部利普希茨的，则完整参数序列也具有相同的收敛速率。这一点在推论 4 和定理 4 中成立，在推论 5 的情形下也成立。在定理 3 中，由于 $\nabla^2_{\theta''\theta''} - \log q(y,\ \theta',\ \cdot) \succeq \epsilon > 0$ 在紧集 $\{\theta'^*\} \cup \{\theta'^{(k)}$ 上的 $\theta'\colon\ k \in \mathbb{N}\}$，这一结论也成立；前提是 M 由足够光滑且可定义的等式约束和不等式约束给出，并且满足 MFCQ，参见 [43]。

在情形 $\phi(s) = s^{1/2}$下可获得线性速度，这对应于 Polyak-Łojasiewicz不等式。遗憾的是，这是一个相当强的假设（这一点也可参见第 8 节）。

7.4. 失败的原因。现在我们可以列出不完整数据的受约束指数族 EM算法可能无法收敛到临界点的原因：

(1) 迭代点可能无界，或趋于 Θ 的边界。一旦排除这些情况：(2) 收敛仍可能失败，因为 $-\log q(y,\ \cdot) + i_M$不具有 KŁ性质。但即使 f 具有 KŁ性质：(3) 仍可能只有精确参数 $\theta'^{(k)}$ 收敛，而备用序列 $\theta''^{(k)}$ 不收敛。如果 $M(\theta'^*)$不是凸的，就可能出现这种情况。(4) 但即使 $M(\theta'^*)$是凸的（包括无约束情形），$\theta''^{(k)}$ 仍可能无法收敛，因为 $-\log q(y,\ \theta'^*,\ \cdot)$不是严格凸的。这可能是因为 $p(x,\ \theta)$不是极小的。这个问题可以在构建问题时避免。

在曲线族 $M = \{\theta(u)\colon\ u \in U\}$ 中，收敛的可能性反而更大，因为部分自由度被消除了。正如我们所见，当存在连续依赖 $\theta'' = \theta''(\theta')$ 时，整个序列必然收敛。即使这种情况过于乐观，毕竟缺失数据所占比例可能小于观测数据所占比例，我们仍然可以得到 $\theta = (\theta',\ \theta'',\ \theta''')$，其中 θ' 是精确参数，$\theta'' = \theta''(\theta')$

一部分剩余参数实际上依赖于 θ'，因此必然收敛；剩下的 θ''' 个剩余坐标则需要额外条件才能收敛。

注19. 由仿射约束导致的降维（引理1）对于 Θ EM参数序列 $\theta^{(k)}$的收敛问题并非关键。充分统计量的非极小性导致的降维则不同（命题3）。

8. 交替Bregman投影

本节通过将EM算法与 [2]的em算法相匹配，找出更多能够保证 $\theta^{(k)}$ 收敛的情形。我们考虑来自一个假定为极小的约束指数族的缺失数据情形，其中 $T(x) = (T_1(x)\ T_2(x)) = (y, z)$， y 为观测变量， z 为隐变量，并相应地对 $\theta = (\theta_y, \theta_z)$进行划分。完全数据族的极小性保证 $\eta = \nabla\psi(\theta) = \mathbb{E}_\theta[T(x)]$ 是一个从 G 到 G^* 的微分同胚，其逆函数为 $(\nabla\psi)^{-1} = \nabla\psi^*$，因此我们可以使用期望参数 η。以同样的方式划分 $\eta = (\eta_y, \eta_z) = (\nabla_{\theta_y}\psi(\theta)\ \nabla_{\theta_z}\psi(\theta))$，我们将数据集定义为

$$D = \{\vartheta \in \Theta : \vartheta = \nabla\psi^*(\eta), \eta_y = y\},$$

其中，我们将数据参数记为 $\vartheta \in D$，将模型参数记为 $\theta \in M$。我们有

引理9。（Amari [2]）。设 $\theta \in \Theta$。则数据集 D 上的右*Bregman*投影 $\vartheta = \vec{P}_D(\theta)$ 是唯一的，满足 $\vartheta_z = \theta_z$，并且 $\mathbb{E}_\theta[z|y] = \mathbb{E}_\vartheta[z|y]$。

在信息几何中， $\vartheta_e = \vec{P}_D(\theta)$ 被称为从 $\theta \in M$ 出发的e步。事实证明，从 $\theta \in M$ 出发的E步也可以表示为数据集中的点 $\vartheta_E \in D$，并且它将下一个M步生成为左Bregman投影 $\theta^+ \in \overleftarrow{P}_M(\vartheta_E)$。

引理10。（Amari [2]）。设 $\eta_E = \nabla\psi(\vartheta_E)$ 和 $\eta_e = \nabla\psi(\vartheta_e)$ 分别为从 $\theta \in M$ 出发的*E*步和*e*步的期望参数。则 $\eta_E = (y, \mathbb{E}_{\vartheta_e}[z|y]) \in \nabla\psi(D)$，以及 $\eta_e = (y, \mathbb{E}_{\vartheta_e}[z]) \in \nabla\psi(D)$。

E步与e-步何时 совпп coincide，可由下述内容回答：

引理11。（Amari [2]）。来自 θ 的*E*步与*e*步 *совп coincide*，当且仅当 $\mathbb{E}_{\vartheta_e}[z|y] = \mathbb{E}_{\vartheta_e}[z]$ 对于 $\vartheta_e = \vec{P}_D(\theta)$ 成立。若始终满足这一条件，则期望最大化算法 与*em*算法生成相同的迭代点。

当Amari 条件 $\mathbb{E}(Z|Y = y) = \mathbb{E}(Z)$ 满足时，我们称 Z 基于对 Y的知识是不可预测的。这是一种介于较强的独立性（of$Z$， Y）与较弱的不相关性（cov(Z， Y）$= 0$）之间的性质。

设Amari 条件得到满足。则E步是将迭代点 $\theta^{(k)} \in M$ 投影到数据集 $\vartheta^{(k+1)} = \vec{P}_D(\theta^{(k)})$ 上的右Bregman投影，而M步是将E步迭代点 $\vartheta^{(k)}$ 投影到模型集 M上的左Bregman投影，即$\theta^{(k)} \in \overleftarrow{P}_M(\vartheta^{(k)})$；Bregman距离由完全数据族 $p(x, \theta)$ 的对数归一化函数 ψ 所诱导。与 [59], 中一样，我们用构件图对此进行可视化

$$\vartheta \xrightarrow[m]{l} \theta \xrightarrow[e]{r} \vartheta^+ \xrightarrow[m]{l} \theta^+ \qquad \vartheta^{(k)} \xrightarrow[m]{l} \theta^{(k)} \xrightarrow[e]{r} \vartheta^{(k+1)} \xrightarrow[m]{l} \theta^{(k+1)} \tag{31}$$

定理5。（通过信息几何实现收敛）。设 *Amari* 条件成立。则针对指数族缺失数据的约束期望最大化算法期望最大化算法会生成 D与 M之间交替*Bregman*投影的序列 $\theta^{(k)} \in M$, $\vartheta^{(k)} \in D$，其中 $\theta^{(k)} \in M$ 由M 步 $\theta^{(k)} \in \overleftarrow{P}_M(\vartheta^{(k)}$生成， $\vartheta^{(k)} = \vec{P}_D(\theta^{(k-1)})$ 则由E 步生成。假设 $p(x, \theta)$ 是极小的， ψ, M是可定义的，并且假设 $M \subset G$ 有界。那么：

(a) 序列 $\vartheta^{(k)}$ 收敛到某个 $\vartheta^* \in D$。(b)序列 $\theta^{(k)}$ 的每个聚点 θ^* 都求解约束不完全数据最大似然估计问题，并满足 $\vartheta^* = \vec{P}_D(\theta^*)$、 $\theta^* \in \overleftarrow{P}_M(\vartheta^*)$。存在 z^*，使得对数 $p(y, z^*, \vartheta^*) = \mathbb{E}_{\theta^*}$(对数 $p(y, z, \theta)|y)$ 对每个 θ^* 都相同。

(c) 假设 $D \cap M \neq \emptyset$。那么 $D \cap M$ 存在一个邻域 U，使得任何进入 U 的期望最大化算法 序列都会收敛到某个点 $\theta^* \in D \cap M$，即 $\theta^{(k)} \to \theta^*$ 且 $\vartheta^{(k)} \to \theta^* = \vartheta^*$。(d) 精确参数序列 $\theta'^{(k)}$ 收敛到某个 θ'^*，条件分布弱收敛，并且每个 θ^* 都会产生相同的 $P\theta^* = \theta'^*$, $\vartheta^* = P_D(\theta^*)$。(e) 当 $M(\theta'^*)$ 为凸时，序列 $\theta^{(k)}$ 收敛到某个 $\theta^* \in M$，且$\vec{P}_D(\theta^*) = \vartheta^*$ 并且 $\overleftarrow{P}_M(\vartheta^*) = \theta^*$。(f) 假设该期望最大化算法 实例无约束。那么每个具有内点性的序列 $\theta^{(k)}$ 都会收敛。

证明： 情形(a)。我们采用 [59], 中的记号 (1.4)，其中 $a_k \xrightarrow{l} b_k \xrightarrow{r} a_{k+1}$ 表示左投影和右投影。与 (31) 对应可得 $A = D$ 和 $B = M$、$\vartheta^k = a_k$ 和 $\theta^{(k)} = b_k$，从而可以应用 [59] 的结果。于是 [59, 定理 8.1] 给出 $a_k = \vartheta^{(k)}$ 的收敛性。这里，lr-角条件 [59, 定义 8.1] 可由 L, M 和 ψ 的可定义性推出。lr-三点不等式可由凸性，通过 [59, 命题 6.4] 推得，其中 $\nabla\psi(A) = \nabla\psi(D)$ 为凸，因为 $\nabla\psi(D)$ 是仿射空间 $L = \{\eta : \eta_y = y\}$ 与 im($\nabla\psi$) 的交集。这就证明了 (a)。

情形(b)。这是一个一般事实：根据 Amari 条件，期望最大化算法与上述 D 和 M 之间的交替 Bregman 投影相一致。

情形(c)。这利用了 [59, 推论7，4], 它借助近端正则性给出了更强的结论（另见命题6）。情形(d)是定理2。情形(e)是定理3。

情形(f)。最后一种情形是约束为Θ，且序列 $\theta^{(k)}$ 有界，并且与其聚点一起包含在 G中。此时，我们可以找到一个闭有界凸可定义集 $M \subset G$，使得 $\theta^{(k)}$ 在 D 和 M 之间交替（见 [59, 第2节2]）。于是$\overleftarrow{\quad}$ $\vec{\quad}$两个投影都是唯一的：P_M 因为 M 是凸的；P_D 因为 $\nabla\psi$（D）$= L \cap$im($\nabla\psi$)，因为仿射子空间与im($\nabla\psi$)的交也是凸的。M 和$\nabla\psi$（D）的凸性也保证了 rl三点不等式和 lr三点不等式成立（见 [59, 命题64]）。ψ 的可定义性意味着 Θ 和 Θ 可定义，因而 M也可定义；同时也意味着$\nabla\psi$可定义，进而im($\nabla\psi$)可定义。又由于 L 作为仿射子空间是代数的，因此我们得到$\nabla\psi$（D）的可定义性。D 的可定义性现在也随之成立，因为 D 是可定义集 $\nabla\psi$（D）在可定义微分同胚 $\nabla\psi^*$下的像，见 [27]。因此，rl角条件和 lr角条件都成立（关于 rl，见 [59, 命题53]；关于 lr，则利用对偶性参见 [59,第5节2] ）。

现在，根据 [59, 定理 7.1], ，序列 $b_k = \theta^{(k)}$ 的收敛性随之得到；根据 [59, 定理 8.1]，序列 $a_k = \vartheta^{(k)}$ 的收敛性也随之得到。一般而言，这些序列收敛到一个间隙（ϑ^*, θ^*），即 $\vartheta^* = \vec{P}_D(\theta^*)$、$\theta^* = \overleftarrow{P}_M(\vartheta^*)$，可能还带有 $\vartheta^* \neq \theta^*$。□

注 20. 1）在信息几何中，[2, 3] $\overleftarrow{P}_M$ 称为 m-投影，$\vec{P}_D$ 称为 e-投影。m-测地线，或在左投影点处垂直于 M 的垂线，在 θ坐标中是弯曲的；而 e-测地线，或在右投影点处垂直于 D 的垂线，在 θ坐标中是直线（见 [59, 13]）。集合 M 是 m-平坦的，当且仅当向 M 作左投影是唯一的；集合 D 是 e-平坦的，当且仅当向 D 作右投影是唯一的。使用 [13], 并假设 ψ 是 1-强制的，m-平坦等价于 M 凸，而 e-平坦等价于 $\nabla\psi$（D）凸。

2) 或者，$D(x, \cdot)$ 的 Legendre性和严格凸性也能保证 C 凸时 $\vec{P}_C$ 的唯一性（参见 [11]），因此这里的凸集也是 e-平坦的。不过，这在当前情形下用处不大，因为凸的是 $\nabla\psi(D)$，而不是 D。

3) 信息几何文献中反复出现过陈述 (f)，但并未假设 ψ 具有可定义性。我们知道有几篇已发表的证明是错误的。我们自己的证明需要 KŁ条件，当然也希望知道能否避免这一条件。注意，我们可以处理 $D \cap M = \emptyset$ 的情形。

4) 情形(f)也可以由定理3推导得出。

5) 情形(f)可能存在局部最小值，或收敛到鞍点，这证实了即使是最简单的期望最大化算法实例，也不属于凸性范畴，尽管 $-\log q(y, \cdot)$和$-\log p(x^+, \cdot)$是凸的。例 5 给出了一个简单情形： $D \cap M$中有两个点，且 D 与 M 之间存在非零间隙。

6) KŁ条件中的情形 $\phi(s) = s^{1/2}$ 会导致线性收敛。例如，在满足 D, M横截相交的点 $\theta^\sharp \in D \cap M$附近，就会出现这种情况。（关于Bregman意义下横截相交的定义，见 [59]）。

注21. 与其要求e-平坦和m-平坦，不如只要求对足够接近的点，投影具有唯一性。这引出了近正则性或正到达性的概念，后者见 [59]。二者的显著区别在于，正到达性在对偶变换下保持不变，即 D 具有正到达性，当且仅当 $\nabla\psi(D)$ 具有正到达性； M也同样如此，而e-平坦和m-平坦则不具有自对偶性。下面给出一个示例结果：

命题6. 考虑满足定理2假设的缺失数据情形。假设集合 M 在序列$\theta^{(k)}$的聚点 θ^* 处具有左*Bregman*到达半径 $r^* > 0$ 。令 $\vartheta^* = \vec{P}_D(\theta^*)$ ，并假设 最小值$_{\theta \in M}\, D(\theta, \vartheta^*) < \frac{1}{2} r^{*2}$。那么序列 $\theta^{(k)}$ 收敛。

证明： 我们知道，数据集序列 $\vartheta^{(k)}$ 收敛于 ϑ^* 和 $\theta^{(k)} \in \overleftarrow{P}_M(\vartheta^{(k)})$。但由于关于 M的左Bregman到达半径的假设，$\overleftarrow{P}_M(\vartheta^*) = \theta^\diamond$是单点集。这显然意味着 $\theta^{(k+1)} \to \theta^\diamond$。□

9. 更一般的族

用于指数族的若干论证可以扩展到更一般的情形。我们考虑 $d\mathbb{P}_\theta = p(T(\cdot)\ \theta)d\mu$的情形，并且对于条件族，$d\mathbb{P}_\theta^{x|y} = k(T(\cdot)|y, \theta)d\mu_y$对于充分统计量 $T(x)$。假设存在一个正交坐标变换 $Q\theta = (\theta', \theta'')$以及一个可能为非线性的、映射到极小充分统计量$T'(x)$的约化，使得该条件族具有如下等价表示

$$d\mathbb{P}_\theta^{x|y} = d\mathbb{P}_{\theta'}^{x|y} = k'(T'(x)|y, \theta')d\mu_y(x),$$

仅取决于 θ'。我们考虑以下性质，它将指数族情形下 T_j 的仿射独立性（定义2）进行了扩展：

(32) 在仿射极小表示 $k'(T'(x)|y, \theta')$ 中， $\mathbb{P}_\theta^{x|y}$不存在 $v \neq 0$ 使得对于 μ_y-几乎所有的 $x \in h^{-1}(y)$，都有 $v \cdot \nabla_{\theta'} \log k'(T'(x)|y, \theta') = 0$。

命题 7. 假设 $\mathbb{P}^{x|y}$具有上述性质，并设 $k'(T'(x)|y, \theta')d\mu_y(x)$ 是 $\mathbb{P}_\theta^{x|y}$ 的一个仿射极小表示。则 $\nabla_{22}^2 \bar{K}_y(\theta'||\theta') \succ 0$。

证明： 为方便记号，令 $k(\cdot|y, \theta)d\mu_y$ 为仿射极小表示。由 (18)，由于 $\nabla_{22}^2 K_y(\theta||\theta) = \mathbb{E}_\theta[\nabla_\theta k(\cdot|y, \theta)\nabla_\theta k(\cdot|y, \theta)^T|y] \succeq 0$， $v \cdot \nabla_{22}^2 K_y(\theta||\theta)v = 0$ 蕴含 $\mathbb{E}_\theta\left[|v \cdot \nabla_\theta k(\cdot|y, \theta)|^2|y\right] = 0$，因此 $v \cdot \nabla_\theta \log k(x|y, \theta) = 0$ μ_y-几乎处处成立；由极小性可知 $v = 0$。因此$\nabla_{22}^2 K_y(\theta||\theta) \succ 0$。□

这意味着我们可以得到此前在指数族中发现的情形。存在一个正交的 $n \times n$-矩阵 Q，使得

$$Q\nabla_{22}^2 K_y(\theta||\theta)Q^T = \left[\begin{array}{c|c} \nabla_{22}^2 \bar{K}_y(\theta'||\theta') & 0 \\ \hline 0 & 0 \end{array}\right], \quad Q\theta = \begin{bmatrix} \theta' \\ \theta'' \end{bmatrix}, \quad \nabla_{22}^2 \bar{K}_y(\theta'||\theta') \succ 0,$$

其中 θ' 是在移除 θ中的仿射依赖后剩余的精确参数。因此，我们可以再次利用部分收敛定理证明精确参数序列 $\theta'^{(k)}$ 收敛。

定理 6. 假设 $-\log q(y,\cdot)+i_M$ 满足 ***KŁ*** 不等式，且 $M\subset G$ 有界。假设该条件族具有上述极小性。令 $\theta^{(k)}$ 由约束***EM***算法生成。则精确参数序列 $\theta'^{(k)}$ 收敛。

最后一步是考察分裂程序（P_∞）。我们得到以下结果，其证明用到了命题 5 的证明。

命题 8. 在定理 6 的假设下。设给定 y 时，完全数据的条件期望费希尔信息矩阵 $\mathcal{I}_c(\theta,y)$ 在 G 上正定。那么，对于固定的 θ'，有$-\nabla^2_{\theta''\theta''}\log q(y,\theta',\theta'')\succ 0$。

这再次使我们能够将 $\theta'^{(k)}$ 的收敛提升为完整参数序列的收敛，例如，当 $M(\theta')$ 为凸，且完全数据族在给定 y 时具有满秩的完全数据条件期望 Fisher 信息，即 $\mathcal{I}_c(\theta,y)\succ 0$ 时。

10. 示例

本节讨论若干极限示例。

例 1. 我们考虑缺失数据情形 $x=(y,\ z)$，其中，$h:(y,\ z)\mapsto y$，y为观测到的数据，z为隐变量，且特殊统计量为 $T(x)=x$。设 $d\mu=m(y,\ z)d\mu_Y\otimes d\mu_Z$，其中 $\mu_Y\otimes\mu_Z$ 是 $X=Y\times Z$ 上的乘积测度。我们有 $p(x,\ \theta)=e^{\theta_y\cdot y+\theta_z\cdot z-\psi(\theta_y,\ \theta_z)}$。按照测度分解的记号，$\nu=\mu_Y$ 且 $d\mu_y=m(y,\ z)d(\delta_{\{y\}}\otimes\mu_Z)$，对于 $y\in Y$，因为

$$\begin{aligned}\int_{Y\times Z}f(y,z)d\mu(y,z)&=\int_Y\left[\int_Z f(y,z)m(y,z)d\mu_Z(z)\right]d\mu_Y(y)\\&=\int_Y\left[\int_{\{y\}\times Z}f(y,z)m(y,z)d(\delta_{\{y\}}\otimes\mu_Z)(y,z)\right]d\mu_Y(y)\\&=\int_Y\left[\int_{h^{-1}(y)}f(y,z)d\mu_y(y,z)\right]d\nu(y).\end{aligned}$$

由此得到

$$\begin{aligned}\psi_y(\theta)&=\log\int_{h^{-1}(y)}e^{\theta_y\cdot y}e^{\theta_z\cdot z}d\mu_y(y,z)\\&=\log\int_{\{y\}\times Z}e^{\theta_y\cdot y}e^{\theta_z\cdot z}m(y,z)d(\delta_{\{y\}}\otimes\mu_Z)(y,z)\\&=\theta_y\cdot y+\log\int_Z e^{\theta_z\cdot z}m(y,z)d\mu_Z(z).\end{aligned}$$

这里 $T(h^{-1}(y))=\{y\}\times Z$，因此我们需要在命题3中进行降维。我们得到

$$\begin{aligned}\log k(y,z|y,\theta_y,\theta_z)&=\theta_y\cdot y+\theta_z\cdot z-\theta_y\cdot y-\log\int_Z e^{\theta_z\cdot z}m(y,z)d\mu_Z(z)\\&=:\theta_z\cdot z-\chi_y(\theta_z)=:\log\bar{k}(z|y,\theta_z)\end{aligned}\tag{33}$$

定义 $\chi_y(\theta_z)=\log\int_Z e^{\theta_z\cdot z}m(y,\ z)d\mu_Z(z)$。因此，条件族 i

on s

$$k(y,z|y,\theta_y,\theta_z)d\mu_y(y,z)=\bar{k}(z|y,\theta_z)d\mu_Z(z)$$

且精确参数为 θ_z，至少当 z 是该条件族的极小参数时如此，剩余参数为 θ_y。这意味着，即使在可定义性条件下，我们也只能期待 θ 的 θ_z 部分收敛。至于 θ_y 部分的收敛，则必须依赖分裂程序（P_∞）。

例2。（续.....）。我们比较（16）中目标函数和正则项的定义域。从 $q(y,\theta_y,\theta_z)=\int_Z e^{\theta_y\cdot y}e^{\theta_z\cdot z}e^{-\psi(\theta_y,\theta_z)}m(y,z)d\mu_Z(z)=e^{\theta_y\cdot y-\psi(\theta_y,\theta_z)}\int_Z e^{\theta_z\cdot z}m(y,z)d\mu_Z(z)$ 可得 $\operatorname{dom}q(y,\cdot)=\operatorname{dom}p(x,\cdot)=\operatorname{dom}(\psi)=\Theta=\{(\theta_y,\theta_z):\ \int_Z e^{\theta_y\cdot y}e^{\theta_z\cdot z}m(y,z)d\mu_y\otimes\mu_Z<\infty\}$,

而 dom $\psi_y = Y \times \{\theta_z : \int_Z e^{\theta_z \cdot z} m(y, z) d\mu_Z(z) < \infty\} =$ dom $\bar{K}_y(\theta^{(k)}||\cdot)$ 更大。尽管如此，Θ 和 dom ψ_y 具有共同的边界点。

例3。摘自 [2]。考虑一个独立正态样本 $x_1, \ldots, x_N$，其统计量为$T(x) = (T_1(x)\ T_2(x)) = (\sum_{i=1}^N x_i/N, \sum_{i=1}^N x_i^2/N)$，并假设观测到 $y = T_1(x)$，而$z = T_2(x)$是隐变量。相应的指数族 $p(x, \mu, \sigma^2)$可写为 $p(x, \theta)$，其中$\theta = (\theta_1, \theta_2) = (N\mu/\sigma^2, -N/2\sigma^2)$，因此 $\mu = -\theta_1/2\theta_2, \sigma^2 = -N/2\theta_2, \Theta = G = \mathbb{R} \times (-\infty, 0)$。

为简单起见，令 $N = 2$，则有 $\psi(\theta_1, \theta_2) = -\theta_1^2/4\theta_2 - \log(-\theta_2)$，定义在$G$上，其勒让德变换 $\psi^*(\eta_1, \eta_2) = -1 - \log(\eta_2 - \eta_1^2)$定义在 $G^* = \{\eta$ 上：$\eta_2 > \eta_1^2\}$、梯度、逆梯度和期望参数分别为

$$\nabla\psi(\theta) = \begin{bmatrix} -\frac{\theta_1}{2\theta_2} \\ \frac{\theta_1^2}{4\theta_2^2} - \frac{1}{\theta_2} \end{bmatrix} =: \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \qquad \theta_1 = -\frac{2\eta_1}{\eta_1^2 - \eta_2}, \theta_2 = \frac{1}{\eta_1^2 - \eta_2}, \qquad \nabla\psi^*(\eta) = \begin{bmatrix} -\frac{2\eta_1}{\eta_1^2 - \eta_2} \\ \frac{1}{\eta_1^2 - \eta_2} \end{bmatrix}$$

以及在原坐标 $\eta_1 = \mu, \eta_2 = \mu^2 + \sigma^2$ 中。我们有 $h(x_1, x_2) = y = \frac{x_1 + x_2}{2}$。该族是二维的，但 $T(h^{-1}(y)) = (y, \frac{1}{2}(x_1^2 + (2y - x_1)^2))$ 包含于一维仿射空间 $T_1 = y$ of (T_1, T_2) 中。因此我们需要进行参数约化。条件族是

$$\begin{aligned} k(x_1, x_2|y, \theta) &= C(\sigma) \exp\{-[(x_1 - \mu)^2 + (x_2 - \mu)^2]/2\sigma^2\} / \exp\{-(y - \mu)^2/\sigma^2\} \\ &= \exp\{-(x_1 - y)^2/\sigma^2\}/2\sqrt{\pi}\sigma =: \bar{k}(x_1|y, \theta_2) \\ &= \exp\left\{-(x_1^2 - 2yx_1)/\sigma^2 - \left[y^2/\sigma^2 - \tfrac{1}{2}\log\sigma^{-2}\right]\right\}/2\sqrt{\pi} \\ &= \exp\left\{\theta_2 \cdot (x_1^2 - 2yx_1) - \left[-y^2\theta_2 - \tfrac{1}{2}\log(-\theta_2)\right]\right\}/2\sqrt{\pi} \end{aligned}$$

仅依赖于 $\theta' = \theta_2$ 的约化族，其中 $\theta' = \theta_2$ 是精确参数。不完整数据族的构造如下：由于 $\mathbb{E}_\theta[(x_1 + x_2)/2] = \mu$ 且 $\mathbb{V}_\theta[(x_1 + x_2)/2] = \sigma^2/2$，我们有 $q(y, \theta) \sim N(\mu, \sigma^2/2)$。

现在我们考虑 M步。我们讨论两种情形：有约束和无约束。在第一种情形中，引入一个曲指数族形式的约束，即 $\mu^2 = \sigma^2$，使得模型族为 $N(\mu, \mu^2)$。用自然参数表示时，这就是 $M = \{\theta : \theta_1^2 = -4\theta_2\}$。这里，定理2保证精确参数序列 $\theta_2^{(k)}$ 收敛。但该约束将 θ_1^2 表示为 θ_2 的函数，因此 $\theta_1^{(k)}$ 也收敛。

在无约束情形下，参数 $\theta_2^{(k)}$ 收敛，因此 $\sigma^{(k)2}$也收敛，而参数 θ_1 是自由的。我们必须考虑 M步程序 (P_k)，其目标是 $\min_{\theta_1} -\log q(y, \theta_1, \theta_2^{(k)})$.由于 $-\log q(y, \theta_1, \theta_2^{(k)}) = (y - \mu)^2/\sigma^{(k)2} + \log C(\sigma^{(k)})$，解总是 $\mu = y = \mu^{(k)}$，这意味着 $\theta_1^{(k)} = 2y/\sigma^{(k)2} = -2y\theta_2^{(k)}$，而它也随之收敛。这很有意思，因为起初可用信息似乎不足以估计 μ和 σ^2。

例 4。（续.....）。正如 [2], 中所观察到的，在这个例子中，EM算法与em-算法之间出现了差异。用我们目前的术语来说，这是因为 $\mathbb{E}_\vartheta[(x_1^2 + x_2^2)/2] = \mu^2 + \sigma^2 = y^2 + \sigma^2$ 和 $\mathbb{E}_\vartheta[(x_1^2 + x_2^2)/2|y] = y^2 + \sigma^2/2$ 并不相同。

只需令 $T_2(x) = \sum_{i=1}^N (x_i - \bar{x})^2/(N - 1)$，即可轻松弥补这一差异，因为此时$T_1, T_2$ 相互独立，因此满足 Amari 条件。

例 5. 考虑 $\mathbb{R}_+^2$ 中给出的 Kullback-Leibler 距离 $K(a||b) = \sum_{i=1}^2 a_i \log \frac{a_i}{b_i} - a_i + b_i$，它是由 $\psi(x) = \sum_{i=1}^2 x_i \log x_i - x_i$ 诱导的一种特殊 Bregman 距离。选取两个点 $p, q \in (-\infty, 0)^2$ 并计算 $a = \exp(p)$，$b = \exp(q)$。令 $p_t = tp + (1 - t)q, t \in [0, 1]$ 为线段 $[p, q]$ 上的点，则 $a_t = \exp(p_t) = \exp(tp)\exp((1 - t)q)$ 在 $(0, 1)^2$, 中形成一条曲线，通常并非直线。令 $A = \{a_t : t \in [0, 1]\}$。则$\nabla\psi(A) = \log(A) = [p, q]$ 这条线段，因此 A 是 e-平坦的。换言之，A 上的 Kullback-Leibler 右投影是唯一的。现在令 $B = [a, b]$ 为连接 a, b 的线段。则B 是凸的，因此 B 上的 Kullback-Leibler 左投影是唯一的，并且 B 是 m-平坦的。我们

在 A 是曲线的情况下，存在 $A \cap B = \{a, b\}$。这意味着 Bregman投影 $\vec{P}_A$,$\overleftarrow{P}_B$ 是唯一的，而且 $P_A \circ P_B$ 有两个吸引点。然而，从中间某处的点 a_t 出发，可以看到迭代点要么向左移动，要么向右朝着 a 或 b 移动。除了$\overrightarrow{\quad}$ $\overleftarrow{\quad}$点对 a', b'（满足 $a' = P_A(b')$、$b' = P_B(a')$），会形成一个间隙或不动点对。如果从 a' 出发，交替投影法将停留在这个不动点对上。

例6。（曲线族的收敛失败）。我们考虑独立高斯随机变量 X_1、X_2、X_3，其 $\mathbb{E}(X_i)$ 为 $= \mu_i$，且 $\mathbb{V}(X_i)$ 为 $= 1$。我们希望基于对 $Y = X_3$ 的观测值 y 来估计 $\mu = (\mu_1, \mu_2, \mu_3)$，其中假设 $\mu_3 = f(\mu_1, \mu_2)$，且已知 f。关于统计量，即 $x = (x_1, x_2, x_3) = (z_1, z_2, y)$，其中 y 为观测值，z 为隐变量。这给出了模型集 $M = \{\mu : \mu_3 = f(\mu_1, \mu_2)\}$，因此该族是曲线族。设 $y = 0$。于是，E步 $x^+ = \mathbb{E}(X|Y = y, \mu)$ 得到 $x_1^+ = \mu_1$、$x_2^+ = \mu_2$、$x_3^+ = 0$。M步为 $\mu^+ \in \operatorname{argmin}_{\mu \in M} -\log p(x^+, \mu)$，其中

$$p(x^+, \mu) = C \exp\{-\tfrac{1}{2}(x_1^+ - \mu_1)^2 - \tfrac{1}{2}(x_2^+ - \mu_2)^2 - \tfrac{1}{2}(x_3^+ - \mu_3)^2\}.$$

因此，E步是将 $\mu \in \mathbb{R}^3$ 正交投影到数据集 $D = \{x_3 = 0\}$上。M步是将 $x^+ = (x_1^+, x_2^+, 0)$正交投影到模型集 $M = \operatorname{graph}(f)$上。该投影的结果是 $\mu^+ \in M$，然后重复这一过程。因此，这里出现了期望最大化算法与em算法相一致的情形。

总的来说，这种方法现在就是在集合 $A = \{(x_1, x_2, x_3) : x_3 = 0\}$ 和 $B = \{(x_1, x_2, x_3) : x_3 = f(x_1, x_2)\}$之间进行交替投影法，并且很容易将其推广到 $x \in \mathbb{R}^n$，其中 $A = \{(x, 0) : x \in \mathbb{R}^n\}$，而 $B = \{(x, f(x)) : x \in \mathbb{R}^n\}$ 是函数 $f : \mathbb{R}^n \to \mathbb{R}$的图像。假设 $f(x) \geq 0$。那么

$$P_A(x_k, f(x_k)) = (x_k, 0), \quad (x_k, f(x_k)) \in P_B(x_{k-1}, 0) \text{ iff } x_{k-1} = x_k + f(x_k)\nabla f(x_k). \tag{34}$$

从无穷小的角度看，该方法沿着最速上升方向反向进行。

回到 $n = 2$，令 $B = M$ 为定义在$x_1^2 + x_2^2 \leq 1$上的墨西哥帽函数 [1] 的图像。与 [1], 中关于无穷小步长下最速下降的论证类似，采用无穷小步长的交替投影也会沿着帽形函数的谷底向下移动，不断绕圈并趋近边界曲线 $x_1^2 + x_2^2 = 1$，此处 $f = 0$。无法保证收敛的症结在于 M不满足KŁ性质，因为它不可定义。图示见 [1]。

例 7.（PPM 与期望最大化算法）。再次考虑情形 $A = \{(x, 0) : x \in \mathbb{R}^n\}$, $B = \{(x, f(x)) : x \in \mathbb{R}^n\}$，其中 $f \geq 0$。考虑近端点步骤

$$x_k \in \operatorname{argmin} \tfrac{1}{2} f(x)^2 + \tfrac{1}{2}\|x - x_{k-1}\|^2. \tag{35}$$

必要的最优性条件为 $0 = f(x_k)\nabla f(x_k) + x_k - x_{k-1}$，这就是上文的交替投影步骤 (34)。这意味着，对于上面的例子，(35) 必须是方案 (4)，亦即命题 1 中对该方案的实现。选择 f 使得交替投影不收敛，我们也就构造出了一个例子：非凸近端点法（固定 $\lambda_k = 1$）不收敛，此时目标函数为$\frac{1}{2} f(x)^2$。这一构造表明，每个目标函数有下界的 PPM 实例都是期望最大化算法的一个特例。

参考文献

[1] Absil, P.A., Mahony, R., Andrews, B：解析代价函数下降法迭代的收敛性。*SIAM* 优化期刊，16(2), 531-547（2005）。[2] Sh.-I. Amari：神经网络中 EM 算法与 em 算法的信息几何。《神经网络》，8(9):1995,1379-1408。[3] Sh.-I. Amari。信息几何及其应用。Springer 应用数学科学丛书 194，2016。[4] F.J. Aragón Artacho, A.L. Donchev, M.H. Geoffrey：度量正则映射的近端点方法的收敛性。*ESAIM*：会议录17:2007,1–8。[5] H. Attouch, J. Bolte：关于涉及解析特征的非光滑函数的近端算法收敛性。数学规划，116(1-2, Ser. B)，5-16（2009）。

[6] H. Attouch, J. Bolte, P. Redont, A. Soubeyran：非凸问题的近端交替最小化与投影方法：基于 Kurdyka-Łojasiewicz 不等式的方法。《运筹学数学》，35(2):2010, 438–457。[7] H. Attouch, J. Bolte, B.F. Svaiter：半代数与温和问题中下降方法的收敛性：近端算法、前向-后向分裂及正则化高斯-赛德尔方法。数学规划 137:2013,91-129。[8] O.E. Barndorff-Nielsen。《统计理论中的信息与指数族》，Wiley，1978。[9] H.H. Bauschke, J.M. Borwein：勒让德函数与随机布雷格曼投影方法。《凸分析杂志》4(1):1997,27-67。[10] H.H. Bauschke, M.N. Dao, S.B. Lindstrom：利用布雷格曼-莫罗包络进行正则化。SIAM 优化期刊 28(4):2018,3208-3228。[11] H.H. Bauschke, D. Noll：前向投影方法。《非线性凸分析杂志》3:2002,191-205。[12] H.H. Bauschke, D. Noll, A. Celler, J.M. Borwein：用于动态单光子发射计算机断层成像的EM算法。《IEEE医学影像汇刊》18(3):1999,252-261。[13] H.H. Bauschke, X. Wang, J. Ye, X. Yuan：布雷格曼距离与切比雪夫集。《逼近理论杂志》159:2009,3-25。[14] P.M. Bentler, J. Liang, M.-L. Tang, K.-H. Yuan：双层均值与协方差结构模型的约束最大似然估计。《教育心理测量》71(2):2011, 325-345。[15] E. Bierstone, P. Milman：半解析集与次解析集。高等科学研究所数学出版物，67:1988, 5-42。[16] J. Bolte, A. Daniilidis, O. Ley, L. Mazet：Łojasiewicz不等式的刻画：次梯度流、谷底曲线与凸性。《美国数学会汇刊》362(6):2009,3319-3363。[17] J. Bolte, A. Daniilidis, A.S. Lewis, M. Shiota：可分层函数的 Clarke次梯度。SIAM 优化期刊 18:2007,556-572。[18] L. D. Brown：统计指数族基础及其在统计决策理论中的应用。IMS讲义与专著丛书，第9辑：1986, , 1986。[19] R. Caprio, A.M. Johansen：对数Sobolev不等式下期望最大化算法的快速收敛。《生物统计学》103:2024,1-18。[20] G. Chen, M. Teboulle：利用布雷格曼函数的类近端最小化算法的收敛性分析。SIOPT 3(3):1993,538–543。[21] S. Chrétien, A.O. Hero：用于最大似然估计的 Kullback近端算法。《IEEE信息论汇刊》46(5):2000,1800–1810。[22]S. Chrétien, A.O. Hero：关于EM算法及其近端推广。ESAIM：概率与统计 12:2008,308–326。[23] S. Chrétien, A.O. Hero：用于最大似然估计的 Kullback近端算法。arXiv:1201.5907v1 [stat.C0] 27 2012年1月[24] S. Chrétien, A. O. Hero：广义近端点算法，技术报告，密歇根大学，1998。[25] S. Chrétien, A.O. Hero：通过近端点迭代加速EM算法。1998年IEEE国际信息论研讨会论文集，麻省理工学院，剑桥，444。[26] F. Clarke。《优化与非光滑分析》。约翰·威利父子出版公司，纽约（1983）[27] M. Coste：o-极小几何导论。RAAD讲义，雷恩研究所，1999。[28] I. Csiszár：概率分布的I-散度几何与最小化问题，《概率年刊》3:1975,146-158。[29] I. Csiszár, G. Tusnády：信息几何与交替最小化过程。《统计与决策》1(1):1984,205-237。[30] A.P. Dempster, N.M. Laird, D.B. Rubin：通过EM算法从不完整数据进行最大似然估计。《英国皇家统计学会杂志B辑》39:1977,1-38。[31] L. van den Dries。《温和拓扑与o-极小结构》。剑桥大学出版社，1998。[32] L. van den Dries, C. Miller：几何范畴与o-极小结构。《杜克数学杂志》85:1996,487-540。[33] L. van den Dries, C. Miller：具有受限解析函数的实指数域。《以色列数学杂志》85, No. 1-3, 19-56（1994）。[34] B. Flury, A. Zoppé：EM练习。《美国统计学家》54:2000,207–209。[35]O. Guler：凸最小化中近端点算法的收敛性。《SIAM控制与优化杂志》29:1991,403–419。[36]R.J. Hathaway：用于单变量正态混合模型的约束em算法。《统计计算与模拟杂志》23:1986,211-230。[37] J. Hoffmann-Jørgensen：条件概率的存在性。《斯堪的纳维亚数学杂志》28:1971,257-264。[38] A.N. Iusem：特定泊松模型中EM算法的简短收敛性证明。《巴西概率与统计杂志》6(1): 1992,

[39] A. Kagan，Z. Landsman。协方差矩阵与Fisher信息矩阵之间的关系。《统计与概率快报》42：1999，7-13。[40]A. Kaplan，R. Tichatschke。近端点方法与非凸优化。《全球优化杂志》13：1998，389-406。[41] D.K. Kim，J.M.G. Taylor。在线性参数约束下进行最大似然估计的受限 EM 算法。《美国统计学会会刊》90：1995，708-716。[42] M.E. Khan，P. Baqué，F. Fleuret，P. Fua。Kullback-Leibler 近端变分推断。预印本，2025。[43] D. Klatte。扰动 C^1，1程序解的上利普希茨行为。《数学规划》B辑88：2000，285-311。[44] F. Kunstner，R. Kumar，M. Schmidt。EM 的同胚不变性：指数族中采用镜像下降的 KL 散度非渐近收敛。第24届国际人工智能与统计会议（AISTATS）论文集，2021，圣迭戈，加利福尼亚州，美国。PMLR：第130卷。arXiv:2011.01170[45] K. Kurdyka。关于 o-极小结构中可定义函数的梯度。《傅里叶研究所年刊》48：1998，769-783。[46] E. L. Lehmann，G. Casella。点估计理论。施普林格统计学丛书，第2版，1998。[47] M. Levine，D. Richards，J. Su。截断多元正态分布的非陡峭性与最大似然估计性质。arXiv:2303.10287v3 [math.ST][48] Ch. Lui，D.B. Rubin。ECME 算法：EM 和 ECM 的一种简单扩展，具有更快的单调收敛性。《生物统计学》81(4)：1994，633-648。[49] F.J. Luque。近端点算法的渐近收敛分析。《SIAM 控制与优化杂志》22，1984，277-293。[50] J. Maeght，S.P. Boyd，D. Noll。动态发射断层成像——正则化与反演。《加拿大数学会会议论文集》27，2000，211–234。[51] B. Martinet。通过逐次逼近对变分不等式进行正则化。《法国自动化、信息学与运筹研究杂志》4：1970，154-159。[52] B. Martinet。伪压缩映射不动点的近似确定。《巴黎科学院院报》274：1972，163-165。[53] G.J McLachlan，T. Krishnan。EM 算法及其扩展。Wiley 概率与统计丛书，第2版，2008。[54] C. Miller。指数化难以避免。《美国数学会会刊》122(1)：1994，257-259。[55] B.S. Mordukhovich。变分分析与广义微分 I：基础理论，II：应用。施普林格出版社，纽约，2006。[56] D. Nettleton。受约束参数空间中 EM 算法的收敛性质。《加拿大统计学杂志》27(3)：1999，639-648。[57] D. Noll。利用 Kurdyka-Łojasiewicz 不等式的非光滑下降方法的收敛性。《优化理论与应用杂志》160(2)：2014，553-572。[58] D. Noll。交替投影及其在 Gerchberg-Saxton 误差缩减中的应用。《集值与变分分析》29：2021，771-802。[59] D. Noll。交替 Bregman投影与 EM 算法的收敛性。《纯粹与应用泛函分析杂志》10(4)：2025，979-1021。[60] D. Noll，A. Rondepierre。关于交替投影法的局部收敛。《计算数学基础》16(2)，2016，425-455。[61] T. Orchard，M.A. Woodbury。缺失信息原理：理论与应用。第6届伯克利数学、统计与概率研讨会论文集，第1卷，伯克利，加利福尼亚州：加利福尼亚大学出版社，697-715页。[62]A. Padgett，P. Speissegger。o-极小结构中复函数的可定义性。arXiv:2506.15119.v1 [math.LO] 18 2025年6月。[63] R.F. Phillips。估计切换回归的受约束最大似然方法。《计量经济学杂志》48(1-2)：1991，241-262[64] T. Pennanen。无需单调性的近端点算法与乘子法的局部收敛。《数学运筹学研究》27(1)：2002，170-191。[65] S.M. Robinson。广义方程及其解。第二部分：非线性规划的应用。《数学规划研究》19：1982，200-221[66] R.T. Rockafellar。凸分析。普林斯顿大学出版社，新泽西州，1970。[67] R.T. Rockafellar。单调算子与近端点算法。《SIAM 控制与优化杂志》14：1976，877–898。[68] R.T. Rockafellar。近端点算法的收敛性与适用范围进展。《非线性与凸分析杂志》22(11)：2021，2347–2374。[69] R.T. Rockafellar，R.J.B. Wets。变分分析。《数学科学基础丛书》，施普林格出版社317：2009。

[70] R. Schoenberg。约束最大似然。计算经济学10(3)：1997，251-266。[71] Ning-Zhong Shi、Shu-Rong Zheng 和 Jianhua Guoa。参数不等式约束下的受限 EM 算法。多元分析杂志92（2005）53-76。[72] J.E.Spingarn。次单调映射与近端点算法。数值泛函分析与优化4(2)：1982，123-150。[73] R. Sundberg。不完整指数族数据的最大似然理论。斯堪的纳维亚统计学杂志1：1974，49-58。[74] M. Shiota。亚解析集与半代数集的几何。Birkhäuser 出版社，1997。[75] P. Tseng。EM 算法与熵类近端点方法分析。数学运筹学研究29(1)：2004，27-44。[76] F. Vaida。EM 和 MM 算法的参数收敛。《中华统计学志》15：2005，831-840。[77] A. Wilkie。利用受限 Pfaff 函数和指数函数扩展实数有序域的模型完备性结果。《美国数学会杂志》9：1996，1051–1094。[78] C.F.J. Wu。关于 EM 算法收敛性质。《统计年鉴》*11*：1983，95-103。

图卢兹大学数学研究所，法国电子邮件地址：
dominikus.noll@math.univ-toulouse.fr