arXiv:2302.01193v1 [cs.LG] 2 Feb 2023

模仿严谨的专家，避免灾难性事件

Jack R. P. Hanslope
系 f C计算机 S
科学 布里斯托大学 布里斯托，英国 jack.hanslope@bristol.ac.uk

Laurence Aitchison
计算机科学系，布里斯托大学，英国布里斯托 laurence.aitchison@bristol.ac.uk

摘要

强化学习正越来越多地用于控制与人类密切互动的机器人系统。这种互动带来了安全强化学习方面的问题：如何确保由强化学习控制的机器人系统绝不会伤害人类，例如。本问题在复杂、真实的环境中尤其具有挑战性，因为在这种环境下，甚至无法明确写出一个能够涵盖这些结果的奖励函数。在这种情况下，也许唯一可行的方法是基于逆向强化学习，从人类的示范中推断奖励。然而，逆向强化学习存在严重的欠定性，因为许多不同的奖励都可能产生相同的最优策略；我们表明，这使得区分灾难性结果（如伤害人类）与单纯的不良结果变得困难。我们的关键洞见是，当灾难性结果有可能发生时，人类确实会表现出不同的行为：他们会变得谨慎得多。我们将谨慎信号纳入逆向强化学习，并发现这些信号确实能够帮助逆向强化学习区分不良结果与灾难性结果，这对于确保未来现实世界中的人机交互安全至关重要。

1 引言

工业界正越来越多地尝试取消机器人与人类之间传统的安全屏障，转而采用紧密协作的方式（Galin 等，2020年；Moreno 等，2020年；Simoes 等，2019年；Michalos 等，2022年；Hjorth & Chrysostomou，2022年）。然而，这种紧密协作也带来了严重风险，因为工业机器人可能会、而且有时确实会严重伤害人类（Henley，2022年）。这引出了一个问题：我们能否设计出由强化学习驱动的机器人控制系统，使其绝不采取会造成灾难性后果的糟糕动作，例如伤害人类？在丰富的环境中，这尤其具有挑战性，因为很难严格定义或模拟所有灾难性结果。在这种情况下，或许唯一可行的方法是采用某种形式的模仿学习（泛指通过观察专家进行学习）（Ho & Ermon，2016年；Paine 等，2018年；Peng 等，2018年a,b）。具体而言，我们考察逆强化学习（IRL），即推断驱动专家行为的奖励函数（Russell，1998年；Ng & Russell，2000年；Ziebart 等，2008年；Finn 等，2016年；Peng 等，2018年b）。

然而，逆向强化学习存在一个根本问题：通常有许多奖励函数都与观察到的行为相符（Ng & Russell，2000年；Ramachandran & Amir，2007年）。很多时候，我们无法区分灾难性事件与仅仅令人不快的事件。

相比之下，人类和其他动物只需少量样本（通常仅一个），并且只需观察例如父母或教师对潜在灾难性事件的反应，无须亲眼目睹灾难性事件本身，就能学会躲避灾难性事件。例如，幼年动物只要观察到父母面对捕食者时表现出恐惧反应，就会学会害怕这种捕食者（它们不必亲眼看到另一只动物遭到捕食）（Mineka & Cook, 1993年；Askew & Field, 2008年；Dunne & Askew, 2013年；Reynolds et al., 2018年；Marin et al., 2020年）。另一种情况是，当机器人面临灾难性结果的风险时，人类操作者会更加谨慎；例如，当机器人在人类附近运行时，他们可能会让机器人减速。

NeurIPS 2022机器人学习研讨会：可信机器人学，线上，线上

我们应该能够利用这些谨慎信号，判断人类操作者何时认为存在灾难性结果的风险。

我们设计了一种融入谨慎信号的逆向强化学习框架，并发现该框架能够区分灾难性结果与不良结果；在实践中，若不考虑谨慎度，这两者便无法区分。我们首先在网格世界中开展了一项玩具实验。该网格世界的一侧有一处悬崖，坠下悬崖代表发生灾难性事件。起初，我们让了解应遵循的最优策略的计算机专家在这个网格世界中进行模拟。随后，我们让一位充分了解环境和最优策略的人类玩家再次进行相同的模拟。我们发现，在这两种情形下，融入谨慎度的框架都能更好地理解坠下悬崖的严重程度。

2 相关工作

关于安全强化学习（Alshiekh et al., 2018; Saunders et al., 2017; Berkenkamp et al., 2017; Achiam et al.,2017; Garcia &Fernández,2012）、安全模仿学习（Menda et al., 2017, 2019）和安全逆向强化学习（Brown et al., 2018; Brown & Niekum, 2018），已有大量研究。然而，这些研究均未利用这一关键洞见：人类的谨慎度或许可以用来区分灾难性动作与仅仅是不良的动作。

3 种方法

网格世界环境。 我们使用一个四行六列的网格世界环境（Sutton & Barto，2018年），其中一条边沿线设有悬崖，悬崖尽头设有目标状态。进入任何悬崖状态都代表发生了灾难性事件，会导致回合终止，并获得非常大的负奖励。到达目标状态也会终止回合，并获得中等正奖励。其余状态的奖励均为 0。每个动作都需要支付少量成本。环境布局如图 1 所示。

在最简单的环境设置中，有四个动作，分别对应智能体可以移动的四个方向。环境中的移动是随机的，这意味着智能体有可能朝随机方向移动，而非朝其选择的方向移动。所有动作的成本都相同。

在更复杂的环境版本中，动作数量更多。动作由方向和谨慎度组成。选择较高的谨慎度会提高智能体遵循指定动作的可能性，但也会增加移动成本。移动成本与谨慎度成正比，而遵循指定动作的概率为 $1-2^{-c}$，其中 c是谨慎度。

简单逆向强化学习智能体。 给定策略 π（或该策略生成的轨迹），我们的目标是找到一个奖励 R，使得 π 在该奖励下为最优策略。我们使用约束优化。具体而言， π 的最优性会对 Q施加约束；由于我们可以用 R表示 Q，因此也就得到了对 R的约束。我们在附录中给出了这些约束的具体形式（这是对 Ng & Russell（2000）中约束的扩展），并使用线性规划求解 R （Anderson 等人，2022年）。不过，请记住，逆向强化学习的关键难题之一是区分这些有效奖励函数。按照我们对 Ng & Russell（2000）的扩展，我们最大化

$$\sum_{s\in S}\left\{Q^{\pi}(s,\pi(s))-\max_{a\in A\setminus\{\pi(s)\}}Q^{\pi}(s,a)\right\}-\lambda||\boldsymbol{R}||_1 \tag{1}$$

这会促使奖励函数尽可能拉大最优策略所选动作与次优动作之间的差距。第二项（$\lambda||\boldsymbol{R}||_1$）则鼓励奖励函数保持稀疏（包含许多零值）。附录中的定理 4 说明了如何仅用 $\boldsymbol{R}$来表示这一目标函数。

用状态奖励来表示奖励。在我们的环境中，奖励可以表示为动作奖励与状态奖励之和，

$$R(s,a)=R_A(a)+R_S(s) \tag{2}$$

这种奖励形式极大地有助于解决逆向强化学习问题，因为我们只需要找出 $|S|+|A|$ 个未知奖励；如果不进行这种分解，我们就需要找出 $|S||A|$ 个未知奖励。

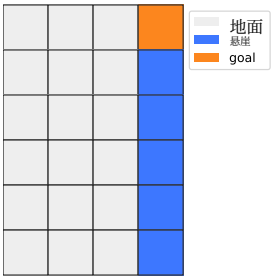

图 1：网格世界的布局

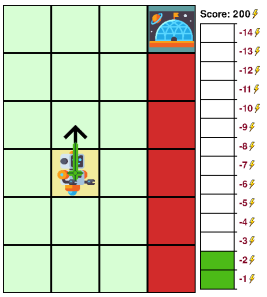

图 2：游戏的布局

游戏与损失逆向强化学习智能体。 为了在人类数据上测试逆向强化学习智能体，我们开发了一个可供人类游玩的网格世界交互版本。玩家可以使用方向键在网格世界中移动，并尝试抵达目标状态，同时避免跌落悬崖。与之前一样，环境具有随机性：游戏角色有一定几率不会朝玩家指定的方向移动。玩家可以通过长按移动键来抵消这一情况。随着长按时间增加，侧边代表谨慎度的进度条会逐渐填满，指示移动方向的箭头也会变长。玩家在回合开始时得分为 200，而更加谨慎需要付出更高代价（代价显示在侧边栏旁）。游戏布局如图 2 所示

毫不意外，人类玩这个游戏时，并不会严格遵循最优策略。这会带来问题，因为上文的简单逆向强化学习智能体假设最优策略是已知且确定的。为了解决这个问题，我们开发了另一种方法，称为 Loss IRL 智能体，

$$\mathcal{L}(R_S) = \sum_{s \in S} \exp\left(\mathrm{ReLU}\left(\sum_{a \in A} \pi(s,a) \max_{a' \in A \setminus \{a\}} Q(s,a') - Q(s,a)\right)\right) \tag{3}$$

我们将最小化损失 $\mathcal{L}(R_S)$。我们预计，对于策略最偏好的动作， Q 会最高。当给定状态下最受偏好的动作产生最大的 Q值时，ReLU 内部的项将为 0；否则，该项将为正。

4 实验结果

最大熵基线 作为基线，我们采用 Kim 等人（2021）提出的不考虑谨慎度的最大熵方法。我们选择这种方法，是因为即使不考虑谨慎度，只要数据量趋于无穷，该方法就能保证恢复奖励。该方法采用确定性转移，其策略为

$$\pi(a|s) = \frac{1}{Z(s)} e^{\beta Q(s,a)} \tag{4}$$

其中， β 是一个超参数， $Z(s)$确保动作的分布归一化为 1。Kim 等人（2021）证明，在这一设定下存在唯一的最优奖励函数。我们从该环境中的一个智能体生成了 10000 条轨迹，并恢复出一个奖励函数，如图5所示。因此，尽管该奖励函数准确地刻画了悬崖和目标状态的存在，但它严重低估了坠落悬崖所受惩罚的严重程度，这很可能表明，要准确估计奖励，需要多得多的样本。

简单逆向强化学习智能体。 这是针对第 3 节所述智能体的实验。我们通过值迭代求解 MDP，从而得到具有14个谨慎程度等级的设定下的最优策略。该策略如图3所示。正如预期，越靠近悬崖，最优动作就越谨慎，而且往往会远离悬崖。我们生成了 100 条轨迹：智能体从随机的“地面”状态（见图1）出发，然后遵循图3所示的最优策略。接着，我们使用这些轨迹，以 1000 为最大奖励，并采用 $\lambda = 0$ 来训练简单逆向强化学习智能体。借助谨慎度，我们得以学习奖励函数的严重程度；见图6。

Loss IRL 智能体。 一位人类专家（作者之一）玩了 10 条游戏轨迹（而基线使用了 10,000 条最优轨迹）。请注意，某些状态可能被多次访问，每次访问时选择的动作也可能不同，因此策略并非确定性的。图 7 展示了每个状态下最常被选择的动作。按照第 3 节所述训练的智能体

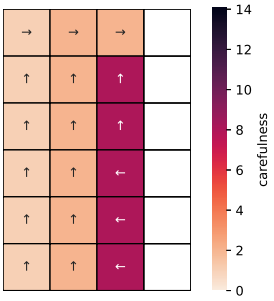

图3：包含14个谨慎程度等级的设置下的最优策略

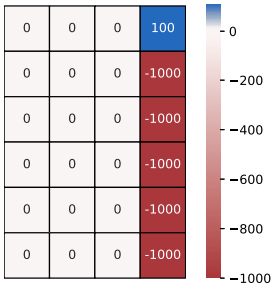

图4：真实奖励

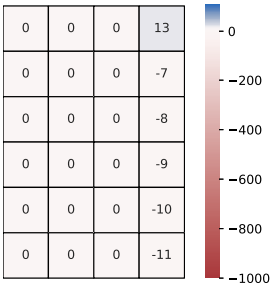

图5：基准设置下使用计算机生成的轨迹学习到的奖励

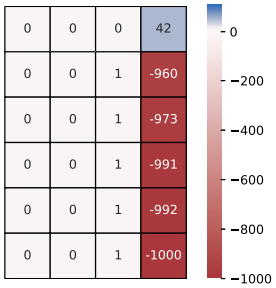

图6：使用计算机生成的轨迹学习到的奖励

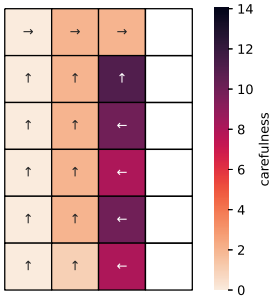

图 7：人类遵循的策略

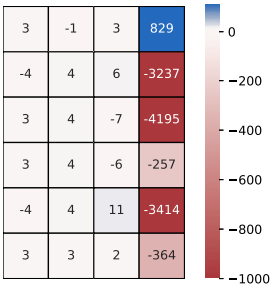

图8：从人类轨迹中学习到的奖励

使用这些轨迹进行训练，得到的奖励如图8所示；请与图4中的真实奖励和图5中的基准奖励进行比较。推断出的奖励存在相当大的差异，这可能是因为只有 10 条轨迹，但我们已经实现了目标，即捕捉掉下悬崖所受惩罚的严重程度。

5 结论与未来工作

我们已经证明，如果逆向强化学习考虑到人类的谨慎度，就能区分灾难性奖励和仅仅不受欢迎的奖励。我们在一个简单的网格世界任务中验证了这一想法，在该任务中，人类会明确输入谨慎程度。未来，重要的是将这些想法扩展到更复杂的场景（例如机器人模拟器），并了解如何从实际的人类动作中推断谨慎度（例如，将动作速度用作代理指标）。

参考文献

Joshua Achiam、David Held、Aviv Tamar 和 Pieter Abbeel。《约束策略优化》。载于 Doina Precup 和 Yee Whye Teh（编），第*34*届国际机器学习会议论文集，机器学习研究论文集第70卷，第22–31页。机器学习研究论文集，2017年8月6日至11日。网址：https://proceedings.mlr.press/v70/achiam17a.html。

Mohammed Alshiekh、Roderick Bloem、Rüdiger Ehlers、Bettina Könighofer、Scott Niekum 和 Ufuk Topcu。《通过屏蔽实现安全强化学习》。载于*AAAI*人工智能会议论文集，第32卷，2018年。

Martin Anderson、Lieven Vandenberghe 和 Joachim Dahl。《CVXOPT》，2022年。网址：https://cvxopt.org。

Chris Askew 和 Andy P Field。《恐惧替代性学习途径：40年后》。《临床心理学评论》《临床心理学评论》，28(7):1249–1265，2008年。

Felix Berkenkamp、Matteo Turchetta、Angela Schoellig 和 Andreas Krause。具有稳定性保证的安全的基于模型的强化学习。载于 I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett（编），《神经信息处理系统进展》，第 30 卷。Curran Associates, Inc.，2017年。网址：https://proceedings.neurips.cc/paper/2017/file/766ebcd59621e305170616ba3d3dac32-Paper.pdf。

Daniel Brown 和 Scott Niekum。逆强化学习的高效概率性能界。*AAAI*人工智能会议论文集，32(1)，2018 年 4 月。doi：10.1609/aaai.v32i1.11755。网址：https://ojs.aaai.org/index.php/AAAI/article/view/ 11755。

Daniel S. Brown、Yuchen Cui 和 Scott Niekum。风险感知主动逆强化学习。载于 Aude Billard、Anca Dragan、Jan Peters 和 Jun Morimoto（编），第二届机器人学习会议论文集，机器学习研究论文集第 87 卷，第 362–372 页。机器学习研究论文集，2018年10月29–31日。网址：https://proceedings.mlr.press/v87/brown18a. html。

Guler Dunne 和 Chris Askew。通过母亲与陌生人模型，在童年时期替代性学习恐惧及消退恐惧。《情绪》，13(5)：974，2013年。

Chelsea Finn、Sergey Levine 和 Pieter Abbeel。引导式代价学习：通过策略优化实现深度逆最优控制。载于国际机器学习会议，第 49–58 页。机器学习研究论文集，2016年6月。网址：https://proceedings.mlr.press/v48/finn16.html。

Rinat Galin、Roman Meshcheryakov、Saniya Kamesheva 和 Anna Samoshina。《协作机器人及其在智能制造中的应用优势》。载于*IOP*会议系列：材料科学与工程，第862卷，第032075页。IOP出版社，2020年。

Javier Garcia 和 Fernando Fernández。《强化学习中对状态与动作空间的安全探索》。《style i d='1'>人工智能研究杂志》，45：515–564，2012年。

Jon Henley。国际象棋机器人抓住并弄伤了七岁对手的手指。《style id='1'>卫报》，2022年。网址：https://www.theguardian.com/sport/2022/jul/24/chess-robot-grabs-and-breaks-finger-of-seven-year-old-opponent-moscow。

Sebastian Hjorth 和 Dimitrios Chrysostomou。《工业环境中的人机协作：关于无损拆卸的文献综述》。《style id='1'>机器人与计算机集成制造》，73：102208，2022年。

Jonathan Ho 和 Stefano Ermon。生成对抗模仿学习。载于神经信息处理系统进展，第29卷。Curran Associates, Inc.，2016年。URL https://papers.nips.cc/paper/2016/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html。

Kuno Kim、Shivam Garg、Kirankumar Shiragur 和 Stefano Ermon。逆强化学习中的奖励识别。载于 Marina Meila 和 Tong Zhang（编），第*38*届国际机器学习会议论文集，机器学习研究论文集第139卷，第5496–5505页。机器学习研究论文集，2021年7月18日至24日。URL https://proceedings.mlr.press/v139/kim21c.html。

Marie-France Marin、Alexe Bilodeau-Houle、Simon Morand-Beaulieu、Alexandra Brouillard、Ryan J Herringa 和 Mohammed R Milad。亲子二元组中的替代性条件性恐惧习得与消退。《科学报告》，10(1)：1–10，2020年。

Kunal Menda、Katherine Driggs-Campbell 和 Mykel J Kochenderfer。《Dropoutdagger：一种用于安全模仿学习的贝叶斯方法》。*arXiv* 预印本 *arXiv:1709.06166*，2017 年。

Kunal Menda、Katherine Driggs-Campbell 和 Mykel J Kochenderfer。EnsembleDAgger：一种用于安全模仿学习的贝叶斯方法。发表于*2019*年 *IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第 5041–5048 页。IEEE，2019年。

George Michalos、Panagiotis Karagiannis、Nikos Dimitropoulos、Dionisis Andronas 和 Sotiris Makris。工业环境中的人机协作。发表于《*21*世纪工业机器人：当工具成为协作者》，第 17–39 页。Springer，2022年。

Susan Mineka 和 Michael Cook。恐惧的观察性条件作用所涉及的机制。《实验心理学杂志：总论》，122(1):23，1993年。

Miguel Ángel Moreno、Diego Carou 和 J Paulo Davim。自主机器人和协作机器人：在制造业中的应用。载于《成功部署工业*4.0*的赋能技术》，第47–66页。CRC出版社，2020年。

Andrew Ng 和 Stuart J Russell。逆强化学习算法。载于国际机器学习大会，200 0.

Tom Le Paine、Sergio Gómez Colmenarejo、Ziyu Wang、Scott Reed、Yusuf Aytar、Tobias Pfaff、Matt W Hoffman、Gabriel Barth-Maron、Serkan Cabi、David Budden 等。单次高保真模仿：使用强化学习训练大规模深度网络。*arXiv* 预印本 *arXiv:1810.05017*，2018年。

Xue Bin Peng、Pieter Abbeel、Sergey Levine 和 Michiel van de Panne。DeepMimic：通过示例引导的深度强化学习习得基于物理的角色技能。《*ACM*图形学汇刊》，37 (4):1–14，2018年8月。ISSN 0730-0301，1557-7368。doi：10.1145/3197517.3201311。

Xue Bin Peng、Angjoo Kanazawa、Sam Toyer、Pieter Abbeel 和 Sergey Levine。变分判别器瓶颈：通过约束信息流改进模仿学习、逆强化学习和生成对抗网络。*arXiv* 预印本 *arXiv:1810.00821*，2018年。

Deepak Ramachandran 和 Eyal Amir。贝叶斯逆强化学习。载于 国际人工智能联合会议，第 7 卷，第 2586–2591 页，2007年。

Gemma Reynolds、Andy P Field 和 Chris Askew。采用积极示范减少儿童替代性习得的回避和心率反应。*Journal of Clinical Child& AdolescentPsychology*，47(4)：555–568，2018年。

Stuart Russell。不确定环境中的学习智能体（扩展摘要）。载于 第十一届计算学习理论年会论文集 - *COLT' 98*，第 101–103 页，美国威斯康星州麦迪逊，1998年。ACM出版社。ISBN 978-1-58113-057-7。doi：10.1145/279943. 279964。URL http://portal.acm.org/citation.cfm?doid=279943.279964。

William Saunders、Girish Sastry、Andreas Stuhlmueller 和 Owain Evans。《无错误试验：通过人工干预实现安全强化学习》。*arXiv*预印本 *arXiv:1707.05173*，2017年。

Ana C Simoes、António Lucas Soares 和 Ana C Barros。《影响制造业情境中协作机器人采用的驱动因素：一项定性研究》。载于国际制造科学技术会议，第203–212页。Springer，2019年。

Richard S Sutton 和 Andrew G Barto。《强化学习：导论》。麻省理工学院出版社，201 8.

Brian D Ziebart、Andrew L Maas、J Andrew Bagnell、Anind K Dey 等。《最大熵逆强化学习》。载于*AAAI*，第8卷，第1433–1438页。美国伊利诺伊州芝加哥，2008年。

背景

一个（有限）马尔可夫决策过程（MDP）定义为元组（S，A，T，γ，R），其中

- S 是一个有限的 N_S 状态集合
- A是一个有限的 N_A 动作集合
- $T(s,\ a,\ s')$ 是采取动作a时从状态 s 转移到状态 s' 的概率
- $\gamma \in [0, 1)$ 是折扣因子
- $R(s, a)$ 是从状态 s 通过动作 a 转移时获得的奖励

策略定义为一个函数 $\pi: S \times A \to [0,1]$。给定策略的价值函数是一个映射$V^\pi: S \to \mathrm{R}$，表示从给定状态遵循策略 π 时的期望折扣奖励。它的递归定义如下:

$$V^\pi(s) = \sum_{a \in A} \pi(s,a) \left(R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') V^\pi(s') \right) \tag{5}$$

另一个相关量是动作价值函数，$Q^\pi: S \times A \to \mathrm{R}$。$Q^\pi(s, a)$ 表示从状态 s 开始，立即执行动作 a，然后遵循策略 π 时的期望折扣奖励。它用 V^π 表示为

$$Q^\pi(s,a) = R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') V^\pi(s')$$

B IRLagent 的证明

定理1。

$$\boldsymbol{V}^\pi = (\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^\pi)^{-1} \hat{\boldsymbol{\pi}} \operatorname{vec}(\boldsymbol{R})$$

其中 $N_S := |S|$, $\boldsymbol{T}^\pi$ 是一个 N_S 乘 N_S 矩阵，表示遵循最优策略时从一个状态转移到另一个状态的概率，而 $\hat{}\boldsymbol{\pi}$ 是一个 N_S 乘 $N_S N_A$ 矩阵，使得$\hat{}\boldsymbol{\pi}\mathrm{vec}(\boldsymbol{R}) = \boldsymbol{R}^\pi$，并且 $\boldsymbol{R}^\pi$ 是遵循 π 时发生状态转移所对应的期望即时奖励。

证明。我们有，$\forall s \in S$

$$V^\pi(s) = \sum_{a \in A} \pi(s,a) \left(R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') V^\pi(s') \right)$$
$$V^\pi(s) = \sum_{a \in A} \pi(s,a) \left(\boldsymbol{R}_{s,a} + \gamma \boldsymbol{T}_{s,a} \boldsymbol{V}^\pi \right)$$

因此

$$\begin{aligned} \boldsymbol{V}^\pi &= \boldsymbol{R}^\pi + \gamma \boldsymbol{T}^\pi \boldsymbol{V}^\pi \\ (\boldsymbol{I} - \gamma \boldsymbol{T}^\pi) \boldsymbol{V}^\pi &= \boldsymbol{R}^\pi \\ \boldsymbol{V}^\pi &= (\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1} \hat{\boldsymbol{\pi}} \operatorname{vec} \boldsymbol{R} \end{aligned}$$

□

定理2。

$$\boldsymbol{Q}^\pi = \left(\hat{\boldsymbol{I}} + \gamma \boldsymbol{T} (\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^\pi)^{-1} \hat{\boldsymbol{\pi}} \right) \operatorname{vec}(\boldsymbol{R})$$

$\hat{}$其中 $\boldsymbol{I}$ 被重塑为 $N_S \times N_A \times N_S N_A$

证明。

$$\begin{aligned} Q^\pi(s,a) &= R(s,a) + \gamma \sum_{s' \in S} \boldsymbol{T}_{s,a}(s') V^\pi(s') \\ \boldsymbol{Q}^\pi &= \boldsymbol{R} + \gamma \boldsymbol{T} \boldsymbol{V}^\pi \\ \boldsymbol{Q}^\pi &= \boldsymbol{R} + \gamma \boldsymbol{T} (\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^\pi)^{-1} \hat{\boldsymbol{\pi}} \operatorname{vec}(\boldsymbol{R}) \\ \boldsymbol{Q}^\pi &= \left(\hat{\boldsymbol{I}} + \gamma \boldsymbol{T} (\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^\pi)^{-1} \hat{\boldsymbol{\pi}} \right) \operatorname{vec}(\boldsymbol{R}) \end{aligned}$$

□

定理3。要使奖励函数 $\boldsymbol{R}$ 有效，我们要求:

$$\left\{ \left[\widetilde{\boldsymbol{I}}_{N_S} + \gamma (\widetilde{\boldsymbol{T}}^\pi - \boldsymbol{T}) (\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^\pi)^{-1} \right] \hat{\boldsymbol{\pi}} - \hat{\boldsymbol{I}} \right\} \operatorname{vec}(\boldsymbol{R}) \succeq 0$$

其中

• 对于一个维度为 $N_S \times N_S$ 的矩阵 $\boldsymbol{X}$， $\boldsymbol{X}$ 是将 $\widetilde{\boldsymbol{X}}$ 堆叠 N_A 份后得到的矩阵，其维度为 $N_S \times N_A \times N_S$

• 对于维度为 $n \times m$ 的矩阵 $\boldsymbol{X}$，以及 $c \in \mathbb{R}$， $\boldsymbol{X} \succeq c$表示对于所有的 $0 \leq i \leq n$ 和 $0 \leq j \leq m$，$\boldsymbol{X}_{i,j} \geq c$

证明。 当且仅当 $\forall s \in S$ 且 $\forall a \in A$时，策略 π 是最优的

$$
\begin{aligned}
& Q^{\pi}(s, \pi(s)) \geq Q^{\pi}(s, a') \\
& \quad \Leftrightarrow \boldsymbol{R}_{s,\pi(s)} + \gamma \left(\boldsymbol{T}^{\pi} \boldsymbol{V}^{\pi}\right)_s \geq \boldsymbol{R}_{s,a} + \gamma \left(\boldsymbol{T}_{:,a,:} \boldsymbol{V}^{\pi}\right)_s \\
& \quad \Leftrightarrow \boldsymbol{R}^{\pi} + \gamma \boldsymbol{T}^{\pi} \boldsymbol{V}^{\pi} \succeq \boldsymbol{R}_{:,a} + \gamma \boldsymbol{T}_{:,a,:} \boldsymbol{V}^{\pi} \\
& \quad \Leftrightarrow \hat{\boldsymbol{\pi}} \operatorname{vec}(\boldsymbol{R}) - \boldsymbol{R}_{:,a} \succeq \gamma (\boldsymbol{T}_{:,a,:} \boldsymbol{V}^{\pi} - \boldsymbol{T}^{\pi} \boldsymbol{V}^{\pi}) \\
& \quad \Leftrightarrow \hat{\boldsymbol{\pi}} \operatorname{vec}(\boldsymbol{R}) - \boldsymbol{R}_{:,a} \succeq \gamma (\boldsymbol{T}_{:,a,:} - \boldsymbol{T}^{\pi}) \boldsymbol{V}^{\pi} \\
& \quad \Leftrightarrow \hat{\boldsymbol{\pi}} \operatorname{vec}(\boldsymbol{R}) - \boldsymbol{R}_{:,a} \succeq \gamma (\boldsymbol{T}_{:,a,:} - \boldsymbol{T}^{\pi}) (\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^{\pi})^{-1} \hat{\boldsymbol{\pi}} \operatorname{vec}(\boldsymbol{R}) \\
& \quad \Leftrightarrow \hat{\boldsymbol{\pi}} \operatorname{vec}(\boldsymbol{R}) - \gamma (\boldsymbol{T}_{:,a,:} - \boldsymbol{T}^{\pi}) (\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^{\pi})^{-1} \hat{\boldsymbol{\pi}} \operatorname{vec}(\boldsymbol{R}) - \boldsymbol{R}_{:,a} \succeq 0 \\
& \quad \Leftrightarrow \left[\boldsymbol{I}_{N_S} + \gamma (\boldsymbol{T}^{\pi} - \boldsymbol{T}_{:,a,:}) (\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^{\pi})^{-1}\right] \hat{\boldsymbol{\pi}} \operatorname{vec}(\boldsymbol{R}) - \boldsymbol{R}_{:,a} \succeq 0 \\
& \quad \Leftrightarrow \left[\widetilde{\boldsymbol{I}}_{N_S} + \gamma (\widetilde{\boldsymbol{T}}^{\pi} - \boldsymbol{T}) (\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^{\pi})^{-1}\right] \hat{\boldsymbol{\pi}} \operatorname{vec}(\boldsymbol{R}) - \boldsymbol{R} \succeq 0
\end{aligned}
$$

□

定理 4。 最大化表达式 *(1)* 等价于最小化

$$
\sum_{s \in S} \max_{a \in A \setminus \{\pi(s)\}} \left(\hat{\boldsymbol{I}} - \left(\widetilde{\boldsymbol{I}}_{N_S} - \gamma \boldsymbol{T}\right) \left(\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^{\pi}\right)^{-1} \hat{\boldsymbol{\pi}}\right)_{s,a} \cdot \operatorname{vec}(\boldsymbol{R}) + \lambda ||\boldsymbol{R}||_1 \tag{6}
$$

证明。 最大化表达式 (1) 等价于最小化

$$
\begin{aligned}
& \sum_{s \in S} \left\{ \max_{a \in A \setminus \{\pi(s)\}} \left(\boldsymbol{Q}^{\pi}_{s,a}\right) - \boldsymbol{Q}^{\pi}_{s,\pi(s)} \right\} + \lambda ||\boldsymbol{R}||_1 \\
& \quad = \sum_{s \in S} \max_{a \in A \setminus \{\pi(s)\}} \left(\boldsymbol{Q}^{\pi}_{s,a} - \boldsymbol{V}^{\pi}_{s}\right) + \lambda ||\boldsymbol{R}||_1 \\
& \quad = \sum_{s \in S} \max_{a \in A \setminus \{\pi(s)\}} \left(\boldsymbol{Q}^{\pi} - \widetilde{\boldsymbol{I}}_{N_S} \boldsymbol{V}\right)_{s,a} + \lambda ||\boldsymbol{R}||_1 \\
& \quad = \sum_{s \in S} \max_{a \in A \setminus \{\pi(s)\}} \left\{ \left(\hat{\boldsymbol{I}} - \left(\widetilde{\boldsymbol{I}}_{N_S} - \gamma \boldsymbol{T}\right) \left(\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^{\pi}\right)^{-1} \hat{\boldsymbol{\pi}}\right)_{s,a} \cdot \operatorname{vec}(\boldsymbol{R}) \right\} \\
& \qquad + \lambda ||\boldsymbol{R}||_1
\end{aligned}
$$

□

在表格设定中，我们可以将方程 (2) 表示为

$$
\operatorname{vec}(\boldsymbol{R}) = \operatorname{vec}(\boldsymbol{R}_A) + \hat{\mathbb{1}}^{\intercal}_{N_s, N_A} \boldsymbol{R}_S
$$

其中 $\hat{\mathbb{1}}_{N_s, N_A}^{\intercal}$ 满足 $\hat{\mathbb{1}}_{N_s, N_A} \boldsymbol{R}_S$等于 N_A 个 $\boldsymbol{R}_S$ 彼此上下堆叠而成。

定理 5. 写作 $\boldsymbol{\Omega} = \hat{\boldsymbol{I}} - \left(\widetilde{\boldsymbol{I}}_{N_S} - \gamma \boldsymbol{T}\right) \left(\boldsymbol{I}_{N_S} - \gamma \boldsymbol{T}^{\pi}\right)^{-1} \hat{\boldsymbol{\pi}}$，最小化表达式 *(6)* 等价于最小化

$$
\sum_{s \in S} \max_{a \in A \setminus \{\pi(s)\}} \boldsymbol{\Omega}_{s,a} \hat{\mathbb{1}}^{\intercal}_{N_s, N_A} \boldsymbol{R}_S + \lambda \left|\left|\boldsymbol{R}_S\right|\right|_1 \tag{7}
$$

证明。 表达式 (6) 为

$$\begin{aligned}
&\sum_{s\in S}\max_{a\in A\setminus\{\pi(s)\}}\boldsymbol{\Omega}_{s,a}\,\mathrm{vec}(\boldsymbol{R})+\lambda||\boldsymbol{R}||_1\\
&\quad=\sum_{s\in S}\max_{a\in A\setminus\{\pi(s)\}}\boldsymbol{\Omega}_{s,a}\left(\mathrm{vec}(\boldsymbol{R}_A)+\hat{\mathbb{1}}^{\intercal}_{N_s,N_A}\boldsymbol{R}_S\right)\\
&\qquad+\lambda\left|\left|\mathrm{vec}(\boldsymbol{R}_A)+\hat{\mathbb{1}}^{\intercal}_{N_s,N_A}\boldsymbol{R}_S\right|\right|_1
\end{aligned}$$

由于第二项是为了在奖励中鼓励产生更多零值而任意选取的，我们可以将其替换为 $||\boldsymbol{R}_S||_1$。因此，我们希望最小化

$$\sum_s\left\{\max_{a\in A\setminus\{\pi(s)\}}\left(\boldsymbol{\Omega}_{s,a}\,\mathrm{vec}(\boldsymbol{R}_A)\right)+\max_{a\in A\setminus\{\pi(s)\}}\left(\boldsymbol{\Omega}_{s,a}\hat{\mathbb{1}}^{\intercal}_{N_s,N_A}\boldsymbol{R}_S\right)\right\}+\lambda||\boldsymbol{R}_S||_1$$

由于 $\boldsymbol{R}_A$ 是固定的，我们得到

$$\sum_{s\in S}\max_{a\in A\setminus\{\pi(s)\}}\boldsymbol{\Omega}_{s,a}\hat{\mathbb{1}}^{\intercal}_{N_s,N_A}\boldsymbol{R}_S+\lambda\,||\boldsymbol{R}_S||_1$$

□