arXiv:2302.03805v2 [cs.LG] 1 Nov 2023

通过比较反馈 eliciting 用户偏好以进行个性化多目标函数决策

Han Shao
TTIC
han@ttic.edu

Lee Cohen TTIC
leecohencs@gmail.com

Avrim Blum
TTIC
avrim@ttic.edu

Yishay Mansour 特拉维夫大学和
Google Research
mansour.yishay@gmail.com

Aadirupa Saha
TTIC
aadirupa@ttic.edu

Matthew R. Walter
TTIC
mwalter@ttic.edu

摘要

在本文中，我们提出了一个多目标决策框架，用于适应用户对目标函数的不同偏好，其中用户偏好通过策略比较进行学习。我们的模型由一个已知的马尔可夫决策过程和一个向量值奖励函数组成，每位用户都有一个未知的偏好向量，用于表示各目标函数的相对重要性。我们的目标是针对给定用户高效计算近似最优策略。我们考虑两种用户反馈模型。首先，我们处理向用户提供两个策略，并由用户返回其偏好策略作为反馈的情形。随后，我们转向另一种用户反馈模型：向用户提供两个由代表性轨迹构成的小型加权集合，由用户选择其中更偏好的集合。在这两种情况下，我们都提出了一种算法，能够通过数量关于目标数量呈拟线性增长的比较查询，为用户找到近似最优策略。

1 引言

许多现实世界中的决策问题都涉及对多个目标进行优化。例如，在设计投资组合时，投资策略需要权衡最大化预期收益与最小化风险。在使用 Google 地图进行导航时，人们会关注多种因素，例如最差和平均预计到达时间、交通状况、路面状况（例如是否铺设了路面），以及沿途的风景。随着技术进步推动个性化机器学习的发展（McAuley, 2022），本文设计了用于个性化多目标决策的高效算法。

以往研究主要致力于近似帕累托最优解集[1]（相关综述可参见 Hayes 等人（2022）和 Roijers 等人（2013）），而我们的目标是为用户找到能够反映其对 k 目标函数未知偏好的最优个性化策略。由于用户偏好未知，我们需要通过请求用户对选定策略提供反馈来 eliciting 用户偏好。Wang 等人（2022）使用一种强大的查询模型研究了偏好 elicitation 问题，该模型针对单个策略的质量提供随机反馈。相比之下，我们关注一种更自然、更直观的查询模型——比较查询，即询问用户对两个选定策略的偏好，例如：“您更偏好使平均预计到达时间最短的策略，还是使转弯次数最少的策略？”我们的目标是找到最优个性化

[1]帕累托最优解集包含针对每种可能的用户偏好的一个最优个性化策略。

第37届神经信息处理系统大会（NeurIPS 2023）。

策略，同时尽可能减少查询次数。据我们所知，我们是首个针对特定的个性化多目标决策问题，通过策略比较提供具有理论保证的算法的研究。

与多目标决策领域的既有研究类似，我们使用有限时间范围的马尔可夫决策过程（MDP）对该问题进行建模，其奖励向量为 k 维，其中每个分量都是代表 k 目标函数之一的非负标量奖励。为考虑用户偏好，我们假设用户由一个具有非负分量的（隐藏的） k 维偏好向量刻画，并且对于每个状态-动作，该用户的个性化奖励是该偏好向量与该状态-动作对应奖励向量之间的内积。我们还区分策略的 k 维价值与策略的个性化价值：前者是在按照该策略选择动作时的期望累积奖励，后者则是在按照该策略选择动作时的期望累积个性化奖励这一标量。马尔可夫决策过程对智能体是已知的，目标是通过策略比较反馈，学习用户个性化奖励函数的最优策略（以下简称最优个性化策略）。

比较反馈。如果人们能够清晰地定义自己对各目标的偏好（例如，“我的偏好向量中，风景目标的权重为 3，交通目标的权重为 2，路面目标的权重为 1”），问题就会很简单——只需将个性化奖励函数用作标量奖励函数，并求解相应的策略即可。具体而言，Wang 等人（2022）研究了一个类似问题：针对单个多目标策略（即从状态到动作的映射）的价值，使用带噪反馈进行学习。由于用户难以给出这种细粒度的偏好反馈，尤其是在需要做出序贯决策的环境中，因此我们将智能体的反馈来源限制为比较反馈。比较反馈在实践中应用广泛。例如，ChatGPT 会要求用户比较两个回答，以提升其性能。与要求用户为 ChatGPT 的回答打数值评分相比，这种方式对用户而言更直观。

事实上，只要知道用户偏好向量，就足以求得其最优个性化策略，因此，挑战在于如何通过尽可能少且易于理解的查询来学习用户的偏好向量。因此，我们聚焦于比较查询。接下来要问的是，究竟应该比较什么？对于前述任务而言，比较状态-动作对或许并非好选择——路线中的两个单独步骤究竟意味着什么？比较单条轨迹（例如 Google 地图中的路线）也不理想。以两个策略为例：一个随机生成（个性化的）好或差轨迹，另一个则始终生成（个性化的）中等轨迹。如果只比较单条轨迹，而不考虑轨迹集合，我们就无法判断用户对这两个策略的偏好。

可解释的策略表示。由于我们希望通过策略比较查询来学习偏好，因此也提出了一种替代方案，即采用更具可解释性的策略表示。具体而言，我们设计了一种算法：给定一个显式策略表示（即从状态到动作分布的映射），该算法会返回一个大小至多为 $k+1$ 的加权轨迹集，使其期望回报与该策略的价值相同。[2] 根据我们的形式化定义，可以立即得出，对于任意用户，加权轨迹集的个性化回报也与该策略的个性化价值相同。

在这项工作中，我们着重探讨以下两个问题：

(1) 如何通过尽可能少的策略比较查询找到最优个性化策略？(2) 如何高效地找到一种更具可解释性的策略表示?

贡献。第2节中，我们形式化地描述了通过比较反馈来 eliciting 用户偏好并寻找最优个性化策略这一问题。作为显式策略表示的替代方案，我们提出使用加权轨迹集作为一种更易于理解的表示。第3节中，我们提供了一种高效算法，用于寻找近似最优个性化策略；其中策略以其形式化表示给出，从而回答问题（1）。第4节中，我们设计了两种高效算法，用于寻找策略的加权轨迹集表示。结合第3节中的算法，我们得到一种算法，可在策略以加权轨迹集表示时寻找近似最优个性化策略，从而回答问题（2）。

[2]轨迹的回报是该轨迹中获得的累积奖励。加权轨迹集的期望回报中的期望是关于权重的期望。

相关工作。近年来，多目标决策受到了广泛关注（综述参见 Roijers 等人（2013）；Hayes 等人（2022））。先前研究探索了多种方法，例如假设偏好为线性，或采用贝叶斯设定，或寻找近似帕累托前沿。然而，融入比较反馈（如在多臂老虎机或主动学习中所做的那样，例如 Bengs 等人（2021）和 Kane 等人（2017））使我们能够以更全面、更系统的方式处理不同类型的用户偏好，并获得（近乎）最优的个性化决策结果。更多相关工作请参见附录 A。

2 问题设定

序贯决策模型。我们考虑一个智能体已知[3] 的马尔可夫决策过程（MDP），用元组 $\langle \mathcal{S}, \mathcal{A}, s_0, P, R, H \rangle$表示，其状态集和动作集分别为有限集 $\mathcal{S}$和 $\mathcal{A}$，初始状态为 $s_0 \in \mathcal{S}$，时间范围为有限时域 $H \in \mathbb{N}$。例如，在 Google 地图示例中，状态是一个交叉路口，动作则是转弯。转移函数 P: $\mathcal{S} \times \mathcal{A} \mapsto$ 单纯形$^{\mathcal{S}}$ 将状态-动作对映射为状态概率分布。为建模多个目标，奖励函数R: $\mathcal{S} \times \mathcal{A} \mapsto [0,1]^k$ 将每个状态-动作对映射为 k维奖励向量，其中每个分量对应多个目标之一（例如路面状况、最差和平均预计到达时间）。轨迹 $\tau = (s_0, a_0, \ldots, s_{H-1}, a_{H-1}, s_H)$ 的回报为$\Phi(\tau) = \sum_{t=0}^{H-1} R(s_t, a_t)$。

策略 π 是从状态到动作分布的映射。我们用 Π表示策略集合。策略 π的价值记作 V^π，即从初始状态 s_0开始执行策略 π 所获得的期望累积奖励。换句话说，π 的价值为 $V^\pi = V^\pi(s_0) = \mathbb{E}_{S_0=s_0}\left[\sum_{t=0}^{H-1} R(S_t, \pi(S_t))\right] \in [0,H]^k$，其中 S_t 是执行π时在时间步 t 的随机状态，期望是对 P和 π的随机性取的。注意，$V^\pi = \mathbb{E}_\tau[\Phi(\tau)],$，其中$\tau = (s_0, \pi(s_0)\ S_1, \ldots, \pi(S_{H-1})\ S_H)$是执行 π生成的随机轨迹。

我们假设存在一个“什么都不做”动作 $a_0 \in \mathcal{A}$，该动作仅在初始状态 s_0下可用，对每个目标函数 $R(s_0, a_0) = \mathbf{0}$ 的奖励均为零，并使系统保持在初始状态，即 $P(s_0 \mid s_0, a_0) = 1$（例如，该动作对应于不通勤或拒绝参加国际象棋比赛。）[4]。我们还定义了（确定性的）“什么都不做”策略 π_0，该策略始终选择动作 a_0，其价值为 $V^{\pi_0} = \mathbf{0}$。从数学角度而言，“什么都不做”的假设确保 0 属于价值向量空间 $\{V^\pi | \pi \in \Pi\}$，这正是我们所需要的。

由于奖励的取值范围为 $[0,1],$，因此对于每个策略 π，都有$1 \le \|V^\pi\|_2 \le \sqrt{k}H$。为方便起见，我们记$C_V = \sqrt{k}H$。我们将 $d \le k$ 记为通过 Π得到的所有价值向量所张成空间的秩，即 $d :=$ rank(span($\{V^\pi | \pi \in \Pi\}$))。

线性偏好。为了纳入用户对目标函数的个性化偏好，我们假设每个用户都由一个未知的 k 维偏好向量 $w^* \in \mathbb{R}_+^k$ 表征，且其范数有界$1 \le \|w^*\|_2 \le C_w$，其中 $C_w \ge 1$未知。为适应一般的线性奖励，我们不假设 $C_w = 1$。请注意，w^* 的大小不会改变个性化最优策略，但会影响反馈模型中的“不可区分性”。不对 $\|w^*\|_2 = 1$进行归一化，使我们能够允许用户具有不同程度的辨别能力。该偏好向量编码了用户对多个目标的偏好，因此决定了用户对策略的偏好。

形式化地，对于由 w^*描述的用户，策略 π 的个性化价值为 $\langle w^*, V^\pi \rangle \in \mathbb{R}_+$。我们将 $\pi^* :=$ arg $\max_{\pi \in \Pi} \langle w^*, V^\pi \rangle$ 和 $v^* := \langle w^*, V^{\pi^*} \rangle$ 分别记为由 w^*描述的用户的最优个性化策略及其对应的最优个性化价值。我们指出，“什么都不做”策略 π_0（始终选择动作 a_0）的价值为 $V^{\pi_0} = \mathbf{0}$，这意味着对于每个 w^*，其个性化价值均为 $\langle w^*, V^{\pi_0} \rangle = 0$。对于任意两个策略 π_1 和 π_2，由 w^* 描述的用户在 $\langle w^*, V^{\pi_1} - V^{\pi_2} \rangle > 0$时更偏好 π_1 而非 π_2。我们的目标是使用尽可能少的交互，为给定用户找到最优个性化策略。

[3]如果马尔可夫决策过程未知，可以对其进行近似，并将本研究的结果应用于近似后的马尔可夫决策过程中。更多讨论见附录 5。[4]我们指出，“什么都不做”动作需要领域知识。例如，在 Google 地图的示例中，“什么都不做”意味着原地不动；在 ChatGPT 的示例中，“什么都不做”意味着不作答。

比较反馈。给定两个策略 π_1, π_2，当 $\langle w^*, V^{\pi_1} - V^{\pi_2} \rangle > \epsilon$ 时，用户会返回 $\pi_1 \succ \pi_2$；否则，用户会返回"无法区分"（即当 $|\langle w^*, V^{\pi_1} - V^{\pi_2} \rangle| \leq \epsilon$ 时）。这里，$\epsilon > 0$ 衡量比较反馈的精度，通常较小。智能体可以通过两种不同类型的策略表示来询问用户的策略偏好：

1. π的显式策略表示：将策略显式表示为映射，$\pi: \mathcal{S} \to$单纯形$^{\mathcal{A}}$。2. π的加权轨迹集表示：一个大小为 κ的轨迹-权重对集合 $\{(p_i, \tau_i)\}_{i=1}^{\kappa}$，其中 $\kappa \leq k+1$ 满足：(i) 权重 $p_1, \ldots, p_\kappa$ 均为非负数且总和为 1；(ii) 集合中的每条轨迹都属于策略 π的支撑集[5]；(iii) 按权重计算，这些轨迹的期望回报与 π的价值相同，即 $V^\pi = \sum_{i=1}^{\kappa} p_i \Phi(\tau_i)$。这种比较对人类来说可能很实用。例如，在 Google 地图的场景中，如果目标是从家前往机场，选择特定路线有 90% 的概率耗时 40 分钟，但发生事故时（概率为 10%）可能需要 3 小时；而乘坐地铁则始终需要 1 小时。

在这两种情况下，反馈完全相同，并且取决于隐藏精度参数 ϵ。因此，ϵ 的值会影响查询次数，以及最终个性化策略的价值与最优个性化价值之间的接近程度。或者，我们可以让智能体预先决定查询次数的上限，这会影响所返回策略的最优性。

技术挑战。在第3节中，我们通过 $\mathcal{O}(\log \frac{1}{\epsilon})$ 次查询找到一个近似最优策略。为此，我们分两步解决该问题。首先，我们找出一组线性无关的策略值，然后利用结合比较反馈的线性规划来估计偏好向量 w^*。w^* 的估计误差通常取决于该线性规划的条件数。因此，我们面临的主要挑战是，如何搜索出能够使条件数较小的线性无关策略值，并为这一估计提供保证。

在第4节中，我们重点探讨如何设计高效算法，以找到策略的加权轨迹集表示。首先，我们采用广为人知的 Carathéodory 定理，由此得到一个效率低下、运行时间可能达到指数级 $|S|^H$ 的算法。我们面临的主要挑战是开发一种高效算法，其运行时间为 $\mathcal{O}(\text{poly}(H, |S|, |\mathcal{A}|))$。基于 Carathéodory 定理的方法将轨迹的回报视为彼此独立的 k 维向量，忽略了它们都由同一个马尔可夫决策过程生成这一事实。为克服这一挑战，我们利用马尔可夫决策过程的内在结构。

3 从显式策略中学习

本节考虑基于显式策略比较查询与用户进行交互的情形。我们设计了一种算法，输出对该用户而言近乎最优的策略。对于多个不同的用户，我们只需反复运行算法中的一部分。为简洁起见，所有证明均放在附录中。

如果已知用户的偏好向量 w^*（允许正比例缩放），就可以高效地计算最优策略及其个性化价值，例如使用有限时域价值迭代算法。在我们的工作中，w^* 未知，因此我们通过比较反馈与用户交互，以学习 w^*。由于结构模型规定，只有当比较的策略值至少相差 ϵ时，才能获得有意义的反馈，因此无法恢复 w^* 的精确值。下面我们将概述如何估计 w^*。

(1) 基策略：我们找出策略 π_1、……、π_d 及其相应的价值 V^{π_1}、……、$V^{\pi_d} \in [0, H]^k$，使其价值彼此线性无关，并且共同张成整个价值向量空间。[6] 这些策略不一定是当前用户的个性化最优策略，而仅作为构建模块，用于估计偏好向量 w^*。在第3.1节中，我们将介绍一种算法，用于为任意给定的马尔可夫决策过程寻找一组基策略。

[5]我们要求该集合中的轨迹属于该策略的支撑集，以避免出现不合乎情理的轨迹，例如在不同且互不连通的状态之间"瞬间移动"的轨迹（比如，在曼哈顿的某个状态以 3 英里/小时通勤，接着在新奥尔良的后续状态以 40 英里/小时通勤）。[6]回顾一下，$d \leq k$ 是所有策略所获得的价值向量张成空间的秩。

(2) 基准比率：对于基策略，记 $\alpha_i > 0$ 为基准策略 π_1 的个性化价值与 π_{i+1} 的个性化价值之比，即：

$$\forall i \in [d-1] : \alpha_i \langle w^*, V^{\pi_1} \rangle = \langle w^*, V^{\pi_{i+1}} \rangle \ . \tag{1}$$

我们将使用比较查询来估计所有 $i \in [d-1]$ 的 α_i 的 $\widehat{\alpha}_i$。估计这些比率的详细算法见第3.2节。为了直观说明，假设我们获得了每个 $i \in [d-1]$ 的精确比率 $\widehat{\alpha}_i = \alpha_i$，那么就可以按如下方式计算向量 $\frac{w^*}{\|w^*\|_1}$。考虑方程（1）中的 $d-1$ 个方程和 $d-1$ 个变量。由于 d 是线性无关的价值向量的最大数量，并且 V^{π_1}、$\ldots V^{\pi_d}$ 构成一个基，加入方程 $\|w\|_1 = 1$ 后，就得到含有 d 个变量的 d 个线性无关方程，因此可以求出 w^*。第3.3节介绍了如何计算 w^* 的估计值。

3.1 策略基的寻找

为了高效地找到 d 个策略，使其价值向量线性无关并张成价值向量空间，人们可能会认为，只需选择 k 个策略，让每个策略分别优化 k 个目标中的一个即可。然而，这种方法可能会失败——在附录N中，我们展示了一个实例：尽管存在 k 个策略，其价值张成的空间的秩为 k，但这些 k 个价值向量仍然线性相关。

此外，我们的目标不只是找到任意一组策略基，而是找到一组满足以下条件的策略基：(1) 基准策略 $\langle w^*, V^{\pi_1} \rangle$ 的个性化价值较大（因此，比率 α_i, $|\widehat{\alpha}_i - \alpha_i|$ 的估计误差较小）；(2) 由这组策略基和基准比率生成的线性规划，能够对 w^* 给出良好的估计值。

对 π_1 的选择。除了价值的线性无关性之外，另一个挑战是找到一组包含基准策略 π_1（不失一般性，令其索引为 1）的策略基，并且该基准策略具有相对较大的个性化价值 $\langle w^*, V^{\pi_1} \rangle$，从而使 $\widehat{\alpha}_i$ 的误差较小（例如，在极端情况下，若 $\langle w^*, V^{\pi_1} \rangle = 0$，我们将无法估计 α_i）。

对于任意 $w \in \mathbb{R}^k$，我们用 π^w 表示使标量奖励 $\langle w, R \rangle$ 最大化的策略，即

$$\pi^w = \arg\max_{\pi \in \Pi} \langle w, V^{\pi} \rangle \ , \tag{2}$$

并用 $v^w = \left\langle w, V^{\pi^w} \right\rangle$ 表示相应的个性化价值。令 $\mathbf{e}_1, \ldots, \mathbf{e}_k$ 表示标准基。为找到个性化价值 $\langle w^*$较大的 $\pi_1, V^{\pi_1} \rangle$，我们针对每个 $j = 1, \ldots, k$，分别找到使第 j 个目标函数最大化的策略 $\pi^{\mathbf{e}_j}$，然后向用户询问，让其对这些策略进行比较，直到找到其中个性化价值（近似）最大的一个 $\pi^{\mathbf{e}_j}$。该策略将作为我们的基准策略，π_1。具体细节见算法 1 的第 1–6 行。

算法 1 基策略识别

1：初始化 $\pi^{\mathbf{e}^*} \leftarrow \pi^{\mathbf{e}_1}$2：对于 $j = 2, \ldots, k$，执行 3：比较 $\pi^{\mathbf{e}^*}$ 和 $\pi^{\mathbf{e}_j}$4：如果 $\pi^{\mathbf{e}_j} \succ \pi^{\mathbf{e}^*}$，则 $\pi^{\mathbf{e}^*} \leftarrow \pi^{\mathbf{e}_j}$5：结束循环 6：$\pi_1 \leftarrow \pi^{\mathbf{e}^*}$ 和 $u_1 \leftarrow \frac{V^{\pi}\mathbf{e}}{\|_{V^{\pi}} \quad \mathbf{e}_*\|_2}$

7：对于 $i = 2, \ldots, k$，执行 8：任意选取一个正交归一基，$\rho_1, \ldots, \rho_{k+1-i}$，作为 $\mathrm{span}(V^{\pi_1}, \ldots, V^{\pi_{i-1}})^{\perp}$的基。9：$j_{\max} \leftarrow \arg\max_{j \in [k+1-i]} \max(|v^{\rho_j}|, |v^{-\rho_j}|)$。10：如果 $\max(|v^{\rho_{j_{\max}}}|, \overline{|v^{-\rho_{j_{\max}}}|}) > 0$，则 11：若 $|v^{\rho_{j_{\max}}}| > |v^{-\rho_{j_{\max}}}|$，则 $\pi_i \leftarrow \pi^{\rho_{j_{\max}}}$；否则 $\pi_i \leftarrow \pi^{-\rho_{j_{\max}}}$。$u_i \leftarrow \rho_{j_{\max}}$12：否则 13：返回 $(\pi_1, \pi_2, \ldots), (u_1, u_2, \ldots)$ 14：结束条件判断 15：结束循环

π_2、……、π_d的选择。在找到 π_1 后，我们接下来依次搜索剩余的 $d-1$ 个策略 π_2、……、π_d（算法 1 的第 8–13 行）。对于 $i = 2$、……、d，我们寻找一个方向 u_i，使得 (i) 向量 {v

1} 与当前价值向量张成的空间 span(V^{π_1}、……、$V^{\pi_{i-1}}$) 正交，且 (ii) 存在一个策略 π_i，使得 V^{π_i} 在 u_i 的方向上有显著分量。条件 (i) 用于保证策略 π_i 的价值向量与 span(V^{π_1}、……、$V^{\pi_{i-1}}$) 线性无关。条件 (ii) 用于应对比率 $\widehat{\alpha}_i$ 的不准确近似所造成的误差。直观而言，当 $\|\alpha_i V^{\pi_1} - V^{\pi_{i+1}}\|_2 \ll \epsilon$ 时，$\widehat{\alpha}_i V^{\pi_1} - V^{\pi_{i+1}}$ 与 $\alpha_i V^{\pi_1} - V^{\pi_{i+1}}$ 之间的角度可能非常大，从而导致沿 $\alpha_i V^{\pi_1} - V^{\pi_{i+1}}$ 方向对 w^* 的估计不准确。例如，若对于 $i = 2$、……、k，$V^{\pi_1} = \mathbf{e}_1$ 且 $V^{\pi_i} = \mathbf{e}_1 + \frac{1}{w_i^*}\epsilon \mathbf{e}_i$，那么 π_1、π_i 就“无法区分”，估计比率 $\widehat{\alpha}_{i-1}$ 可以为 1。通过求解方程 (1) 中的线性方程得到的 w^* 估计值为 (1, 0, …… , 0)，这可能与真实的 w^* 相差甚远。找到那些策略值具有较大分量的 u_i，有助于解决这个问题。

算法 1 对该过程进行了更详细的描述。请注意，如果有 n 个不同的用户，我们最多运行算法 1 k 次，而不是 n 次。原因是，算法 1 仅在搜索 π_1 时使用偏好比较（第 1–6 行），而搜索 π_2、……、π_k 时则不使用（这使计算复杂度增加了 k^2 倍）。由于 π_1 最多有 k 个候选项，即 π^{e_1}、……、π^{e_k}，我们将算法 1 的第 1–6 行执行 n 轮，而第 8–13 行仅执行 k 轮。

3.2 基准比率的计算

如前所述，仅比较基策略并不能精确计算 α_i 比率，因为比较 π_1、π_i 只能揭示哪一个更好，却无法说明好多少。为此，我们将使用“什么都不做”策略，通过对参数 $\widehat{\alpha}_i \in [0, C_\alpha]$ 进行二分搜索，并使用策略 π_{i+1} 与策略 $\widehat{\alpha}_i \pi_1 + (1 - \widehat{\alpha}_i)$ 进行比较查询（如果 $\widehat{\alpha}_i \leq 1$，则为 π_0），以近似每个比率 α_i，使其加性误差不超过 $|\widehat{\alpha}_i - \alpha_i|$。其中 $C_\alpha \geq 1$ 将在后文确定；如果 $\widehat{\alpha}_i > 1$，则改为比较 π_1 和 $\frac{1}{\widehat{\alpha}_i}\pi_{i+1} + (1 - \frac{1}{\widehat{\alpha}_i})\pi_0$。[7] 请注意，$\widehat{\alpha}_i \pi_1 + (1 - \widehat{\alpha}_i)\pi_0$ 的个性化价值，与 π_1 的个性化价值乘以 $\widehat{\alpha}_i$ 完全相同。当用户返回“无法区分”时，我们就停止。停止时，这两个策略的个性化价值大致相同，

$$\text{if } \widehat{\alpha}_i \leq 1, |\widehat{\alpha}_i \langle w^*, V^{\pi_1}\rangle - \langle w^*, V^{\pi_{i+1}}\rangle| \leq \epsilon; \text{ if } \widehat{\alpha}_i > 1, \left|\langle w^*, V^{\pi_1}\rangle - \frac{1}{\widehat{\alpha}_i}\langle w^*, V^{\pi_{i+1}}\rangle\right| \leq \epsilon. \quad (3)$$

由式 (1) 与上述不等式可得 $|\widehat{\alpha}_i - \alpha_i| \langle w^*, V^{\pi_1}\rangle \leq C_\alpha \epsilon$。因此，每个比率的近似误差均被限制在 $|\widehat{\alpha}_i - \alpha_i| \leq \frac{C_\alpha \epsilon}{\langle w^*, V^{\pi_1}\rangle}$。为确保该过程能够终止，~~我们需要~~设置 $C_\alpha \geq \frac{v^*}{\langle w^*, V^{\pi_1}\rangle}$，因为 α_i 必须位于区间 $[0, \frac{v^*}{\langle w^*, V^{\pi_1}\rangle}]$ ~~内~~。二分搜索在式 (3) 成立时停止，估计所有 α_i 最多需要 $\mathcal{O}(d\log(C_\alpha \langle w^*, V^{\pi_1}\rangle / \epsilon))$ 次比较查询。

由于算法 1 中精心选取了 π_1，我们可以将 $\frac{v^*}{\langle w^*, V^{\pi_1}\rangle}$ 的上界~~限定为~~ $2k$，并通过选择 $C_\alpha = 2k$ 得出 $|\widehat{\alpha}_i - \alpha_i|$ 的上界。

引理 1. 当 $\epsilon \leq \frac{v^*}{2k}$ 时，我们有 $\frac{v^*}{\langle w^*, V^{\pi_1}\rangle} \leq 2k$，并且对于每个 $i \in [d]$，都有 $|\widehat{\alpha}_i - \alpha_i| \leq \frac{4k^2\epsilon}{v^*}$。

接下来，我们设定 $C_\alpha = 2k$。上述估计 α_i 的过程的伪代码见附录 C 中的算法 3。

3.3 偏好近似与个性化策略

接下来，我们将介绍一种算法，用于估计 w^* 并计算近乎最优的个性化策略。给定算法 1 返回的 π_i和算法 3 返回的 $\widehat{\alpha}_i$，考虑一个矩阵 $\widehat{A} \in \mathbb{R}^{d \times k}$，其第一行为 $V^{\pi_1 \top}$，第 i行为 $(\widehat{\alpha}_{i-1} V^{\pi_1} - V^{\pi_i})^\top$，对每个 $i = 2$，……，d均如此。令$\widehat{w}$ 为 $\widehat{A}x = \mathbf{e}_1$的一个解。我们将证明，$\widehat{w}$ 是 w' 的良好估计值：$= \frac{w^*}{\langle w^*, V^{\pi_1}\rangle}$ 并且 $\pi^{\widehat{w}}$ 是近乎最优的个性化策略。具体而言，当 ϵ 较小~~时~~，~~对~~于每个策略 π，都有 $|\langle \widehat{w}, V^\pi\rangle - \langle w', V^\pi\rangle| = \mathcal{O}(\epsilon^{\frac{1}{3}})$。综合以上结果，我们得到以下定理。

[7] 我们用 $\widehat{\alpha}_i \pi_1 + (1 - \widehat{\alpha}_i)\pi_0$ 表示以 $\widehat{\alpha}_i$ 的概率使用 π_1，并以 $1 - \widehat{\alpha}_i$ 的概率使用 π_0。

$\widehat{}$ $\widehat{}$定理 1。考虑用于计算 A 的算法，以及 $Ax = \mathbf{e}_1$ 的任意解 $\widehat{w}$，并输出策略 $\pi^{\widehat{w}} = \arg\max_{\pi\in\Pi} \langle \widehat{w},\ V^{\pi}\rangle$；该策略是偏好向量 $\widehat{w}$对应的最优个性化策略。则输出策略 $\pi^{\widehat{w}}$满足 $v^* - \left\langle w^*,\ V^{\pi^{\widehat{w}}}\right\rangle \leq \mathcal{O}\left(\left(\sqrt{k}+1\right)^{d+\frac{14}{3}} \epsilon^{\frac{1}{3}}\right)$

使用 $\mathcal{O}(k\log(k/\epsilon))$ 次比较查询。

计算复杂度 我们指出，算法 1 至多求解 $\mathcal{O}(k^2)$ 次标量奖励马尔可夫决策过程中的最优策略，即求解式 (2)。例如，使用有限时域值迭代求解最优策略需要 $\mathcal{O}(H|\mathcal{S}|^2|\mathcal{A}|)$ 步。然而，尽管为单个用户返回最优策略所需的时间复杂度为 $\mathcal{O}(k^2H|\mathcal{S}|^2|\mathcal{A}| + k\log(\frac{k}{\epsilon}))$，但若考虑 n 个不同用户而非一个用户，总体时间复杂度将达到 $\mathcal{O}((k^3+n)H|\mathcal{S}|^2|\mathcal{A}| + nk\log(\frac{k}{\epsilon}))$。

证明技巧。标准技巧通常先推导出 $\|\widehat{w} - w^*\|_2$ 的一个上界，再利用该上界为 $\sup_\pi |\langle \widehat{w}, V^\pi\rangle - \langle w^*, V^\pi\rangle|$ 给出上界，其中 $C_V \|\widehat{w} - w^*\|_2$。然而，当两个基策略彼此接近时，这种方法无法得到非平凡的界。例如，考虑返回的基策略值：$V^{\pi_1} = (1, 0, 0)$、$V^{\pi_2} = (1, 1, 0)$ 和 $V^{\pi_3} = (1, 0, \eta)$，其中 $\eta > 0$。当 η 极小时，估计值 $\widehat{w}$ 在 (0, 0, 1) 方向上会变得极不准确，从而导致较大的 $\|\widehat{w} - w^*\|_2$。即使在这种情况下，我们仍然可以获得非平凡的保证，因为选择 V^{π_3}（算法 1 的第 11 行）意味着 Π 中没有任何策略在 (0, 0, 1) 方向上的分量大于 π_3。

证明概要。定理 1 的分析分为两部分。首先，如第 3.1 节所述，当$\|\alpha_i V^{\pi_1} - V^{\pi_{i+1}}\|_2 \ll \epsilon$时，$\widehat{\alpha}_{i+1}$ 的误差可能导致对方向$\alpha_i V^{\pi_1} - V^{\pi_{i+1}}$上的 w^* 产生不准确的估计值。因此，我们考虑仅基于一些 π_{i+1}的另一种 w^* 估计值，这些 π_{i+1}具有相对较大的 $\|\alpha_i V^{\pi_1} - V^{\pi_{i+1}}\|_2$。具体而言，对于任意 $\delta > 0$，令 $d_\delta := \min_{i\geq 2:\max(|v^{u_i}|, |v^{-u_i}|)\leq\delta} i -$

1. 也就是说，对于 $i = 2$、……、d_δ，策略 π_i 满足 $\langle u_i$、$V^{\pi_i}\rangle > \delta$；对于任意策略 π，都有$\langle u_{d_\delta+1}$、$V^\pi\rangle \leq \delta$。于是，对于任意策略 π 以及 span(V^{π_1}、……、$V^{\pi_{d_\delta}}$)$^\perp$中的任意单位向量 $\xi \in$，都有$\langle \xi$、$V^\pi\rangle \leq \sqrt{k}\delta$。这是因为在第 $d_\delta + 1$轮，我们选取 span(V^{π_1}、……、$V^{\pi_{d_\delta}}$)$^\perp$ 的一组标准正交基 ρ_1、……、ρ_{k-d_δ}（算法 1 第 8 行），并按照第 9 行所述，选取其中存在某个策略具有最大分量的 $u_{d_\delta+1}$。因此，对于所有 $j \in [k - d_\delta]$，都有 $|\langle \rho_j$、$V^\pi\rangle| \leq \delta$。于是，根据柯西-施瓦茨不等式可得 $\langle \xi$、$V^\pi\rangle = \sum_{j=1}^{k-d_\delta} \langle \xi$、$\rho_j\rangle \langle \rho_j$、$V^\pi\rangle \leq \sqrt{k}\delta$。令$\widehat{A}^{(\delta)} \in \mathbb{R}^{d_\delta\times k}$ 为由 $\widehat{A}$的前 d_δ 行组成的子矩阵。接着，我们考虑另一种估计值 $\widehat{w}^{(\delta)} = \arg\min_{x:\widehat{A}^{(\delta)}x=\mathbf{e}_1} \|x\|_2$，即 x 到 $\widehat{A}^{(\delta)}x = \mathbf{e}_1$的最小范数解。接着，我们在引理 2 中给出 $\sup_\pi \left|\left\langle \widehat{w}^{(\delta)}\text{、}\ V^\pi\right\rangle - \langle w'\text{、}\ V^\pi\rangle\right|$ 的上界，并在引理 3 中给出 $\sup_\pi \left|\langle \widehat{w}\text{、}\ V^\pi\rangle - \left\langle \widehat{w}^{(\delta)}\text{、}\ V^\pi\right\rangle\right|$ 的上界。至此，定理 1 的证明完成。

引理 2. 如果对于所有 $i \in [d-1]$，$|\widehat{\alpha}_i - \alpha_i| \leq \epsilon_\alpha$和 $\alpha_i \leq C_\alpha$，那么对于每个 $\delta \geq 4C_\alpha^{\frac{2}{3}} C_V d^{\frac{1}{3}} \epsilon_\alpha^{\frac{1}{3}}$，我们有 $\left|\left\langle \widehat{w}^{(\delta)}, V^\pi\right\rangle - \langle w', V^\pi\rangle\right| \leq \mathcal{O}(\frac{C_\alpha C_V^4 d_\delta^{\frac{3}{2}} \|w'\|_2^2 \epsilon_\alpha}{\delta^2} + \sqrt{k}\delta \|w'\|_2)$，对所有 π 均成立，其中 $w' = \frac{w^*}{\langle w^*, V^{\pi_1}\rangle}$。

$\widehat{}$由于我们只移除 A 中对应于 u_i 的行，而这些 u_i 位于不存在任何策略的价值具有较大分量的子空间中，因此就 $\sup_\pi \left|\langle \widehat{w}, V^\pi\rangle - \left\langle \widehat{w}^{(\delta)}, V^\pi\right\rangle\right|$ 而言，$\widehat{w}$ 和 $\widehat{w}^{(\delta)}$ 彼此接近。

引理 3. 若对所有 $i \in [d-1]$，$|\widehat{\alpha}_i - \alpha_i| \leq \epsilon_\alpha$ 且 $\alpha_i \leq C_\alpha$，则对于每个策略 π 和每个 $\delta \geq 4C_\alpha^{\frac{2}{3}} C_V d^{\frac{1}{3}} \epsilon_\alpha^{\frac{1}{3}}$，我们有 $\left|\widehat{w}\cdot V^\pi - \widehat{w}^{(\delta)}\cdot V^\pi\right| \leq \mathcal{O}((\sqrt{k}+1)^{d-d_\delta} C_\alpha \epsilon^{(\delta)})$，其中 $\epsilon^{(\delta)} = \frac{C_\alpha C_V^4 d_\delta^{\frac{3}{2}} \|w'\|_2^2 \epsilon_\alpha}{\delta^2} + \sqrt{k}\delta \|w'\|_2$ 是引理 2 中的上界。

注意，定理 1 的结果中包含因子 $k^{\frac{d}{2}}$，它对 d呈指数依赖。通常，我们考虑 $k = \mathcal{O}(1)$ 较小的情况，因此 $k^d = \mathcal{O}(1)$ 也较小。通过应用 $\widehat{w}^{(\delta)}$ 直接估计 w^*，我们消除了对 d 的指数依赖；这要求我们预先设定 δ 的值。下面的定理可由在引理 2 中为 δ 赋予最优值直接推出。

定理 2。考虑计算 $\widehat{A}$ 的算法，以及任意一个解 $\widehat{w}^{(\delta)}$ 对于 $\widehat{A}^{(\delta)}x = \mathbf{e}_1$ 而言，$\delta = k^{\frac{5}{3}}\epsilon^{\frac{1}{3}}$ 并输出策略 $\pi^{\widehat{w}^{(\delta)}} = \arg\max_{\pi\in\Pi} \left\langle \widehat{w}^{(\delta)},\ V^\pi\right\rangle$。那么，策略 $\pi^{\widehat{w}^{(\delta)}}$满足 $v^* - \left\langle w^*,\ V^{\pi^{\widehat{w}}{}_{(\delta)}}\right\rangle \leq \mathcal{O}\left(k^{\frac{13}{6}}\epsilon^{\frac{1}{3}}\right)$.

请注意，定理 2 中的算法需要预先设置超参数 δ，而定理 1 中则无需设置任何超参数。δ 的取值不当可能会降低算法的性能。不过，我们可以通过二分搜索 $\eta \in [0,1]$，并将 π_1 与其缩放版本（$1-\eta$）π_1 进行比较，从而近似估计 ϵ，直到找到一个 η，使用户无法区分 π_1 和（$1-\eta$）π_1。然后，我们可以得到 ϵ 的估计值，并用该估计值来设置超参数。

我们指出，尽管我们认为 k 是一个较小的数，但定理 1 和定理 2 中对 ϵ 的依赖是否达到最优尚不明确。关于 ϵ 的紧确依赖仍是一个开放问题。我们在附录 I 中简要讨论了改进这一界限的一个潜在方向。

4 从加权轨迹集表示中学习

在上一节中，我们使用策略的显式形式——状态-动作映射——来表示策略。然而，用户可能难以理解这种表示。例如，仅用一列 $|\mathcal{S}|$ 个状态和动作（如“左转”）描述的汽车有多安全？本节中，我们设计了一些算法，它们返回一种更易于理解的策略表示——加权轨迹集。

回顾第 2 节中的定义，加权轨迹集是一个较小的集合，其中包含来自策略支撑集的轨迹及其对应权重，并且具有如下性质：该集合中轨迹的期望回报（按权重计算）恰好等于策略的价值（下文称为精确价值性质）。[8]由于这些集合保留了策略多目标价值的所有信息，因此可将其用作算法 3 中策略比较查询的策略表示，而不会降低反馈质量。因此，使用这些表示可以在第 3 节中获得关于所返回策略的相同最优性保证（但计算这些集合需要额外的计算时间）。

本节算法基于两个关键观察：

(1) 每个策略 π 都会诱导出一个轨迹分布。令 $q^\pi(\tau)$ 表示按照 π选择动作时采样到轨迹 τ 的概率。在 q^π 下所有轨迹的期望回报与策略价值相同，即 $V^\pi = \sum_\tau q^\pi(\tau)\Phi(\tau)$。具体而言，策略价值是其支撑集内各轨迹回报的凸组合。然而，我们避免使用这种凸组合来表示策略，因为策略支撑集中的轨迹数量可能随状态和动作数量呈指数级增长。

(2) Carathéodory 定理蕴含小型加权轨迹集的存在。也就是说，由于策略价值本身是其支撑集中各轨迹回报的凸组合，Carathéodory 定理表明，策略的支撑集中存在 $k+1$ 条轨迹及其对应的权重，使得这些轨迹回报的凸组合等于策略价值。这样的（$k+1$）大小的集合将作为我们算法的输出。

我们可以运用 Carathéodory 定理证明背后的思路，按如下方式压缩轨迹。对于任意一个大小为 $k+2$、由 k 维向量 $\{\mu_1, \ldots, \mu_{k+2}\}$ 构成的集合，以及这些向量的任意凸组合 $\mu = \sum_{i=1}^{k+2} p_i \mu_i$，我们总能找到一个大小为 $k+1$ 的子集，通过求解一个线性方程，将 μ 表示为该子集的凸组合。给定一个定义在由 k 维向量 M 构成的集合上的概率分布 p作为输入，我们从 M中选取 $k+2$ 个向量，并通过上述过程至少消去其中一个向量。重复此步骤，直到剩下至多 $k+1$个向量。我们将此算法称为 C4（使用Carathéodory定理压缩凸组合）。伪代码见算法4；由于篇幅所限，相关内容放在附录J中。该算法的构造可立即推出以下引理。

引理 4. 给定一组 k 维向量 $M \subset \mathbb{R}^k$ 以及一个定义在 M 上的分布 p，C4(M,p) 在时间 $\mathcal{O}(|M|\,k^3)$ 内输出 $M' \subset M$，满足 $|M'| \le k+1$，且分布 $q \in$ 属于单纯形$^{M'}$，并满足 $\mathbb{E}_{\mu\sim q}[\mu] = \mathbb{E}_{\mu\sim p}[\mu]$。

[8]如果不要求满足精确价值性质，我们可以直接从策略中采样 $\mathcal{O}(\log k/(\epsilon')^2)$ 条轨迹，并赋予均匀权重。以较高概率，每个目标函数的期望回报都与其期望值相差不超过 ϵ'。问题在于，这组轨迹不一定能捕捉到罕见事件。例如，如果根据策略不同，任何一辆汽车发生碰撞的概率介于 $(\epsilon')^4$ 和 $(\epsilon')^2$, 之间，那么只关心安全性（即不发生碰撞）的用户很可能完全观察不到任何“不安全”的轨迹，从而导致我们错失有价值的反馈。

因此，我们现在知道如何将一组轨迹压缩到所需的大小。主要挑战是如何高效地完成这件事（在时间 $\mathcal{O}(\text{poly}(H\,|S|\,|\mathcal{A}|))$ 内）。也就是说，由于策略支撑集中的所有轨迹回报构成的集合大小可能达到 $\Omega(|\mathcal{S}|^H)$，将其作为 C4 算法的输入效率很低。相反，我们只会在轨迹数量较少时将 C4 用作子程序。

我们提出两种高效的方法来寻找加权轨迹表示。这两种方法都利用了这样一个性质：所有轨迹都是由同一马尔可夫决策过程上的同一策略生成的。第一种方法从一组长度为 1 的小规模轨迹开始，对其进行扩展和压缩，并重复这一过程，直到得到轨迹长度集合 H。另一种方法基于分层图的构造，其中策略对应于该图中的流；我们证明，寻找代表性轨迹等价于流分解。

在下一小节中，我们将介绍扩展与压缩方法。考虑到篇幅，我们将在附录 K 中介绍基于流分解的方法。值得指出的是，流分解方法的运行时间为 $\mathcal{O}(H^2\,|\mathcal{S}|^2 + k^3 H\,|\mathcal{S}|^2)$（详见附录），只要满足 $|\mathcal{S}|H + |\mathcal{S}|k^3 = \omega(k^4 + k|\mathcal{S}|)$，其性能就不如扩展与压缩方法（见定理 3）。为便于说明，以下我们只考虑确定性策略。我们的方法可以轻松扩展到随机策略。[9]

扩展与压缩方法。

基本思路是先寻找长度为 1 的 $k+1$ 轨迹来表示 V^π，然后在不增加轨迹数量的情况下延长轨迹。对于策略 π，令 $V^\pi(s,h) = \mathbb{E}_{S_0=s}\left[\sum_{t=0}^{h-1} R(S_t, \pi(S_t))\right]$ 表示初始状态为 $S_0 = s$、时间范围为 h时 π 的价值。由于本节研究的是固定策略 π 下的表示，我们稍微滥用一下记号，用 $\tau^\pi = (s, s_1, \ldots, s_H)$ 表示一条轨迹。我们将轨迹 τ 在时间 t 的状态记为 $s_t^\tau = s_t$。对于 π 的一条轨迹前缀 $\tau = (s, s_1, \ldots, s_h)$，其初始状态为 s，并包含 $h \le H$ 个后续状态，τ的回报为 $\Phi(\tau) = R(s, \pi(s)) + \sum_{t=1}^{h-1} R(s_t, \pi(s_t))$。令 $J(\tau)$表示长度为 H且前缀为 τ的轨迹的期望回报，即，

$$J(\tau) := \Phi(\tau) + V(s_h^\tau, H-h)\,.$$

对于任意 $s \in \mathcal{S}$，令 $\tau \circ s$表示将 s追加到 τ后得到的轨迹。我们可以通过动态规划在时间 $\mathcal{O}(kH\,|\mathcal{S}|^2)$ 内求解所有$s \in \mathcal{S}$、$h \in [H]$ 对应的 $V^\pi(s,h)$。具体而言，根据定义，有 $V^\pi(s,1) = R(s,\pi(s))$，并且

$$V^\pi(s, h+1) = R(s,\pi(s)) + \sum_{s' \in \mathcal{S}} P(s'|s,\pi(s)) V^\pi(s', h)\,. \tag{4}$$

因此，我们可以将 V^π 表示为

$$V^\pi = R(s_0,\pi(s_0)) + \sum_{s\in\mathcal{S}} P(s|s_0,\pi(s_0)) V^\pi(s, H-1) = \sum_{s\in\mathcal{S}} P(s|s_0,\pi(s_0)) J(s_0, s)\,.$$

通过应用 C4，我们可以找到一组长度为 1 的代表性轨迹，$F^{(1)} \subset \{(s_0, s)|s \in \mathcal{S}\}$，以及满足 $\left|F^{(1)}\right| \le k+1$ 的权重 $\beta^{(1)} \in$，其中 $^{F^{(1)}}$ 属于单纯形

$$V^\pi = \sum_{\tau \in F^{(1)}} \beta^{(1)}(\tau) J(\tau)\,. \tag{5}$$

假设给定一组长度为 t、权重为 $\beta^{(t)}$ 的轨迹 $F^{(t)}$，满足 $V^\pi = \sum_{\tau\in F^{(t)}} \beta^{(t)}(\tau) J(\tau)$，我们可以先通过公式（4）将轨迹长度增加1，从而得到 $\{\tau \circ s | \tau \in F^{(t)}, s \in \mathcal{S}\}$的一个子集，其中的轨迹长度为 $t+1$。具体而言，我们有

$$V^\pi = \sum_{\tau \in F^{(t)}, s\in\mathcal{S}} \beta^{(t)}(\tau) P(s|s_t^\tau, \pi(s_t^\tau)) J(\tau \circ s)\,. \tag{6}$$

[9]在第 3 节中，我们考虑一种特殊类型的随机策略，它是某个确定性策略（通过求解标量奖励马尔可夫决策过程最优策略的算法所得到的输出）与“什么都不做”策略的混合策略。对于这种特定的随机策略，我们可以分别为这两种策略找到加权轨迹表示，然后应用算法4来压缩该表示。

接下来，由于计算时间有限，我们希望通过C4压缩上述凸组合，以便最多保留 $k+1$ 条长度为 $t+1$ 的轨迹。更形式化地，令$J_{F^{(t)}} := \{J(\tau \circ s) | \tau \in F(t)\, s \in \mathcal{S}\}$ 为期望回报的集合，$p_{F^{(t)}}$、$\beta^{(t)} \in$为公式（6）中出现的单纯形$^{F^{(t)} \times \mathcal{S}}$，其中$p_F(t)\ _{\beta^{(t)}}(\tau \circ s) = \beta^{(t)}(\tau)P(s|s_t^\tau, \pi(s_t^\tau))$为权重。这里，$p_F(t)\ _{\beta^{(t)}}$定义了一个关于$J_F(t)$的分布，抽取 $J(\tau \circ s)$的概率为$p_F(t)\ _{\beta^{(t)}}(\tau \circ s)$。然后，我们可以对$(J_F(t)\ p_F(t)\ _{\beta^{(t)}})$应用C4，并压缩代表性轨迹 $\{\tau \circ s | \tau \in F(t)\, s \in \mathcal{S}\}$。我们从长度为1的轨迹开始，反复进行扩展和压缩，直至得到长度为 H的轨迹。具体过程见算法2。

算法2 扩展与压缩轨迹

1: 根据公式（4），通过动态规划计算所有 $s \in \mathcal{S}, h \in [H]$ 对应的 $V^\pi(s,h)$
2: $F^{(0)} = \{(s_0)\}$ 和 $\beta^{(0)}(s_0) = 1$
3: 对于 $t = 0, \ldots, H-1$ 执行
4: $J_{F^{(t)}} \leftarrow \{J(\tau \circ s) | \tau \in F^{(t)},\ s \in \mathcal{S}\}$和 $p_{F^{(t)}\ \beta^{(t)}}(\tau \circ s) \leftarrow \beta^{(t)}(\tau)\ P(s|s_t^\tau,\ \pi(s_t^\tau))$ 用于 $\tau \in F^{(t)},\ s \in \mathcal{S}$ // 扩展步骤
5: $(J^{(t+1)}, \beta^{(t+1)}) \leftarrow$C4$(J_{F^{(t)}}\ p_{F^{(t)}\ \beta^{(t)}})$ 和 $F_{(t+1)} \leftarrow \{\tau | J(\tau) \in J^{(t+1)}\}$ // 压缩 step
6: 结束循环
7: 输出 $F^{(H)}$ 和 $\beta^{(H)}$

定理 3. 算法2 输出 $F^{(H)}$ 和 $\beta^{(H)}$，满足$\left|F^{(H)}\right| \leq k+1$ 和$\sum_{\tau \in F^{(H)}} \beta^{(H)}(\tau)\Phi(\tau) = V^\pi$，耗时 $\mathcal{O}(k^4 H\,|\mathcal{S}| + kH\,|\mathcal{S}|^2)$。

定理 3 的证明可由算法的构造直接得出。根据式 (5)，我们有 $V^\pi = \sum_{\tau \in F^{(1)}} \beta^{(1)}(\tau)J(\tau)$。然后，我们可以对代表性轨迹的长度进行归纳，证明算法2的输出是一个有效的加权轨迹集。C4 保证对于所有 $t = 1, \ldots, H$，$\left|F^{(t)}\right| \leq k+1$，因此，我们在每一步最多只需跟踪 $k+1$ 条轨迹，即可满足定理中的计算保证。结合定理 1，我们得到以下推论。

推论1. 使用算法2返回的加权轨迹集表示运行定理 1 中的算法，可在 $\mathcal{O}(k^2 H|\mathcal{S}|^2|\mathcal{A}| + (k^5 H\,|\mathcal{S}| + k^2 H\,|\mathcal{S}|^2)\log(\frac{k}{\epsilon}))$ 时间内获得与定理 1 相同的保证。

5 讨论

本文设计了高效算法，用于通过比较反馈学习用户对多个目标的偏好。这种高效性体现在运行时间和查询数量上（两者均关于 H、$|\mathcal{S}|$、$|\mathcal{A}|$、k 为多项式，并关于 $1/\epsilon$ 为对数级）。随后，可以利用学到的用户偏好，将为该用户寻找个性化最优策略的问题，转化为一个（有限时域的）单一标量奖励马尔可夫决策过程；这一问题已有已知的高效解法。由于我们的重点是尽量减少策略比较查询，因此所设计的算法基于多项式时间的预处理计算，从而为用户节省宝贵的比较时间。

第3节中的结果本身也具有独立价值，并可应用于更一般的学习场景：对于某个未知的线性参数 w^*，给定一组点 X，并可对任意两个点进行比较查询，目标是学习 $\arg\max_{x \in X} \langle w^*, x \rangle$。例如，在依据咖啡供应商描述的咖啡特征（醇厚度、香气、咖啡脂、烘焙程度.....）为咖啡豆提供个性化推荐时，用户可能无法描述自己理想的咖啡豆特征；采用第3节的方法，则可以通过比较为用户找到理想的咖啡豆（其中，与“什么都不做”混合的情形可通过加水稀释咖啡来实现，而对于给定特征而言，最优咖啡就是与之最接近的咖啡）。

当我们从将策略明确表示为从状态到动作的映射，转向更自然的加权轨迹集策略表示时，在查询数量方面仍获得了相同的最优性保证。策略表示还可以有其他形式（例如，常见状态的一个小子集），而我们的加权轨迹

集表示的一项优势在于，它通过 $\mathcal{O}$ 条轨迹及其权重，清晰地体现了策略的多目标价值。第4节提供的算法是独立的，也可能对可解释强化学习（Alharin 等人，2020）具有独立的参考价值。例如，为了展示通用机器人吸尘器的多目标性能（如果我们只有例如 3 台吸尘器，这会很有用——我们可以应用第4节中的算法生成加权轨迹集表示，并直接进行比较，无需经过第3节中的算法）。

未来研究的一个有趣方向是放宽“马尔可夫决策过程预先已知”这一假设。一种直接的方法是先学习模型（采用基于模型的强化学习），然后在学习得到的马尔可夫决策过程中应用我们的算法。此时，所返回策略的次优性将取决于模型的估计误差以及我们的算法引入的误差（后者取决于学习所得模型中的参数）。

致谢

本研究部分得到了美国国家科学基金会的资助（项目编号 CCF-2212968 和 ECCS-2216899），以及美国国防高级研究计划局根据合作协议 HR00112020003 提供的资助。本研究中表达的观点不一定反映政府的立场或政策，也不应据此推断政府给予了官方认可。本文已获准公开发布；分发不受限制。
本项目部分得到了欧洲研究理事会（ERC）通过欧盟“地平线2020”研究与创新计划提供的资助（资助协议编号 882396）、以色列科学基金会（项目编号 993/17）、特拉维夫大学人工智能与数据科学中心（TAD）、埃里克与温迪·施密特基金，以及特拉维夫大学的 Yandex 机器学习倡议的资助。
We 谨向所有匿名审稿人，尤其是 Gnoo 审稿人，表示感谢，感谢他们提出的建设性 e co意见。

参考文献

Ailon, N.（2012）。一种基于成对偏好进行排序的主动学习算法，具有近乎最优的查询复杂度。《机器学习研究杂志》，13(1)：137–164。

Ailon, N.、Karnin, Z. S. 和 Joachims, T.（2014）。将对决老虎机归约为基数老虎机。载于 ICML，第 32 卷，第 856–864 页。

Akrour, R.、Schoenauer, M. 和 Sebag, M.（2012）。APRIL：基于主动偏好学习的强化学习。载于《欧洲机器学习与数据库知识发现联合会议论文集》，第 116–131 页。

Alharin, A., Doan, T.-N. 和 Sartipi, M.（2020）。强化学习解释方法：综述。IEEE Access，8：171058–171077。

Balcan, M.、Blum, A. 和 Vempala, S. S.（2015）。终身学习和自编码的高效表示。载于学习理论年度会议（COLT）论文集，第 191–210 页。

Balcan, M.-F.、Vitercik, E. 和 White, C.（2016）。通过成对比较学习组合函数。载于学习理论年度会议（COLT）论文集。

Barrett, L. 和 Narayanan, S.（2008）。通过多重标准学习所有最优策略。载于国际机器学习会议（ICML）论文集，第 41–47 页。

Bengs, V.、Busa-Fekete, R.、El Mesaoudi-Paul, A. 和 Hüllermeier, E.（2021）。基于偏好的在线学习与决斗赌博机：综述。机器学习研究期刊。

Bhatia, K., Pananjady, A., Bartlett, P., Dragan, A., and Wainwright, M. J. (2020). 多重标准下的偏好学习：博弈论视角。载于《神经信息处理系统进展》（NeurIPS），第7413–7424页。

Biyik, E. and Sadigh, D. (2018). 奖励函数的批量主动偏好学习。载于《机器人学习会议》（CoRL）论文集，第519–528页。Blum, A. and Shao, H. (2020). 基于主损失和次损失的在线学习。《神经信息处理系统进展》，33:20427–20436。Chatterjee, K. (2007). 具有多个长期平均目标的马尔可夫决策过程。载于《国际软件技术基础与理论计算机科学会议》（FSTTCS）论文集，第473–484页。Chatterjee, K., Majumdar, R., and Henzinger, T. A. (2006). 具有多个目标的马尔可夫决策过程。载于《理论计算机科学年度研讨会》（STACS）论文集，第325–336页。Chen, X., Ghadirzadeh, A., Björkman, M., and Jensfelt, P. (2019). 面向多目标强化学习的元学习。载于《IEEE/RSJ智能机器人与系统国际会议》（IROS）论文集，第977–983页。Cheng, W., Fürnkranz, J., Hüllermeier, E., and Park, S.-H. (2011). 基于偏好的策略迭代：利用偏好学习进行强化学习。载于《欧洲机器学习与数据库知识发现联合会议》（ECML PKDD）论文集，第312–327页。Cheung, W. C. (2019). 基于向量反馈和复杂目标的强化学习遗憾最小化。载于《神经信息处理系统进展》（NeurIPS）。Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). 从人类偏好中进行深度强化学习。载于《神经信息处理系统进展》（NeurIPS）。Doumpos, M. and Zopounidis, C. (2007). 多准则决策中的偏好分解正则化估计。《计算优化与应用》，38(1):61–80。Fürnkranz, J., Hüllermeier, E., Cheng, W., and Park, S.-H. (2012). 基于偏好的强化学习：形式化框架与策略迭代算法。《机器学习》，89(1):123–156。Hayes, C. F., Radulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A. A., Mannion, P., Nowé, A., de Oliveira Ramos, G., Restelli, M., Vamplew, P., and Roijers, D. M. (2022). 多目标强化学习与规划实用指南。《自主智能体与多智能体系统》。Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., and Amodei, D. (2018). 在 Atari 中从人类偏好和演示中学习奖励。载于《神经信息处理系统进展》（NeurIPS）。Jain, A., Sharma, S., Joachims, T., and Saxena, A. (2015). 通过在线协作反馈学习操作任务中的偏好。《国际机器人研究杂志》，34(10):1296–1313。Jamieson, K. G. and Nowak, R. (2011). 使用成对比较进行主动排序。《神经信息处理系统进展》，24。Kane, D. M., Lovett, S., Moran, S., and Zhang, J. (2017). 基于比较查询的主动分类。载于《IEEE第58届计算机科学基础年度研讨会》（FOCS），第355–366页。IEEE。Knox, W. B., Hatgis-Kessell, S., Booth, S., Niekum, S., Stone, P., and Allievi, A. (2022). 用于学习奖励函数的人类偏好模型。arXiv预印本 arXiv:2206.02231。Lee, K., Smith, L., and Abbeel, P. (2021). PEBBLE：通过经验重标记和无监督预训练实现反馈高效的交互式强化学习。arXiv预印本 arXiv:2106.05091。

Mannor, S., Perchet, V., and Stoltz, G. (2014). 未知博弈中的可达性：在线学习与多目标优化相遇。载于《学习理论会议》，第339–355页。PMLR。Mannor, S. and Shimkin, N. (2001). 多准则强化学习的引导方法。载于《神经信息处理系统进展》（NeurIPS）。Mannor, S. and Shimkin, N. (2004). 多准则强化学习的几何方法。《机器学习研究杂志》，5:325–360。Marinescu, R., Razak, A., and Wilson, N. (2017). 具有可能最优策略的多目标影响图。AAAI人工智能会议论文集。McAuley, J. (2022). 个性化机器学习。剑桥大学出版社。即将出版。Pacchiano, A., Saha, A., and Lee, J. (2022). 对决式强化学习：基于轨迹偏好的强化学习。Ren, W., Liu, J., and Shroff, N. B. (2018). 基于成对和列表式查询的PAC排序：下界与上界。arXiv预印本 arXiv:1806.02970。Roijers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R. (2013). 多目标序贯决策综述。《人工智能研究期刊》，48:67–113。Rothkopf, C. A. and Dimitrakakis, C. (2011). 偏好 elicitation与逆向强化学习。载于《欧洲机器学习与数据库知识发现联合会议》论文集，第34–48页。Sadigh, D., Dragan, A. D., Sastry, S., and Seshia, S. A. (2017). 奖励函数的主动偏好学习。载于《机器人学：科学与系统》（RSS）论文集。Saha, A. and Gopalan, A. (2018). 赌博机对决。载于《人工智能中的不确定性》。Saha, A. and Gopalan, A. (2019). Plackett-Luce模型中的PAC对抗赌博机。载于《算法学习理论》，第700–737页。Saha, A., Koren, T., and Mansour, Y. (2021). 对决式凸优化。载于《机器学习国际会议》，第9245–9254页。PMLR。Sui, Y., Zhuang, V., Burdick, J., and Yue, Y. (2017). 具有相关臂的多重对决赌博机。载于《人工智能不确定性会议》，UAI'17。Sui, Y., Zoghi, M., Hofmann, K., and Yue, Y. (2018). 对决赌博机的进展。载于 IJCAI，第5502–5510页。Wang, N., Wang, H., Karimzadehgan, M., Kveton, B., and Boutilier, C. (2022). Imoˆ 3：交互式多目标离策略优化。载于第三十一届国际人工智能联合会议论文集，IJCAI-22。Wilson, A., Fern, A., and Tadepalli, P. (2012). 从轨迹偏好查询中进行策略学习的贝叶斯方法。载于《神经信息处理系统进展》（NeurIPS）。Wirth, C., Akrour, R., Neumann, G., and Fürnkranz, J. (2017a). 基于偏好的强化学习方法综述。《机器学习研究期刊》。Wirth, C., Akrour, R., Neumann, G., and Fürnkranz, J. (2017b). 基于偏好的强化学习方法综述。《机器学习研究杂志》，18(136):1–46。Wirth, C., Fürnkranz, J., and Neumann, G. (2016). 无模型基于偏好的强化学习。载于全国人工智能会议（AAAI）论文集。Yona, G., Moran, S., Elidan, G., and Globerson, A. (2022). 使用标签比较进行主动学习。载于《人工智能不确定性会议》（UAI）论文集。Zoghi, M., Whiteson, S., Munos, R., Rijke, M. d., et al. (2014). 针对 k臂对决赌博机问题的相对上置信界。载于 JMLR研讨会与会议论文集，第32卷，第10–18页。JMLR。

A 相关工作

多目标序贯决策 多目标序贯决策领域已有悠久的研究历史（Roijers 等人，2013），其中一个重点是实现高效算法以近似帕累托前沿（Chatterjee 等人，2006；Chatterjee，2007；Marinescu 等人，2017）。我们不去寻找可能最优的策略，而是专注于特定用户的偏好，并为该用户寻找最优策略。就像一个人理想的汽车可能是雪佛兰 Spark（小型车），而另一个人的理想汽车则可能是福特 Ranger（皮卡）。

在多目标强化学习的背景下（Hayes 等人，2022），目标可以表述为学习一种策略，使其平均回报向量属于某个目标集合（因此称为“多准则”强化学习）；已有研究将此问题视为随机博弈（Mannor 和 Shimkin，2001，2004）。其他研究则试图最大化（期望意义上的）奖励标量化值，该值可能对应于多个目标的加权和（Barrett 和 Narayanan，2008；Chen 等人，2019），正如本文所考虑的那样，也可能是目标的非线性函数（Cheung，2019）。多目标在线学习也已有研究，例如参见 Mannor 等人（2014）以及 Blum 和 Shao（2020）。

定义该标量化函数的参数（例如，各目标的相对权重）往往未知，并且会随任务设置或用户而变化。在这种情况下，通常使用偏好学习（Wirth 等人，2017a）来获取这些参数的取值。Doumpos 和 Zopounidis（2007）介绍了一种在多目标决策背景下获取用户相对权重的方法。Bhatia 等人（2020）利用博弈论方法，通过成对查询学习用户对多个目标的偏好，以确定最优随机化策略。在同时涉及标量目标和向量值目标的强化学习中，用户偏好查询提供了一种替代方案，可替代从示范中学习（这可能难以由人们提供，例如对于具有高度自由度的机器人），或明确指定奖励（Cheng 等人，2011；Rothkopf 和 Dimitrakakis，2011；Akrour 等人，2012；Wilson 等人，2012；Fürnkranz 等人，2012；Jain 等人，2015；Wirth 等人，2016；Christiano 等人，2017；Wirth 等人，2017b；Ibarz 等人，2018；Lee 等人，2021）。这些研究通常假设，人类对两条轨迹的含噪偏好与其效用之差相关（即，奖励作为潜在项，可用于预测偏好）。许多当代方法通过最小化基于奖励的预测与人类提供的偏好之间的交叉熵损失来估计潜在奖励（即，寻找使观测到的偏好似然最大的奖励）（Christiano 等人，2017；Ibarz 等人，2018；Lee 等人，2021；Knox 等人，2022；Pacchiano 等人，2022）。

Wilson 等人（2012）描述了一种用于策略学习的贝叶斯方法：通过询问用户在一对轨迹之间的偏好，并利用这些偏好来维护潜在策略参数的后验分布。由于轨迹空间是连续的，选择信息量最大的查询颇具挑战，而且通常是 NP 困难的（Ailon，2012）。因此，他们假设可以获得一个轨迹分布，该分布能够反映轨迹的可行性及其与目标策略的相关性，并描述了两种基于该分布选择轨迹查询的启发式方法。最后，Sadigh 等人（2017）描述了一种在连续状态与动作空间中进行主动偏好学习的方法。他们的工作核心在于能够合成动态可行的轨迹查询。Biyik 和 Sadigh（2018）将这一方法扩展到了批量查询设置。

其他问题中的比较反馈 学习理论中的其他问题也研究了比较反馈，例如组合函数（Balcan 等人，2016）。一个密切相关的问题是利用成对比较进行主动排序/学习（Jamieson 和 Nowak，2011；Kane 等人，2017；Saha 等人，2021；Yona 等人，2022）。这些研究通常考虑给定的有限点集。Kane 等人（2017）表明，即使这些点满足我们在本文中假设的线性结构约束，比较次数的下界仍与该集合的基数成线性关系。在我们的工作中，这些点是通过在相同的马尔可夫决策过程（MDP）下运行不同策略而生成的价值向量，因此具有特定结构。此外，我们允许比较不属于策略集的策略。因此，我们能够将查询复杂度降至策略数量的次线性。另一个使用比较/偏好反馈的相关问题是对决老虎机，其目标是通过成对反馈进行学习（Ailon 等人，2014；Zoghi 等人，2014）（综述参见 Bengs 等人（2021）；Sui 等人（2018）），或更一般地使用任意子集式反馈（Sui 等人，2017；Saha 和 Gopalan，2018，2019；Ren 等人，2018）。然而，与对决老虎机不同，我们考虑的是无噪声的比较反馈。

B 引理 1 的证明

引理 1。当 $\epsilon \leq \frac{v^*}{2k}$ 时，对于每个 $i \in [d]$，我们有 $\langle w^*, V^{\pi_1} \rangle \leq 2k$ 和 $|\widehat{\alpha}_i - \alpha_i| \leq \frac{4k^2\epsilon}{v^*}$。

证明。我们首先说明，根据我们的算法（算法 1 的第 1-6 行），返回的 π_1 满足

$$\langle w^*, V^{\pi_1} \rangle \geq \max_{i \in [k]} \left\langle w^*, V^{\pi^{\mathbf{e}_i}} \right\rangle - \epsilon \,.$$

这可以通过对 k 进行归纳法证明。在 $k = 2$ 的基础情形中，很容易看出返回的 π_1 满足上述不等式。假设上述不等式对任意 $k \leq n-1$ 都成立，我们将证明它对 $k = n$ 也成立。在对 $j = 2, \ldots, k-1$ 运行算法后（第 2 行），返回的策略 π_{e^*} 满足

$$\left\langle w^*, V^{\pi^{e^*}} \right\rangle \geq \max_{i \in [k-1]} \left\langle w^*, V^{\pi^{\mathbf{e}_i}} \right\rangle - \epsilon \,.$$

那么有两种情况:

- If $\left\langle w^*, V^{\pi^{e^*}} \right\rangle < \left\langle w^*, V^{\pi^{\mathbf{e}_k}} \right\rangle - \epsilon$, we will return $\pi_1 = \pi^{\mathbf{e}_k}$ and also, $\left\langle w^*, V^{\pi^{\mathbf{e}_k}} \right\rangle \geq \left\langle w^*, V^{\pi^{\mathbf{e}_i}} \right\rangle - \epsilon$ for all $i \in [k]$.
- If $\left\langle w^*, V^{\pi^{e^*}} \right\rangle \geq \left\langle w^*, V^{\pi^{\mathbf{e}_k}} \right\rangle - \epsilon$, then we will return π^{e^*} and it satisfies $\left\langle w^*, V^{\pi^{e^*}} \right\rangle \geq \max_{i \in [k]} \left\langle w^*, V^{\pi^{\mathbf{e}_i}} \right\rangle - \epsilon$.

由于当用户偏好向量为 $\mathbf{e}_i$ 时，$\pi^{\mathbf{e}_i}$ 是最优个性化策略，因此我们有

$$v^* = \left\langle w^*, V^{\pi^*} \right\rangle = \sum_{i=1}^{k} w_i^* \left\langle V^{\pi^*}, \mathbf{e}_i \right\rangle \leq \sum_{i=1}^{k} w_i^* \left\langle V^{\pi^{\mathbf{e}_i}}, \mathbf{e}_i \right\rangle \leq \left\langle w^*, \sum_{i=1}^{k} V^{\pi^{\mathbf{e}_i}} \right\rangle ,$$

其中，最后一个不等式成立，是因为 $V^{\pi^{\mathbf{e}_i}}$ 和 w^* 的各项均为非负数。因此，存在 $i \in [k]$，使得 $\left\langle w^*, V^{\pi^{\mathbf{e}_i}} \right\rangle \geq \frac{1}{k} \left\langle w^*, V^{\pi^*} \right\rangle = \frac{1}{k} v^*$。
于是我们得到

$$\langle w^*, V^{\pi_1} \rangle \geq \max_{i \in [k]} \left\langle w^*, V^{\pi^{\mathbf{e}_i}} \right\rangle - \epsilon \geq \frac{1}{k} v^* - \epsilon \geq \frac{1}{2k} v^* \,,$$

当 $\epsilon \leq \frac{v^*}{2k}$ 时。重新排列各项，可得 $\langle w^*, V^{\pi_1} \rangle \leq 2k$。

令 $C_\alpha = 2k$，可得 $|\widehat{\alpha}_i - \alpha_i| \langle w^*, V^{\pi_1} \rangle \leq C_\alpha \epsilon = 2k\epsilon$ 因此，$|\widehat{\alpha}_i - \alpha_i| \leq \frac{4k^2\epsilon}{v^*}$。 □

C 基准比率计算的伪代码

搜索 $\widehat{\alpha}_i$ 的伪代码如算法 3 所示。

算法 3 基比率计算 1：输入：$(V^{\pi_1},\ldots,V^{\pi_d})$ 和 C_α2：对于 $i=1,\ldots,d-1$ 执行 3：令 $l=0, h=2C_\alpha$ 和 $\widehat{\alpha}_i=C_\alpha$4：当 True 时执行 5：如果 $\widehat{\alpha}_i>1$，则 6：比较 π_1 和 $\frac{1}{\widehat{\alpha}_i}\pi_{i+1}+(1-\frac{1}{\widehat{\alpha}_i})\pi_0$；如果 $\pi_1\succ\frac{1}{\widehat{\alpha}_i}\pi_{i+1}+(1-\frac{1}{\widehat{\alpha}_i})\pi_0$，则 $h\leftarrow\widehat{\alpha}_i,\widehat{\alpha}_i\leftarrow\frac{l+h}{2}$；如果 $\pi_1\prec\frac{1}{\widehat{\alpha}_i}\pi_{i+1}+(1-\frac{1}{\widehat{\alpha}_i})\pi_0$，则 $l\leftarrow\widehat{\alpha}_i,\widehat{\alpha}_i\leftarrow\frac{l+h}{2}$7：否则 8：比较 π_{i+1} 和 $\widehat{\alpha}_i\pi_1+(1-\widehat{\alpha}_i)\pi_0$；如果 $\widehat{\alpha}_i\pi_1+(1-\widehat{\alpha}_i)\pi_0\succ\pi_{i+1}$，则 $h\leftarrow\widehat{\alpha}_i,\widehat{\alpha}_i\leftarrow\frac{l+h}{2}$；如果 $\widehat{\alpha}_i\pi_1+(1-\widehat{\alpha}_i)\pi_0\prec\pi_{i+1}$，则 $l\leftarrow\widehat{\alpha}_i,\widehat{\alpha}_i\leftarrow\frac{l+h}{2}$9：结束 if 10：如果回报“无法区分” 11：跳出 12：结束 if 13：结束 while 14：结束 for 15：输出：$(\widehat{\alpha}_1,\ldots,\widehat{\alpha}_{d-1})$

D 定理 1 的证明

$\widehat{\ }$ $\widehat{\ }$定理 1。考虑计算 A 的算法，以及 $\widehat{w}$ 到 $Ax=\mathbf{e}_1$ 的任意解，并输出策略 $\pi^{\widehat{w}}=\arg\max_{\pi\in\Pi}\langle\widehat{w},V^\pi\rangle$；对于偏好向量 $\widehat{w}$，则输出策略 $\pi^{\widehat{w}}$ 满足 $v^*-\left\langle w^*,V^{\pi^{\widehat{w}}}\right\rangle\leq\mathcal{O}\left(\left(\sqrt{k}+1\right)^{d+\frac{14}{31}\epsilon^{\frac{1}{3}}}\right)$

使用 $\mathcal{O}(k\log(k/\epsilon))$ 个比较查询。

证明。定理 1 可通过将 $\epsilon_\alpha=\frac{4k^2\epsilon}{v^*}$ 和 $C_\alpha=2k$ 设为引理 1 中所示的形式，并结合引理 2 和 3 的结果与三角不等式得到。

具体而言，对于任意策略 π，我们有

$$\begin{aligned}|\langle\widehat{w},V^\pi\rangle-\langle w',V^\pi\rangle|\leq&\left|\langle\widehat{w},V^\pi\rangle-\left\langle\widehat{w}^{(\delta)},V^\pi\right\rangle\right|+\left|\left\langle\widehat{w}^{(\delta)},V^\pi\right\rangle-\langle w',V^\pi\rangle\right|\\ \leq&\mathcal{O}((\sqrt{k}+1)^{d-d_\delta}C_\alpha(\frac{C_\alpha C_V^4 d_\delta^{\frac{3}{2}}\left\|w'\right\|_2^2\epsilon_\alpha}{\delta^2}+\sqrt{k}\delta\left\|w'\right\|_2))\\ \leq&\mathcal{O}((\sqrt{k}+1)^{d-d_\delta+3}\left\|w'\right\|_2(\frac{C_V^4k^4\left\|w'\right\|_2\epsilon}{v^*\delta^2}+\delta))\,.\end{aligned}\tag{7}$$

由于 $\|w'\|=\frac{\|w^*\|}{\langle w^*,V^{\pi_1}\rangle}$ 且 $\langle w^*,\ V^{\pi_1}\rangle\geq\frac{v^*}{2k}$根据引理 1，我们推得

$$\begin{aligned}&v^*-\left\langle w^*,V^{\pi^{\widehat{w}}}\right\rangle=\langle w^*,V^{\pi_1}\rangle\left(\left\langle w',V^{\pi^*}\right\rangle-\left\langle w',V^{\pi^{\widehat{w}}}\right\rangle\right)\\ \leq&\langle w^*,V^{\pi_1}\rangle\left(\left\langle\widehat{w},V^{\pi^*}\right\rangle-\left\langle\widehat{w},V^{\pi^{\widehat{w}}}\right\rangle+\mathcal{O}((\sqrt{k}+1)^{d-d_\delta+3}\left\|w'\right\|_2(\frac{C_V^4k^4\left\|w'\right\|_2\epsilon}{v^*\delta^2}+\delta))\right)\\ \leq&\mathcal{O}\left(\langle w^*,V^{\pi_1}\rangle(\sqrt{k}+1)^{d-d_\delta+3}\left\|w'\right\|_2(\frac{C_V^4k^4\left\|w'\right\|_2\epsilon}{v^*\delta^2}+\delta)\right)\\ =&\mathcal{O}\left((\sqrt{k}+1)^{d-d_\delta+3}\left\|w^*\right\|_2(\frac{C_V^4k^5\left\|w^*\right\|_2\epsilon}{v^{*2}\delta^2}+\delta)\right)\\ =&\mathcal{O}\left((\frac{C_V^2\left\|w^*\right\|_2^2}{v^*})^{\frac{2}{3}}(\sqrt{k}+1)^{d+\frac{16}{3}}\epsilon^{\frac{1}{3}}\right)\,.\end{aligned}$$

第一个不等式由下式得出

$$\left\langle w',V^{\pi^*}\right\rangle-\left\langle w',V^{\pi^{\widehat{w}}}\right\rangle=\left\langle w',V^{\pi^*}\right\rangle-\left\langle w',V^{\pi^{\widehat{w}}}\right\rangle$$

$$+\left(\left\langle \widehat{w}, V^{\pi^*}\right\rangle - \left\langle \widehat{w}, V^{\pi^*}\right\rangle\right) + \left(\left\langle \widehat{w}, V^{\pi^{\widehat{w}}}\right\rangle - \left\langle \widehat{w}, V^{\pi^{\widehat{w}}}\right\rangle\right),$$

$\widehat{}$并将（7）应用两次：一次用于 π^*，一次用于 π^w。最后一个不等式通过令 $\delta = \left(\frac{C_V^4 k^5 \|w^*\|_2 \epsilon}{v^{*2}}\right)^{\frac{1}{3}}$ 得出。 □

E 引理 2 的证明

引理 2. 如果对于所有 $i \in [d-1]$，$|\widehat{\alpha}_i - \alpha_i| \le \epsilon_\alpha$ 和 $\alpha_i \le C_\alpha$，则对于每个 $\delta \ge 4C_\alpha^{\frac{2}{3}} C_V d^{\frac{1}{3}} \epsilon_\alpha^{\frac{1}{3}}$，我们有 $\left|\left\langle \widehat{w}^{(\delta)}, V^\pi\right\rangle - \left\langle w', V^\pi\right\rangle\right| \le \mathcal{O}(\frac{C_\alpha C_V^4 d_\delta^{\frac{3}{2}} \|w'\|_2^2 \epsilon_\alpha}{\delta^2} + \sqrt{k}\delta \|w'\|_2)$ 对所有 π成立，其中 $w' = \frac{w^*}{\langle w^*, V^{\pi_1}\rangle}$。

为了证明引理 2，我们首先定义一个矩阵，$A^{(\text{full})}$。

Giv在算法 1 的输出 $(V^{\pi_1}, \ldots, V^{\pi_d})$ 中，我们有 $\text{rank}(\text{span}(\{V^{\pi_1}, \ldots, V^{\pi_{d_\delta}}\})) = d_\delta$。

令 $b_1, \ldots, b_{d-d_\delta}$ 为一组正交归一向量，它们与 $\text{span}(V^{\pi_1}, \ldots, V^{\pi_{d_\delta}})$ 正交，并且与 $V^{\pi_1}, \ldots, V^{\pi_{d_\delta}}$ 一起构成 $\text{span}(\{V^\pi | \pi \in \Pi\})$ 的一组基。

我们将 $A^{(\text{full})} \in \mathbb{R}^{d \times k}$ 定义为这样一个矩阵：用 $b_1, \ldots, b_{d-d_\delta}$ 替换 A 的最后 $d - d_\delta$ 行，即

$$A^{(\text{full})} = \begin{pmatrix} V^{\pi_1 \top} \\ (\alpha_1 V^{\pi_1} - V^{\pi_2})^\top \\ \vdots \\ (\alpha_{d_\delta - 1} V^{\pi_1} - V^{\pi_{d_\delta}})^\top \\ b_1^\top \\ \vdots \\ b_{d-d_\delta}^\top \end{pmatrix}.$$

观察 1. 我们有 $\text{span}(A^{(full)}) = \text{span}(\{V^\pi | \pi \in \Pi\})$，且 $\text{rank}(A^{(full)}) = d$。

引理 5. 对于所有满足 $A^{(full)} w = \mathbf{e}_1$，的 $w \in \mathbb{R}^k$，我们有 $|w \cdot V^\pi - w' \cdot V^\pi| \le \sqrt{k}\delta \|w'\|_2$，对所有 π 均成立。

接下来我们证明，存在一个满足 $A^{(\text{full})} w = \mathbf{e}_1$ 的 $w \in \mathbb{R}^k$，使得对于所有 $\pi \in \Pi$，$\left|\widehat{w}^{(\delta)} \cdot V^\pi - w \cdot V^\pi\right|$ 都很小。

引理 6. 若对于所有 $i \in [d-1]$，$|\widehat{\alpha}_i - \alpha_i| \le \epsilon_\alpha$ 和 $\alpha_i \le C_\alpha$，则对于每个 $\delta \ge 4C_\alpha^{\frac{2}{3}} C_V d^{\frac{1}{3}} \epsilon_\alpha^{\frac{1}{3}}$，都存在一个满足 $A^{(full)} w = \mathbf{e}_1$ s.t. $\left|\widehat{w}^{(\delta)} \cdot V^\pi - w \cdot V^\pi\right| \le \mathcal{O}(\frac{C_\alpha C_V^4 d_\delta^{\frac{3}{2}} \|w'\|_2^2 \epsilon_\alpha}{\delta^2})$（对所有 π）的 $w \in \mathbb{R}^k$。

现在我们利用上述两个引理推导引理 2。

引理 2 的证明。令 w 如引理 6 中所定义。则对于任意策略 π，我们有

$$\left|\widehat{w}^{(\delta)} \cdot V^\pi - w' \cdot V^\pi\right| \le \left|\widehat{w}^{(\delta)} \cdot V^\pi - w \cdot V^\pi\right| + |w \cdot V^\pi - w' \cdot V^\pi|$$

$$\le \mathcal{O}(\frac{C_\alpha C_V^4 d_\delta^{\frac{3}{2}} \|w'\|_2^2 \epsilon_\alpha}{\delta^2} + \sqrt{k}\delta \|w'\|_2),$$

通过应用引理 5 和引理 6。 □

F 引理 5 和引理 6 的证明

引理 5. 对于所有满足 $A^{(full)} w = \mathbf{e}_1$，的 $w \in \mathbb{R}^k$，我们有 $|w \cdot V^\pi - w' \cdot V^\pi| \le \sqrt{k}\delta \|w'\|_2$，对所有 π 均成立。

引理 5 的证明。由于 $\text{span}(A^{(\text{full})}) = \text{span}(\{V^\pi | \pi \in \Pi\})$，对于每个策略 π，价值向量都可以表示为 $A^{(\text{full})}$ 的行向量的线性组合，即存在$a = (a_1, .., a_d) \in \mathbb{R}^d$，使得

$$V^\pi = \sum_{i=1}^{d} a_i A_i^{(\text{full})} = A^{(\text{full})\top} a \,. \tag{8}$$

现在，对于任意属于 $\text{span}(b_1, \ldots, b_{d-d_\delta})$ 的单位向量 $\xi \in$，我们有 $\langle V^\pi, \xi \rangle \le \sqrt{k}\delta$。

原因在于，在每一轮 $d_\delta + 1$ 中，我们选取 $\text{span}(V^{\pi_1}, \ldots, V^{\pi_{d_\delta}})$ 的一组标准正交基 $\rho_1, .., \rho_{k-d_\delta}{}^\perp$（算法 1 第 8 行），并按照第 9 行所述，选取 $u_{d_\delta+1}$，使得其中存在一个具有最大分量的策略。因此，对于所有 $j \in [k - d_\delta]$，都有 $|\langle \rho_j, V^\pi \rangle| \le \delta$。

根据柯西-施瓦茨不等式可得 $\langle \xi, V^\pi \rangle = \sum_{j=1}^{k-d_\delta} \langle \xi, \rho_j \rangle \langle \rho_j, V^\pi \rangle \le \sqrt{k}\delta$。

结合以下观察：$b_1, \ldots, b_{d-d_\delta}$ 两两正交，且它们各自都与 $\text{span}(V^{\pi_1}, \ldots, V^{\pi_{d_\delta}})$ 正交，我们有

$$\sum_{i=d_\delta+1}^{d} a_i^2 = \left| \left\langle V^\pi, \sum_{i=d_\delta+1}^{d} a_i b_{i-d_\delta} \right\rangle \right| \le \sqrt{\sum_{i=d_\delta+1}^{d} a_i^2}\sqrt{k}\delta \,,$$

这意味着

$$\sqrt{\sum_{i=d_\delta+1}^{d} a_i^2} \le \sqrt{k}\delta \,. \tag{9}$$

由于 w' 满足 $Aw' = \mathsf{e}_1,$，我们有

$$A^{(\text{full})} w' = (1, 0, \ldots, 0, \langle b_1, w' \rangle, \ldots, \langle b_{d-d_\delta}, w' \rangle) \,.$$

对于任意满足 $A^{(\text{full})} w = \mathsf{e}_1$，的 $w \in \mathbb{R}^k$，考虑 $\widetilde{w} = w + \sum_{i=1}^{d-d_\delta} \langle b_i, w' \rangle b_i$。于是我们有 $A^{(\text{full})} \widetilde{w} = A^{(\text{full})} w'$。

因此，将（8）应用两次，可得

$$\widetilde{w} \cdot V^\pi = \widetilde{w}^\top A^{(\text{full})\top} a = w'^\top A^{(\text{full})\top} a = w' \cdot V^\pi \,.$$

因此，

$$|w \cdot V^\pi - w' \cdot V^\pi| = |w \cdot V^\pi - \widetilde{w} \cdot V^\pi| \overset{(a)}{=} \left| \sum_{i=1}^{d} a_i (w - \widetilde{w}) \cdot A_i^{(\text{full})} \right|$$

$$= \left| \sum_{i=d_\delta+1}^{d} a_i \langle b_{i-d_\delta}, w' \rangle \right| \overset{(b)}{\le} \sqrt{\sum_{i=d_\delta+1}^{d} a_i^2} \, \|w'\|_2 \overset{(c)}{\le} \sqrt{k}\delta \, \|w'\|_2 \,,$$

其中，等式（a）由（8）得出，不等式（b）由柯西-施瓦茨不等式得出，不等式（c）由应用（9）得出。

□

引理 6. 若对所有 $i \in [d-1]$，$|\widehat{\alpha}_i - \alpha_i| \le \epsilon_\alpha$ 和 $\alpha_i \le C_\alpha$，则对于每个 $\delta \ge 4C_\alpha^{\frac{3}{2}} C_V^{\frac{1}{2}} d^{\frac{1}{3}} \epsilon_\alpha^{\frac{1}{3}}$，都存在一个满足 $A^{(full)} w = \mathsf{e}_1$ 的 $w \in \mathbb{R}^k$，使得对所有 π，$\left| \widehat{w}^{(\delta)} \cdot V^\pi - w \cdot V^\pi \right| \le \mathcal{O}\left(\frac{C_\alpha C_V^4 d_\delta^2 \|w'\|_2^2 \epsilon_\alpha}{\delta^2} \right)$。

在证明引理 6 之前，我们先介绍一些记号和一个断言。

- 对于任意 x、$y \in \mathbb{R}^k$，令 $\theta(x, y)$ 表示 x 与 y之间的角度。
- 对于任意子空间 $U \subset \mathbb{R}^k$，令 $\theta(x, U) := \min_{y \in U} \theta(x, y)$。
- 对于任意两个子空间 $U, U' \subset \mathbb{R}^k$，我们将 $\theta(U, U')$ 定义为 $\theta(U, U') = \max_{x \in U} \min_{y \in U'} \theta(x, y)$。

•对于任意矩阵 M，令 M_i表示 M 的第 i 个行向量 $M_{i:\ j}$表示由第 $i, i+1, \ldots, j$ 行组成的 M 的子矩阵，而 $M_{i:}$ 表示由所有行$j \geq i$组成的子矩阵。

•令 span(M) 表示 M 的各行所张成的空间。

回想一下，$\widehat{A} \in \mathbb{R}^{d\times k}$ 定义为

$$\widehat{A} = \begin{pmatrix} V^{\pi_1\top} \\ (\widehat{\alpha}_1 V^{\pi_1} - V^{\pi_2})^\top \\ \vdots \\ (\widehat{\alpha}_{d-1} V^{\pi_1} - V^{\pi_d})^\top \end{pmatrix},$$

它是由 α_i 的真实值所定义的矩阵 $A \in \mathbb{R}^{d\times k}$ 的近似，即

$$A = \begin{pmatrix} V^{\pi_1\top} \\ (\alpha_1 V^{\pi_1} - V^{\pi_2})^\top \\ \vdots \\ (\alpha_{d-1} V^{\pi_1} - V^{\pi_d})^\top \end{pmatrix}.$$

我们分别用 $\widehat{A}^{(\delta)} = \widehat{A}_{1:d_\delta}$、$A^{(\delta)} = A_{1:d_\delta} \in \mathbb{R}^{d_\delta\times k}$ 表示由 A 和 A 的前 d_δ 行 $\widehat{}$ 构成的子矩阵。

主张1. 若对所有 $i \in [d-1]$，$|\widehat{\alpha}_i - \alpha_i| \leq \epsilon_\alpha$ 和 $\alpha_i \leq C_\alpha$，则对于每个 $\delta \geq 4C_\alpha^{\frac{2}{3}} C_V d^{\frac{1}{3}} \epsilon_\alpha^{\frac{1}{3}}$，我们有

$$\theta(\mathrm{span}(A^{(\delta)}_{2:}), \mathrm{span}(\widehat{A}^{(\delta)}_{2:})) \leq \eta_{\epsilon_\alpha,\delta}\,, \tag{10}$$

and

$$\theta(\mathrm{span}(\widehat{A}^{(\delta)}_{2:}), \mathrm{span}(A^{(\delta)}_{2:})) \leq \eta_{\epsilon_\alpha,\delta}\,, \tag{11}$$

其中 $\eta_{\epsilon_\alpha,\ \delta} = \frac{4C_\alpha C_V^2 d_\delta \epsilon_\alpha}{\delta^2}$。

为了证明上述断言，我们使用 Balcan 等人（2015）提出的以下引理。

$\widehat{}$ $\widehat{}$ $\widehat{}$引理 7（Balcan 等人（2015）的引理 3）。设 $U_l = \mathrm{span}(\xi_1, \ldots, \xi_l)$ 且 $U_l = \mathrm{span}(\xi_1, \ldots, \xi_l)$。设 ϵ_{acc}、$\gamma_{new} \geq 0$ 和 $\epsilon_{acc} \leq \gamma_{new}^2/(10l)$，并假设对于 $i = 2, \ldots, l$，$\theta(\widehat{\xi}_i, \widehat{U}_{i-1}) \geq \gamma_{new}$，且对于 $i = 1, \ldots, l$，$\widehat{}$ 满足 $\theta(\xi_i, \xi_i) \leq \epsilon_{acc}$。

那么，

$$\theta(U_l, \widehat{U}_l) \leq 2l \frac{\epsilon_{acc}}{\gamma_{new}}\,.$$

主张1的证明。对于所有 $2 \leq i \leq d_\delta$，我们有

$$\begin{aligned}\theta(\widehat{A}^{(\delta)}_i, \mathrm{span}(\widehat{A}^{(\delta)}_{2:i-1})) \geq& \theta(\widehat{A}^{(\delta)}_i, \mathrm{span}(\widehat{A}^{(\delta)}_{1:i-1})) \geq \sin(\theta(\widehat{A}^{(\delta)}_i, \mathrm{span}(\widehat{A}^{(\delta)}_{1:i-1}))) \\ \overset{(a)}{\geq}& \frac{\left|\widehat{A}^{(\delta)}_i \cdot u_i\right|}{\left\|\widehat{A}^{(\delta)}_i\right\|_2} \overset{(b)}{\geq} \frac{\delta}{\left\|\widehat{A}^{(\delta)}_i\right\|_2} = \frac{\delta}{\|\widehat{\alpha}_{i-1} V^{\pi_1} - V^{\pi_i}\|_2} \geq \frac{\delta}{(C_\alpha+1)C_V}\,,\end{aligned}$$

其中，根据算法 1 第 8 行，Ineq (a) 成立，因为 u_i 与 span($\widehat{A}^{(\delta)}_{1\ :\ i-1}$) 正交；Ineq (b) 成立是因为 $\left|\widehat{A}^{(\delta)}_i \cdot u_i\right| = |V^{\pi_i} \cdot u_i| \geq \delta$。最后一个不等式成立是因为

$\|\widehat{\alpha}_{i-1} V^{\pi_1} - V^{\pi_i}\|_2 \leq \widehat{\alpha}_{i-1} \|V^{\pi_1}\|_2 + \|V_i\|_2 \leq (C_\alpha + 1) C_V$。

同样，我们也有

$$\theta(A^{(\delta)}_i, \mathrm{span}(A^{(\delta)}_{2:i-1})) \geq \frac{\delta}{(C_\alpha+1)C_V}\,.$$

We 继续将V^{π_i} 分解为沿V^{π_1} 方向和垂直于V^{π_1}的方向。

为便于说明，我们记为 $v_i^{\|} := V^{\pi_i} \cdot \frac{V^{\pi_1}}{\|V^{\pi_1}\|_2}, V_i^{\|} := v_i^{\|} \frac{V^{\pi_1}}{\|V^{\pi_1}\|_2}, V_i^{\perp} := V^{\pi_i} - V_i^{\|}$ 和 $v_i^{\perp} := \left\|V_i^{\perp}\right\|_2$.

那么我们有

$\theta(A_i^{(\delta)}, \widehat{A}_i^{(\delta)}) = \theta(\alpha_{i-1}V^{\pi_1} - V^{\pi_i}, \widehat{\alpha}_{i-1}V^{\pi_1} - V^{\pi_i}) = \theta(\alpha_{i-1}V^{\pi_1} - V_i^{\|} - V_i^{\perp}, \widehat{\alpha}_{i-1}V^{\pi_1} - V_i^{\|} - V_i^{\perp})$。

如果$(\widehat{\alpha}_{i-1}V^{\pi_1} - V_i^{\|}) \cdot (\alpha_{i-1}V^{\pi_1} - V_i^{\|}) \geq 0$，也就是说，$\widehat{\alpha}_{i-1}V^{\pi_1} - V_i^{\|}$ 和 $\alpha_{i-1}V^{\pi_1} - V_i^{\|}$ 方向相同，那么

$$\begin{aligned}
\theta(A_i^{(\delta)}, \widehat{A}_i^{(\delta)}) &= \left| \arctan \frac{\left\| \widehat{\alpha}_{i-1}V^{\pi_1} - V_i^{\|} \right\|_2}{v_i^{\perp}} - \arctan \frac{\left\| \alpha_{i-1}V^{\pi_1} - V_i^{\|} \right\|_2}{v_i^{\perp}} \right| \\
&\leq \left| \frac{\left\| \widehat{\alpha}_{i-1}V^{\pi_1} - V_i^{\|} \right\|_2}{v_i^{\perp}} - \frac{\left\| \alpha_{i-1}V^{\pi_1} - V_i^{\|} \right\|_2}{v_i^{\perp}} \right| \qquad (12) \\
&= \frac{|\widehat{\alpha}_{i-1} - \alpha_{i-1}| \left\| V^{\pi_1} \right\|_2}{v_i^{\perp}} \\
&\leq \frac{\epsilon_\alpha C_V}{\delta}, \qquad (13)
\end{aligned}$$

其中，不等式（12）由反正切函数的导数至多为 1 这一事实推出，即 $\partial \text{反正切函数} / \partial x \Big|_{x=} = \lim_{a \to x} \frac{\text{反正切函数}_{a=} - x}{a-x} = \frac{1}{1+x^2} \leq 1$反正切函数。

不等式（13）成立，因为 $v_i^{\perp} \geq |\langle V^{\pi_i}, u_i \rangle| \geq \delta$。

若 $(\widehat{\alpha}_{i-1}V^{\pi_1} - V_i^{\|}) \cdot (\alpha_{i-1}V^{\pi_1} - V_i^{\|}) < 0$，即 $\widehat{\alpha}_{i-1}V^{\pi_1} - V_i^{\|}$ 和 $\alpha_{i-1}V^{\pi_1} - V_i^{\|}$ 方向相反，则我们有 $\left\| \widehat{\alpha}_{i-1}V^{\pi_1} - V_i^{\|} \right\|_2 + \left\| \alpha_{i-1}V^{\pi_1} - V_i^{\|} \right\|_2 = \|(\widehat{\alpha}_{i-1} - \alpha_{i-1})V^{\pi_1}\|_2 \leq \epsilon_\alpha \left\| V^{\pi_1} \right\|_2$。

同理，我们有

$$\begin{aligned}
\theta(\widehat{A}_i^{(\delta)}, A_i^{(\delta)}) &= \left| \arctan \frac{\left\| \widehat{\alpha}_{i-1}V^{\pi_1} - V_i^{\|} \right\|_2}{v_i^{\perp}} + \arctan \frac{\left\| \alpha_{i-1}V^{\pi_1} - V_i^{\|} \right\|_2}{v_i^{\perp}} \right| \\
&\leq \left| \frac{\left\| \widehat{\alpha}_{i-1}V^{\pi_1} - V_i^{\|} \right\|_2}{v_i^{\perp}} + \frac{\left\| \alpha_{i-1}V^{\pi_1} - V_i^{\|} \right\|_2}{v_i^{\perp}} \right| \\
&\leq \frac{\epsilon_\alpha \left\| V^{\pi_1} \right\|_2}{\left| v_i^{\perp} \right|} \\
&\leq \frac{\epsilon_\alpha C_V}{\delta}.
\end{aligned}$$

将引理 7 应用于 $\epsilon_{\text{acc}} = \frac{\epsilon_\alpha C_V}{\delta}$, $\gamma_{\text{new}} = \delta / (C_{\alpha+1} C_V)$, $(\xi_i, \widehat{\xi}_i) = (A_{i+1}, \widehat{A}_{i+1})$（以及 $(\xi_i, \widehat{\xi}_i) = (\widehat{A}_{i+1}, A_{i+1})$）后可知，当 $\delta \geq 10^{\frac{1}{3}}(C_\alpha + 1)^{\frac{2}{3}} C_V d_\delta^{\frac{1}{3}} \epsilon_\alpha^{\frac{1}{3}}$ 时，

$$\theta(\text{span}(A_{2:}^{(\delta)}), \text{span}(\widehat{A}_{2:}^{(\delta)})) \leq \frac{2d_\delta (C_\alpha + 1) C_V^2 \epsilon_\alpha}{\delta^2} = \eta_{\epsilon_\alpha, \delta},$$

and

$$\theta(\text{span}(\widehat{A}_{2:}^{(\delta)}), \text{span}(A_{2:}^{(\delta)})) \leq \eta_{\epsilon_\alpha, \delta}.$$

由于 $C_\alpha \geq 1$，这就完成了主张1的证明。 □

引理 6 的证明。回顾一下，$\widehat{w}^{(\delta)} = \text{argmin}_{\widehat{A}^{(\delta)} x = \mathbf{e}_1} \|x\|_2$ 是 $\widehat{A}^{(\delta)} x = \mathbf{e}_1$ 的最小范数解。

因此，对于所有 $i \in [d - d_\delta]$，$\langle \widehat{w}^{(\delta)}, b_i \rangle = 0$。

令 λ_1、...、$\lambda_{d_\delta - 1}$ 为张成空间 $A_{2:}^{(\delta)}$ 的任意标准正交基。

我们通过移除 $\widehat{w}^{(\delta)}$ 在张成空间 $A_2^{(\delta)}:$ 中的分量并重新缩放，构造满足 $A^{(\text{full})}w=\mathbf{e}_1$ 的向量 w。

形式化地,

$$w:=\frac{\widehat{w}^{(\delta)}-\sum_{i=1}^{d_\delta-1}\left\langle\widehat{w}^{(\delta)},\lambda_i\right\rangle\lambda_i}{1-V^{\pi_1}\cdot(\sum_{i=1}^{d_\delta-1}\left\langle\widehat{w}^{(\delta)},\lambda_i\right\rangle\lambda_i)}. \tag{14}$$

可以直接验证，对于 $i=2,\ A_1\cdot w=V^{\pi_1}\cdot w=1$ 和 $A_i\cdot w=0$，....., d_δ。因此，$A^{(\delta)}w=\mathbf{e}_1$。

结合这样一个事实：对于所有 $i\in[d-d_\delta]$，$\widehat{w}^{(\delta)}$ 在 b_i 中分量为零，我们得到 $A^{(\text{full})}w=\mathbf{e}_1$。根据主张1，我们有

$$\theta(\text{span}(A_{2:}^{(\delta)}),\text{span}(\widehat{A}_{2:}^{(\delta)}))\le\eta_{\epsilon_\alpha,\delta}.$$

因此，存在单位向量 $\widetilde{\lambda}_1$，.....，$\widetilde{\lambda}_{d_\delta-1}\in$ 张成空间$(\widehat{A}_2^{(\delta)}:)$，使得 $\theta(\lambda_i,\widetilde{\lambda}_i)\le\eta_{\epsilon_\alpha,\ \delta}$。由于 $\widehat{A}^{(\delta)}\widehat{w}^{(\delta)}=\mathbf{e}_1$,，我们有：对于所有 $i=1$，.....，$d_\delta-1$，$\widehat{w}^{(\delta)}\cdot\widetilde{\lambda}_i=0$，因此，

$$\left|\widehat{w}^{(\delta)}\cdot\lambda_i\right|=\left|\widehat{w}^{(\delta)}\cdot(\lambda_i-\widetilde{\lambda}_i)\right|\le\left\|\widehat{w}^{(\delta)}\right\|_2\eta_{\epsilon_\alpha,\delta}.$$

这意味着对于任意策略 π,

$$\left|V^\pi\cdot\sum_{i=1}^{d_\delta-1}(\widehat{w}^{(\delta)}\cdot\lambda_i)\lambda_i\right|\le\|V^\pi\|_2\sqrt{d_\delta}\left\|\widehat{w}^{(\delta)}\right\|_2\eta_{\epsilon_\alpha,\delta}\le C_V\sqrt{d_\delta}\left\|\widehat{w}^{(\delta)}\right\|_2\eta_{\epsilon_\alpha,\delta}.$$

记为 $\gamma=V^{\pi_1}\cdot(\sum_{i=1}^{d_\delta-1}\left\langle\widehat{w}^{(\delta)},\lambda_i\right\rangle\lambda_i)$，其不大于 $C_V\sqrt{d_\delta}\left\|\widehat{w}^{(\delta)}\right\|_2\eta_{\epsilon_\alpha,\ \delta}$。我们有

$$\begin{aligned}\left|\widehat{w}^{(\delta)}\cdot V^\pi-w\cdot V^\pi\right|&\le\left|\widehat{w}^{(\delta)}\cdot V^\pi-\frac{1}{1-\gamma}\widehat{w}^{(\delta)}\cdot V^\pi\right|+\left|\frac{1}{1-\gamma}\widehat{w}^{(\delta)}\cdot V^\pi-w\cdot V^\pi\right|\\&\le\frac{\gamma\left\|\widehat{w}^{(\delta)}\right\|_2C_V}{1-\gamma}+\frac{1}{1-\gamma}\left|\sum_{i=1}^{d_\delta-1}(\widehat{w}^{(\delta)}\cdot\lambda_i)\lambda_i\cdot V^\pi\right|\\&\le\frac{\gamma\left\|\widehat{w}^{(\delta)}\right\|_2C_V}{1-\gamma}+\frac{C_V\sqrt{d_\delta}\left\|\widehat{w}^{(\delta)}\right\|_2\eta_{\epsilon_\alpha,\delta}}{1-\gamma}\\&=2(C_V\left\|\widehat{w}^{(\delta)}\right\|_2+1)C_V\sqrt{d_\delta}\left\|\widehat{w}^{(\delta)}\right\|_2\eta_{\epsilon_\alpha,\delta},\end{aligned} \tag{15}$$

其中，当 $C_V\sqrt{d_\delta}\left\|\widehat{w}^{(\delta)}\right\|_2\eta_{\epsilon_\alpha,\ \delta}\le\frac{1}{2}$ 时，最后一个等式成立。

现在我们证明，对于某个常数 C，有 $\left\|\widehat{w}^{(\delta)}\right\|_2\le C\left\|w'\right\|_2$。

由于 $\widehat{w}^{(\delta)}$ 是 $\widehat{A}^{(\delta)}x=\mathbf{e}_1$，的最小范数解，我们将按照与式 (14) 中构造类似的方式，构造 $\widehat{A}^{(\delta)}x=\mathbf{e}_1$，的另一个解，记为 $\widehat{w}_0$，并证明$\|\widehat{w}_0\|_2\le C\left\|w'\right\|$。

令 $\xi_1,\ldots,\xi_{d_\delta-1}$为张成空间$(\widehat{A}_2^{(\delta)}:)$的任意标准正交基。

我们通过从 w' 中移除其在张成空间$(\widehat{A}_2^{(\delta)}:)$中的分量并重新缩放，构造一个 $\widehat{w}_0$，使得 $\widehat{A}^{(\delta)}\widehat{w}_0=\mathbf{e}_1$。具体而言，令

$$\widehat{w}_0=\frac{w'-\sum_{i=1}^{d_\delta-1}\left\langle w',\xi_i\right\rangle\xi_i}{1-\left\langle V^{\pi_1},(\sum_{i=1}^{d_\delta-1}\left\langle w',\xi_i\right\rangle\xi_i)\right\rangle}. \tag{16}$$

由于 $A^{(\delta)}w'=\mathbf{e}_1$,，可直接推出 $\left\langle\widehat{A}_1,\widehat{w}_0\right\rangle=\langle V^{\pi_1},\widehat{w}_0\rangle=1$，并且对于$i=2$，有 $\left\langle\widehat{A}_i,\widehat{w}_0\right\rangle=0$，.....，$d_\delta$，即 $\widehat{A}^{(\delta)}\widehat{w}_0=\mathbf{e}_1$。

由于主张1意味着 $\theta(\text{span}(\widehat{A}_2^{(\delta)}:)\ ,\text{span}(A_2^{(\delta)}:))\leq\ \eta_{\epsilon_\alpha,\ \delta}$，因此存在单位向量$\widetilde{\xi}_1$，.....，$\widetilde{\xi}_{d_\delta-1}\in\text{span}(A_2^{(\delta)}:)$，使得 $\theta(\xi_i,\widetilde{\xi}_i)\leq\eta_{\epsilon_\alpha,\ \delta}$。

由于 w' 在张成空间$(A_2^{(\delta)}:)$中没有分量，因此 w' 在张成空间$(\widehat{A}_2^{(\delta)}:)$中应具有较小的分量。特别地，

$$|\langle w',\xi_i\rangle|=\left|\left\langle w',\xi_i-\widetilde{\xi}_i\right\rangle\right|\leq\|w'\|_2\,\eta_{\epsilon_\alpha,\delta}\,,$$

这意味着

$$\left\|\sum_{i=1}^{d_\delta-1}\langle w',\xi_i\rangle\,\xi_i\right\|_2\leq\sqrt{d_\delta}\,\|w'\|_2\,\eta_{\epsilon_\alpha,\delta}\,.$$

因此

$$\left|\left\langle V^{\pi_1},(\sum_{i=1}^{d_\delta-1}\langle w',\xi_i\rangle\,\xi_i)\right\rangle\right|\leq C_V\sqrt{d_\delta}\,\|w'\|_2\,\eta_{\epsilon_\alpha,\delta}\,.$$

因此，当 $C_V\sqrt{d_\delta}\,\|w'\|_2\,\eta_{\epsilon_\alpha,\ \delta}\ \leq\ \frac{1}{3}$ 时，$\|\widehat{w}_0\|_2\ \leq\ \frac{3}{2}\,\|w'\|_2$，当 ϵ_α 足够小时这一点成立。

根据引理 1，$\epsilon_\alpha\leq\frac{4k^2\epsilon}{v^*}$，当 $\epsilon\to0$ 时，$\epsilon_\alpha\to0$。

因此，我们有 $\left\|\widehat{w}^{(\delta)}\right\|_2\leq\|\widehat{w}_0\|_2\leq\frac{3}{2}\,\|w'\|_2$，以及 $C_V\sqrt{d_\delta}\left\|\widehat{w}^{(\delta)}\right\|_2\eta_{\epsilon_\alpha,\ \delta}\leq\frac{1}{2}$。结合公式（15），可得

$$\left|\widehat{w}^{(\delta)}\cdot V^\pi-w\cdot V^\pi\right|=\mathcal{O}\left((C_V\,\|w'\|_2+1)C_V\sqrt{d_\delta}\,\|w'\|_2\,\eta_{\epsilon_\alpha,\delta}\right)\,.$$

Since $C_V\,\|w'\|_2\geq|\langle V^{\pi_1},w'\rangle|=1$，将 $\eta_{\epsilon_\alpha,\delta}=\frac{4C_\alpha C_V^2 d_\delta\epsilon_\alpha}{\delta^2}$ 代入上述方程后，我们有

$$\left|\widehat{w}^{(\delta)}\cdot V^\pi-w\cdot V^\pi\right|=\mathcal{O}\left(\frac{C_\alpha C_V^4 d_\delta^{\frac{3}{2}}\,\|w'\|_2^2\,\epsilon_\alpha}{\delta^2}\right)\,,$$

至此，证明完成。 □

G 引理 3 的证明

引理 3. 若对于所有 $i\in[d-1]$，$|\widehat{\alpha}_i-\alpha_i|\leq\epsilon_\alpha$ 和 $\alpha_i\leq C_\alpha$ 成立，则对于每个策略 π 和每个 $\delta\geq\frac{4}{3}C_\alpha^{\frac{2}{3}}C_V d^{\frac{1}{3}}\epsilon_\alpha^{\frac{1}{3}}$，我们有 $\left|\widehat{w}\cdot V^\pi-\widehat{w}^{(\delta)}\cdot V^\pi\right|\leq\mathcal{O}((\sqrt{k}+1)^{d-d_\delta}C_\alpha\epsilon^{(\delta)})$，其中$\epsilon^{(\delta)}=\frac{C_\alpha C_V^4 d_\delta^{\frac{3}{2}}\|w'\|_2^2\epsilon_\alpha}{\delta^2}+\sqrt{k}\delta\,\|w'\|_2$ 是引理 2 中的上界。

证明。给定算法 1 的输出 $(V^{\pi_1},..,V^{\pi_d})$，我们有 $\text{rank}(\text{span}(\{V^{\pi_1},\ldots,V^{\pi_{d_\delta}}\}))=d_\delta$。

对于 $i=d_\delta+1,\ldots,d$，令 ψ_i 为 V^{π_i} 在 $\text{span}(V^{\pi_1},\ldots,V^{\pi_{i-1}})^\perp$ 上的投影的归一化向量，其中 $\|\psi_i\|_2=1$。

那么我们有 $\text{span}(V^{\pi_1},\ldots,V^{\pi_{i-1}},\psi_i)=\text{span}(V^{\pi_1},\ldots,V^{\pi_i})$，并且 $\{\psi_i|i=d_\delta+1,\ldots,d\}$ 是正交归一的。

对于每个策略 π，其价值向量都可以表示为 $\widehat{A}_1,\ldots,\widehat{A}_{d_\delta},\psi_{d_\delta+1},\ldots,\psi_d$ 的线性组合，即存在唯一的 $a=(a_1,\ldots,a_d)\in\mathbb{R}^d$，使得 $V^\pi=\sum_{i=1}^{d_\delta}a_i\widehat{A}_i+\sum_{i=d_\delta+1}^{d}a_i\psi_i$。

由于对于所有 $j\neq i$，ψ_i 都与 ψ_j 正交，且 ψ_i 与 $\text{span}(A_1,\ldots,A_{d_\delta})$ 正交，因此对于 $i\geq d_\delta+1$，我们有$a_i=\langle V^\pi,\psi_i\rangle$。

这意味着

$$\left|\langle \widehat{w}, V^{\pi}\rangle - \left\langle \widehat{w}^{(\delta)}, V^{\pi}\right\rangle\right| \leq \underbrace{\left|\sum_{i=1}^{d_\delta} a_i \left(\left\langle \widehat{w}, \widehat{A}_i\right\rangle - \left\langle w^{(\delta)}, \widehat{A}_i\right\rangle\right)\right|}_{(a)} + \underbrace{\left|\sum_{i=d_\delta+1}^{d} a_i \left\langle \widehat{w}, \psi_i\right\rangle\right|}_{(b)}$$
$$+ \underbrace{\left|\sum_{i=d_\delta+1}^{d} a_i \left\langle \widehat{w}^{(\delta)}, \psi_i\right\rangle\right|}_{(c)}.$$

由于 $\widehat{A}^{(\delta)}\widehat{w} = \widehat{A}^{(\delta)}\widehat{w}^{(\delta)} = \mathbf{e}_1,$ ，我们有项 $(a) = 0$。

接下来，我们对项 (c) 进行界定。

注意，向量 $\{\psi_i | i = d_\delta + 1, \ldots, d\}$ 与张成空间 span$(V^{\pi_1}, \ldots, V^{\pi_{d_\delta}})$ 正交，并且它们与 $V^{\pi_1}, \ldots,$ $V^{\pi_{d_\delta}}$ 一起构成 span$(\{V^{\pi} | \pi \in \Pi\})$ 的一组基。

因此，在引理 2 的证明中，我们可以令 b_i 为 ψ_{i+d_δ}。

此外，$\{b_i | i \in [d - d_\delta]\}$ 的所有性质同样适用于 $\{\psi_i | i = d_\delta + 1, \ldots, d\}$。因此，类似于式 (9)，

$$\sqrt{\sum_{i=d_\delta+1}^{d} a_i^2} \leq \sqrt{k}\delta\,.$$

因此，我们可以将项 (c) 限定为

$$(c) \leq \sqrt{k}\delta \left\|\widehat{w}^{(\delta)}\right\|_2 = \frac{3}{2}\sqrt{k}\delta \left\|w'\right\|_2$$

因为 $\left\|\widehat{w}^{(\delta)}\right\|_2 \leq \frac{3}{2}\left\|w'\right\|_2$ 当 $C_V\sqrt{d_\delta}\left\|w'\right\|_2 \eta_{\epsilon_\alpha},\ \delta \leq \frac{1}{3}$ 如引理 6 的证明中所述。

现在只需对项 (b) 进行界定。

我们无法像界定项 (c) 那样界定项 (b)，因为 $\|\widehat{w}\|_2$ 不一定有 $\|w'\|_2$ 作为上界。

对于 $i = d_\delta + 1,$ ....., $d,$ 我们定义

$$\epsilon_i := \left|\left\langle \psi_i, \widehat{A}_i\right\rangle\right|.$$

对于任意 $i,\ j = d_\delta + 1,$ ....., $d,$ ψ_i 垂直于 V^{π_1}，因此 $\left|\left\langle \psi_i,\ \widehat{A}_j\right\rangle\right| = |\langle \psi_i,$ $\widehat{\alpha}_{j-1}V^{\pi_1} - V^{\pi_j}\rangle| = |\langle \psi_i,\ V^{\pi_j}\rangle|$。特别地，我们有

$$\epsilon_i = \left|\left\langle \psi_i, \widehat{A}_i\right\rangle\right| = |\langle \psi_i, V^{\pi_i}\rangle|\,.$$

令 $\widehat{A}_i^{\parallel} := \widehat{A}_i - \sum_{j=d_\delta+1}^{d}\left\langle \widehat{A}_i, \psi_j\right\rangle \psi_j$ 表示 $\widehat{A}_i$ 在 span$(\widehat{A}_1, \ldots, \widehat{A}_{d_\delta})$ 上的投影。

由于对于 $j > i,$ $\widehat{A}_i$ 在方向 ψ_j 上没有分量，因此我们有 $\widehat{A}_i^{\parallel} = \widehat{A}_i - \sum_{j=d_\delta+1}^{i}\left\langle \widehat{A}_i, \psi_j\right\rangle \psi_j$。

于是，我们有

$$0 = \left\langle \widehat{w}, \widehat{A}_i\right\rangle = \widehat{w}\cdot\widehat{A}_i^{\parallel} + \widehat{w}\cdot\sum_{j=d_\delta+1}^{i}\left\langle \widehat{A}_i, \psi_j\right\rangle\psi_j = \widehat{w}\cdot\widehat{A}_i^{\parallel} - \sum_{j=d_\delta+1}^{i}\langle V^{\pi_i}, \psi_j\rangle\langle \widehat{w}, \psi_j\rangle\,,$$

其中，第一个等式成立是因为 $A\widehat{w} = \mathbf{e}_1$。

重新排列各项后，我们得到

$$\langle V^{\pi_i}, \psi_i\rangle\langle \widehat{w}, \psi_i\rangle = \widehat{w}\cdot\widehat{A}_i^{\parallel} - \sum_{j=d_\delta+1}^{i-1}\langle V^{\pi_i}, \psi_j\rangle\langle \widehat{w}, \psi_j\rangle\,. \tag{17}$$

回顾一下，在算法 1 的第 j 次迭代中，我们在第 8 行选取了张成空间 $\text{span}(V^{\pi_1},\dots,V^{\pi_{j-1}})^\perp$ 的标准正交基 $\rho_1,\dots,\rho_{k+1-j}$。根据 ψ_j 的定义，由于 ψ_j 属于张成空间 $\text{span}(V^{\pi_1},\dots,V^{\pi_{j-1}})^\perp$，因此 $|\langle V^{\pi_i},\psi_j\rangle|$ 不大于 V^{π_i} 投影到张成空间 $\text{span}(V^{\pi_1},\dots,V^{\pi_{j-1}})^\perp$ 上的范数。
因此，我们有

$$|\langle V^{\pi_i},\psi_j\rangle| \le \sqrt{k}\max_{l\in[k+1-j]}|\langle V^{\pi_i},\rho_l\rangle| \overset{(d)}{\le} \sqrt{k}\max_{l\in[k+1-j]}\max(\left|\left\langle V^{\pi^{\rho_l}},\rho_l\right\rangle\right|,\left|\left\langle V^{\pi^{-\rho_l}},-\rho_l\right\rangle\right|)$$

$$\overset{(e)}{=}\sqrt{k}\,|\langle V^{\pi_j},u_j\rangle| \overset{(f)}{\le} \sqrt{k}\,|\langle V^{\pi_j},\psi_j\rangle| = \sqrt{k}\epsilon_j\,, \tag{18}$$

其中，不等式 (d) 成立，是因为 π^{ρ_l} 是关于偏好向量 ρ_l 的最优个性化策略；等式 (e) 则根据 u_j 的定义成立（算法 1 的第 9 行）。不等式 (f) 成立，是因为 $\langle V^{\pi_j},\psi_j\rangle$ 是 V^{π_j} 在张成空间 $\text{span}(V^{\pi_1},\dots,V^{\pi_{j-1}})^\perp$ 上的投影的范数，且 u_j 属于张成空间 $\text{span}(V^{\pi_1},\dots,V^{\pi_{j-1}})^\perp$。

对等式 (17) 两边取绝对值，可得

$$\epsilon_i\,|\langle\widehat{w},\psi_i\rangle| = \left|\widehat{w}\cdot\widehat{A}_i^{\parallel} - \sum_{j=d_\delta+1}^{i-1}\langle V^{\pi_i},\psi_j\rangle\,\langle\widehat{w},\psi_j\rangle\right| \le \left|\widehat{w}\cdot\widehat{A}_i^{\parallel}\right| + \sqrt{k}\sum_{j=d_\delta+1}^{i-1}\epsilon_j\,|\langle\widehat{w},\psi_j\rangle|\,. \tag{19}$$

现在我们可以如下界定 $\left|\widehat{w}\cdot\widehat{A}_i^{\parallel}\right|$。

$$\begin{aligned}
\left|\widehat{w}\cdot\widehat{A}_i^{\parallel}\right| &= \left|\widehat{w}^{(\delta)}\cdot\widehat{A}_i^{\parallel}\right| &(20)\\
&= \left|\widehat{w}^{(\delta)}\cdot(\widehat{A}_i - \sum_{j=d_\delta+1}^{d}\left\langle\widehat{A}_i,\psi_j\right\rangle\psi_j)\right| = \left|\widehat{w}^{(\delta)}\cdot\widehat{A}_i\right| &(21)\\
&\le \left|\widehat{w}^{(\delta)}\cdot A_i\right| + \left|\widehat{w}^{(\delta)}\cdot(\widehat{A}_i - A_i)\right|\\
&\le |w'\cdot A_i| + \left|(\widehat{w}^{(\delta)}-w')\cdot A_i\right| + \left|\widehat{w}^{(\delta)}\cdot(\widehat{A}_i - A_i)\right|\\
&\le 0 + (C_\alpha+1)\sup_\pi\left|(\widehat{w}^{(\delta)}-w')\cdot V^\pi\right| + C_V\left\|\widehat{w}^{(\delta)}\right\|_2\epsilon_\alpha\\
&\le C'C_\alpha\epsilon^{(\delta)}\,,
\end{aligned}$$

其中 $C'>0$ 是某个常数。
等式 (20) 成立，是因为 $\widehat{A}^{(\delta)}\widehat{w} = \widehat{A}^{(\delta)}\widehat{w}^{(\delta)} = \mathsf{e}_1$，且 $\widehat{A}_i^{\parallel}$ 属于张成空间 $\text{span}(\widehat{A}^{(\delta)})$。等式 (21) 成立，是因为 $\widehat{w}^{(\delta)}$ 是 $\widehat{A}^{(\delta)}x = \mathsf{e}_1$，的最小范数解，这意味着 $\widehat{w}^{(\delta)}\cdot\psi_i = 0$。最后一个不等式由引理 2 得出。
我们将通过对 $i = d_\delta+1,\dots,d$ 进行归纳，给出 $\epsilon_i\,|\langle\widehat{w},\psi_i\rangle|$ 的上界。

在 $i = d_\delta+1$ 的基础情形下，

$$\epsilon_{d_\delta+1}\,|\langle\widehat{w},\psi_{d_\delta+1}\rangle| \le \left|\widehat{w}\cdot\widehat{A}_{d_\delta+1}^{\parallel}\right| \le C'C_\alpha\epsilon^{(\delta)}\,.$$

然后，通过对式 (19) 进行归纳，对于 $i = d_\delta+2,\dots,d$，我们有：

$$\epsilon_i\,|\langle\widehat{w},\psi_i\rangle| \le (\sqrt{k}+1)^{i-d_\delta-1}C'C_\alpha\epsilon^{(\delta)}\,.$$

与式 (18) 中的偏差类似，我们在算法 1 第 8 行选取张成空间 $\text{span}(V^{\pi_1},\dots,V^{\pi_{i-1}})^\perp$ 的一组正交归一基 $\rho_1,\dots,\rho_{k+1-i}$，于是，对于任意策略 π，有：

$$|\langle V^\pi,\psi_i\rangle| \le \sqrt{k}\max_{l\in[k+1-i]}|\langle V^\pi,\rho_l\rangle| \le \sqrt{k}\,|\langle V^{\pi_i},u_i\rangle| \le \sqrt{k}\,|\langle V^{\pi_i},\psi_i\rangle| = \sqrt{k}\epsilon_i\,.$$

那么，我们有项 (b) 的上界为

$$(b) = \left|\sum_{i=d_\delta+1}^{d}\langle V^\pi,\psi_i\rangle\,\langle\widehat{w},\psi_i\rangle\right| \le \sum_{i=d_\delta+1}^{d}|\langle V^\pi,\psi_i\rangle|\cdot|\langle\widehat{w},\psi_i\rangle|$$

$$\leq \sqrt{k} \sum_{i=d_\delta+1}^{d} \epsilon_i \left|\langle \widehat{w}, \psi_i \rangle\right| \leq (\sqrt{k}+1)^{d-d_\delta} C' C_\alpha \epsilon^{(\delta)} .$$

因此，对于任意策略 π，我们有

$$\left|\langle \widehat{w}, V^\pi \rangle - \left\langle \widehat{w}^{(\delta)}, V^\pi \right\rangle\right| \leq (\sqrt{k}+1)^{d-d_\delta} C' C_\alpha \epsilon^{(\delta)} + \frac{3}{2}\sqrt{k}\delta \left\|w'\right\|_2 .$$

□

H 定理 2 的证明

定理 2. 考虑计算 $\widehat{A}$ 的算法，以及任意一个解 $\widehat{w}^{(\delta)}$，其属于 $\widehat{A}^{(\delta)} x = \mathbf{e}_1$，针对 $\delta = k^{\frac{5}{3}} \epsilon^{\frac{1}{3}}$并输出策略 $\pi^{\widehat{w}^{(\delta)}} = \arg\max_{\pi \in \Pi} \langle \widehat{w}^{(\delta)}, V^\pi \rangle$。则该策略 $\pi^{\widehat{w}^{(\delta)}}$满足 $v^* - \left\langle w^*, V^{\pi^{\widehat{w}_{(\delta)}}} \right\rangle \leq \mathcal{O}\left(k^{\frac{13}{6}} \epsilon^{\frac{1}{3}}\right)$.

定理 2 的证明。如引理 1 所示，我们设定 $C_\alpha = 2k$ 并且有 $\epsilon_\alpha = \frac{4k^2\epsilon}{v^*}$。我们有 $\|w'\| = \frac{\|w^*\|}{\langle w^*, V^{\pi_1} \rangle}$ 并且在引理 1 的证明中证明了 $\langle w^*, V^{\pi_1} \rangle \geq \frac{v^*}{2k}$。
应用引理 2 并令 $\delta = \left(\frac{C_V^4 k^5 \|w^*\|_2 \epsilon}{v^{*2}}\right)^{\frac{1}{3}}$，我们得到

$$\begin{aligned}
&v^* - \left\langle w^*, V^{\pi^{\widehat{w}^{(\delta)}}} \right\rangle = \langle w^*, V^{\pi_1} \rangle \left(\left\langle w', V^{\pi^*} \right\rangle - \left\langle w', V^{\pi^{\widehat{w}^{(\delta)}}} \right\rangle \right) \\
\leq &\langle w^*, V^{\pi_1} \rangle \left(\left\langle \widehat{w}^{(\delta)}, V^{\pi^*} \right\rangle - \left\langle \widehat{w}^{(\delta)}, V^{\pi^{\widehat{w}^{(\delta)}}} \right\rangle + \mathcal{O}(\sqrt{k} \left\|w'\right\|_2 (\frac{C_V^4 k^5 \left\|w^*\right\|_2 \epsilon}{v^{*2}})^{\frac{1}{3}}) \right) \\
= &\mathcal{O}(\sqrt{k} \left\|w^*\right\|_2 (\frac{C_V^4 k^5 \left\|w^*\right\|_2 \epsilon}{v^{*2}})^{\frac{1}{3}}) .
\end{aligned}$$

□

I 对 ϵ 的依赖

In t本节将讨论一种可能的方法，以提高对 ϵ i 的依赖 n
The定理 1 和 2。
考虑一个玩具示例，其中返回的三个基策略分别是 π_1，其 $V^{\pi_1} = (1, 0, 0)$； π_2，其$V^{\pi_2} = (1, 1, 1)$; 以及 π_3，其 $V^{\pi_3} = (1, \eta, -\eta)$，其中某个 $\eta > 0$，以及 $w^* = (1, w_2, w_3)$，其中某个w_2, w_3。

估计比率 $\widehat{\alpha}_1$ 位于 $[1 + w_2 + w_3 - \epsilon, 1 + w_2 + w_3 + \epsilon]$，而 $\widehat{\alpha}_2$ 位于 $[1 + \eta w_2 - \eta w_3 - \epsilon$, $1 + \eta w_2 - \eta w_3 + \epsilon]$。假设 $\widehat{\alpha}_1 = 1 + w_2 + w_3 + \epsilon$ 和 $\widehat{\alpha}_2 = 1 + \eta w_2 - \eta w_3 + \epsilon$。
通过求解

$$\begin{pmatrix} 1 & 0 & 0 \\ w_2 + w_3 + \epsilon & -1 & -1 \\ \eta w_2 - \eta w_3 + \epsilon & -\eta & \eta \end{pmatrix} \widehat{w} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

我们可以推导出 $\widehat{w}_2 = w_2 + \frac{\epsilon}{2}(1 + \frac{1}{\eta})$ 和 $\widehat{w}_3 = w_3 + \frac{\epsilon}{2}(1 - \frac{1}{\eta})$。

我们关注的次优性度量是 $\sup_\pi \left|\langle \widehat{w}, V^\pi \rangle - \langle w^*, V^\pi \rangle\right|$，其上界为 $C_V \left\|\widehat{w} - w^*\right\|_2$。但是， ℓ_2 与 $\widehat{w}$ 之间的距离取决于 A的条件数，当 η 较小时，该条件数会很大。为了在第3节中得到一个非平凡的上界，我们引入另一个基于 $\widehat{A}$ 截断版本的估计值 $\widehat{w}^{(\delta)}$，然后分别对 $\left\|\widehat{w}^{(\delta)} - w^*\right\|_2$和 $\sup_\pi \left|\langle \widehat{w}, V^\pi \rangle - \left\langle \widehat{w}^{(\delta)}, V^\pi \right\rangle\right|$ 求上界。

然而，目前尚不清楚 $\sup_\pi \left|\langle \widehat{w}, V^\pi \rangle - \langle w^*, V^\pi \rangle\right|$ 是否取决于 A 的条件数。由于算法 1 的构造方式，我们可以获得有关所有策略值集合的一些额外信息。

首先，由于我们在 π_3 之前找到 π_2， η 必须不大于 1。根据算法，当偏好向量为 u_2 时，V^{π_2} 是最优策略（ u_2 的定义见算法 1 第 11 行）；当偏好向量为 $(0, 1, -1)$ 时， V^{π_3} 是最优策略。注意，根据 u_2 的定义， u_2 与 V^{π_2} 之间的角度不大于 45 度。因此，所有策略的价值只能落在小盒子 $B = \{x \in \mathbb{R}^3 | \left|u_2^\top x\right| \leq |\langle u_2, V^{\pi_2}\rangle|, \left|u_3^\top x\right| \leq |\langle u_3, V^{\pi_3}\rangle|\}$ 中。容易验证，对于任意 $x \in B, |\langle \widehat{w}, x\rangle - \langle w^*, x\rangle| < (1+\sqrt{2})\epsilon$。这个示例说明，即使 A 的条件数很大，$\sup_\pi |\langle \widehat{w}, V^\pi\rangle - \langle w^*, V^\pi\rangle|$ 也可能很小。目前尚不清楚这一结论是否普遍成立。直接利用这些附加信息为 $\sup_\pi |\langle \widehat{w}, V^\pi\rangle - \langle w^*, V^\pi\rangle|$ 求上界，而不是通过为 $C_V \left\|\widehat{w} - w^*\right\|_2$ 求界，可能是改进项 $\epsilon^{\frac{1}{3}}$ 的一种方法。

J C4的描述与引理 4 的证明

算法4 C4：使用 Carathéodory 定理压缩凸组合

1：输入一组 k 维向量 $M \subset \mathbb{R}^k$ 和一个分布 $p \in$ 单纯形 ^{M}
2：当 $|M| > k+1$ 时
3：　从 M 中任意选取 $k+2$ 个向量 μ_1、……、 μ_{k+2}
4：　求解 $x \in \mathbb{R}^{k+2}$，使得 $\sum_{i=1}^{k+2} x_i(\mu_i \circ 1) = \mathbf{0}$，其中 $\mu \circ 1$ 表示在向量后附加 1 to μ
5：　$i_0 \leftarrow \operatorname{argmax}_{i\in[k+2]} \frac{|x_i|}{p\ \mu_i}()$
6：　如果 $x_{i_0} < 0$，那么 $x \leftarrow -x$
7：　$\gamma \leftarrow \frac{p(\mu_{i_0})}{x_{i_0}}$ 以及 $\forall i \in [k+2], p(\mu_i) \leftarrow p(\mu_i) - \gamma x_i$
8：　从 M 中移除 μ_i，其中包含 $p(\mu_i) = 0$
9：结束 while 循环
10：输出 M 和 p

引理 4. 给定一组 k 维向量 $M \subset \mathbb{R}^k$ 以及在 M 上的分布 p，C4(M, p) 输出 $M' \subset M$，其中 $|M'| \leq k+1$，并给出一个分布 $q \in$ 单纯形$^{M'}$，满足 $\mathbb{E}_{\mu\sim q}[\mu] = \mathbb{E}_{\mu\sim p}[\mu]$，耗时 $\mathcal{O}(|M|\, k^3)$。

证明。该证明与 Carathéodory 定理的证明类似。给定算法4第 3 行选取的向量 μ_1、……、μ_{k+2} 及其概率质量 $p(\mu_i)$，我们在算法中求解 $x \in \mathbb{R}^{k+2}$，使得 $\sum_{i=1}^{k+2} x_i(\mu_i \circ 1) = \mathbf{0}$。

注意，由于 $\{\mu_i \circ 1 | i \in [k+2]\}$ 线性相关， x 存在非零解。此外， x 满足 $\sum_{i=1}^{d+2} x_i = 0$。

因此，

$$\sum_{i=1}^{d+2}(p(\mu_i) - \gamma x_i) = \sum_{i=1}^{d+2} p(\mu_i).$$

对于所有 i，若 $x_i < 0, p(\mu_i) - \gamma x_i \geq 0$ 如 $\gamma > 0$；若 $x_i > 0,$ 则 $p\frac{x_i}{(\mu_i)} \leq p^{x_{i_0}}(\mu_{i_0}) = \frac{1}{\gamma}$，因此$p(\mu_i) - \gamma x_i \geq 0$。

因此，经过一次迭代后，更新后的 p 仍然是 M 上的概率分布（即，对于所有 $\mu \in M$， $p(\mu) \geq 0$，且 $\sum_{\mu\in M} p(M) = 1$）。此外， $\sum_{i=1}^{d+2}(p(\mu_i) - \gamma x_i)\mu_i = \sum_{i=1}^{d+2} p(\mu_i)\mu_i - \gamma\sum_{i=1}^{d+2} x_i\mu_i = \sum_{i=1}^{d+2} p(\mu_i)\mu_i$。
因此，经过一次迭代后，期望值 $\mathbb{E}_{\mu\sim p}[\mu]$ 保持不变。
当我们最终输出（M', q）时，有 q是 M上的分布，且 $\mathbb{E}_{\mu\sim q}[\mu] = \mathbb{E}_{\mu\sim p}[\mu]$。

Due 根据算法第 6 行，我们知道 $x_{i_0} > 0$。因此 $p(\mu_{i_0}) - \gamma x_{i_0} = p(\mu_{i_0}) - {}^{p(\mu_{i_0})\, x_{i_0}}\underline{\qquad x_{i_0}} = 0.$

我们从 M 中至少移除一个向量 μ_{i_0}，并且最多运行 $|M|$ 次迭代。最后，求解 x 需要 $\mathcal{O}(k^3)$ 的时间，因此算法 4 的总耗时为 $\mathcal{O}(|M|\,k^3)$。 □

基于流分解的 K 方法

我们首先介绍一种基于流分解思想的算法。

为此，我们构造一个分层图 $G=((L^{(0)}\cup\ldots\cup L^{(H+1)})\ E)$，其中包含 $H+2$ 个两两不相交的层 $L^{(0)}$、.....、$L^{(H+1)}$；每一层 $t\le H$ 都包含一组顶点，这些顶点标记为相应时间步 t 可达的状态（可能有重复），即 $\{s\in\mathcal{S}|\Pr(S_t=s|S_0=s_0)\ >0\}$。

我们将 $L^{(t)}$ 中标记为状态 s 的顶点记作 $x_s^{(t)}$。

第 $L^{(H+1)}=\{x_*^{(H+1)}\}$ 层仅包含一个人工顶点 $x_*^{(H+1)}$，其标记为人工状态 $*$。对于 $t=0$、.....、$H-1$，对于每个 $x_s^{(t)}\in L^{(t)}$，$x_{s'}^{(t+1)}\in L^{(t+1)}$，如果 $P(s'|s,\ \pi(s))\ >0$，则我们用一条标记为 $(s,\ s')$ 的边连接 $x_s^{(t)}$ 和 $x_{s'}^{(t+1)}$。第 H 层中的每个顶点 $x_s^{(H)}$ 都通过一条边与 $x_*^{(H+1)}$ 相连，该边标记为 $(s,\ *)$。

我们用 $E^{(t)}$ 表示 $L^{(t)}$ 和 $L^{(t+1)}$之间的边。注意，每条轨迹 $\tau\ =(s_0,s_1,\ldots,s_H)$都对应一条从$x_{s_0}^{(0)}$ 到 $x_*^{(H+1)}$、长度为 $H+2$ 的唯一路径 $(x_{s_0}^{(0)},x_{s_1}^{(1)},\ldots,x_{s_H}^{(H)},x_*^{(H+1)})$。

这是一一映射，接下来我们将路径和轨迹互换使用。

该策略对应于图 G中的一个 $(x_{s_0}^{(0)},x_*^{(H+1)})$-流，其流值为 1。具体而言，该流定义如下。

当上下文已明确指出层 t 时，我们实际上会用顶点 s来指代顶点 $x_s^{(t)}$.

对于 $t=0,\ldots,H-1$，对于任意边 $(s,s')\ \in E^{(t)}$，令 $f\colon\ E\to\mathbb{R}^+$ 定义为

$$f(s,s')=\sum_{\tau:(s_t^\tau,s_{t+1}^\tau)=(s,s')}q^\pi(\tau)\,, \tag{22}$$

其中，$q^\pi(\tau)$ 是对 τ 进行采样的概率。

对于任意边 $(s,*)\ \in E^{(H)}$，令 $f(s,*)=\sum_{(s',s)\in E^{(H-1)}}f(s',s)$。不难验证，函数 f 是定义良好的流。因此，我们可以通过动态规划计算 f。

对于所有 $(s_0,s)\ \in E^{(0)}$，有 $f(s_0,s)\ =P(s|s_0,\pi(s_0))$；对于 $(s,s')\ \in E^{(t)}$，

$$f(s,s')=P(s'|s,\pi(s))\sum_{s'':(s'',s)\in E^{(t-1)}}f(s'',s)\,. \tag{23}$$

现在，我们可以通过在算法 5 中分解 f 来介绍我们的算法。

算法 5 的每次迭代都会将至少一条边归零，因此该算法将在$|E|$ 轮内停止。

算法 5 基于流分解的方法

1: 初始化 $Q\leftarrow\emptyset$。 2: 根据式 (23)，通过动态规划计算所有边 $e\in E$ 对应的 $f(e)$。 3: 当 $\exists e\in E$ 满足 $f(e)\ >0$ 时，执行循环。 4: 选择一条路径 $\tau=(s_0,s_1,\ldots,s_H,*)$，使得 $f(s_i,s_{i+1})\ >0\ \forall i\ge 05$: $f_\tau\leftarrow\min_e$于 $_\tau f(e)$。 6: 对于 τ 中的 e，$Q\leftarrow Q\cup\{(\tau,f_\tau)\}$, $f(e)\ \leftarrow f(e)\ -f_\tau$。 7: 结束循环。 8: 输出 Q

定理 4。算法 5 在时间 $\mathcal{O}(H^2\,|\mathcal{S}|^2)$ 内输出满足 $\sum_{(\tau,f_\tau)\in Q} f_\tau \Phi(\tau) = V^\pi$ 的 Q。

该证明的核心思想是，对于任意边（s，s'）$\in E^{(t)}$，（s，s'）上的流捕获了 $S_t = s \wedge S_{t+1} = s'$ 的概率，因此，策略 V^π 的价值关于 $\{f(e)\,|e \in E\}$是线性的。
输出 Q 至多包含 $|E|$ 条加权路径（轨迹）。我们还可以通过 C4 进一步压缩该表示，其耗时为 O（$|Q|\,k^3$）。
推论 2. 先执行算法 5，得到输出 Q，然后在{（Φ（τ） f_τ）|（τ，f_τ）上运行 C4，即可在$\mathcal{O}$（$H^2\,|\mathcal{S}|^2 + k^3 H\,|\mathcal{S}|^2$）时间内返回一个大小为（$k+1$）的加权轨迹表示。

我们指出，当 $|\mathcal{S}|H + |\mathcal{S}|k^3 = \omega(k^4 + k|\mathcal{S}|)$ 时，该流分解方法的运行时间劣于扩展与压缩方法（见定理 3）。

L 定理 4 的证明

The 定理 4。算法 5 输出满足 Q 的 $\sum_{(\tau,\ f_\tau)\in Q} f_\tau \Phi(\tau) = V^\pi$，时间为 $\mathcal{O}(H^2\,|\mathcal{S}|^2)$。

证明。正确性：由公式 (22) 定义的函数 f 是一个定义良好的流，因为对于所有 t =1，....., H，以及所有 $s \in L^{(t)}$，我们有

$$\sum_{s'\in L^{(t+1)}:(s,s')\in E^{(t)}} f(s,s') = \sum_{s'\in L^{(t+1)}:(s,s')\in E^{(t)}} \sum_{\tau:(s_t^\tau,s_{t+1}^\tau)=(s,s')} q^\pi(\tau) = \sum_{\tau:s_t^\tau=s} q^\pi(\tau)$$

$$= \sum_{s''\in L^{(t-1)}:(s'',s)\in E^{(t-1)}} f(s'',s)\,.$$

下文中，我们首先证明，对于所有 $e \in E$，算法 5 将终止，并满足 $f(e) = 0$。

首先，在每次迭代之后，f 仍然是一个可行的（$x^{(0)}$，$x^{(H+1)}$）-流，可行流从 $x^{(0)}$ 流出的总流量减少了 f_τ。此外，对于在开始时具有 $f(e) > 0$ 的边 e，在整个算法运行过程中，我们都有 $f(e) \geq 0$，因为我们从未将 $f(e)$ 减少超过 $f(e)$ 的量。

于是，由于 f是一个 $(x^{(0)}, x^{(H+1)})$-流，并且对于所有 $e \in E$，$f(e) \geq 0$，我们总能在算法 5 的第 4 行找到一条路径 τ。
否则，从 $x^{(0)}$ 出发，沿着流量为正的边可到达的顶点集合不包含$x^{(H+1)}$，并且流出该集合的流量等于流出 $x^{(0)}$的总流量。但由于其他顶点均不可达，该集合没有流流出，这就产生了矛盾。

在第 6 行，至少存在一条边 e，使得 $f(e) > 0$ 被减至 0。因此，该算法最多运行 $|E|$ 次迭代，并在对于所有 $e \in E$，$f(e) = 0$ 时终止。

Thus we have that for any $(s,s') \in E^{(t)}$, $f(s,s') = \sum_{(\tau,f_\tau)\in Q:(s_t^\tau,s_{t+1}^\tau)=(s,s')} f_\tau$.

那么我们有

$$\begin{aligned}
V^\pi &= \sum_\tau q^\pi(\tau)\Phi(\tau) = \sum_\tau q^\pi(\tau)\left(\sum_{t=0}^{H-1} R(s_t^\tau, \pi(s_t^\tau))\right) \\
&= \sum_{t=0}^{H-1}\sum_\tau q^\pi(\tau) R(s_t^\tau, \pi(s_t^\tau)) \\
&= \sum_{t=0}^{H-1}\sum_{(s,s')\in E^{(t)}} R(s,\pi(s)) \sum_{\tau:(s_t^\tau,s_{t+1}^\tau)=(s,s')} q^\pi(\tau) \\
&= \sum_{t=0}^{H-1}\sum_{(s,s')\in E^{(t)}} R(s,\pi(s)) f(s,s')
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{t=0}^{H-1} \sum_{(s,s')\in E^{(t)}} R(s,\pi(s)) \sum_{(\tau,f_\tau)\in Q:(s_t^\tau,s_{t+1}^\tau)=(s,s')} f_\tau \\
&= \sum_{(\tau,f_\tau)\in Q} f_\tau \Big(\sum_{t=0}^{H-1} R(s_t^\tau,\pi(s_t^\tau))\Big) \\
&= \sum_{(\tau,f_\tau)\in Q} f_\tau \Phi(\tau)\,.
\end{aligned}
$$

计算复杂度：求解 f 需要 $\mathcal{O}(|E|)$ 时间。该算法将运行 $\mathcal{O}(|E|)$ 次迭代，每次迭代需要 $\mathcal{O}(H)$ 时间。由于 $|E| = \mathcal{O}(|\mathcal{S}|^2 H)$，算法 5 的总运行时间为 $\mathcal{O}(|\mathcal{S}|^2 H^2)$。C4 需要 $\mathcal{O}(k^3 |E|)$ 时间。

□

定理 3 的证明

定理 3. 算法2输出 $F^{(H)}$ 和 $\beta^{(H)}$，满足 $\left|F^{(H)}\right| \le k+1$ 和$\sum_{\tau\in F^{(H)}} \beta^{(H)}(\tau)\Phi(\tau) = V^\pi$，且所需时间为 $\mathcal{O}(k^4 H|\mathcal{S}| + kH|\mathcal{S}|^2)$。

证明。正确性：C4 保证 $\left|F^{(H)}\right| \le k+1$。
我们将对 $t = 1,\dots,H$ 使用归纳法，证明 $\sum_{\tau\in F^{(H)}} \beta_\tau^{(H)}\Phi(\tau) = V^\pi$。

回忆一下，对于任意长度为 h 的轨迹 τ，$J(\tau) = \Phi(\tau)+V(s_h^\tau, H-h)$ 定义为前缀为 τ 的轨迹（长度为 H）的期望回报。
此外，回想一下，$J_{F^{(t)}} = \{J(\tau\circ s)|\tau \in F_{(t)}\, s\in\mathcal{S}\}$和 $p_{F^{(t)},\beta^{(t)}}$是通过令$p_{F^{(t)}\ \beta^{(t)}}(\tau\circ s) = \beta^{(t)}(\tau)P(s|s_t^\tau,\pi(s_t^\tau))$定义的。
对于基础情形，我们有，在 $t=1$

$$
\begin{aligned}
V^\pi &= R(s_0,\pi(s_0)) + \sum_{s\in\mathcal{S}} P(s|s_0,\pi(s_0))V^\pi(s,H-1) = \sum_{s\in\mathcal{S}} P(s|s_0,\pi(s_0))J((s_0)\circ s) \\
&= \sum_{s\in\mathcal{S}} p_{F^{(0)},\beta^{(0)}}((s_0)\circ s)J((s_0)\circ s) = \sum_{\tau\in F^{(1)}} \beta^{(1)}(\tau)J(\tau)\,.
\end{aligned}
$$

假设 $V^\pi = \sum_{\tau'\in F^{(t)}} \beta^{(t)}(\tau')J(\tau')$ 在时刻 t成立，那么我们证明该命题在时刻 $t+1$成立。

$$
\begin{aligned}
V^\pi &= \sum_{\tau'\in F^{(t)}} \beta^{(t)}(\tau')J(\tau') = \sum_{\tau'\in F^{(t)}} \beta^{(t)}(\tau')(\Phi(\tau') + V^\pi(s_t^{\tau'}, H-t)) \\
&= \sum_{\tau'\in F^{(t)}} \beta^{(t)}(\tau') \left(\Phi(\tau') + \sum_{s\in\mathcal{S}} P(s|s_t^{\tau'},\pi(s_t^{\tau'}))\left(R(s_t^{\tau'},\pi(s_t^{\tau'})) + V^\pi(s,H-t)\right)\right) \\
&= \sum_{\tau'\in F^{(t)}} \sum_{s\in\mathcal{S}} \beta^{(t)}(\tau')P(s|s_t^{\tau'},\pi(s_t^{\tau'}))\left(\Phi(\tau') + R(s_t^{\tau'},\pi(s_t^{\tau'})) + V^\pi(s,H-t)\right) \\
&= \sum_{\tau'\in F^{(t)}} \sum_{s\in\mathcal{S}} \beta^{(t)}(\tau')P(s|s_t^{\tau'},\pi(s_t^{\tau'}))\left(\Phi(\tau'\circ s) + V^\pi(s,H-t)\right) \\
&= \sum_{\tau'\in F^{(t)}} \sum_{s\in\mathcal{S}} p_{F^{(t)},\beta^{(t)}}(\tau'\circ s)J(\tau'\circ s) \\
&= \sum_{\tau\in F^{(t+1)}} \beta^{(t+1)}(\tau)J(\tau)\,.
\end{aligned}
$$

根据归纳法，该陈述在 $t = H$通过归纳法成立，即 $V^\pi = \sum_{\tau\in F^{(H)}} \beta_\tau^{(H)} J(\tau) = \sum_{\tau\in F^{(H)}} \beta_\tau^{(H)}\Phi(\tau)$。

计算复杂度：对所有 $s \in \mathcal{S}, h \in [H]$，求解 $V^{\pi}(s,h)$ 需要 $\mathcal{O}(kH\,|\mathcal{S}|^2)$ 时间。每一轮中，我们需要对 $\leq (k+1)\,|S|$ 个向量调用 C4，这需要 $\mathcal{O}(k^4\,|\mathcal{S}|)$ 时间。因此，总共需要 $\mathcal{O}(k^4 H\,|\mathcal{S}| + kH\,|\mathcal{S}|^2)$ 时间。 □

N 最大化单个目标函数的示例

观察2。假设存在 $k > 2$ 个策略，它们共同构成 k 个线性无关的价值向量。考虑 k 个不同的策略 π_1^*、.....、π_k^*，每个策略都能最大化目标函数$i \in [k]$。那么，它们各自对应的价值向量 V_1^*、.....、V_k^* 不一定线性无关。此外，如果 V_1^*、.....、V_k^* 线性相关，并不意味着不存在 k 个线性无关的价值向量。

证明。为简单起见，我们展示一个时间范围为 $H = 1$ 的示例，但结果可以扩展到任意 $H \geq 1$。我们将展示一个示例，其中有 4 个不同的价值向量，其中 3 个由最大化 3 个目标函数的策略 $k = 3$ 得到，且这 3 个向量线性相关。
考虑一个只有一个状态的马尔可夫决策过程（也称为多臂老虎机），其中有 4 个动作，其奖励向量是确定的（在这种情况下，它们也是 4 个可能策略的期望值）：

$$r(1) = \begin{pmatrix} 8 \\ 4 \\ 2 \end{pmatrix}, \quad r(2) = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad r(3) = \begin{pmatrix} 85/12 \\ 25/6 \\ 35/12 \end{pmatrix} \approx \begin{pmatrix} 7.083 \\ 4.167 \\ 2.9167 \end{pmatrix}, \quad r(4) = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}.$$

将 π^a 记为始终选择动作 a 的固定策略。显然，策略 π^1 使第一个目标函数最大化，策略 π^2 使第三个目标函数最大化，而策略 π^3 使第二个目标函数最大化（π^4 均不使任何目标函数最大化）。然而，

- $r(3)$ 线性依赖于 $r(1)$ 和 $r(2)$，因为

$$\frac{5}{6} r(1) + \frac{5}{12} r(2) = r(3).$$

- 此外， $r(4)$ 在线性无关的 $r(1)$ $r(2)$ 中：反设不然，则存在 $\beta_1, \beta_2 \in \mathbb{R}$ 使得：

$$\beta_1 \cdot r(1) + \beta_2 \cdot r(2) = \begin{pmatrix} 8\beta_1 + \beta_2 \\ 4\beta_1 + 2\beta_2 \\ 2\beta_1 + 3\beta_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} = r(4).$$

因此，前两个方程推出 $\beta_2 = 1 - 8\beta_1$，而 $4\beta_1 + 2 - 16\beta_1 = 3$，于是 $\beta_1 = -\frac{1}{12}$ 和 $\beta_2 = \frac{5}{3}$。将其代入第三个方程，得到 $-\frac{1}{6} + 5 = 2$，这与假设矛盾。

□