多目标强化学习在认知雷达资源管理中的应用

Ziyang Lu, Subodh Kalia, M. Cenk Gursoy, Chilukuri K. Mohan, Pramod K. Varshney
Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse NY, 13066
{zlu112, skalia, mcgursoy, ckmohan, varshney}@syr.edu

摘要—多功能认知雷达系统中的时间分配问题集中在新出现目标的扫描与之前检测到的目标跟踪之间的权衡。我们将此问题形式化为一个多目标优化问题，并采用深度强化学习来寻找帕累托最优解，并比较了深度确定性策略梯度（DDPG）和软演员-评论家（SAC）算法。我们的结果展示了这两种算法在适应各种场景方面的有效性，其中 SAC 相较于 DDPG 显示出了更好的稳定性和样本效率。我们进一步采用了 NSGA-II 算法来估计所考虑问题的帕累托前沿上的上限。本研究促进了更高效、更具自适应性的认知雷达系统的发展，这些系统能够平衡动态环境中的多个竞争目标。

***Index Terms*—认知雷达，多目标优化，约束深度强化学习，帕累托前沿，DDPG，SAC**

I. 介绍

认知雷达系统 [1] 已经成为一种有前景的技术，用于在复杂和动态环境中提升雷达性能，并且最近的进展总结于 [2]、[3] 和 [4] 中。这些系统可以根据当前情况实时调整其操作参数，从而提高检测和跟踪性能。然而，在各种功能之间确定最优时间分配策略仍然是认知雷达系统中的关键挑战，由于存在冲突的目标 [2]，尤其是在需要同时进行大范围监视和多个目标的精确跟踪时尤为重要。在 [5] 中提出的作者建议了多功能相控阵雷达系统的调度方法，并且在 [6] 中提出的作者将多雷达系统中的功率分配问题建模为一个非合作博弈，进行了纳什均衡及其收敛性的分析。

近期的研究展示了深度强化学习（DRL）在适应动态环境中的雷达操作参数的适用性，例如，在拥挤频谱环境中进行雷达检测和跟踪 [7]。在 [8] 中，作者使用 DRL 进行场景自适应雷达跟踪。[9] 的工作提出了一种基于服务质量的雷达资源管理的 DRL 方法，有效地平衡了多个性能指标。[10] 的研究将强化学习应用于多功能雷达中的重复访问间隔选择问题，表明 Q 学习方法能够实现比传统方法更低的跟踪负载，同时保持可比较的丢失概率。

我们将雷达资源分配问题表述为一个多目标优化问题，解决扫描与追踪之间的矛盾。先前的研究试图通过一组固定的权重来优化多个目标函数的加权和，但这种方法对于实际应用来说是不充分的，在实际应用中，不同目标的重要性可能根据任务需求和操作环境的不同而变化。相反，我们通过调整扫描与追踪目标之间的折衷系数来生成并分析帕累托前沿。我们评估了两种最先进的深度强化学习算法，即深度确定性策略梯度（DDPG）[11] 和软演员批评（SAC）[12]，以确定帕累托最优解。我们也使用了一种著名的多目标进化算法（NSGA-II）[13] 来建立所考虑问题的帕累托前沿的上限。我们对学习到的策略进行了详细分析，展示了资源分配策略如何适应扫描与追踪目标的不同优先级。

第二节详细描述了所考虑的雷达时间分配问题的系统模型。第三节介绍了所考虑的方法，包括 DDPG 和 SAC 算法以及生成 Pareto 前沿的过程。在第五节中，我们展示了并分析了数值结果，并在第六节中得出结论。

II. 系统模型

我们考虑一个认知雷达系统，该系统将时间分配在搜索潜在目标和跟踪已检测到的目标之间。系统模型基于我们的前期工作 [14]，我们在下面对其进行总结。

A. 目标运动模型

目标在时间 t 的状态定义为 $\boldsymbol{x}_t = [x_t, y_t, \dot{x}_t, \dot{y}_t]^T$，其中 (x_t, y_t) 是目标的位置，$(\dot{x}_t, \dot{y}_t)$ 是其水平和垂直速度。目标运动被建模为

$$\boldsymbol{x}_{t+1} = \boldsymbol{F}_t \boldsymbol{x}_t + \boldsymbol{w}_t \tag{1}$$

其中 $\boldsymbol{F}_t$ 是状态转移矩阵，$\boldsymbol{w}_t$ 是具有协方差 $\boldsymbol{Q}_t$ 的高斯机动噪声。

B. 测量模型

雷达获得目标距离 r 和方位角 θ 的噪声测量值：

$$\boldsymbol{z_t} = h(\boldsymbol{x_t}) + \boldsymbol{v_t} = \left[\sqrt{x_t^2 + y_t^2}, \quad \tan^{-1}\left(\frac{y_t}{x_t}\right)\right]^T + \boldsymbol{v_t} \tag{2}$$

其中 $\boldsymbol{v}_t$ 是具有协方差 $\boldsymbol{R}_t$ 的高斯测量噪声。

C. 跟踪模型

我们采用扩展卡尔曼滤波器（EKF），一种用于雷达跟踪应用的非线性估计技术 [15]。每个目标在时间 t 的跟踪性能通过一个成本函数 c_t 来量化，定义为

$$c_t(\tau_t) = \text{trace}(\boldsymbol{E}\boldsymbol{P}_{t|t}\boldsymbol{E}^T) \tag{3}$$

其中 $\boldsymbol{P}_{t|t} \in \mathbb{R}^{4\times 4}$ 是后验状态估计的协方差矩阵，由 EKF 获得。$\boldsymbol{E}$ 是一个投影矩阵，用于从状态向量中提取位置分量。τ_t 是驻留时间，即雷达在一个测量周期内

聚焦于单个目标的时间长度。增加驻留时间可以提高状态估计的准确性并减少不确定性。我们将测量周期的持续时间 T_t 固定为 T_0。增加一个目标的驻留时间 τ_t 可以提高其跟踪精度，但会牺牲其他目标和扫描性能。

D. 扫描模型

采用均匀圆环阵 (UCA) 雷达进行扫描。信噪比 (SNR) 在扫描期间由下式给出

$$\mathrm{SNR}_{\mathrm{scan}} = \frac{P_t \tau_{\mathrm{beam}} G_t G_r \lambda_r^2 \sigma}{(4\pi)^3 r^4 L k T_s} \tag{4}$$

其中 τ_{beam} 是波束持续时间，P_t 是发射功率，G_t 和 G_r 分别是发射天线增益和接收天线增益，λ_r 是雷达信号的波长，σ 是目标的雷达截面，r 是雷达与目标的距离，L 是一个损耗因子，k 是玻尔兹曼常数，T_s 是系统温度。τ_{beam} 定义为

$$\tau_s = \frac{360^\circ}{\phi} \tau_{beam} \tag{5}$$

其中 τ_s 表示分配给扫描的总时间，ϕ 是相邻雷达波束之间的相位延迟，而 τ_{beam} 是每个波束的时间持续期。

给定检测概率（P_d）和虚警概率（P_f），我们可以确定所需的最小信噪比。使用这个 $\mathrm{SNR}_{\mathrm{min}}$，我们可以通过设定 $\mathrm{SNR}_{\mathrm{scan}} = \mathrm{SNR}_{\mathrm{min}}$ 并求解 r 来得出最大可检测范围 $r_{\max}$：

$$r_{\max} = \left(\frac{P_t \tau_{\mathrm{beam}} G_t G_r \lambda_r^2 \sigma}{(4\pi)^3 L k T_s \cdot \mathrm{SNR}_{\mathrm{min}}} \right)^{1/4}. \tag{6}$$

此 $r_{\max}$ 表示雷达能够以所需的 P_d 和 P_f 检测目标的最大范围。对于距离 $r \le r_{\max}$，实际检测概率将达到或超过所需的 P_d 同时保持期望的 P_f。

为了量化我们认知雷达系统的扫描性能，我们引入了度量标准 Γ，定义为最大可探测区域与参考区域的比值：

$$\Gamma = \frac{A_{\max}}{A_{\mathrm{ref}}} = \frac{\pi r_{\max}^2}{\pi r_0^2} = \left(\frac{r_{\max}}{r_0} \right)^2 \tag{7}$$

其中 $r_{\max}$ 是给定的最大可探测范围（在 6中给出）。r_0 是参考距离，通常设置为雷达的默认工作范围。$A_{\max} = \pi r_{\max}^2$ 是最大可探测区域，而 $A_{\mathrm{ref}} = \pi r_0^2$ 是参考区域。

增加分配给扫描的时间会导致 Γ 增加，即提高检测新目标的能力、其覆盖区域以及在给定距离下的检测概率。

E. 轨迹初始化

跟踪初始化是雷达系统捕获新目标的过程。最常用的跟踪初始化策略是 M-of-N 模型。我们使用的是 3-of-4 跟踪初始化模型，如果在 4 连续扫描内获得 3 个相关的测量/检测，则初始化一个跟踪。我们使用全局最近邻（GNN）方法定义相关测量 [15]。具体而言，如果两个测量之间的欧几里得距离低于预定义的阈值，则认为一个测量与之前的测量相关。

III. 问题表述

在本节中，我们将认知雷达资源分配问题公式化为一个约束优化任务，在现有目标的跟踪性能和新目标的扫描之间寻求平衡。

A. 效用函数

使用多目标优化中的线性标量化方法，我们定义一个结合跟踪和扫描性能的效用函数如下：

$$U_t(\{\tau_t^n\}_{n=1}^N) = -\sum_{n=1}^{N} c_t^n(\tau_t^n) + \beta\Gamma \tag{8}$$

其中

- $c_t^n(\tau_t^n)$ 是在第二节 C 部分定义的目标 $n \in \{1,\dots,N\}$ 在时间 t 的跟踪成本；
- τ_t^n 是分配给跟踪目标 n 的驻留时间；
- Γ 是在第 II.D 节中定义的扫描指标；
- β 是平衡扫描与跟踪重要性的权衡系数。

B. 优化问题

我们的目标是找到一个最优的时间分配策略 π，该策略在时间预算约束下最大化随时间的预期折现效用总和：

$$\begin{aligned} \max_{\pi} \quad & \sum_{m=0}^{\infty} \left[\gamma^m U_{t+m}^n(\tau_{t+m}^n) \right] \\ \text{s.t.} \quad & \sum_{m=0}^{\infty} \gamma^m \left(\sum_{n=1}^{N} \frac{\tau_{t+m}^n}{T_0} - \Theta_{max} \right) \le 0. \end{aligned} \tag{9}$$

其中，$\gamma \in (0,1]$ 是折扣因子，T_0 是雷达测量周期的持续时间，Θ_{max} 是跟踪的总时间预算。

该约束确保分配给跟踪的总时间不超过预定义的预算，剩余时间 $\tau_s = T_0 - \sum_{n=1}^{N} \tau_t^n$ 被分配用于扫描任务。

通过引入一个对偶变量 λ，该问题可以被松弛为一个无约束优化问题：

$$\min_{\lambda_t \ge 0} \max_{\pi} \sum_{m=0}^{\infty} \gamma^m \left[U_{t+m}^n(\tau_{t+m}^n) - \lambda_t \left(\sum_{n=1}^{N} \frac{\tau_{t+m}^n}{T_0} - \Theta_{max} \right) \right]. \tag{10}$$

C. 帕累托前沿生成

为了探讨跟踪与扫描性能之间的权衡，我们改变权衡系数 β 并为每个值求解优化问题。这种方法使我们能够生成帕累托前沿，提供有关可实现的跟踪精度和扫描效果组合的见解。

对于每个 β 值，我们确定最大化组合目标函数的最佳策略 π^*。通过改变 β 并求解优化问题，我们生成一组帕累托最优解。这种方法刻画了跟踪精度与扫描性能之间的权衡边界。得到的帕累托前沿为雷达操作员提供了一个关于可实现性能组合的整体视图，帮助他们根据特定的任务目标和约束选择最合适的操作点。

IV. 受约束的深度强化学习框架

提出了一种约束深度强化学习（CDRL）框架来解决约束优化问题（9）。

A. 状态

在时间 t 处的状态 s_t 定义为

$$s_t = [\{c_{t-1}^n(\tau_{t-1}^n)\}_{n=1}^N, \{\tau_{t-1}^n\}_{n=1}^N, \lambda_{t-1}, \beta] \tag{11}$$

其中 $\{c_{t-1}^n(\tau_{t-1}^n)\}_{n=1}^N$ 是所有 N 目标的跟踪成本，$\{\tau_{t-1}^n\}_{n=1}^N$ 是先前时间段分配的停留时间，λ_{t-1} 是对偶变量，β 是权衡系数。状态的大小是 $2N+2$。

1) 动作: 动作 a_t 是分配给每个目标的停留时间集合：

$$a_t = \{\tau_t^n\}_{n=1}^N, \quad \text{where } \tau_t^n \in [0, T_0] \tag{12}$$

2) 奖励: 为了求解无约束优化问题 (10)，定义奖励函数 r_t 为

$$r_t = U_t(\{\tau_t^n\}_{n=1}^N) - \lambda_t \left(\sum_{n=1}^N \frac{\tau_t^n}{T_0} - \Theta_{max} \right) \tag{13}$$

其中 $U_t(\{\tau_t^n\}_{n=1}^N)$ 是在 (8) 中定义的效用函数，第二项是违反时间预算约束的惩罚。

在提出的 CDRL 框架中，对偶变量 λ 与神经网络参数同时更新为

$$\lambda_{t+1} = \max\left(0, \lambda_t + \alpha_\lambda \left(\sum_{n=1}^N \frac{\tau_t^n}{T_0} - \Theta_{max} \right)\right) \tag{14}$$

其中 α_λ 是对偶变量的学习率。λ 的更新动态调整约束违规的惩罚，引导学习过程趋向可行解。

在本文中，我们考虑使用 DDPG 和 SAC 算法来寻找多目标雷达资源分配问题的 Pareto 前沿。虽然 DDPG 提供了一种将状态直接映射到动作的确定性策略，但 SAC 提供了另一种随机策略，并且还考虑了最大化一个熵项。SAC 中的这种熵最大化鼓励探索并通过学习一组多样化的有效策略来提高鲁棒性。具体来说，SAC 旨在在最大化预期回报的同时也最大化策略的熵：

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t (r_t + \alpha H(\pi(\cdot|s_t))) \right] \tag{15}$$

其中 $J(\pi)$ 是策略 π 的预期回报，$H(\pi(\cdot|s_t))$ 是状态 s_t 下策略的熵，而 α 是一个温度参数，用于平衡奖励最大化和熵最大化之间的关系。这种表述使 SAC 有可能找到 Pareto 前沿上更加多样化的解决方案，尤其是在多个分配可能产生类似性能的区域。确实，SAC 被认为是凹增广帕累托 Q 学习算法在 [16] 中提出的一种实现。通过比较 DDPG 和 SAC，我们旨在提供关于不同 CDRL 实现在解决复杂的多目标雷达资源分配问题中的有效性见解。

表 I
模拟参数

N	5
$\sigma_{r,0}^2$ (m^2)	16
$\sigma_{\theta,0}^2$ (rad^2)	1e-6
σ_w $((m/s^2)^2)$	16
Measurement Cycle T_0 (s)	2.5
Time Budget for Tracking (Θ_{max})	0.9
Required Probability of Detection (P_d)	0.9
Required Probability of False Alarm (P_f)	1e-3
DRL Discount Factor γ	0.9
Initial Dual Variable (λ_0)	5000
Step Size of Dual Variable (α_λ)	15000

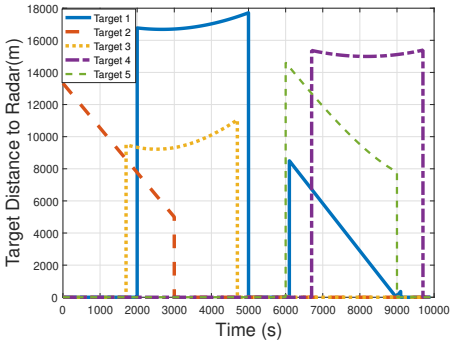

图 1. 目标到雷达的距离

V. 数值结果与分析

A. 超参数

关键超参数的值在表 I中给出。在 DDPG 实现中，评论网络有两个前馈层，每个包含 100 个神经元，并且演员网络也有两个前馈层，分别包含 256 和 128 个神经元。所有网络的学习率都设置为 0.0001。在 SAC 实现中，评论网络有每层包含 100 个神经元的两个前馈层，而演员网络则有两个每层包含 128 个神经元的前馈层。在 SAC 中，所有网络的学习率同样设置为 0.0001。训练过程中使用了 Adam 优化器。模拟是在配备有 Intel Core i7-8700 CPU（3.20GHz，6 核，12 逻辑处理器）的台式计算机上进行的。

B. 动态雷达资源管理与 CDRL

图 1展示了某一事件中目标与雷达的距离，其中目标以随机初始位置、速度和生成时间产生。目标的数量也会随时间变化。对于带有 $\beta = 150000$ 的 CDRL-SAC，图 2说明了当跟踪更多目标时，框架会分配更多时间用于跟踪任务，例如，在有三个被跟踪的目标时，即 $t \in [7000, 9000]$。此外，CDRL-SAC 还会给远距离目标分配更多时间，这些目标从额外资源中受益，例如在 $t \in [7000, 9000]$ 期间，以最大化整体奖励函数。这展示了所提出的 CDRL-SAC 框架平衡多个目标并适应动态雷达环境的能力。

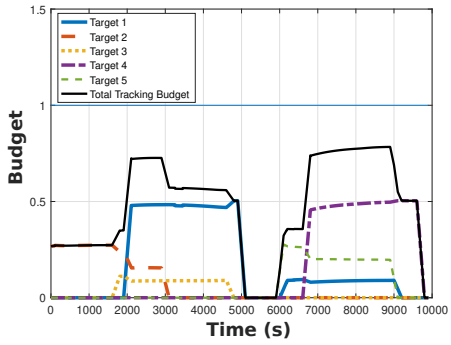

图 2. 时间分配策略由 CDRL-SAC 决定

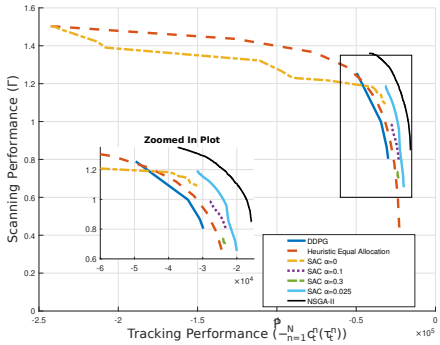

图 3. 帕累托前沿的比较

C. 帕累托前沿的比较

我们使用 DDPG 和 SAC 算法评估我们的方法，将其与等分配策略（将总跟踪时间平均分配给目标）以及 NSGA-II [13] 进行比较，在汇总跟踪性能 $obj_t = -\sum_{n=1}^{N} c_t^n(\tau_t^n)$ 和扫描性能 $obj_s = \Gamma$ 方面进行了评估，同时改变权衡系数 $\beta \in [0, 300000]$。非支配 (obj_t, obj_s) 对构成了帕累托前沿。NSGA-2 的实现使用了 120 个时间分配决策变量（用于扫描和跟踪 5 个潜在目标，在 20 个决策点处），在整个过程中有 10000 个时间槽。每个决策点的时间分配总和受约束为 $\leq T_0$。图 3比较了所达到的帕累托前沿，得出以下观察结果：

- 跟踪性能的提高通常会导致扫描性能的下降。
- CDRL-SAC 与 $\alpha = 0.025$ 达到了所有基于学习方案中最佳的帕累托前沿。其操作点涵盖了广泛的运行区域，并且始终占优势。
- 过度强调熵（即更高的 α）可能导致性能不佳，因为过分优先考虑探索而非开发。
- 需要进行一些探索，正如 CDRL-SAC 使用 $\alpha = 0$ 和 DDPG 算法时表现出的不足所示，这些算法仅专注于奖励最大化。
- 在使用了 60 个不同的 β 值的情况下，CDRL-SAC 需要大约 48 分钟来在整个包含 10000 时间步长的时期内生成帕累托前沿。
- NSGA-II 使用了锦标赛选择、模拟二进制交叉（参数为 $\eta = 15$，概率为 0.9）和多项式变异（参数为 $\eta = 20$），种群规模为 2400，进化代数为 1000。计算分布在十个节点上进行，每个节点配备有 128 个核心。NSGA-II 提供了可实现的帕累托前沿的有用上限，但具有较高的计算需求（我们的仿真中耗时 11.5 小时）。

VI. 结论

在本文中，我们提出了一种用于在时间预算约束下确定多功能雷达系统动态时间分配的 Pareto 最优解的受约束深度强化学习（CDRL）框架。我们的方法成功地在整个目标空间中找到了多样且高质量的解决方案，超越了启发式等量分配的方法。我们的结果表明，在多目标优化中平衡探索对于发现多样化解决方案至关重要。CDRL 框架为雷达操作员提供了一种灵活的工具，通过调整所提出的深度学习框架中的参数来动态调节不同任务之间的权衡，从而实现实时资源分配优化以响应动态的任务优先级。

参考文献

[1] S. Haykin, "Cognitive radar: a way of the future," *IEEE signal processing magazine*, vol. 23, no. 1, pp. 30–40, 2006.

[2] A. Charlish, F. Hoffmann, C. Degen, and I. Schlangen, "The development from adaptive to cognitive radar resource management," *IEEE Aerospace and Electronic Systems Magazine*, vol. 35, no. 6, pp. 8–19, 2020.

[3] M. S. Greco, F. Gini, P. Stinco, and K. Bell, "Cognitive radars: On the road to reality: Progress thus far and possibilities for the future," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 112–125, 2018.

[4] S. Z. Gurbuz, H. D. Griffiths, A. Charlish, M. Rangaswamy, M. S. Greco, and K. Bell, "An overview of cognitive radar: Past, present, and future," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 12, pp. 6–18, 2019.

[5] A. Orman, C. N. Potts, A. Shahani, and A. Moore, "Scheduling for a multifunction phased array radar system," *European Journal of operational research*, vol. 90, no. 1, pp. 13–25, 1996.

[6] A. Deligiannis, A. Panoui, S. Lambotharan, and J. A. Chambers, "Game-theoretic power allocation and the nash equilibrium analysis for a multistatic mimo radar network," *IEEE Transactions on Signal Processing*, vol. 65, no. 24, pp. 6397–6408, 2017.

[7] C. E. Thornton, M. A. Kozy, R. M. Buehrer, A. F. Martone, and K. D. Sherbondy, "Deep reinforcement learning control for radar detection and tracking in congested spectral environments," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 4, pp. 1335–1349, 2020.

[8] M. Stephan, L. Servadei, J. Arjona-Medina, A. Santra, R. Wille, and G. Fischer, "Scene-adaptive radar tracking with deep reinforcement learning," *Machine Learning with Applications*, vol. 8, p. 100284, 2022.

[9] S. Durst and S. Brüggenwirth, "Quality of service based radar resource management using deep reinforcement learning," in *2021 IEEE Radar Conference (RadarConf21)*. IEEE, 2021, pp. 1–6.

[10] P. Pulkkinen, T. Aittomäki, A. Ström, and V. Koivunen, "Time budget management in multifunction radars using reinforcement learning," in *2021 IEEE Radar Conference (RadarConf21)*, 2021, pp. 1–6.

[11] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[12] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[13] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[14] Z. Lu and M. C. Gursoy, "Resource allocation for multi-target radar tracking via constrained deep reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 6, pp. 1677–1690, 2023.

[15] Y. Bar-Shalom, P. K. Willett, and X. Tian, *Tracking and data fusion*. YBS publishing Storrs, CT, USA:, 2011, vol. 11.

[16] H. Lu, D. Herman, and Y. Yu, "Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=TjEzIsyEsQ6