arXiv:2301.08811v1 [cs.MA] 20 Jan 2023

合作式多智能体规划中的差分隐私

Bo Chen[1, *]、Calvin Hawkins[1, *]、Mustafa O. Karabag[2, *]、Cyrus Neary[2, *]、Matthew Hale[1] 和 Ufuk Topcu[2]

[1]佛罗里达大学[2]德克萨斯大学奥斯汀分校

{bo.chen、calvin.hawkins、matthewhale}@ufl.edu, {karabag、cneary、utopcu}@utexas.edu

摘要

隐私感知多智能体系统必须保护智能体的敏感数据，同时确保智能体实现共同目标。为此，我们提出一个框架，用于在合作式多智能体决策问题中对智能体间通信进行隐私化处理。我们研究以具有到达-避障目标的合作马尔可夫博弈建模的序贯决策问题。我们应用差分隐私机制对智能体传递的符号状态轨迹进行隐私化处理，并分析隐私强度与团队性能之间的权衡。研究表明，在给定隐私级别下，这种权衡在很大程度上取决于智能体状态-动作过程之间的总相关性。我们通过降低总相关性的值，合成能够抵御隐私化影响的策略。数值实验表明，与非隐私化的通信实现相比，采用这些策略时，团队性能仅下降 3%。相较之下，使用忽略总相关性、仅优化团队性能的基线策略时，团队性能约下降 86%。

1 引言

在合作式多智能体系统中，决策智能体团队与共享环境交互，以实现共同目标 [Cao et al., 2013; Parker et al., 2016]。在这些系统中，智能体间通信通常是团队成功协调的必要条件；每个智能体在做出自身决策时，往往依赖与队友相关的信息。然而，这些通信信息可能涉及敏感内容。例如，在解决多车辆路径规划问题时，自动驾驶车辆共享位置数据可能有益，但这些数据也会泄露乘客的敏感位置数据。因此，隐私感知多智能体系统应保护智能体的敏感数据，同时确保智能体能够实现共同目标。

在这项工作中，我们开发了此类隐私感知多智能体系统。具体而言，我们研究以到达-避障目标为目标、建模为合作马尔可夫博弈的序贯多智能体决策问题。我们假设使用可信的中央聚合器，预先为智能体团队合成一组局部策略。智能体的局部策略是从智能体联合状态空间到该智能体局部动作空间的映射。然而，在策略执行期间，智能体希望对队友和潜在窃听者保密各自的状态轨迹（聚合器不参与运行时过程）。如数值结果所示，如果局部策略未考虑隐私，那么在私密通信条件下，其性能可能会大幅下降。因此，我们构建了一个框架，用于保护执行策略所需的智能体间通信，并合成在私密通信条件下仍能保持良好性能的策略。

我们采用差分隐私 [Dwork 等人， 2014] 构建一个框架，为多智能体系统提供形式化的隐私保证。在马尔可夫博弈中，每个智能体都由一个马尔可夫决策过程（MDP）建模，我们关注的是对这些MDP的状态轨迹进行隐私保护。我们使用 [Chen 等人， 2022]提出的马尔可夫链在线机制来实现差分隐私。该机制为MDP生成的符号状态轨迹提供差分隐私保证，为智能体实时生成隐私状态提供高效方法，并确保隐私轨迹符合MDP的底层动力学。每个智能体都可以调整这些隐私保证的强度。

本研究的具体贡献如下：

1. 多智能体系统差分隐私框架。我们提出了一种适用于多智能体规划问题的差分隐私框架。该框架支持在智能体间隐私通信的条件下去中心化执行局部策略。
2. 理论结果：分析隐私与性能之间的权衡。我们根据隐私强度以及智能体状态—动作过程的总相关性，对采用隐私通信时团队的成功概率给出了界限。3. 综合策略以平衡隐私与性能。通过最小化这一总相关性值，

[*] 表示贡献相同。

我们利用隐私与性能之间的权衡，为多智能体系统综合策略，使其在高强度隐私保护下仍能实现高性能。

数值实验表明，即使在通信经过隐私处理的情况下，综合策略仍表现出强劲的性能。我们观察到，在通信经过隐私处理时：1）所提出的最小依赖策略比仅在真实通信下优化团队性能、且忽略总相关性的基线策略性能高出 80%；2）随着总相关性降低，团队性能随之提升；3）最小依赖策略的性能对于隐私机制所实施的隐私级别具有稳健性。

尽管隐私在多智能体系统中十分重要，[Such 等人，2014], 现有的多智能体规划与学习算法通常不会考察隐私与团队性能之间的权衡，而且许多算法完全不考虑隐私。[Nissim 和 Brafman，2014年；Brafman， 2015] 提出了确定性问题中多智能体规划的强隐私概念。这些研究开发了相关算法，确保智能体在执行分布式规划算法时不会共享敏感状态或动作。[Ye 等人， 2022] 将强隐私概念替换为差分隐私，以便在确定性场景下对去中心化规划所需的信息进行隐私处理。[Hefner 等人，2022] 将强隐私概念扩展至随机系统，并开发了分布式价值迭代算法。这些研究关注隐藏每个智能体状态中的私有部分，但没有考虑智能体通过改变共享信息来实现隐私的机制。相比之下，我们研究一种差分隐私机制，该机制会在随机环境中的多智能体通信期间改变智能体的状态轨迹。

与此同时，差分隐私也在马尔可夫决策过程的规划与强化学习领域得到了研究 [Garcelon 等人，2020；Qiao 和 Wang，2022；Gohari 等人， 2021]。然而，这些研究探讨的是单智能体问题，主要关注价值函数、奖励值或转移概率的隐私保护。相较之下，我们研究多智能体设置，并针对符号状态轨迹定义差分隐私。具体而言，我们将 [Chen 等人， 2022] 提出的差分隐私机制扩展到多智能体规划问题，并研究隐私对团队性能的影响。

去中心化策略执行在多智能体马尔可夫决策过程的规划与强化学习中日益受到关注 [Becker 等人，2003年；Rashid 等，2018；Son 等人，2019；Oliehoek 和 Amato，2016；Karabag 等人， 2022]。作为去中心化策略执行的副产品，这些算法可能实现隐私，即智能体不会通信本地可用的信息。然而，这些研究并未明确考虑隐私或提供隐私保证。我们则允许通信，并将总相关性用作一种软性去中心化度量，从而能够综合出在私密通信条件下表现良好的策略。

2 预备知识

离散随机变量 Y 在支撑集 $\mathcal{Y}$ 上的熵为 $H(Y) = -\sum_{y\in\mathcal{Y}} \Pr(Y=y)\log(\Pr(Y=y))$。

2.1 合作马尔可夫博弈

给定一个由有限个智能体构成的集合 N，并将其索引为 $i \in \{1,2,\dots,N\}$，我们使用马尔可夫决策过程 $\mathcal{M}^i$ 对智能体 i 的动力学建模。马尔可夫决策过程是一个元组 $\mathcal{M}^i = (\mathcal{S}^i, s_I^i, \mathcal{A}^i, \mathcal{T}^i)$，其中 $\mathcal{S}^i$ 是智能体 i 的有限局部状态集合，$s_I^i \in \mathcal{S}^i$ 是一个初始状态，$\mathcal{A}^i$ 是智能体 i 的有限局部动作集合，而 $\mathcal{T}^i$：$\mathcal{S}^i \times \mathcal{A}^i \to \Delta(\mathcal{S}^i)$ 是一个转移概率函数，其中 $\Delta(\mathcal{S}^i)$ 表示状态空间 $\mathcal{S}^i$. 上的概率分布集合。为简洁起见，我们用 $\mathcal{T}^i(s^i, a^i, y^i)$ 表示由分布 $\mathcal{T}^i(s^i, a^i)$ 给出的 y^i 的概率。若存在某个动作 $a^i \in \mathcal{A}^i$，使得 $\mathcal{T}^i(s_k^i, a^i, s_j^i) > 0$，则称状态 $s_j^i \in \mathcal{S}^i$ 是另一状态 $s_k^i \in \mathcal{S}^i$ 的可行状态。

给定这样一组智能体，我们将团队决策问题表述为一个合作马尔可夫博弈 M。由 N 个智能体组成的合作马尔可夫博弈中，每个智能体都由一个马尔可夫决策过程 $\mathcal{M}^i = (\mathcal{S}^i, s_I^i, \mathcal{A}^i, \mathcal{T}^i)$ 建模，该博弈由元组 $\mathrm{M} = (\mathrm{S}\ \boldsymbol{s_I}, \mathrm{A}, \mathrm{T})$ 给出。其中，$\mathrm{S} = \mathcal{S}^1 \times \cdots \times \mathcal{S}^N$ 是联合状态空间，$\boldsymbol{s_I} = (s_I^1, \dots, s_I^N)$ 是联合初始状态，$\mathrm{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$ 是联合动作空间，T 是联合转移概率函数。为简洁起见，我们用 $\mathrm{T}(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{y})$ 表示给定分布 $\mathrm{T}(\boldsymbol{s}, \boldsymbol{a})$ 时 $\boldsymbol{y}$ 的概率。令 $\boldsymbol{s} \in \mathrm{S}$ 和 $\boldsymbol{a} \in \mathrm{A}$ 分别表示联合状态和联合动作。对于所有 $\boldsymbol{s} = (s^1, \dots, s^N) \in \mathrm{S}$ $\boldsymbol{y} = (y^1, \dots, y^N) \in \mathrm{S}$ 以及 $\boldsymbol{a} = (a^1, \dots, a^N) \in \mathrm{A}$，T 定义为 $\mathrm{T}(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{y}) = \prod_{i=1}^N \mathcal{T}^i(s^i, a^i, y^i)$。

为方便记号，我们用 $\boldsymbol{s}^{-i} \in \mathcal{S}^1 \times \ldots \times \mathcal{S}^{i-1} \times \mathcal{S}^{i+1} \times \ldots \times \mathcal{S}^N$ 表示智能体 i 的队友（不包括智能体 i 自身）的状态。用 S 表示其可取值的集合。类似地，$\boldsymbol{a}^{-i}$ 和 $\mathrm{A}^{-i} = \mathcal{S}^1 \times \ldots \times \mathcal{S}^{i-1} \times \boldsymbol{s}^{-i}$ 分别表示智能体 i 的队友所采取的动作，以及队友所有可能动作的集合。

智能体 i 的（平稳）局部策略 $\pi^i : \mathrm{S} \to \Delta(\mathcal{A}^i)$ 是一个映射，将特定的联合状态映射为智能体 i 动作上的概率分布。给定团队处于联合状态 $\boldsymbol{s}$，$\pi^i(\boldsymbol{s}, a^i)$ 表示智能体 i 选择动作 a^i 的概率，其中该动作由 π^i 选出。我们将（平稳）联合策略 $\boldsymbol{\pi}$ 定义为一组局部策略，即 $\{\pi^i\}_{i=1}^N$。

在真实通信设定中，每个时间步，每个智能体 i 都会观察其局部状态 s_t^i，并将此信息传达给所有队友。随后，每个智能体利用队友传达的信息构建团队的联合状态 $\boldsymbol{s} \in \mathrm{S}$，然后使用其局部策略 $\pi^i(\boldsymbol{s})$采样一个动作 $a^i \in \mathcal{A}^i$ 并执行。

在本文中，我们研究团队到达-避障问题。也就是说，团队的目标是共同到达由状态组成的目标集 $\mathrm{S}_{\mathcal{T}} \subseteq \mathrm{S}$，同时避开由状态组成的集合 $\mathrm{S}_{\mathcal{A}} \subseteq \mathrm{S}$。因此，集中式规划问题就是求解一组局部策略 $\{\pi^i\}_{i=1}^N$，使团队从初始联合状态 $\boldsymbol{s_I}$出发，到达 $\mathrm{S}_{\mathcal{T}}$ 的概率最大，同时避开 $\mathrm{S}_{\mathcal{A}}$。我们将这一概率值称为成功概率。更正式地说，如果存在某个时刻 M，使得 $\boldsymbol{s}_M \in \mathrm{S}_{\mathcal{T}}$，且对所有

$t < M$, $\boldsymbol{s}_t \notin \mathsf{S}_{\mathcal{A}}$。尽管我们专注于到达-避开问题，我们的框架也可以应用于具有通用奖励的场景。

我们用 $x_{\boldsymbol{s},\boldsymbol{a}}$ 表示状态-动作对（$\boldsymbol{s}$， $\boldsymbol{a}$）的占用度量，即在状态 $\boldsymbol{s}$采取动作 $\boldsymbol{a}$ 的预期次数。类似地，x_{s^i,a^i} 表示智能体 i在状态-动作对 $(s^i、\ a^i)$ 上的占用度量，其中 $x_{s^i,a^i} = \sum_{\boldsymbol{s}^{-i}\in\boldsymbol{\mathcal{S}}^{-i}} \sum_{\boldsymbol{a}^{-i}\in\boldsymbol{\mathcal{A}}^{-i}} x_{(s^i,\boldsymbol{s}^{-i}),(a^i,\boldsymbol{a}^{-i})}$。令 $\mathsf{S}_{\mathcal{D}}$ 表示在任意局部策略组合下，到达 $\mathsf{S}_{\mathcal{T}}$的概率为 0 的状态。以下假设确保每条轨迹都会在有限时间内满足或违反可达性规范。

假设1。状态 $\mathsf{S} \setminus (\mathsf{S}_{\mathcal{T}} \cup \mathsf{S}_{\mathcal{D}})$处的总占用度量有限，即 $\sum_{\boldsymbol{s}\in\boldsymbol{\mathcal{S}}\setminus(\boldsymbol{\mathcal{S}}_{\mathcal{T}}\cup\boldsymbol{\mathcal{S}}_{\mathcal{D}})} \sum_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} x_{\boldsymbol{s},\boldsymbol{a}} < \infty$。

马尔可夫决策过程 $\mathcal{M}^i$ 的一条状态-动作轨迹 ξ^i 是一个序列 $\xi^i = s_0^i a_0^i s_1^i a_1^i \ldots$，满足对所有 $t = 0,\ 1,$ ……，$\mathcal{T}\ s_t^i\ a_t^i\ s_{t+1}^i > 0.$ $\boldsymbol{\xi} = \boldsymbol{s}_0\boldsymbol{a}_0\boldsymbol{s}_1\ldots$ (，，)。我们用 来表示所有智能体的联合状态-动作轨迹，用 $\boldsymbol{\xi}^{-i} = \boldsymbol{s}_0^{-i}\boldsymbol{a}_0^{-i}\boldsymbol{s}_1^{-i}\boldsymbol{a}_1^{-i}\ldots$表示排除智能体 i后的联合状态-动作轨迹。注意， $\boldsymbol{\xi}$和 $\boldsymbol{\xi}^{-i}$都是向量字符串。我们将字符串 len（$\boldsymbol{\xi}$ $=\boldsymbol{s}_0\boldsymbol{a}_0\ldots$）的有效长度定义为 $=\min\{t+1|\boldsymbol{s}_t \in \mathsf{S}_{\mathcal{T}}\cup\mathsf{S}_{\mathcal{D}}\}$。令智能体 i截至时间 t的状态轨迹为 $h_t^i = s_0^i s_1^i \ldots s_t^i$.。我们关注 h_t^i的隐私，以便智能体能够在不泄露敏感信息的情况下执行其策略。

2.2 差分隐私

差分隐私通过一种机制来实施，该机制是一种随机映射。我们以逐代理的方式实施差分隐私，这种方法有时称为“局部差分隐私”。对于相近的局部状态轨迹，机制必须生成近似不可区分的隐私轨迹。“相近”的定义由邻接关系给出，该邻接关系使用 [Schulz和Mihov， 2003] 记作 $d(w,v)$ 的汉明距离；这是一种度量，用于衡量对轨迹 v 进行替换以将其转换为 w 所需的最少替换次数。

定义1（邻接）。固定长度 $T \in \mathbb{N}^+$ 和邻接参数 $k \in \mathbb{N}^+$。对于状态空间为 $\mathcal{S}^i$ 的马尔可夫决策过程，$(\mathcal{S}^i)^T$ 上的邻接关系为 $\mathsf{Adj}_{T,k} = \{(v,w) \in (\mathcal{S}^i)^T \times (\mathcal{S}^i)^T \mid d(v,w) \le k\}.$

该邻接关系规定了哪些轨迹彼此“接近”，因此也规定了差分隐私必须使哪些轨迹对近似不可区分。在（$\mathcal{S}^i$）中，两个长度为 T–^{T} 的局部轨迹在汉明距离小于或等于 k 时相邻。接下来，我们介绍词级差分隐私的定义。该定义保证，在给定一条隐私轨迹的情况下，接收方很难有把握地区分其背后的敏感轨迹与其他相邻轨迹。

定义 2（词级差分隐私 [Chen 等人, 2022]）。固定一个概率空间 $(\Omega\ \mathcal{F}, \mathbb{P})$、一个邻接参数 $k \in \mathbb{N}^+$、一个长度 $T \in \mathbb{N}^+$ 和一个隐私参数$\epsilon > 0$。对于状态空间为 $\mathcal{S}^i$的马尔可夫决策过程，若机制$M\colon (\mathcal{S}^i)^T \times \Omega \to \Delta((\mathcal{S}^i)^T)$ 满足以下条件，则称其为 ϵ-词级差分隐私：对于所有轨迹 $(v,w) \in \mathsf{Adj}_{T,k}$ 以及所有 $L \subseteq (\mathcal{S}^i)^T$，均有 $\mathbb{P}[M(v) \in L] \le e^{\epsilon}\mathbb{P}[M(w) \in L].$

隐私参数 ϵ 控制隐私强度， ϵ 越小意味着隐私性越强。在文献中， ϵ通常取值范围为 0.01 到 10 [Hsu 等人，2014]。

3 问题表述与假设

本节将阐述并分析合作马尔可夫博弈中智能体间通信的隐私保护问题。我们首先给出问题陈述。考虑 N 个智能体参与一个具有到达-避障目标的合作马尔可夫博弈，如 §2.1 所述。

问题 1. 设计一种在线隐私机制，使其在实时情况下为智能体 i 的状态轨迹 $h_t^i = s_1^i s_2^i \ldots s_t^i$ 提供 ϵ-词级差分隐私（定义 2），即在时刻 t. 不需要了解 s_{t+1}^i、s_{t+2}^i、……。该机制应确保隐私轨迹仍符合底层马尔可夫决策过程 $\mathcal{M}^i$ 的动力学。

问题 2. 定义一种算法，用于在私密通信下以分散方式执行局部策略 $\{\pi^i\}_{i=1}^N$。

问题3. 给定一组局部策略 $\{\pi^i\}_{i=1}^N$，求出私密通信 $\boldsymbol{v}^{pr}$ 下成功概率的一个界。利用此界分析多智能体系统中隐私与性能之间的权衡。

问题 4. 综合考虑问题3中分析的权衡，为多智能体系统综合出在强隐私保护下实现高性能的策略。

隐私假设：我们将形式化说明智能体向中央规划器提供哪些信息，以及它们隐藏哪些信息。随后，我们通过一个示例来说明这一设定。

我们假设每个智能体都信任中央规划器为其设计局部策略。每个智能体允许规划器访问其个体马尔可夫决策过程，记为 $\mathcal{M}^i$，对应每个 $i \in [N]$。规划器还了解博弈目标，该目标可以由到达集与避障集 $\mathsf{S}_{\mathcal{T}}$和 $\mathsf{S}_{\mathcal{A}}$ 或奖励函数指定。中央规划器为每个智能体提供一个局部策略 π^i，其中“局部”指从联合状态空间 S 到局部动作空间 $\mathcal{A}^i$的映射。我们假设这些局部策略是平稳的，并且在给定联合状态的情况下，某个智能体的动作分布独立于其他智能体的动作。这意味着中央规划器不会综合出损害隐私的策略，即智能体 i 不会通过采样自己的局部策略 π^i. 而获知其他任何智能体的动作。此外，我们假设初始联合状态 $\boldsymbol{s_I}$是公开信息。

我们还假设智能体彼此并非完全信任。了解其他智能体的转移概率 $\mathcal{T}^i$、动作a^i或奖励，可能会损害该智能体的隐私。不过，每个智能体只需获取其他智能体的状态信息，即可执行其局部策略。因此，我们假设智能体无法获取彼此的转移概率或动作。智能体也无法观察到是否满足或违反了可达-避开规范。每个智能体只能从网络中的其余部分接收私有状态信息，以及来自中央规划器的策略。这可以防止智能体控制其他智能体的敏感状态轨迹。我们还指出，智能体可以了解其他智能体的状态空间和可行状态转移，即

（$\mathcal{T}^i(s^i, a^i)$的支撑集），而不会损害隐私。例如，两名网约车司机都知道对方可能出现的位置，以及彼此如何在这些位置之间移动，但这些知识并不会妨碍司机相互保护自己的位置信息。

最后，本文提出的方法也适用于每个智能体具有不同隐私级别的情况，即 ϵ 的取值不同。不过，为方便起见，我们假设每个智能体具有相同的隐私参数 ϵ.

示例1。在此示例中，敏感信息是两位网约车司机 Alice 和 Bob 的位置。中央规划器为 Alice 和 Bob 生成局部策略，以优化其自身目标。为了保护隐私，Alice 和 Bob 使用差分隐私相互传递各自的位置。例如，他们可以在分享位置数据前对其进行随机化，从而避免泄露真实位置。Alice 和 Bob 无需了解彼此的偏好或约束，只需了解对方传递的私有状态信息。

有了这些私有信息，Alice 和 Bob 随后便可执行中央规划器综合得出的局部策略。然而，由于他们分享的是扰动后的位置数据，局部策略的执行效率可能不如他们能够获取彼此真实位置时那么高。

为减轻这种性能损失，中央规划器可以采用我们论文中提出的方法，综合出一组局部策略，并将隐私对性能的影响纳入考量。这样一来，他们就能在隐私需求与高效执行策略的需求之间取得平衡，并确保尽可能快速地接载乘客，同时保护司机的隐私。

4 使用私密通信实施局部策略

本节将解决问题 1 和问题 2。具体而言，在 §4.1 中，我们修改了 [Chen 等人 2022] 提出的马尔可夫链在线机制，以对马尔可夫决策过程的状态轨迹进行隐私保护。随后，在 §4.2 中，我们详细说明每个智能体如何利用其他智能体的私有状态信息来执行其局部策略。

4.1 实施差分隐私

我们以逐代理的方式实施隐私保护。也就是说，我们为智能体 i 开发一种机制，使其能够实时共享局部状态轨 $h_t^i = s_1^i s_2^i \dots s_t^i \in (\mathcal{S}^i)^t$ 迹，同时满足定义 3 中的 ϵ-词级差分隐私。这里的“实时”是指私有字符串将逐个符号生成。为此，智能体 i 只会共享一条私有状态轨迹 $\tilde{h}_t^i = \tilde{s}_1^i \tilde{s}_2^i \dots \tilde{s}_t^i \in (\mathcal{S}^i)^t$.。为了实时生成 $\tilde{h}_t^i$，智能体 i 会使用在线机制 $M_{h_t^i}$，在每个时间步 t.生成一个单个私有状态 $\tilde{s}_t^i$

每个智能体都需要在每个时间步 t 与其他所有智能体共享其私有状态，t 以便智能体能够执行其策略。然而，定义 2 中的差分隐私保证适用于整个 $T-$长度的状态轨迹。这意味着，尽管智能体在每个时间步都进行通信，我们仍为其整个 $T-$长度的轨迹提供隐私保护。现在，我们定义能够确保状态轨迹满足差分隐私的在线隐私机制。

定义 3（在线机制 [Chen 等人， 2022]）。固定一个概率空间 $(\Omega\ \mathcal{F}, \mathbb{P})$ 和一个马尔可夫决策过程 $\mathcal{M}^i = (\mathcal{S}^i, s_I^i, \mathcal{A}^i, \mathcal{T}^i)$。给定状态轨迹 $h_t^i = s_1^i s_2^i \dots s_t^i \in (\mathcal{S}^i)^t$ 及初始状态 s_I^i，定义在线机制$M_{h_t^i}$，使其生成私有轨迹 $\tilde{h}_t^i = \tilde{s}_1^i \tilde{s}_2^i \dots \tilde{s}_t^i \in (\mathcal{S}^i{}^t \quad \tilde{s}_t^i \quad \mathbb{P}[\tilde{s}_t^i] =)$，该 $\mu_\epsilon^i\ \tilde{s}_t^i | s_t^i\ \tilde{s}_{t-1}^i \quad \mu_\epsilon^i \quad I$ 轨迹从分布 (,) 中采样，其中由算法 计算得出。

算法 1：在线机制构造

输入：真实转移概率 τ_ϵ

输出：μ_ϵ^i

对于$(s_t^i, \tilde{s}_{t-1}^i, \tilde{s}_t^i) \in \mathcal{S}^i \times \mathcal{S}^i \times \mathcal{S}^i$ 执行

如果 $s_t^i = \tilde{s}_t^i$ 且 $\beta(\tilde{s}^i{}_t, \tilde{s}_{t-1}^i) = 1$则

$\mu_\epsilon^i(\tilde{s}_t^i \mid s_t^i, \tilde{s}_{t-1}^i) = \tau_\epsilon(\tilde{s}_{t-1}^i)$.

否则，如果 $s_t^i \neq \tilde{s}_t^i$ 且 $\beta(\tilde{s}_t^i, \tilde{s}_{t-1}^i) = 1$ 则

$\mu_\epsilon^i(\tilde{s}_t^i \mid s_t^i, \tilde{s}_{t-1}^i) = \frac{1-\tau_\epsilon(\tilde{s}_{t-1}^i)\beta(s_t^i, \tilde{s}_{t-1}^i)}{\rho(\tilde{s}_{t-1}^i)-\beta(s_t^i, \tilde{s}_{t-1}^i)}$.

else

$\mu_\epsilon^i(\tilde{s}_t^i \mid s_t^i, \tilde{s}^i{}_{t-1}) = 0$.

在算法 1 中，可行性指示函数 β 针对所有 $s^i, y^i \in \mathcal{S}^i$ 定义为

$$\beta(s^i, y^i) = \begin{cases} 1, \text{ if } \exists a^i \in \mathcal{A}^i\ s.t.\ \mathcal{T}^i(y^i, a^i, s^i) > 0, \\ 0,\ otherwise, \end{cases}$$

出度 ρ 针对每个状态 $s^i \in \mathcal{S}$ 定义为$\rho(s^i) = |\{y^i \in \mathcal{S}^i \mid \exists a^i \in \mathcal{A}^i\ s.t.\ \mathcal{T}^i(s^i, a^i, y^i) > 0\}|$.

定义 3 和算法 1 将隐私机制定义为条件概率分布 μ_ϵ^i.。为了实现该机制，智能体 i 在每个时间步 t. 从概率分布 $\mu_\epsilon^i(\cdot \mid s_t^i, \tilde{s}_{t-1}^i)$中采样得到隐私输出 $\tilde{s}_t^i$。该机制的构造方式确保：当从最近的私有状态 $\tilde{s}_{t-1}^i$出发，$\tilde{s}_t^i$ 可行时，概率$\mu_\epsilon^i(\tilde{s}_t^i \mid s_t^i, \tilde{s}_{t-1}^i)$为正；否则为 0。这样可以防止该机制输出相对于 $\mathcal{M}^i$.动力学而言不可行的隐私轨迹。

当真实的敏感状态 s_t^i 可由先前的隐私输出 $\tilde{s}_{t-1}^i$到达时，机制以 $\tau_\epsilon(\tilde{s}_{t-1}^i)$ 的概率输出 s_t^i，并以均匀概率输出其他任一可行状态，这些概率之和等于 $1 - \tau_\epsilon(\tilde{s}_{t-1}^i)$。我们将时间 t 输出敏感状态 s_t^i 这一事件称为“真实转移”，并将 $\tau_\epsilon(\tilde{s}_{t-1}^i)$ 称为“真实转移概率”。在 §5 中，我们将确立该机制实现 ϵ 词级差分隐私所需满足的要求。

4.2 隐私策略执行

本节将求解问题 2，并定义一种在隐私通信下去中心化执行局部策略 $\{\pi^i\}_{i=1}^N$ 的算法（算法 2）。

由于智能体传递的信息可能是虚假的，没有任何智能体真正了解网络的真实联合状态。因此，在通信内容为私有信息时，智能体网络无法精确执行局部策略。为解决这一问题，每个智能体都会维护对联合状态的估计，并根据该估计自行做出动作决策。在

算法 2：通过私密通信进行策略执行—信

每个智能体 i 的输入：局部策略 π^i
设置 $\tilde{s}_0^i = s_I^i$ 对所有 $i \in [N]$。
对于 $t = 0, 1, \ldots$ 执行
  每个智能体 i 并行执行
    设置 $\hat{s}_{t,i} = (\tilde{s}_t^1, \ldots, \tilde{s}_t^{i-1}, s_t^i, \tilde{s}_t^{i+1} \ldots, \tilde{s}_t^N)$。
    采样一个动作 $a_t^i \sim \pi^i(\hat{s}_{t,i})$.
    执行 a_t^i 并转换为 $s_{t+1}^i \sim \mathcal{T}^i(s_t^i, a_t^i)$。
    共享 $\tilde{s}_{t+1}^i \sim \mu_\epsilon^i(\cdot|s_{t+1}^i, \tilde{s}_t^i)$ 与其他智能体通信。

这项工作中，我们假设每个智能体都将私有信息视为事实，即每个智能体将其接收到的私有信息作为对联合状态的估计。具体而言，在时间步 t，智能体 i 知道自己的局部状态 s_t^i，以及其他每个智能体 j 的私有状态 $\tilde{s}_t^j$。智能体 i 对联合状态的估计表示为 $\hat{s}_{t,i} = (\tilde{s}_t^1, \ldots, \tilde{s}_t^{i-1}, s_t^i, \tilde{s}_t^{i+1}, \ldots, \tilde{s}_t^N)$。由于智能体 i 不知道真实的联合状态 $(s_t^1, \ldots, s_t^N)$，智能体 i 使用其状态估计 $\hat{s}_{t,i}$，从 π^i 中为自己采样一个动作 a_t^i。需要注意的是，在动作选择阶段，智能体之间不会进行通信，因为给定联合状态后，局部策略彼此独立。选定动作 a_t^i 后，智能体 i 执行该动作，并转移到下一个状态 s_{t+1}^i。在下一个时间步 $t+1$，智能体 i 使用 μ_ϵ^i 采样一个私有状态 $\tilde{s}_t^i$，并与其他智能体共享该私有状态。随后，智能体再次采样并执行各自的局部动作。

5 隐私与性能权衡

本节将讨论问题3，并分析使用私密通信执行一组局部策略时性能与隐私之间的权衡。

我们给出以下引理，该引理引自 [Chen 等人， 2022]并证明了由在线机制生成的智能体状态轨迹满足 ϵ-词级差分隐私。

引理1（[Chen 等人， 2022]）。固定长度 $T \in \mathbb{N}^+$、邻接参数 $k \in \mathbb{N}^+$和隐私参数 $\epsilon \geq 0$。定义 3 中的在线机制相对于定义 1 中的邻接关系 $\text{Adj}_{T,k}$ 满足 $\tau_\epsilon(\tilde{s}_{t-1}^i)$ 时，是 ϵ-词级差分隐私的（定义 2）。

$$\tau_\epsilon(\tilde{s}_{t-1}^i) = 1/((\rho(\tilde{s}_{t-1}^i) - 1)e^{-\epsilon/k+1}).$$

在证明算法 2 的差分隐私保证后，我们现在关注性能保证。为了在私密通信下成功，智能体的局部策略应尽可能不受其他智能体状态的影响。换句话说，应使智能体的行为几乎彼此独立。

局部策略的集合会导出一个联合策略 $\boldsymbol{\pi} = \{\pi^i\}_{i=1}^N$。为衡量智能体之间的依赖关系，我们使用一个称为联合策略 [的“总相关性”的量 2022]Karabag等人， 2022]。令 $\boldsymbol{S}_t$ 为随机变量，表示在无隐私化的联合策略 $\boldsymbol{\pi}$ 下，智能体在时间 t 的联合状态；令 $\boldsymbol{A}_t$ 为随机变量，表示智能体在时间 t 的联合动作；令 S_t^i 为随机变量，表示智能体 i 在时间 t 的状态；并令 A_t^i 为一个随机变量，表示智能体 i 在时间 t 的动作。联合策略 $\boldsymbol{\pi} = \{\pi^i\}_{i=1}^N$ 的总相关性 $C_{\boldsymbol{\pi}}$ 为

$$C_{\boldsymbol{\pi}} = \Sigma_{i=1}^N H(S_0^i A_0^i \ldots S_\eta^i) - H(\boldsymbol{S}_0 \boldsymbol{A}_0 \ldots \boldsymbol{S}_\eta) \quad (1)$$

其中 η 表示到达 $\mathsf{S}_\mathcal{T} \cup \mathsf{S}_\mathcal{D}$ 的随机命中时间，也就是根据到达-避障规范 [Karabag 等人， 2022]所定义的轨迹实际终点。

我们得到以下结果，将私密通信下的成功概率与真实通信（即不使用隐私保护）下的成功概率联系起来。该定理的证明见补充材料。

定理 1. 固定隐私参数 $\epsilon > 0$ 和邻接参数 k. 。给定 N 个智能体，它们根据算法 2，通过私密通信执行一组局部策略 $\boldsymbol{\pi} = \{\pi^i\}_{i=1}^N$；设 $\boldsymbol{v}^{pr}$ 为私密通信下的成功概率， $\boldsymbol{v}^{tr}$ 为真实通信（即不使用隐私保护）下的成功概率。那么，

$$\boldsymbol{v}^{pr} \geq \boldsymbol{v}^{tr} - \sqrt{1 - e^{-C_{\boldsymbol{\pi}}} \left((\rho_m - 1)e^{-\epsilon/k} + 1\right)^{Nl^{tr}}}, \quad (2)$$

其中， $C_{\boldsymbol{\pi}}$ 定义见式 (1)， $\rho_m = \max_{i \in [N], s^i \in \mathcal{S}^i} \rho(s^i)$为最大出度，$l^{tr} = \mathbb{E}_{\boldsymbol{\xi} \sim \boldsymbol{\Gamma}^{tr}}[len(\boldsymbol{\xi})]$ 表示在不使用隐私保护的情况下执行 $\boldsymbol{\pi}$ 时的期望联合轨迹长度， Γ^{tr}表示不使用隐私保护时由执行联合策略所诱导的联合轨迹上的概率分布。

定理 1 中的术语 $\left((\rho_m - 1)e^{-\epsilon/k} + 1\right)^{Nl^{tr}}$ 表示私有状态轨迹与真实状态轨迹相同的事件的概率。定理 1 中的术语 $e^{-C_{\boldsymbol{\pi}}}$是一个代理量，用于计入私有状态轨迹与真实状态轨迹不同的事件。在这些事件中，如果局部策略与其他智能体的状态相互独立，智能体仍然可以成功。总相关性越低，智能体之间的依赖性就越低，智能体也就越有可能成功。我们注意到，当智能体如实通信，即 $\epsilon = \infty$，且每个智能体的行为完全独立于其他智能体，即 $C_{\boldsymbol{\pi}} = 0$ 时，式 (2) 中的等号成立。

6 策略综合

在前面的章节中，我们讨论了固定的一组局部策略的执行，并分析了这些策略在私密通信下的性能。现在，我们将介绍一组在私密通信下仍能保持良好性能的局部策略 $\boldsymbol{\pi} = \{\pi^i\}_{i=1}^N$ 的综合方法。

我们的目标是通过最小化定理 1 中给出的 $\boldsymbol{v}^{pr}$下界，在私有通信条件下最大化到达-避障概率。由于该界本身较为复杂，且是其变量的单调函数，因此我们转而求解以下优化问题：

$$\sup_{\boldsymbol{\pi}} \boldsymbol{v}^{tr} - \delta l^{tr} - \beta C_{\boldsymbol{\pi}} \quad (3)$$

其中， $\delta > 0$ 和 $\beta > 0$ 为常数。

利用 $\boldsymbol{\pi}$的平稳性，式（3）给出的优化问题可以用联合状态-动作空间 [的占用度量变量$x_{\boldsymbol{s},\boldsymbol{a}}$ 表示 2022]Karabag等人。我们建议感兴趣的读者参阅 [Karabag等人 2022] ，了解该

优化问题的详细信息。式（3）的目标函数包含占用度量变量的凸函数和凹函数，可以使用凹凸过程求得局部最优解[Lanckriet 和 Sriperumbudur，2009；Yuille 和 Rangarajan，2001]。

求解出占用测度变量的最优 $x^*_{s,\,a}$ 后，我们计算局部策略。回顾一下，我们在 §3 中假设，智能体获得的局部策略 π^i 在给定联合状态时具有相互独立的动作分布。为了计算局部策略 $\pi^i(s, a^i)$，我们对联合占用测度进行边缘化。形式上，有

$$\pi^i(\boldsymbol{s}, a^i) = \left(\Sigma_{\boldsymbol{a}^{-1}\in\boldsymbol{\mathcal{A}}^{-1}}\, x^*_{\boldsymbol{s},(a^i,\boldsymbol{a}^{-i})}\right) / \left(\Sigma_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\, x^*_{\boldsymbol{s},\boldsymbol{a}}\right).$$

值得注意的是，我们也可以在综合过程中强制局部策略在给定联合状态时相互独立，而不是对联合占用变量进行后处理。该过程的详细信息见补充材料。

7 数值实验

数值实验表明，使用 §6 所述过程综合得到的策略具有抵御私有通信的鲁棒性。在每个实验中，我们求解 (3)，为团队中的智能体综合最小依赖局部策略 $\{\pi^i_{MD}\}_{i=1}^N$。我们用 $\boldsymbol{\pi}_{MD}$ 表示这些局部策略并发执行后得到的联合策略，具体说明见 §4.2。

我们将 $\boldsymbol{\pi}_{MD}$ 的性能与一组基准局部策略 $\{\pi^i_{base}\}_{i=1}^N$ 进行比较。这些策略是在不考虑团队总相关性值的情况下，通过优化团队性能合成的。也就是说，通过将 δ 和 β 设为零来求解 (3)。我们用 $\boldsymbol{\pi}_{base}$ 指代由 $\{\pi^i_{base}\}_{i=1}^N$ 并发执行所产生的联合策略。

用于复现所有实验和分析的代码可在 https://github.com/cyrusneary/differential privacy in mas 获取。

7.1 双智能体导航示例

我们首先考虑 [Karabag等人，2022]介绍的多智能体导航示例。两个智能体在共同环境中行动，各自必须在避开队友、避免碰撞的同时，导航至目标位置。为了到达目标位置，智能体必须穿过两个狭窄走廊之一，彼此交错通过。要在不发生碰撞的情况下共同穿过走廊，就需要智能体之间进行协调。

环境由网格单元格构成，每个单元格对应一个独立的局部状态。在任意给定的时间步，每个智能体都会从五种不同的动作中选择一种：向左移动、向右移动、向上移动、向下移动或留在原地。每次执行动作时，智能体都有 0.05 的概率发生滑移，从而移动到一个相邻状态，而非预期的目标状态。

综合 $\boldsymbol{\pi}_{MD}$ 时，我们将式 (3) 中系数 δ 和 β 的值分别设为 0.01 和 0.4。这些取值旨在平衡优化目标中相互竞争的三个项。在构建所有实验中使用的差分隐私机制时，我们将邻接参数固定为 $k = 3$。

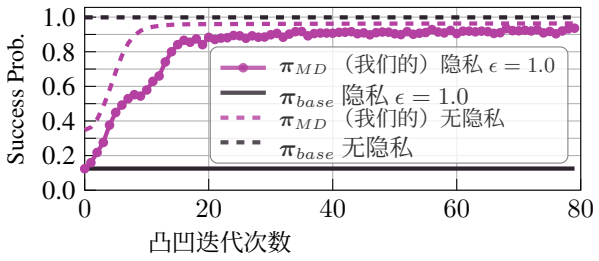

图1：在双智能体导航实验中，任务成功概率随策略综合过程迭代次数的变化。与基线策略相比，$\boldsymbol{\pi}_{MD}$ 即使在私密通信条件下也能取得较高的成功概率。

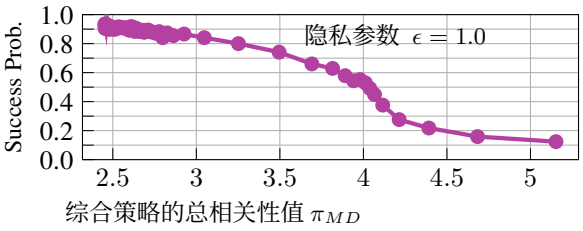

图 2：在私密通信条件下，团队成功概率随综合策略总相关性的变化。

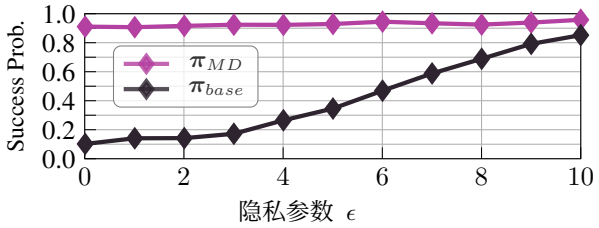

图 3：不同隐私级别下的团队成功概率。隐私参数 ϵ 的值越小，对应的隐私级别越高。

在私密通信条件下，最小依赖策略的性能比基线高 80%。图1展示了 $\boldsymbol{\pi}_{MD}$ 在策略综合过程中的成功概率。我们绘制了私密通信（$\epsilon = 1.0$）和非私密通信（智能体不对其状态轨迹应用隐私机制）两种实现方式下的成功概率。作为对比，我们也绘制了 $\boldsymbol{\pi}_{base}$ 的结果。我们通过模拟每次迭代时策略的 1,000 条运行轨迹，并计算团队到达其目标集的经验成功率，来估计图中的成功概率值。基线策略在非私密通信下的成功概率为 0.99，而在私密通信下则降至 0.13。相比之下，即使在私密通信下，$\boldsymbol{\pi}_{MD}$ 的成功概率仍为 0.92。

在私密通信条件下，总相关性越低，成功概率越高。图2展示了策略综合过程中获得的各个联合策略对应的团队成功概率和总相关性

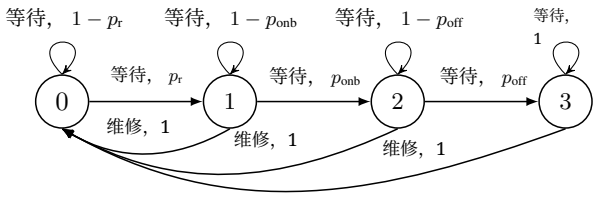

图4：SysAdmin 示例的局部转移动态。标签（a, p）表示在动作 a 下以 p 的概率发生的转移。

。我们观察到，在策略综合过程中，$\boldsymbol{\pi}_{MD}$的总相关性不断降低时，该策略在私密通信条件下的性能会显著提高。这一结果为在策略综合过程中使用总相关性作为正则项提供了有力的实证依据。

$\boldsymbol{\pi}_{MD}$ 的性能对差分隐私机制所强制的隐私级别具有鲁棒性。请注意，参数 ϵ 控制差分隐私机制所强制的隐私强度。ϵ 值越低，对应的隐私级别越强——该机制越有可能扰动智能体的状态轨迹。在图 3 中，我们观察到，无论 ϵ 的值如何，$\boldsymbol{\pi}_{MD}$ 的性能始终保持较高水平。相比之下，$\boldsymbol{\pi}_{base}$ 的性能对 ϵ 高度敏感；在中等到强隐私级别下，其性能会显著下降。

7.2 四智能体系统管理示例

现在我们来考虑一个多智能体系统管理示例的变体，该示例来自 [Guestrin 等人，2003年；Choudhury 等人，2021]。一组服务器必须协调行动，以提供稳定一致的服务，同时执行必要的维护。每台服务器都被建模为一个具有四种局部状态的独立智能体：正常运行 $s^i = 1$、需要维修 $s^i = 2$、正在维修 $s^i = 0$ 和离线 $s^i = 3$。在任意时间步，每个智能体都可以选择继续运行，或启动维修。我们假设智能体的局部转移动力学相互独立，如图4所示。

团队的任务是到达一个目标联合状态，其中所有服务器都处于正常运行状态。然而，我们还施加了额外约束：在运行期间的任意时刻，团队最多允许有两台离线服务器，且最多允许有两台服务器处于维修状态。如果违反其中任一约束，团队就会任务失败。

在此示例中，我们将 $p_{\text{r}} = 0.9$、$p_{\text{onb}} = 0.1$ 和 $p_{\text{off}} =$ 设为 0.1，将策略综合系数 δ和 β 的值分别设为 0.001 和 0.1，并在差分隐私机制中使用 $k = 1$ 的邻接参数。

$\boldsymbol{\pi}_{MD}$ 在各种初始系统配置和隐私级别下的表现始终优于 $\boldsymbol{\pi}_{base}$。图5比较了所提出的最小依赖策略 $\boldsymbol{\pi}_{MD}$与基线 $\boldsymbol{\pi}_{base}$所达到的成功概率。我们在不同隐私级别和各种初始系统配置下测试了团队的性能。我们观察到，在真实通信条件下（实心条形），$\boldsymbol{\pi}_{MD}$ 和 $\boldsymbol{\pi}_{base}$的性能都近乎完美。然而，当通信内容受到隐私保护时，

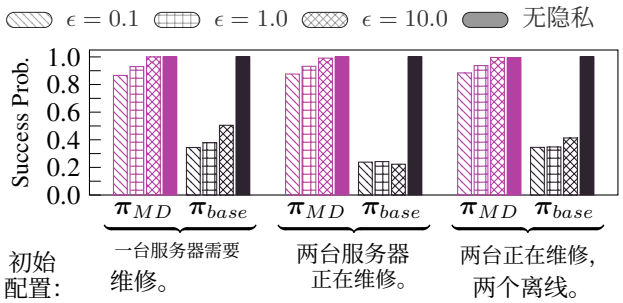

图5：在各种初始系统配置和隐私级别下，系统管理员示例中的成功概率。

$\boldsymbol{\pi}_{MD}$ 的表现始终优于 $\boldsymbol{\pi}_{base}$。在所考虑的初始配置下，即使在隐私级别最高的情况下，$\boldsymbol{\pi}_{MD}$ 的成功概率仍超过87%。

7.3 补充讨论

除了 $\boldsymbol{\pi}_{MD}$ 和 $\boldsymbol{\pi}_{base}$下团队成功概率的差异外，我们还观察到，由这些策略产生的轨迹的预期长度发生了显著变化。例如，在 SysAdmin 实验中采用真实通信时，$\boldsymbol{\pi}_{base}$ 所诱导轨迹的预期长度为 30 至 40 个时间步，具体取决于系统的初始配置。对于 $\boldsymbol{\pi}_{MD}$，这些值则为 3 至 6 个时间步。

这一观察揭示了策略在定性行为上的差异。$\boldsymbol{\pi}_{base}$ 会采取保守行为，通过要求智能体在执行某些动作之前等待特定的联合状态，从而最大限度地提高团队的成功概率；每个智能体的动作都高度依赖于队友所处的确切状态。另一方面，$\boldsymbol{\pi}_{MD}$ 取得的成功概率几乎与 $\boldsymbol{\pi}_{base}$ 相同，但智能体会迅速采取行动，并接受一定程度的风险，以减少其动作对队友状态的依赖。

将总相关性作为正则化项，可以避免策略综合过程为了略微提高成功概率而使智能体之间高度相互依赖。当智能体间通信不完善时，这种权衡就尤为重要，而这在隐私保护多智能体系统中是不可避免的。

最后，我们指出，在某些情况下，可能不存在一组彼此高度独立且能实现高性能的策略。在这种情况下，在隐私通信下，$\boldsymbol{\pi}_{MD}$ 与 $\boldsymbol{\pi}_{base}$ 之间的性能差距可能并不大。不过，即使在这些情况下，总相关性的数值也可以作为一种指标，表明同时实现出色的性能和隐私保护并不可行。

8 结论

本文提出了一个框架，用于在合作式多智能体决策问题中实现智能体间通信的隐私保护。具体而言，我们采用差分隐私机制来保护智能体的符号状态轨迹。我们提供了理论结果，用以分析隐私强度与

团队性能之间的权衡。我们通过降低智能体之间的总相关性，为智能体综合设计鲁棒策略。数值结果表明，在强隐私保护下，最小依赖策略能够实现较高性能；而忽略总相关性的基线策略在私密通信下，团队性能则会大幅下降。

致谢

本研究部分由 AFRL FA9550- 19-1-0169、AFRL FA8651-23-F-A008、ARL ACC-APG- RTP W911NF1920333、ARO W911NF-20-1-0140、NASA 80NSSC21M0071、NSF 1943275 和 ONR N00014-21-1- 2502 资助。

参考文献

[Becker 等人， 2003] Raphen Becker、Shlomo Zilberstein、Victor Lesser 和 Claudia V Goldman。与转移无关的分散式马尔可夫决策过程。载于第 2 届自主智能体与多智能体系统国际会议论文集，第 41–48 页，2003年。[Brafman， 2015] Ronen Israel Brafman。一种保护隐私的多智能体规划与搜索算法。载于第二十四届国际人工智能联合会议，2015年。[Bretagnolle 和 Huber， 1979] Jean Bretagnolle 和 Catherine Huber。密度估计：极小极大风险。《概率论及相关领域》，47(2)：119–137，1979。[Cao 等人， 2013] Yongcan Cao、Wenwu Yu、Wei Ren 和 Guanrong Chen。分布式多智能体协同研究的最新进展综述。《IEEE工业信息学汇刊》，9(1)：427–438，2013。[Chen 等人， 2022] Bo Chen、Kevin Leahy、Austin Jones 和 Matthew Hale。应用于马尔可夫链的符号系统差分隐私。arXiv 预印本 arXiv:2202.03325，2022年。[Choudhury 等人， 2021] Shushman Choudhury、Jayesh K Gupta、Peter Morales 和 Mykel J Kochenderfer。面向多智能体马尔可夫决策过程的可扩展随时规划。arXiv 预印本 arXiv:2101.04788，2021年。

[Cover和Thomas， 1991] Thomas M Cover和Joy A Thomas。《信息论基础》。John Wiley & Sons，纽约，1991年。

[Dwork 等人， 2014] Cynthia Dwork、Aaron Roth 等人。《差分隐私的算法基础》。Foundations and Trends® in Theoretical Computer Science，9(3– 4):211–407，2014年。[Garcelon 等人， 2020] Evrard Garcelon、Vianney Perchet、Ciara Pike-Burke 和 Matteo Pirotta。《强化学习中用于遗憾最小化的局部差分隐私》。arXiv 预印本 arXiv:2010.07778，2020年。[Gohari 等人， 2021] Parham Gohari、Bo Wu、Calvin Hawkins、Matthew Hale 和 Ufuk Topcu。《通过狄利克雷机制实现单位单纯形上的差分隐私》

。IEEE 信息取证与安全汇刊，16:2326–2340，2021年。[Guestrin 等人， 2003] Carlos Guestrin、Daphne Koller、Ronald Parr 和 Shobha Venkataraman。《因子化马尔可夫决策过程的高效求解算法》。《人工智能研究杂志》，19:399–468，2003年。[Hefner 等人， 2022] Tommy Hefner、Guy Shani 和 Roni Stern。《多智能体随机环境中的隐私保护规划》。《自主智能体与多智能体系统》，36(1):1–27，2022年。

[Hsu等人， 2014] Justin Hsu、Marco Gaboardi、Andreas Haeberlen、Sanjeev Khanna、Arjun Narayan、Benjamin C. Pierce和Aaron Roth。差分隐私：一种选择隐私预算参数的经济学方法。2014年IEEE第27届计算机安全基础研讨会，页码398–410，2014年。

[Karabag等人， 2022] Mustafa O. Karabag、Cyrus Neary 和Ufuk Topcu。计划不进行通信：能够抵御通信丢失的多智能体系统。载于第21届国际自主智能体与多智能体系统会议论文集，AAMAS '22，页码705–713，南卡罗来纳州里奇兰，2022年。国际自主智能体与多智能体系统基金会。

[Lanckriet 和 Sriperumbudur， 2009] Gert Lanckriet 和 Bharath K. Sriperumbudur。《关于凹凸过程的收敛性》。载于 Y. Bengio、D. Schuurmans、J. Lafferty、C. Williams 和 A. Culotta 编，《神经信息处理系统进展》第 22 卷。Curran Associates, Inc.，2009 年。

[Nissim 和 Brafman， 2014] Raz Nissim 和 Ronen Brafman。《用于多智能体规划的分布式启发式前向搜索》。《人工智能研究杂志》，51：293–332，2014年。

[Oliehoek 和 Amato， 2016] Frans A Oliehoek 和 Christopher Amato。《分散式部分可观测马尔可夫决策过程简介》。施普林格，Cham，2016 年。

[Parker 等人， 2016] Lynne E. Parker、Daniela Rus 和 Gaurav S. Sukhatme。《多移动机器人系统》，第 1335–1384 页。施普林格国际出版公司，Cham，2016 年。

[Qiao 和 Wang， 2022] Dan Qiao 和 Yu-Xiang Wang。具有差分隐私的离线强化学习。arXiv 预印本 arXiv:2206.00810，2022年。

[Rashid 等人， 2018] Tabish Rashid、Mikayel Samvelyan、Christian Schroeder、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。Qmix：用于深度多智能体强化学习的单调价值函数分解。载于《第 35 届国际机器学习会议论文集》，第 4295–4304 页，2018年。

[Schulz和Mihov， 2003] Klaus U. Schulz 和 Stoyan Mihov。使用 Levenshtein 自动机进行快速字符串纠错。发表于《国际文档分析与识别杂志》(IJDAR)，5(1):67–85，2003年。

[Son 等人， 2019] Kyunghwan Son、Daewoo Kim、Wan Ju Kang、David Earl Hostallero 和 Yung Yi。Qtran：通过变换学习因子分解，以实现协作式多智能体强化学习。发表于第 36 届国际机器学习大会论文集，第 5887–5896 页，2019 年。[Such 等人， 2014] Jose M. Such、Agust´ın Espinosa 和 Ana Garc´ıa-Fornes。多智能体系统中的隐私综述。《知识工程评论》，29(3):314–344，2014年。[Ye 等人， 2022] Dayong Ye、Tianqing Zhu、Sheng Shen、Wanlei Zhou 和 Philip S. Yu。面向物流类问题的差分隐私多智能体规划。IEEE《可信与安全计算汇刊》，19(2):1212–1226，2022年。[Yuille 和 Rangarajan， 2001] Alan L Yuille 和 Anand Rangarajan。凹凸过程（CCCP）。收录于 T. Dietterich、S. Becker 和 Z. Ghahramani 编，《神经信息处理系统进展》第 14 卷。麻省理工学院出版社，2001 年。

合作式多智能体规划中的差分隐私：补充材料

A 理论结果的证明

离散概率分布 Q^1 和 Q^2 的 Kullback-Leibler（KL）散度 [Cover和Thomas, 1991]，其支撑集分别为 $\boldsymbol{Q}^1$ 和 $\boldsymbol{Q}^2$，定义为

$$KL\left(Q^1 \| Q^2\right)=\sum_{q \in \boldsymbol{Q}^1} Q^1(q) \log\left(\frac{Q^1(q)}{Q^2(q)}\right).$$

记号 我们首先定义一些证明中会用到的记号。设 $\boldsymbol{S}_t$ 为一个随机变量，表示在未进行隐私化的联合策略下，智能体在时间 t 的联合状态； $\boldsymbol{A}_t$ 为一个随机变量，表示智能体在时间 t 的联合动作； S_t^i 为一个随机变量，表示智能体 i 在时间 t 的状态； A_t^i 为一个随机变量，表示智能体 i 在时间 t 的动作。 $\boldsymbol{S}_t^{-i}$ 为一个随机变量，表示智能体 i 的队友（不包括智能体 i 本身）在时间 t 的状态； $\boldsymbol{A}_t^{-i}$ 为一个随机变量，表示智能体 i 的队友（不包括智能体 i 本身）在时间 t 的动作。联合策略 $\boldsymbol{\pi}$ 的总相关性 $C_{\boldsymbol{\pi}}$ 为

$$C_{\boldsymbol{\pi}}=\sum_{i=1}^N H(S_0^i A_0^i \ldots S_\eta^i)-H(\boldsymbol{S}_0 \boldsymbol{A}_0 \ldots \boldsymbol{S}_\eta)$$

其中， η 表示随机 命中时间，即到达性规约意义下轨迹的有效终点

规约 [Karabag等人，2022]。

令 W表示所有终止于 $\mathrm{S}_{\mathcal{T}} \cup \mathrm{S}_{\mathcal{D}}$ 中某个状态的轨迹片段，即 $W=\{\boldsymbol{w}=\boldsymbol{s_0 a_0} \ldots \boldsymbol{s_T} | \boldsymbol{s_T} \in \mathrm{S}_{\mathcal{T}} \cup \mathrm{S}_{\mathcal{D}}$和 $\forall t<T, \boldsymbol{s_t} \notin \mathrm{S}_{\mathcal{T}} \cup \mathrm{S}_{\mathcal{D}}\}$，并令 $W^{\prime}$表示所有永远不会到达 $\mathrm{S}_{\mathcal{T}} \cup \mathrm{S}_{\mathcal{D}}$ 的轨迹，即 $W^{\prime}=\{w=\boldsymbol{w}=\boldsymbol{s_0 a_0} \ldots | \forall t \geq 0, \boldsymbol{s_t} \notin \mathrm{S}_{\mathcal{T}} \cup \mathrm{S}_{\mathcal{D}}\}$。注意，每条轨迹要么以 W 中的一个轨迹片段开始，要么属于 $W^{\prime}$。此外，令 $R \subseteq W \cup W^{\prime}$表示所有终止于 $\mathrm{S}_{\mathcal{T}}$ 中某个状态的轨迹片段，即 $R=\{\boldsymbol{w}=\boldsymbol{s_0 a_0} \ldots \boldsymbol{s_T} | \boldsymbol{s_T} \in \mathrm{S}_{\mathcal{T}}$和 $\forall t<T, \boldsymbol{s_t} \notin \mathrm{S}_{\mathcal{T}} \cup \mathrm{S}_{\mathcal{A}}\}$。

令 Γ^{tr} 表示在真实通信（即无隐私）下执行联合策略所诱导的联合轨迹分布。另令 Γ^{pr} 表示在实施隐私保护时的联合轨迹分布。令 v^{tr} 表示真实通信下的成功概率，令 v^{pr} 表示私密通信下的成功概率。

我们用 μ^{tr} 表示真实通信下联合策略所对应的实际（有限或无限）状态-动作过程上的概率测度。 μ^{pr} 表示私密通信下联合策略所对应的实际（有限或无限）状态-动作过程上的概率测度。为简化记法，我们也用 μ_ϵ 表示给定实际状态轨迹时，私有状态轨迹上的条件概率测度。

令 $\boldsymbol{w}=\boldsymbol{s}_0 \boldsymbol{a}_0 \boldsymbol{s}_1 \boldsymbol{a}_1 \ldots \boldsymbol{s}_T \in(\mathrm{S} \times \mathrm{A})^T$表示一个联合轨迹片段，令 $\tilde{\boldsymbol{w}}=\tilde{\boldsymbol{s}}_0 \tilde{\boldsymbol{s}}_1 \ldots \tilde{\boldsymbol{s}}_T \in \mathrm{S}^T$表示一个私有联合状态轨迹片段。我们用 $\hat{\boldsymbol{s}}_{\boldsymbol{t}}^{\boldsymbol{j}}=\{\tilde{s}_t^0, \ldots, \tilde{s}_t^{j-1}, s_t^j, \tilde{s}_t^{j+1}, \ldots, \tilde{s}_t^N\}$ 表示智能体 j对私有联合状态的副本。

对符号集合 V应用克林闭包所得的集合 $V^*=\cup_{i \geq 0} V^i$，是所有有限长度词的集合，其中 $V^0=\{\Lambda\}$，且 Λ 是空字符串。所有无限长度词的集合记作 V^ω。W

e introduce t he following lemma, to be used in the proof of other theoretical results.

引理 2。

$$\mathbb{E}_{\boldsymbol{w} \sim \mu^{tr}}\left[\log \mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w} | \boldsymbol{w})\right] \geq-N \log\left(\left(\rho_{\max }-1\right) \exp(-\frac{\epsilon}{k})+1\right) l^{tr}$$

其中， $\rho(s_{t-1}^i)$ 是 s_{t-1}^i 的出度，且 $\rho_{\max }=\max_{s \in \cup_{i=1}^N \mathcal{S}^i} \rho(s)$。

引理 2 的证明。由于在线隐私机制（算法 1）的马尔可夫性以及智能体之间的独立性，我们有

$$\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w} | \boldsymbol{w}=\boldsymbol{s}_0 \ldots \boldsymbol{s}_T)=\prod_{t=0}^{T-1} \prod_{i=1}^N \mu_\epsilon(\tilde{s}_t^i=s_t^i | s_t^i, \tilde{s}_{t-1}^i)$$

注意，如果 $\tilde{\boldsymbol{w}}=\boldsymbol{w}$，那么对于所有 $t \geq 0$ 和 $j \in[N]$，都有 $\hat{\boldsymbol{s}}_t^j=\boldsymbol{s}_t$，即每个智能体的私有状态副本始终与实际的联合状态一致。因此，

$$\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w} | \boldsymbol{w}=\boldsymbol{s}_0 \ldots \boldsymbol{s}_T)=\prod_{t=0}^{T-1} \prod_{i=1}^N \mu_\epsilon(\tilde{s}_t^i=s_t^i | s_t^i, \tilde{s}_{t-1}^i=s_{t-1}^i)$$

由 [Chen 等人，2022年，定理 7], 可得

$$\mu_\epsilon(\tilde{s}_t^i=s_t^i | s_t^i, \tilde{s}_{t-1}^i=s_{t-1}^i)=\frac{1}{\left(\rho(s_{t-1}^i)-1\right) \exp(-\frac{\epsilon}{k})+1},$$

其中， $\rho(s_{t-1}^i)$ 是 s_{t-1}^i. 的出度。令 $\rho_{\max} = \max^{s \in \cup_{i=1}^N \mathcal{S}^i} \rho(s)$，则有

$$\mu_\epsilon(\tilde{s}_t^i = s_t^i | s_t^i, \tilde{s}_{t-1}^i = s_{t-1}^i) \geq \frac{1}{(\rho_{\max} - 1)\exp(-\frac{\epsilon}{k}) + 1}.$$

由此可得

$$\begin{aligned}
\log \mu_\epsilon(\tilde{\boldsymbol{w}} = \boldsymbol{w} | \boldsymbol{w} = s_0 \dots s_T) &= \log \left(\prod_{t=0}^{T-1} \prod_{i=1}^{N} \mu_\epsilon(s_t^i | s_t^i, \tilde{s}_{t-1}^i = s_{t-1}^i) \right) \\
&= \sum_{t=0}^{T-1} \sum_{i=1}^{N} \log \mu_\epsilon(s_t^i | s_t^i, \tilde{s}_{t-1}^i = s_{t-1}^i) \\
&\geq \sum_{t=0}^{T-1} \sum_{i=1}^{N} \log \frac{1}{(\rho_{\max} - 1)\exp(-\frac{\epsilon}{k}) + 1} \\
&= \sum_{t=0}^{T-1} -N \log \left((\rho_{\max} - 1)\exp(-\frac{\epsilon}{k}) + 1 \right).
\end{aligned}$$

因此，

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{w} \sim \mu^{tr}} \left[\log \mu_\epsilon(\tilde{\boldsymbol{w}} = \boldsymbol{w} | \boldsymbol{w}) \right] &= \sum_{\boldsymbol{w} \in W} \mu^{tr}(\boldsymbol{w}) \log \left(\prod_{t=0}^{T-1} \prod_{i=1}^{N} \mu_\epsilon(s_t^i | s_t^i, s_{t-1}^i) \right) \\
&\geq \sum_{\boldsymbol{w} \in W} -\mu^{tr}(\boldsymbol{w}) \sum_{t=0}^{T-1} N \log \left((\rho_{\max} - 1)\exp(-\frac{\epsilon}{k}) + 1 \right) \\
&= E \left[\sum_{t=0}^{\tau-1} -N \log \left((\rho_{\max} - 1)\exp(-\frac{\epsilon}{k}) + 1 \right) | \mu^{tr} \right] \\
&= -N \log \left((\rho_{\max} - 1)\exp(-\frac{\epsilon}{k}) + 1 \right) E \left[\sum_{t=0}^{\tau-1} 1 | \mu^{tr} \right] \\
&= -N \log \left((\rho_{\max} - 1)\exp(-\frac{\epsilon}{k}) + 1 \right) l^{tr}.
\end{aligned}$$

□

定理 1 的证明。由于唯一机制（算法 1）和联合策略执行（算法 2）具有因果性，我们有

$$\begin{aligned}
\mu^{pr}(\boldsymbol{w}) &= \sum_{\tilde{\boldsymbol{w}} \in \boldsymbol{\mathcal{S}}^T} \mu^{pr}(\boldsymbol{w}, \tilde{\boldsymbol{w}}) \\
&= \sum_{\tilde{\boldsymbol{w}} \in \boldsymbol{\mathcal{S}}^T} \prod_{t=0}^{T-1} \Pr(\boldsymbol{a}_t \boldsymbol{s}_{t+1}, \tilde{\boldsymbol{s}}_t | \boldsymbol{a}_{t-1} \boldsymbol{s}_t \dots \boldsymbol{a}_0 \boldsymbol{s}_1, \tilde{\boldsymbol{s}}_{t-1} \dots \tilde{\boldsymbol{s}}_0),
\end{aligned}$$

其中,

$$
\begin{aligned}
\Pr(\boldsymbol{a}_t\boldsymbol{s}_{t+1}, \tilde{\boldsymbol{s}}_t|\boldsymbol{a}_{t-1}\boldsymbol{s}_t \dots \boldsymbol{a}_0\boldsymbol{s}_1, \tilde{\boldsymbol{s}}_{t-1} \dots \tilde{\boldsymbol{s}}_0) &= \Pr(\boldsymbol{a}_t\boldsymbol{s}_{t+1}, \tilde{\boldsymbol{s}}_t|\boldsymbol{s}_t, \tilde{\boldsymbol{s}}_{t-1}) && (4)\\
&= \prod_{i=1}^{N} \Pr(a_t^i s_{t+1}^i, \tilde{\boldsymbol{s}}_t|\boldsymbol{s}_t, \tilde{\boldsymbol{s}}_{t-1}) && (5)\\
&= \prod_{i=1}^{N} \Pr(a_t^i s_{t+1}^i|\boldsymbol{s}_t, \tilde{\boldsymbol{s}}_t, \tilde{\boldsymbol{s}}_{t-1}) \Pr(\tilde{\boldsymbol{s}}_t|\boldsymbol{s}_t, \tilde{\boldsymbol{s}}_{t-1})\\
&= \prod_{i=1}^{N} \Pr(a_t^i s_{t+1}^i|\boldsymbol{s}_t, \tilde{\boldsymbol{s}}_t, \tilde{\boldsymbol{s}}_{t-1}) \left(\prod_{k=1}^{N} \mu_\epsilon(\tilde{s}_t^k|s_t^k, \tilde{s}_{t-1}^k)\right) && (6)\\
&= \prod_{i=1}^{N} \Pr(a_t^i s_{t+1}^i|\hat{\boldsymbol{s}}_t^i) \left(\prod_{k=1}^{N} \mu_\epsilon(\tilde{s}_t^k|s_t^k, \tilde{s}_{t-1}^k)\right) && (7)\\
&= \prod_{i=1}^{N} \mathcal{T}(s_t^i, a_t^i, s_{t+1}^i)\pi^i(\hat{\boldsymbol{s}}_t^i, a_t^i) \left(\prod_{k=1}^{N} \mu_\epsilon(\tilde{s}_t^k|s_t^k, \tilde{s}_{t-1}^k)\right).
\end{aligned}
$$

公式 (4) 源于马尔可夫性。公式 (5) 源于每个智能体都在独立地选择其下一动作和状态。公式 (6) 源于每个状态都在独立地产生其私有状态。公式 (7) 源于对于每个智能体 i 而言，其真实下一状态 $\tilde{s}_{t+1}^i$ 与其他状态的真实状态以及私有状态本身相互独立。

因此,

$$
\begin{aligned}
\mu^{pr}(\boldsymbol{w}) &= \sum_{\tilde{\boldsymbol{w}} \in \boldsymbol{\mathcal{S}}^T} \prod_{t=0}^{T-1} \prod_{i=1}^{N} \mathcal{T}(s_t^i, a_t^i, s_{t+1}^i)\pi^i(\hat{\mathbf{s}}_t^i, a_t^i) \left(\prod_{i=1}^{N} \mu_\epsilon(\tilde{s}_t^i|s_t^i, \tilde{s}_{t-1}^i)\right)\\
&\geq \prod_{t=0}^{T-1} \prod_{i=1}^{N} \mathcal{T}(s_t^i, a_t^i, s_{t+1}^i)\pi^i(\hat{\mathbf{s}}_t^i, a_t^i) \left(\prod_{i=1}^{N} \mu_\epsilon(\tilde{s}_t^i|s_t^i, \tilde{s}_{t-1}^i)\right), \forall \tilde{\boldsymbol{w}} \in \boldsymbol{\mathcal{S}}^T, \qquad (8)
\end{aligned}
$$

其中，等式 (8) 成立是因为所有可能的私有状态轨迹的概率之和必然大于任意单条私有状态轨迹的概率。我们只考虑 $\tilde{\boldsymbol{s}}_t = \boldsymbol{s}_t$ 的情况，这意味着私有在线机制将在每个时刻 t 都做出正确决策。因此,

$$
\begin{aligned}
\mu^{pr}(\boldsymbol{w}) &\geq \prod_{t=0}^{T-1} \prod_{i=1}^{N} \mathcal{T}(s_t^i, a_t^i, s_{t+1}^i)\pi^i(\boldsymbol{s}_t, a_t^i) \left(\prod_{i=1}^{N} \mu_\epsilon(s_t^i|s_t^i, s_{t-1}^i)\right)\\
&= \mu^{tr}(\boldsymbol{w}) \left(\prod_{i=1}^{N} \mu_\epsilon(s_t^i|s_t^i, s_{t-1}^i)\right)
\end{aligned}
$$

现在我们来看以下 KL 散度：

$$
\begin{aligned}
KL(\mathbf{\Gamma}^{tr}||\mathbf{\Gamma}^{pr}) &= \sum_{\boldsymbol{w}\in W\cup W'} \mu^{tr}(\boldsymbol{w}) \log\left(\frac{\mu^{tr}(\boldsymbol{w})}{\mu^{pr}(\boldsymbol{w})}\right) \\
&= \sum_{\boldsymbol{w}\in W} \mu^{tr}(\boldsymbol{w}) \log\left(\frac{\mu^{tr}(\boldsymbol{w})}{\mu^{pr}(\boldsymbol{w})}\right) \qquad (9)\\
&\leq \sum_{\boldsymbol{w}\in W} \mu^{tr}(\boldsymbol{w}) \log\left(\frac{\mu^{tr}(\boldsymbol{w})}{\mu^{tr}(\boldsymbol{w})\left(\prod_{t=0}^{\infty}\prod_{i=1}^{N}\mu_\epsilon(s_t^i|s_t^i,s_{t-1}^i)\right)}\right) \\
&= \sum_{\boldsymbol{w}\in W} \mu^{tr}(\boldsymbol{w}) \log(\mu^{tr}(\boldsymbol{w})) - \sum_{\boldsymbol{w}\in W} \mu^{tr}(\boldsymbol{w}) \log(\mu^{tr}(\boldsymbol{w})) \\
&\quad - \sum_{\boldsymbol{w}\in W} \mu^{tr}(\boldsymbol{w}) \log\left(\prod_{t=0}^{\infty}\prod_{i=1}^{N}\mu_\epsilon(s_t^i|s_t^i,s_{t-1}^i)\right) \\
&= H(\boldsymbol{S}_0\boldsymbol{A}_0\ldots\boldsymbol{S}_\eta) - H(\boldsymbol{S}_0\boldsymbol{A}_0\ldots\boldsymbol{S}_\eta) - \sum_{\boldsymbol{w}\in W} \mu^{tr}(\boldsymbol{w}) \log\left(\prod_{t=0}^{T-1}\prod_{i=1}^{N}\mu_\epsilon(s_t^i|s_t^i,s_{t-1}^i)\right) \\
&\leq \sum_{i=1}^{N} H(S_0^iA_0^i\ldots S_\eta^i) - H(\boldsymbol{S}_0\boldsymbol{A}_0\ldots\boldsymbol{S}_\eta) - \sum_{\boldsymbol{w}\in W} \mu^{tr}(\boldsymbol{w}) \log\left(\prod_{t=0}^{T-1}\prod_{i=1}^{N}\mu_\epsilon(s_t^i|s_t^i,s_{t-1}^i)\right) \qquad (10)\\
&= C_{\boldsymbol{\pi}} - \sum_{\boldsymbol{w}\in W} \mu^{tr}(\boldsymbol{w}) \log\left(\prod_{t=0}^{T-1}\prod_{i=1}^{N}\mu_\epsilon(s_t^i|s_t^i,s_{t-1}^i)\right) \qquad (11)\\
&= C_{\boldsymbol{\pi}} - \mathbb{E}_{\boldsymbol{w}\sim\mu^{tr}}\left[\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w})\right]
\end{aligned}
$$

其中，(9) 成立是因为 $\sum_{\boldsymbol{w}\in W'}\mu^{tr}(\boldsymbol{w}) = 0$，(10) 成立是因为熵的次可加性，(11) 成立是根据 $C_{\boldsymbol{\pi}}$ 的定义。在 (11) 中使用引理 2 可得

$$
KL(\mathbf{\Gamma}^{tr}||\mathbf{\Gamma}^{pr}) \leq C_{\boldsymbol{\pi}} + N\log\left((\rho_{\max}-1)\exp(-\frac{\epsilon}{\ell})+1\right)l^{tr}. \qquad (12)
$$

最后，我们证明 $v^{pr} \geq v^{tr} - 1 + \exp(-C_{\boldsymbol{\pi}})\left((\rho_{max}-1)\exp\left(-\frac{\epsilon}{\ell}\right)+1\right)^{Nl^{tr}}/2$. 令 $R' \subseteq W\cup W'$ 为任意集合。

$$
\begin{aligned}
\boldsymbol{v}^{tr} - \boldsymbol{v}^{pr} &= \sum_{\boldsymbol{w}\in R} \mu^{tr}(\boldsymbol{w}) - \mu^{pr}(\boldsymbol{w}) \\
&\leq \left|\sum_{\boldsymbol{w}\in R} \mu^{tr}(\boldsymbol{w}) - \mu^{pr}(\boldsymbol{w})\right| \\
&\leq \sup_{R'}\left|\sum_{\boldsymbol{w}\in R'} \mu^{tr}(\boldsymbol{w}) - \mu^{pr}(\boldsymbol{w})\right| \\
&\leq \sqrt{1-\exp(-KL(\mathbf{\Gamma}^{tr}||\mathbf{\Gamma}^{pr}))} \qquad (13a)
\end{aligned}
$$

其中，(13a) 根据 Bretagnolle-Huber 不等式 [Bretagnolle 和 Huber, 1979] 得出。重新整理 (13a) 中的各项并利用 (12)，即可得到所需结果。

□

我们注意到，除了定理 1 外，我们还可以推导出一个关于 v^{pr} 的更紧下界。

定理2。给定 $\epsilon > 0$，对于 N 个智能体，我们有

$$
v^{pr} \geq v^{tr} - 1 + \left((\rho_{\max}-1)\exp(-\frac{\epsilon}{k})+1\right)^{Nl^{tr}}. \qquad (14)
$$

定理 2 的证明。如定理 1 的证明所示，我们有

$$\begin{aligned} v^{pr} &= \sum_{\boldsymbol{w}=\boldsymbol{s}_0\boldsymbol{a}_0\boldsymbol{s}_1\boldsymbol{a}_1\ldots\boldsymbol{s}_T\in W} \mu^{pr}(\boldsymbol{w})\mathbb{1}(\boldsymbol{w}\in R) \\ &\geq \sum_{\boldsymbol{w}=\boldsymbol{s}_0\boldsymbol{a}_0\boldsymbol{s}_1\boldsymbol{a}_1\ldots\boldsymbol{s}_T\in W} \mu^{tr}(\boldsymbol{w})\mathbb{1}(\boldsymbol{w}\in R)\left(\prod_{t=0}^{T-1}\prod_{k=1}^{N}\mu_\epsilon(s_t^k|s_t^k,s_{t-1}^k)\right) \\ &= \Pr(\boldsymbol{w}\in R\wedge\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w}\sim\mu^{tr},\tilde{\boldsymbol{w}}\sim\mu^{\epsilon}(\cdot|\boldsymbol{w})). \end{aligned}$$

根据并集界，我们有

$$\begin{aligned} v^{pr} &\geq \Pr(\boldsymbol{w}\in R|\boldsymbol{w}\sim\mu^{tr})+\mathbb{E}_{\boldsymbol{w}\sim\mu^{tr}}\left[\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w})\right]-1 \\ &= v^{tr}+\mathbb{E}_{\boldsymbol{w}\sim\mu^{tr}}\left[\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w})\right]-1 \end{aligned}$$

接着，结合

$$\mathbb{E}_{\boldsymbol{w}\sim\mu^{tr}}\left[\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w})\right]=\sum_{\boldsymbol{w}\in W}\mu^{tr}(\boldsymbol{w})\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w})$$

詹森不等式，我们有

$$\begin{aligned} \mathbb{E}_{\boldsymbol{w}\sim\mu^{tr}}\left[\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w})\right] &= \exp\left(\log\sum_{\boldsymbol{w}\in W}\mu^{tr}(\boldsymbol{w})\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w})\right) \\ &\geq \exp\left(\sum_{\boldsymbol{w}\in W}\mu^{tr}(\boldsymbol{w})\log\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w})\right) \\ &= \exp\left(\mathbb{E}_{\boldsymbol{w}\sim\mu^{tr}}\left[\log\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w})\right]\right). \end{aligned}$$

利用引理 2，可得

$$\begin{aligned} v^{pr} &\geq v^{tr}-1+\exp\left(\mathbb{E}_{\boldsymbol{w}\sim\mu^{tr}}\left[\log\mu_\epsilon(\tilde{\boldsymbol{w}}=\boldsymbol{w}|\boldsymbol{w})\right]\right) \\ &\geq v^{tr}-1+\left((\rho_{\max}-1)\exp(-\frac{\epsilon}{k})+1\right)^{Nl^{tr}}, \end{aligned}$$

由此完成证明。

□

与 (2) 相比，(14) 未将总相关性 $C_{\boldsymbol{\pi}}$ 考虑在内，仅关注私有状态轨迹与原始状态轨迹相同时的成功概率。因此，通过最小化 (14) 中的下界合成的联合策略 $\boldsymbol{\pi}=\{\pi^i\}_{i=1}^N$ 并不具备最小化 (2) 所带来的鲁棒性。在目标函数中纳入总相关性，可提升私密通信下的团队性能，因为智能体的策略对彼此的状态轨迹不那么敏感。

B 局部策略独立性假设的详细说明

在本文中，我们假设给定联合状态后，智能体的局部策略彼此独立。可以在合成过程中通过以下约束来强制满足这一假设，

$$\sum_{i=1}^{N}H(A_t^i|\boldsymbol{S}_t)=H(\boldsymbol{A}_t^i|\boldsymbol{S}_t)\text{ for all }t=0,1,\ldots,\eta. \tag{15}$$

该约束意味着，给定联合状态后，智能体的动作分布彼此独立。由于策略具有平稳性，我们可以将式 (15) 重写为

$$\sum_{i=1}^{N}\sum_{a^i\in\mathcal{A}^i}x_{\boldsymbol{s},a^i}\log\left(\frac{\sum_{b^i\in\mathcal{A}^i}x_{\boldsymbol{s},b^i}}{x_{\boldsymbol{s},a^i}}\right)=\sum_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}x_{\boldsymbol{s},\boldsymbol{a}}\log\left(\frac{\sum_{\boldsymbol{b}\in\boldsymbol{\mathcal{A}}}x_{\boldsymbol{s},\boldsymbol{b}}}{x_{\boldsymbol{s},\boldsymbol{a}}}\right)$$

对所有 $\boldsymbol{s}\in\mathsf{S}$ 均成立。我们注意到，等式两边都是占用度量变量的凹函数。与我们所考虑的目标函数类似，我们可以采用凸凹过程来处理这一约束。