强化学习方法在自主水下机器人控制任务中的应用

许雅筑, 武辉, 游科友*, 宋士吉

清华大学自动化系与北京信息科学与技术国家研究中心, 北京 100084

* 通信作者. E-mail: youky@tsinghua.edu.cn

收稿日期: 2020-03-19; 修回日期: 2020-04-11; 接受日期: 2020-04-23; 网络出版日期: 2020-10-20

国家重点研发计划 (批准号: 2016YFC0300801) 和国家自然科学基金 (批准号: 41576101, 41427806) 资助项目

摘要 近年来, 强化学习方法得到了广泛的关注, 并已在控制领域取得了一定成果. 自主水下机器人 (AUV) 能完成复杂多样的水下任务, 是海洋探索开发任务中的关键设备, 强化学习技术的发展为 AUV 的智能化控制提供了新的思路. 本文着眼于强化学习与 AUV 控制的交叉领域, 重点介绍了强化学习在镇定控制和跟踪控制等底层控制任务方面的应用. 为此, 我们首先对强化学习框架进行了简要介绍, 并在其后对强化学习在 AUV 控制领域面临的挑战以及最新的进展进行了概述. 最后, 我们详细介绍了两种针对 AUV 的无模型强化学习控制方法, 希望能够为此类控制器的设计和使用提供参考.

关键词 强化学习, 自主水下机器人, 底层控制, 自主控制, 无模型

1 引言

自 1957 年科学家为了探索北极水域开始研发自主水下机器人 (autonomous underwater vehicle, AUV) 以来, AUV 的研发便备受重视, 其性能也随着相关技术的发展而不断完善, 给海洋研究领域带来了一场革命. 确保水下运动可控是水下机器人顺利执行水下任务的基本前提, 近年来, 随着自主运动控制的发展, AUV 已被成功用于场源搜索^[1]、管道检修^[2]、海底测绘^[3] 等各种复杂的水下任务中.

基于强化学习的控制器通过与环境的交互最大化给定任务中获得的累计奖励, 并能实现最优意义下的自主控制^[4]. 强化学习具有的面向未来、考虑长远收益的特性, 已经助其在控制领域取得了不少成果. 在无人机控制领域, 多位研究者^[5~7] 分别利用强化学习方法训练出了能胜过传统控制器或训练有素的飞行员的控制器. Abbeel 等^[8] 通过结合学徒式学习策略, 成功获得能执行任意高难度飞行特技的控制器. 在无人驾驶领域, Hester 等^[9] 在实际车辆上进行实时强化学习, 通过控制油门和刹车控制车速, 实验中控制器仅需经过 3 min 的学习便能使车速稳定在设定速度. Kendall 等^[10] 同样在实际车辆上展开实验, 并实现了更高级别的自动驾驶. 通过搭配车载电脑实时训练, 控制器从零开始学习, 在 30 min 内即获得了跟随道路自动行驶的能力.

引用格式: 许雅筑, 武辉, 游科友, 等. 强化学习方法在自主水下机器人控制任务中的应用. 中国科学: 信息科学, 2020, 50: 1798–1816, doi: 10.1360/SSI-2020-0065
Hsu Y C, Wu H, You K Y, et al. A selected review of reinforcement learning-based control for autonomous underwater vehicles (in Chinese). Sci Sin Inform, 2020, 50: 1798–1816, doi: 10.1360/SSI-2020-0065

图 1 智能体 - 环境交互

Figure 1 The agent-environment interaction

强化学习在这些领域的成功暗示了其在控制 AUV 方面的潜力. 然而, 水下环境远比空中和陆地环境更加复杂, 一方面, 水下环境中光的散射吸收, 高频信号的迅速衰减等特性导致水下环境难以探测, 机器人之间难以通信, 水下控制问题的建模形式与水上控制问题相比具有较大差异; 另一方面, 水体的密度远大于空气, 而与机器人密度较为接近, 导致 AUV 的动力学更为复杂, 基于模型的算法难以直接应用, 无模型算法也更难收敛; 另外, 水下实验成本较高, 因此对于算法的安全性, 采样效率有更高的要求, 这些都极大地增加了控制问题的难度. 本文简述了 AUV 底层控制任务中强化学习方法的应用进展, 希望能对此领域的研究有所启发. 本文的后续章节安排如下: 第 2 节简要介绍了强化学习和基本算法; 第 3 节讨论了应用强化学习方法控制 AUV 面临的挑战; 第 4 节介绍了此领域的研究进展; 第 5 节较为详细地介绍了两个不同的水下任务及其对应的基于强化学习的控制器, 以更好地说明这种控制器的优势; 第 6 节讨论了未来研究的方向; 最后一节对本文进行了总结.

2 强化学习介绍

本节简单介绍了强化学习中的基础概念和经典算法. 由于水下环境具有不确定性, 本节主要考虑无模型的算法.

2.1 Markov 决策过程

本小节的内容主要基于 Sutton 在文献 [11] 中的介绍. 强化学习旨在解决基于 Markov 决策过程的问题, 该类问题由 4 个基本元素组成: 状态集合 $\mathcal{S}$ (以及起始状态分布 ρ_0), 动作集合 $\mathcal{A}$, 奖励函数 $r(s, a)$ 和转移概率函数 $p(s', r|s, a)$. 其中, 状态的转移具有 Markov 性, 即下一个状态的生成只与当前的状态和动作有关, 而与历史数据无关.

智能体通过强化学习方法, 从与环境的交互中掌握环境状态与控制输出之间的关系, 从而优化其控制性能. 这使得强化学习方法能应对传统动态规划方法 [12] 不适用的情况, 即转移概率函数 p 未知的情况. 在每一次交互时, 智能体会观察其所处的状态并根据此观测选择一个动作. 接着, 环境会根据当前的状态和刚选择的动作改变, 并生成一个全新的状态, 同时产生一个奖励信号表示对此次动作的评价. 图 1 描绘了这种智能体 - 环境交互过程, 该循环会生成一系列状态、动作和奖励: $\tau = (s_0, a_0, r_1, s_1, a_1, r_2, \dots)$. 智能体的最终目标是最大化获得的回报 $R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t$, 即在循环中获得的累计奖励, 其中 $\gamma \in (0, 1]$ 是用于降低未来奖励的权重的折扣因子.

2.2 基于价值函数的强化学习方法

智能体遵循的选择动作的规则称为策略, 一般用 π 表示. 从某一状态 s 开始采用固定策略行动能

获得的未来回报的期望称为 s 的价值. 对于 Markov 决策过程, 将状态价值函数和状态 - 动作价值函数定义如下:

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s], \quad (1)$$

$$Q^\pi(s, a) = \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s, a_0 = a], \quad (2)$$

其中 τ 表示一条轨迹, s_0, a_0 分别为其初始状态和初始动作. 若一个策略能使在各种状态下获得的未来回报的期望最大化, 我们称之为最优策略, 一种构造最优策略的方法即选择能够使状态 - 动作价值函数 $Q^\pi(s, a)$ 最大化的动作. 如果一个价值函数对应的策略为最优策略, 我们称之为最优价值函数, 可以表示为

$$V^*(s) = \max_{\pi} \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s], \quad (3)$$

$$Q^*(s, a) = \max_{\pi} \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s, a_0 = a]. \quad (4)$$

价值函数能够衡量某个状态或状态 - 动作对的优劣, 在最优策略的搜索中起引导作用, 因此在强化学习算法设计中十分重要. 用于估计价值函数的方法一般从 Bellman 方程导出:

$$V^\pi(s) = \mathbb{E}_{s' \sim P, a \sim \pi} [r(s, a) + \gamma V^\pi(s')], \quad (5)$$

$$Q^\pi(s, a) = \mathbb{E}_{s' \sim P} [r(s, a) + \gamma \mathbb{E}_{a' \sim \pi} [Q^\pi(s', a')]], \quad (6)$$

$$V^*(s) = \max_a \mathbb{E}_{s' \sim P} [r(s, a) + \gamma V^*(s')], \quad (7)$$

$$Q^*(s, a) = \mathbb{E}_{s' \sim P} \left[r(s, a) + \gamma \max_{a'} Q^*(s', a') \right], \quad (8)$$

其中 P 表示状态转化概率 $p(s', r | s, a)$. 大多数无模型强化学习方法可以抽象为策略评估和策略改进的迭代, 策略评估时对当前策略对应的价值函数进行估计, 策略改进时根据估计出的价值函数改进策略. 这种迭代过程会在策略与价值函数相对应时稳定, 此时即满足最优价值和最优策略的 Bellman 方程. 这种思路被称为广义策略迭代 (GPI).

Bellman 方程的自洽性说明价值函数能够使用自举方法估计, 在基于价值函数的强化学习方法中, 时序差分 (TD) 算法便是一类继承该思想的常用算法. TD 算法通过最小化 TD 误差 δ 更新对价值函数的估计:

$$Q^\pi(s_t, a_t) \leftarrow Q^\pi(s_t, a_t) + \alpha \delta, \quad (9)$$

其中 α 表示学习率, $\delta = Y - Q^\pi(s_t, a_t)$ 为目标价值 Y 以及其估计价值之间的误差. 在不同算法中, 目标价值的取值不同, 如在线策略 SARSA 中采用 $Y = r_t + \gamma Q^\pi(s_{t+1}, a_{t+1})$, 而离线策略 Q 学习采用 $Y = r_t + \gamma \max_a Q^\pi(s_{t+1}, a)$. 在线策略指采样时用的策略与正在更新的策略一致, 而离线策略则会使用不同的策略与环境交互.

除了自举之外, 在回合式任务中还可以采用 Monte Carlo (MC) 方法更新价值函数的估计. 在这类方法中, 通过对已获得的回报取平均以估计价值函数. MC 类方法和 TD 类方法可以通过采取多步自举统一, 从而结合两种方法的优点, 使算法的性能得到提升, 但同时也加大了计算量.

2.3 基于策略的强化学习方法

基于价值的算法最大的缺陷在于需要对价值函数进行最大化操作, 寻找能产生最大价值的动作, 因此其难以处理具有连续动作空间的问题. 要解决这个问题, 可以将策略参数化, 即采用 π_θ . 这种随

机策略与之前的贪婪策略相比, 在处理输入信息不完整的情况时更具有优势. 此外, 若是利用神经网络作为非线性函数逼近器, 这类方法还能够处理高维的输入. 这类策略的性能函数 $J(\pi_\theta)$ 可以定义为累计折扣奖励或是平均奖励的期望, 如:

$$J(\pi_\theta) = V^{\pi_\theta}(s_0). \quad (10)$$

通过采用梯度下降法更新策略的参数可以优化性能

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\pi_{\theta_k}). \quad (11)$$

Sutton 提出的策略梯度定理^[13] 为这类算法打下了坚实的理论基础:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s \sim \rho^\pi, a \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s, a)], \quad (12)$$

其中 ρ^π 表示状态分布. 该式表明性能函数相对策略参数的梯度与策略改变导致的状态分布改变无关, 即只要保证期望无偏, 该梯度就可以通过采样的方式估计. REINFORCE 是该思想的一种 Monte Carlo 实现, 通过对回报采样估计动作 - 价值函数^[14].

另一类常用的算法是动作 - 评价 (actor-critic) 类方法. 这类方法取价值函数作为 critic 辅助策略参数的更新, 而 actor 指用来更新策略的部分, actor 在计算梯度更新策略时会借助 critic 估计的价值函数 $Q^\omega(s, a)$. 在一定约束下, 采用这样的近似值并不会导致偏差^[13]:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s \sim \rho^\pi, a \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(a|s) Q^\omega(s, a)]. \quad (13)$$

当随机策略的方差变为零时, 随机策略会退化为确定性策略. Silver 等^[15] 开创了一类使用确定性策略的新的策略梯度方法, 他们证明了当方差趋于零时, 随机策略的梯度收敛到如下形式的确定性梯度:

$$\nabla_\theta J(\mu_\theta) = \mathbb{E}_{s \sim \rho^\mu} [\nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s, a)|_{a=\mu_\theta(s)}]. \quad (14)$$

这种方法可以理解为通过参数化的策略近似价值函数的最大化操作:

$$\max_a Q(s, a) \approx Q(s, \mu(s)). \quad (15)$$

确定性策略梯度 (DPG) 方法无需在动作空间采样积分, 因此在处理具有高维动作空间的问题时具有较大优势, 从该算法中衍生出的其他算法也表现不俗^[16, 17].

策略梯度方法一般在参数空间中度量策略之间的距离, 而在概率流形中度量距离能够更好地确保每一步更新的效果. 研究者基于该思想提出了自然策略梯度方法^[18], 该类方法通过采用不同的替代目标函数优化策略^[19, 20].

2.4 基于模型的强化学习方法

无模型的强化学习方法能直接从交互中学习, 而基于模型的强化学习则需要利用到环境模型, 即状态转移概率 $p(s', r|s, a)$. 该模型的准确程度对基于模型的强化学习方法的性能好坏能造成决定性的影响. 一般来说, 基于模型的强化学习方法主要关注两个问题, 即模型的获取方式以及模型的利用方式.

图 2 (网络版彩图) AUV 的 6 个自由度
Figure 2 (Color online) The six DOFs of the AUV

在一些简单的任务中, 可以基于先验知识建立准确的数学模型, 对于其他情况, 可以通过与环境的交互学习模型. 若已知模型的形式, 可以通过线性回归等方法推断未知参数, 否则可以通过 Gauss 过程 (GP) 估计转移函数的分布^[21]. 在建立环境模型之后, 可以通过诸如动态规划、启发式搜索、穷举搜索等方法模拟接下来的情况甚至所有可能遭遇的情景. 除此之外, 还可以将生成的模型与无模型算法相结合, 如将基于模型的规划方法作为指导策略更新的专家^[22], 利用基于模型的规划给策略提供辅助状态信息^[23], 利用模型生成数据加强样本^[24]等.

近年来, 有关强化学习的研究不断发展, 并涌现了一大批优秀的算法, 本文将不再对此进行更详细的介绍. 在下面的章节中, 将重点介绍强化学习和 AUV 控制领域的交叉部分.

3 采用强化学习控制 AUV 面临的挑战

基于强化学习的控制方法与包括比例 - 积分 - 微分 (PID) 控制^[25]、自适应控制^[26]、反步控制^[27]、滑模控制^[28]等在内的经典控制方法最大的区别在于其能从与环境的交互中学习. 从运动学的角度看, 控制 AUV 相当于控制一个在六维空间内自由漂浮的刚体 (图 2), 然而水下环境复杂的动力学使得该问题的难度大大提升^[29]. 对大多数的经典控制器而言, 无法获得精确的动力学模型成为了解决该控制问题最大的障碍. 另外, 许多控制器为了简化问题, 采用线性模型或对模型进行解耦, 这种设定没能充分考虑水下环境具有的不确定性和特殊干扰, 也就无法完成对 AUV 的精确自主控制. 而基于强化学习的控制器可以脱离模型进行训练, 则不存在此问题, 这也是该方法的一大优势. 相对地, 利用强化学习方法可能面临其他的挑战, 下面讨论了一些需要注意的方面.

3.1 采样效率

应用强化学习算法要求 AUV 与环境交互获取经验, 然而进行水下实验不仅需要花费大量的时间, 还需消耗巨额的人力和物力. 一方面, AUV 的制造成本十分昂贵, 并且需要仔细维护以减少磨损, 避免碰撞; 另一方面, 建造能进行实验的水箱或寻找合适的水场也并非易事. 即使所有的准备工作都顺利完成, 收集实验数据的过程也十分耗时. 因此, 提高采样效率, 减少实际实验中所需的交互已成为一个突出的问题, 其重要性甚至超过了限制内存消耗和计算复杂性.

在机器人控制中经常采用基于模型的方法, 通过模型生成数据来减少实际中所需的采样, 对于难以得到准确模型的 AUV, 可以利用简化模型进行仿真预训练, 从而加速算法在实际水下环境中的收

敛. 基于离线策略的方法也能有效缓解采样难的问题, 通过将收集到的数据存储在经验池中重复利用, 能够提高样本的利用率, 减少采样的需求.

3.2 开发与探索的权衡

在学习过程中, 开发 (exploitation) 意味着选择具有最大的估计价值的动作, 这样能够使该步的预期回报最大化; 而通过选择其余动作进行探索 (exploration) 则有机会在未来获得更多的回报^[11], 并且有利于访问更多类型的状态从而获得更鲁棒的策略.

如何确保充足的探索是强化学习算法中的固有问题, 而 AUV 需要一个足够鲁棒的控制策略帮助其应对多变的水下环境, 因此该问题尤为重要. 在强化学习算法中, 常通过选择随机动作来探索未知的状态, 而尽管在训练中通过强制试错探索适用于大多数问题, 但是考虑到 AUV 损坏后高昂的维修费用和漫长的返修过程, 要提高强化学习算法的实用性, 需要在保证探索的基础上重视安全问题, 而这方面的考虑在强化学习领域经常被忽略^[30]. 一种解决方案是在人工水池等可控的水下环境进行训练, 确保 AUV 已学习到足够安全的策略再在未知的水体环境中进一步实验. 另一种可能的改善方案是提前准备好应急控制方案, 当 AUV 在训练过程中遭遇危险时, 如过于接近障碍物时, 则提高该控制方案的优先级^[31].

3.3 模型不确定性

常见的 AUV 动力学模型一般是从刚体的牛顿-欧拉方程推导出的, 其中考虑了惯性广义力、流体力学、重力、浮力和推进器的影响^[29]. 但是, 这些影响因素或是有着高度非线性、时变性, 或是与 AUV 的具体结构细节密切相关, 因此很难建立出通用而精确的 AUV 动力学模型. 尽管使用精确的模型能够回避在现实中难以采集数据的问题, 但是只要模型中存在偏差, 就可能导致控制效果大打折扣, 甚至毁坏 AUV. 对于采用模型的算法和利用仿真环境训练的无模型算法, 需要想办法跨越这种现实落差 (reality gap).

要克服不确定性带来的困难, 一方面要提高算法的鲁棒性, 在训练过程中寻找对噪声更不敏感的策略, 而不是单纯追求获得高奖励或高准确率; 另一方面则需要发挥强化学习算法与环境交互的特点, 在水下实时学习, 若需要进行精密的控制, 则甚至是在任务行进中也需要不断对最优的控制参数进行微调, 才能适应不断变化的水下环境.

3.4 状态部分可观

大多数的强化学习算法都是基于可以获得环境的完全观测的假设, 然而水下环境具有高度不确定性且存在多种干扰, 这为 AUV 采集有用信息带来了巨大挑战.

以水下视觉为例, 用于收集水下视觉信息的设备主要分为基于光学的和基于声纳的两种^[32]. 光学传感器能够获得包含颜色信息在内的高分辨率数据, 但是在浑浊的水中, 水分子、溶解于水的有机物、无机物, 以及各种类型的悬浮颗粒会导致光的散射和吸收, 导致生成度图像亮度和对比度较低, 可视性较差. 另外, 由于光在水中的衰减率与波长有关, 还会出现颜色失真的情况^[33]. 声纳能够获得更远处的信息, 但是其提供的低分辨率图像不足以识别物体.

至于定位问题, 与在空中和地上不同, 在水下高频信号迅速衰减, 因此无线电、扩频通信, 以及 GPS 都无法使用. 通过基于声学的传感器和通信方法能够对 AUV 进行定位, 但这种方式受到带宽、传播路径、声速等的限制^[34]. 在多数情况下, AUV 的控制设计人员需要在获取更多信息和降低能源消耗中进行取舍, 避免缩短 AUV 可执行任务的时间. 在实际任务中, AUV 一般只能获取十分有限

的信息.

这些传感器的限制使得 AUV 的控制研究无法简单地仿照无人机控制和无人驾驶的相关研究, 需要在状态设计和控制器结构等方面进行改进. 基于随机策略的方法能更好地应对此种情况, 另外可以通过采用长短期记忆网络来从历史数据中提取更多的状态信息, 弥补传感器无法获得的部分信息缺失.

4 强化学习方法在 AUV 中的应用

本节简要介绍了近年来应用强化学习方法控制 AUV 的研究. 强化学习也被应用在 AUV 的一些高层控制任务中, 如路径规划^[35~37]等, 但这些内容不包括在本文的讨论范围内.

4.1 Markov 决策过程建模

在应用强化学习方法之前, 需要将控制问题建模成 Markov 决策过程, 如前文所述, 需要对 4 个元素 $[S, \mathcal{A}, r(s, a), p(s', r|s, a)]$ 进行定义. 对于我们考虑的底层控制问题, 动作即为 AUV 的控制输入, 水下的状态转移概率是未知的, 因此问题建模的关键是以合适的方式表示状态并构造适当的奖励函数.

4.1.1 状态设计

一种最直观的定义方式便是将原始观测数据, 如图像、速度等直接作为状态. 这么做的好处是能保留最多的有效信息, 但是没有考虑维度诅咒的问题, 当状态空间维数增加, 则会需要更庞大的样本量和计算量来确保算法收敛. 因此, 状态的表述应当在不造成感知混淆 (一组变量可能对应不同的状态) 的前提下包含尽量少的变量. Wu 等^[38]仔细地研究了 3 种不同的深度控制问题并分别设计了合适的状态表示, 在第 5 节中将会对此工作进行更详细的说明. 另外, 一个好的状态设计还能有效提高算法的鲁棒性, 从而获得更为通用的策略, 例如在文献^[39, 40]中采用了目标导向型的状态设计, 训练出的策略能够很好地适应不同的控制目标值.

4.1.2 奖励函数设计

另一方面, 奖励函数的设计也不容忽视. 在处理多个有冲突的目标时, 一种常见的处理方法是给定权重将其组合在一起, 这种方法易于操作并且能获得单一的最优解. 在 AUV 控制问题中, 一般奖励函数由两部分组成, 一部分用于评估与目标的误差, 另一部分用于限制推进器的使用或避免控制量突变. Carlucho 等^[39]指出了这些项的重要性, 前者能确保 AUV 达到控制目标, 后者有利于节约能源, 防止推进器输出震荡. 然而, 各项的权重通常根据目标的相对重要性凭经验选择, 但权重细微的调整也可能导致学到的策略产生很大差异. 文献^[41]中采用了多目标强化学习方法解决权重难以调试的问题, 该方法可以同时生成多个最优策略, 每个策略分别具有不同的优势, 如路径最短、最终速度最小、航向误差最小等, 可以从其中选择最符合需求的策略.

分段函数也是奖励函数中一种常见的形式, 研究者可以根据 AUV 对当前所处状态的优劣分级给出奖励. 在文献^[40]中建议可以通过在训练初期放宽分级的条件帮助 AUV 获得更多有用的经验. 例如, 若 AUV 在到达指定目标点的球形邻域时可获得正向奖励, 那么在训练初期应该适当增大该邻域的半径.

4.2 强化学习在传统 AUV 中的应用

传统的 AUV 通常由螺旋桨、尾舵等控制, 它们大多数有鱼雷形状的船身以获得更好的水下性能 (请参考文献^[42]中的图 “The REMUS family”). 下面根据控制任务的类型介绍在此类 AUV 上开展

的强化学习应用研究.

4.2.1 镇定控制任务

镇定控制是 AUV 控制中最基本的任务. Fernandez-Gauna 等^[43]使用连续 actor-critic 自动学习机 (CACLA) 进行了 AUV 速度控制实验. CACLA 与其他策略梯度方法不同的是, 其只有在评价为严格正值时才在动作空间更新其策略. 此外, 此研究验证了在训练初期采用 PID 的输出代替随机选取的动作有助于使学习到的策略更接近最优策略. Walters 等^[44]采用了基于模型的动态规划方法在现实环境下开展了 AUV 的镇定控制实验, 其中动力学模型是在实验中实时学习获得的. 此实验主要关注时变无旋流的干扰, 并通过 Lyapunov 理论论证了所用方法能够收敛到目标值并获得最优策略. Carlucho 等^[39]借鉴深度确定性策略梯度方法 (DDPG) 对 AUV 进行控制, 并应用了目标导向型控制框架. 与前人的实验相比, 他们的实体 AUV 控制实验包括了 6 个维度, 而不仅仅是在平面内操控. Knudsen 等^[45]采用了双重控制器, 其中基于深度确定性策略梯度方法的控制器负责水平面内的运动, 而 PD 控制器负责竖直平面内的运动. 在实际水下测试中, 每隔一段时间对目标值进行更换, AUV 都能迅速作出反应, 稳定至新的坐标与姿态.

4.2.2 目标点跟踪任务

目标点跟踪可以看作是介于镇定控制和目标跟踪之间的一项过渡任务, AUV 需要依次到达指定目标点, 但点与点之间的路径并未限定. 文献 [46] 通过划分网格将该连续控制问题离散化, 采用了一种简化的表格 Q 学习方法并在仿真环境和现实环境中都进行了实验. Carlucho 等^[40]致力于提高策略的容错性, 提高策略应对推进器失效情况的能力, 这意味着训练出的策略需能同时控制过驱动和欠驱动的 AUV. 他们采用了与文献 [39] 中相似的方法完成此任务. 文献 [47] 中借助协方差分析提高了基于模型的直接策略搜索算法获得的策略的鲁棒性. 其通过在 AUV 的推进器输入中加入 Gauss 噪声测试策略对噪声的敏感程度, 实验结果表明无噪声环境下策略的性能及其鲁棒性之间并无显著关系, 因此可以通过协方差分析筛选性能优秀且对噪声鲁棒的策略.

Istituto Italiano di Tecnologia 的研究者们在此课题上开展了一系列实验^[41, 48, 49]. 在 2013 年, 他们设计了一种线上训练的利用基于模型的策略搜索框架的控制器. 然而, 虽然控制器在仿真环境中运行良好, 但由于策略是时间的开环函数, 所以无法在实际水域中应用. 2014 年, 这个问题通过引入状态反馈闭环环路得到了解决, 该策略成功移植到实体 AUV 上进行了验证实验, 并且还开展了一系列推进器失效情况下的实验. 这些方法需要用到 AUV 的水下动力学模型, 因此实际中应用效果和应用场景有限.

4.2.3 跟踪控制任务

AUV 的跟踪控制问题也是一个相当热门的研究课题, 此问题中 AUV 需要沿着既定路径行进, 若该路径为时间的函数则称之为轨迹跟踪问题, 否则称作路径跟随问题.

Palomeras^[50]提出了一种用于 AUV 控制的结构体系, 并在反应层中使用了一种 actor-critic 类强化学习算法, 采用该控制结构的 AUV 成功在现实中完成了基于视觉的电缆跟随问题. 作者提出了一种二段式训练方法, 即先在仿真环境中从零开始进行训练, 将结果作为在实际环境中的初始策略重新训练, 从而提高算法的收敛能力, 文献 [39] 在实验中也采用了这种训练方法. 在电缆跟随任务中, AUV 的不同运动被解耦并分别训练对应策略, 其输入状态包括从采集的图像中获得电缆的位置和角度信息. 文献 [31] 中扩充了该实验的内容来检验所获策略的鲁棒性, 一组新增的实验重新设计了不同的电缆摆

放方式并在没有重新训练策略的情况下检验原策略的性能, 另一组则是在训练过程中改变 AUV 和电缆的相对距离. 两组实验的结果共同表明了该策略具有较高的适应能力.

由于跟踪问题更为复杂, 因此大多数算法仍停留在仿真实验的阶段, 并未在实际场景中得到充分验证. Sun 等在文献 [51] 中使用正则化的极限学习机代替 Q 学习中的查找表进行实验, 然而该文章中对于实验设置和结果的表述较为模糊, 因此难以借鉴. Shi 等 [52] 改进了计算确定性策略梯度算法中用于更新 critic 的目标值的方法. 这种改进的方法称为伪平均 Q 学习方法, 它采用了多个 actor, 并对多个历史动作估计值取平均, 通过减少目标估计误差的方差使学习过程更加稳定. 在文献 [53] 中, 该方法改进为同时利用多个 actor 和多个 critic, 其中这些 critic 通过使用期望绝对 Bellman 误差加速训练过程. 这些算法都在 AUV 的水下轨迹跟踪仿真实验中得到了验证. 文献 [54] 研究了交互式强化学习方法的应用, 探讨将研究人员的经验知识通过人工给予奖励信号的方式与强化学习方法相结合的影响. 在控制尾舵角度来实现轨迹跟踪的仿真实验中, 结合了人工离散奖励信号后深度 Q 学习方法能收敛得更快.

为了验证算法的有效性, 一些研究人员还对控制系统的稳定性进行了严格的理论分析. 文献 [55] 利用深度确定性策略梯度方法解决 AUV 的轨迹跟踪问题, 并通过理论分析证明了采用合适的奖励函数设计时该系统稳定. Cui 等 [56] 在 2014 年提出了一种基于模型的自适应控制算法, 该算法采用了 actor-critic 结构, 其中 actor 网络可以对动力学方程中的不确定性进行补偿, 而 critic 网络能评估跟踪性能. 在 2017 年, 算法将包括推进器死区、饱和区、名义力和实际力之间的关系等输入非线性性纳入考虑 [57]. 而 2019 年, 研究者针对连续时间系统改进了算法并且新算法能够处理模型完全未知的情况 [58]. Nussbaum 类型的函数被引入来处理控制方向未知的问题. 与文献 [57] 中基于离散时间模型的算法相比, 改进的算法能够在 AUV 到达稳态后避免控制量震荡. 另一个仿真实验结果表明该算法能达到与利用了模型的通用神经网络控制相同水平的结果. 另一方面, 在文献 [59] 中研究了一种事件触发的强化学习自适应跟踪控制算法, 该方法同时考虑了长远性能, 未知的动力学模型和外部干扰, 并且与传统的时间触发方法相比较, 能减少控制器的更新频率, 从而显著地减少计算量和 AUV 能源消耗.

4.3 强化学习在仿生 AUV 中的应用

仿生 AUV 通过模仿水下生物的游动模式获得更长的水下续航时间, 减少系统噪声并提高控制机动性. 与传统 AUV 相比, 它们的运动模式更为复杂, 其控制器的设计也需要投入大量心血.

对于仿鱼 AUV, 由鳍类提供的推进力和力矩取决于其运动波形、波长、幅度和频率等因素的综合影响. Lin 等 [60] 通过使用在线 Q 学习方法选择合适的频率控制一种仿鱼 AUV 的两个波浪状鱼鳍完成航向控制任务. 该实验的结果差强人意, 其中偏航角的误差较大. 在 2010 年, Lin 等 [61] 采用了改进的 Q 学习方法在躯体更扁平的 AUV 上完成了相同的任务, 并成功使该 AUV 平稳地游动. 该改进的方法将获得的数据存在缓冲区内重复利用, 并根据计算的数据相似度淘汰旧的数据, 另外在实验中也借助了 PID 避免算法从零开始学习时, 在初期可能面临的学习率过低的情况. 文献 [62] 中的研究表明采用分层的强化学习结构能够在这种运动控制问题中提高 Q 学习算法的收敛速度. 文献 [63] 中采用了一种对抗式模式控制结构, 利用进化算法调节中枢模式发生器模型的参数, 生成规律控制信号控制一种仿鱼 AUV 尾部的 4 个关节, 并采用基于柔性 actor-critic 的方法生成补偿量来辅助控制. 该方法能在仿真环境中准确控制仿鱼 AUV 到达目标, 同时降低此过程所耗能量.

Aqua 改进自一种六足步行机器人, 为能够完成水下作业的仿生 AUV (请参考文献 [64] 中的图 11). Meger 等 [65] 通过一种策略搜索方法 PILCO (probabilistic inference for learning control) 学习 AUV 的 6 个鳍的运动模式, 其中动力学模型是根据采集的数据拟合的. 在真实的 AUV 上执行的 U 型

转弯、螺旋前进等 6 个不同的固定深度任务中, 有 5 个任务在 7 个训练迭代内即获得了令人满意的结果. 另外一组实验通过共享仿真实验中的数据, 再一次验证了不正确的模型会降低算法的性能. 为了解决该方法计算量较大的问题, Higuera 等^[66]提出了一种结合神经网络控制器的改进的深度 PILCO 方法. Zhang 等^[67]研究了一种仿蛇 AUV (请参考文献 [67] 中的图 1), 该 AUV 上有两个滑翔翼, 因此继承了水下滑翔机的部分优异性能. 研究中应用了 REINFORCE 算法取得了良好的实验结果. 文献 [68] 中采用深度 Q 学习方法控制水下滑翔机的滚转角, 能够在利用南海海流数据模拟的仿真水流环境中, 获得比传统的 PID 控制方法更优的行进路径.

5 案例分析

本节将详细介绍两个使用强化学习方法控制 AUV 的案例. 第 1 个例子是 AUV 海底跟踪控制问题, 展示了一种通用的水下控制问题建模思路, 并可为如何解决强化学习方法应用时遇见的困难提供参考. 该例子采用低阶系统特征作为输入, 通过保留部分历史数据减小状态仅部分可观带来的不良影响, 采用确定性强化学习算法, 配合优先经验回放技巧提高了采样效率, 能够在模型未知的情况下取得良好跟踪控制结果. 第 2 个例子充分利用了强化学习能处理高维输入的优势, 采用了端到端的学习方法, 设计了能提取有效信息的感知网络, 无需位置信息便能实现管道跟踪控制, 展现了强化学习方法在水下自主控制领域更多的可能性.

5.1 海底跟踪问题

5.1.1 海底跟踪问题描述

在海底跟踪控制问题中, AUV 需要在匀速前进的同时维持与海床的距离为定值 z_r . 为了简化问题, 仅考虑 AUV 在竖直平面内的运动, 并假设行进速度为一固定常值, 则该问题中动作即为控制俯仰的相关推进器的输入, 状态可设置为 $\chi = [z, \theta, w, q]^T$, 其中 z 为深度坐标, w 为升沉速度, θ 和 q 分别为俯仰角和角速度. 为了避免角度的周期特性造成的混乱, 可以采用 $[\cos(\theta), \sin(\theta)]^T$ 代替 θ . 另外, 通过使用目标导向的表达式 $\Delta z \doteq z - z_r$ 替代 z 能够有效增加所学策略的通用性. 因此, 我们将状态定义为

$$s = [\Delta z, \cos(\theta), \sin(\theta), w, q]^T. \quad (16)$$

然而, 如图 3 所示, 若只考虑当前的深度, 则未来地势的走向将难以预测, 给决策带来许多困难, 因此我们考虑通过保留部分历史观测数据改善这个问题, 即用 $[\Delta z_{t-N+1}, \dots, \Delta z_{t-1}, \Delta z_t]$ 替换 Δz , 其中 N 表示保留的历史序列的长度. 综上所述, 将海底跟踪问题的状态设计为

$$s = [\Delta z_{t-N+1}, \dots, \Delta z_{t-1}, \Delta z_t, \cos(\theta), \sin(\theta), w, q]^T. \quad (17)$$

奖励函数的设计与第 4 节中提到的基本一致:

$$r = \rho_1 \Delta z_t^2 + \rho_2 w^2 + \rho_3 q^2 + u^T R u, \quad (18)$$

其中第 1 项旨在最小化深度控制的误差, 而另外几项的作用是最小化 AUV 能源消耗, 系数则起到了协调各项目标的作用.

图 3 (网络版彩图) 深度控制问题中的感知混淆
Figure 3 (Color online) Perceptual aliasing in the depth control problem

5.1.2 深度控制器设计

此问题可以通过采用 DPG 算法解决. 如前文所述, 该算法沿性能函数的梯度方向 $\nabla_{\theta} J(\pi_{\theta})$ 更新参数化策略 π_{θ} , 该梯度可以近似为

$$\nabla_{\theta} J(\theta) \approx \frac{1}{M} \sum_{i=1}^M \nabla_{\theta} \pi(s_i | \theta) \nabla_{u_i} Q(s_i, u_i | \omega), \quad (19)$$

其中 (s_k, u_k, s_{k+1}) 为 k 时间点的一组变换, $Q(s, u | \omega)$ 为值函数的一个参数化估计. 如图 4 所示, 策略和值函数分别采用一个 3 层的和一个 4 层的神经网络近似, 并应用了激活函数 ReLU 提高收敛速度. 为了提高采样效率, 我们采用优先经验回放技巧^[69], 根据存储样本的优先级重复利用历史数据. 由于一组样本的 TD 误差可以用来衡量该组样本待学习的信息量, 因此定义每组样本的优先级如下:

$$\text{PRI}_k = |r_k + \gamma Q(s_{k+1}, \pi(s_{k+1} | \theta) | \omega) - Q(s_k, u_k | \omega)|. \quad (20)$$

在训练过程中, 优先级越高的样本越有可能被选中.

5.1.3 海底跟踪仿真结果

我们从中国科学院沈阳自动化所提供的南海海底 (23°06'N, 120°07'E) 处采样生成的真实数据集中生成一条路径进行仿真实验. 经过预实验, 我们将保留的历史 Δz 设为 3. 如图 5 可以看到训练过后的控制器在测试中表现良好, 此方法无需知晓 AUV 的动力学模型, 但能达到与非线性模型预测控制方法 (NMPC) 相当的性能.

5.2 端到端控制问题

5.2.1 管道跟踪问题描述

深度学习具有从未经处理的高维输入中学习的能力, 因此能实现端到端的控制策略, 省去复杂的系统间调试. 然而, 这种高维输入会使得算法难以收敛, 在处理涉及复杂动力学模型的底层控制问题时尤为严重. 在本小节中我们展示一种端到端的控制策略解决简单的管道跟踪问题. 该问题同时使用了运动学信息和视觉信息两种输入, AUV 需要在自身位置以及管道位置未知的条件下, 沿着管道行进并保持管道在相机视野中 (图 6(a) 和 (b)). 输入的视觉信息为一张 $84 \times 84 \times 3$ 的图片, 运动学信息为方向矢量和速度矢量.

我们利用原始图像信息直接对控制器进行训练, 而图像处理技术仅用来辅助奖励信号的生成. 经过一系列处理步骤, 可以提取出图像中管道的中心线, 接着可以计算出中心线与视野中心的距离 d_c 以

图 4 (网络版彩图) 评价网络 (a) 和策略网络 (b) 的结构

Figure 4 (Color online) Structure of the evaluation network (a) and policy network (b)

及偏差角度 θ_c (如图 6(c)). 我们设定动作为与前进和偏航操作有关的推进器的输入, 奖励函数为:

$$r = u \cdot (|\cos \theta_c| - d_c d_{\max}^{-1}), \quad (21)$$

其中 u 为 AUV 的行进速度, $d_{\max}$ 为视相机视图对角线长度的一半.

5.2.2 端到端控制器设计

输入的视觉信息和运动学信息分别由两个不同的编码网络处理, 前者为一个 4 层的 CNN, 后者为一个长短期记忆 (LSTM) 网络. 如图 7 所示, 两者的输出共同连入一个全连接层, 接着连向一个价值网络和一个策略网络. 我们采用一种自然策略梯度方法训练此网络, 即近端策略优化 (PPO) 方法. 该方法采用一种特殊的目标函数替代常见的性能函数:

$$L^{\text{PPO}} = L^{\text{CLIP}} - \lambda_1 \hat{\mathbb{E}}_t [(V_\theta(s_t) - V_t^{\text{targ}})^2] + \lambda_2 \hat{\mathbb{E}}_t [H(\pi_\theta(\cdot|s_t))], \quad (22)$$

式中第 2 项的目的是使值函数的误差最小化, 以更好地估计优势函数 $\hat{A}_t$, 最后一项则是为了鼓励探索. 优势函数能够衡量一个动作的相对优势, 一般取作 $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$, V_t^{targ} 可以取作累积回报, $\hat{\mathbb{E}}_t$ 表示对一批样本取均值, H 用于计算分布的熵, 而 L^{CLIP} 则是一个替代目标:

$$L^{\text{CLIP}} = \hat{\mathbb{E}}_t [\min(w_t(\theta) \hat{A}_t, \text{clip}(w_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t)], \quad (23)$$

图 5 (网络版彩图) NNDPG, NMPC 方法的跟踪轨迹以及实际海床 [38]

Figure 5 (Color online) Tracking trajectory of NNDPG, NMPC, and the realistic seafloor [38] @Copyright 2018 IEEE

图 6 (网络版彩图) 管道跟踪仿真场景 (a), 相机视野 (b), 以及 d_c 和 θ_c 示意图 (c)

Figure 6 (Color online) Simulation scene for pipe following (a), view of the camera (b), and illustration of d_c and θ_c (c)

其中函数 $\text{clip}(x, a, b)$ 将 x 限制在 $[a, b]$ 内, $w_t(\theta)$ 是用来衡量策略之间的差异的权重:

$$w_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}. \quad (24)$$

5.2.3 管道跟踪仿真结果

我们训练出的控制器在仿真环境中能够很好地完成管道跟踪任务, AUV 能够在位置信息和动力学模型未知的情况下沿着直线管道行进. 我们设计了两个额外的实验分别展示端到端控制框架的优势以及所获策略的鲁棒性. 在第 1 个实验中, 用从图像中提取的特征 θ_c 和 d_c 代替 CNN 的输出, 其余设定维持不变. 从图 8 的结果可以看出使用 CNN 的时候能获取更高的奖励, 即采用端到端的处理方式能更加充分地利用获得的信息, 避免主观信息提取过程中有效信息的流失. 第 2 个实验将输入的图像改为现实中的水下管线图像, 并利用训练的策略输出动作 (图 9). 结果表示, 在 30 幅图像中, 有

图 7 网络的结构

Figure 7 Structure of the network

图 8 (网络版彩图) 采用提取特征和原始信息的结果比较 [70]

Figure 8 (Color online) Comparison between using hand-designed image features and using the CNN encoder [70] @Copyright 2019 IEEE

21 幅图像对应的输出动作能够带动 AUV 转向正确的方向. 虽然转向的幅度仍有待改进, 但该结果仍展现了算法在实际应用中的潜力.

6 研究展望

我们认为, 在未来的研究中需要考虑以下问题:

(1) 如何获得更优的奖励函数形式? 在实际任务执行中需要调节许多矛盾, 如控制的准确性、响应的速度、能源消耗量、对噪声的鲁棒性等不同目标之间的矛盾, 仅靠人工设计奖励函数的形式, 往往难以平衡各个目标的需求. 尤其是当目标较为复杂时, 若没能全面考虑各种可能的情况, 会导致智能体过度追求高奖励, 呈现“病态”的运动模式.

(2) 如何结合专家知识? 此处的“专家”可以是人类研究员, 简化的动力学模型, 或是其他控制器等. 仅靠与环境的交互训练出控制策略是强化学习方法的优势, 然而完全不利用其他专家的知识并不是最佳的选择, 特别是对于许多奖励信号稀疏的问题. 除了参考这些知识进行状态和奖励设计之外, 现

图 9 (网络版彩图) 利用所学策略处理实际图像: 当左侧的推进器 (蓝条) 输出大于右侧 (红条) 时, AUV 向右转, 反之亦然^[70]

Figure 9 (Color online) The inference of the learnt policy on several realistic underwater pipeline images: when the output of the left thruster (blue bar) is larger than that of the right thruster (red bar), the AUV tends to turn right; vice versa^[70] ©Copyright 2019 IEEE

有的应用中利用 PID 控制器、仿真动力学模型对算法进行预训练提高准确率也是一种专家知识的利用方法, 未来也应该继续开发更多样的有效利用途径。

(3) 如何实现端到端的控制? 许多强化学习算法应用中依赖其他子算法来从感知信息中提取状态, 使得整体控制算法庞大而复杂。尽管深度学习的发展赋予了强化学习处理高维感知信息的能力, 但强化学习方法处理不完整、有延迟或含有噪声的未经加工的传感器数据的能力未得到充分验证, 在此方面仍有许多值得探究的课题。

(4) 如何实现多 AUV 合作? 随着水下研究的发展, 许多任务可能涉及多个 AUV 的合作交互, 需要共同优化一个目标, 如两台 AUV 相互配合共同搬运物件, 多台 AUV 共同搜索同一片海域等。在这种情况下, 每个 AUV 除了考虑自身的状态外, 还要观察其他 AUV 行动给环境带来的变化, 原有的许多强化学习方法也不再适用。

目前深度强化学习算法已经初步被应用在各种水下控制任务中, 往后的研究可以考虑采用分层强化学习、模仿学习、元强化学习、多智能体强化学习等方法进一步提升 AUV 的自主控制能力。

7 总结

本文从不同运动控制任务的角度概述了现有文献中使用强化学习控制 AUV 的研究成果, 虽然将强化学习与 AUV 控制结合仍面临着许多的挑战, 但可以看到该领域已经取得了令人鼓舞的进展。此外, 我们通过两个案例展示了强化学习在水下控制领域的可行性和潜力。我们相信强化学习方法的应用能提高 AUV 控制器的智能水平, 推动海洋研究与探索工作进一步发展。

参考文献

- 1 Li Z, You K Y, Song S J. AUV based source seeking with estimated gradients. *J Syst Sci Complex*, 2018, 31: 262–275
- 2 Xiang X B, Jouvencel B, Parodi O. Coordinated formation control of multiple autonomous underwater vehicles for pipeline inspection. *Int J Adv Robot Syst*, 2010, 7: 3
- 3 Ribas D, Palomeras N, Ridao P, et al. Girona 500 AUV: from survey to intervention. *IEEE/ASME Trans Mech*, 2012, 17: 46–53
- 4 Kiumarsi B, Vamvoudakis K G, Modares H, et al. Optimal and autonomous control using reinforcement learning: a survey. *IEEE Trans Neural Netw Learn Syst*, 2018, 29: 2042–2062
- 5 Kim H J, Jordan M I, Sastry S, et al. Autonomous helicopter flight via reinforcement learning. In: *Proceedings of the 16th International Conference on Neural Information Processing Systems*, Vancouver, 2003. 799–806
- 6 Bagnell J A, Schneider J G. Autonomous helicopter control using reinforcement learning policy search methods. In: *Proceedings of IEEE International Conference on Robotics and Automation*, Seoul, 2001. 1615–1620
- 7 Waslander S L, Hoffmann G M, Jang J S, et al. Multi-agent quadrotor testbed control design: integral sliding mode vs. reinforcement learning. In: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, 2005. 3712–3717
- 8 Abbeel P, Coates A, Ng A Y. Autonomous helicopter aerobatics through apprenticeship learning. *Int J Robot Res*, 2010, 29: 1608–1639
- 9 Hester T, Quinlan M, Stone P. A real-time model-based reinforcement learning architecture for robot control. 2011. ArXiv:1105.1749
- 10 Kendall A, Hawke J, Janz D, et al. Learning to drive in a day. In: *Proceedings of International Conference on Robotics and Automation*, Montreal, 2019. 8248–8254
- 11 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge: MIT Press, 2018
- 12 Bertsekas D P. *Dynamic Programming and Optimal Control*. Nashua: Athena Scientific Belmont, 1995
- 13 Sutton R S, McAllester D A, Singh S P, et al. Policy gradient methods for reinforcement learning with function approximation. In: *Proceedings of the 12th International Conference on Neural Information Processing Systems*, Denver, 1999. 1057–1063
- 14 Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach Learn*, 1992, 8: 229–256
- 15 Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms. In: *Proceedings of the 31st International Conference on Machine Learning*, Beijing, 2014. 387–395
- 16 Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. 2015. ArXiv:1509.02971
- 17 Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods. In: *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, 2018. 1582–1591
- 18 Amari S. Natural gradient works efficiently in learning. *Neural Comput*, 1998, 10: 251–276
- 19 Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization. In: *Proceedings of the 32nd International Conference on Machine Learning*, Lille, 2015. 1889–1897
- 20 Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. 2017. ArXiv:1707.06347
- 21 Polydoros A S, Nalpantidis L. Survey of model-based reinforcement learning: applications on robotics. *J Intell Robot Syst*, 2017, 86: 153–173
- 22 Anthony T, Tian Z, Barber D. Thinking fast and slow with deep learning and tree search. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Long Beach, 2017. 5360–5370
- 23 Racanière S, Weber T, Reichert D, et al. Imagination-augmented agents for deep reinforcement learning. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Long Beach, 2017. 5690–5701
- 24 Feinberg V, Wan A, Stoica I, et al. Model-based value estimation for efficient model-free reinforcement learning. 2018. ArXiv:1803.00101
- 25 Khodayari M H, Balochian S. Modeling and control of autonomous underwater vehicle (AUV) in heading and depth attitude via self-adaptive fuzzy PID controller. *J Mar Sci Technol*, 2015, 20: 559–578
- 26 Narasimhan M, Singh S N. Adaptive optimal control of an autonomous underwater vehicle in the dive plane using dorsal fins. *Ocean Eng*, 2006, 33: 404–416
- 27 Lapiere L, Soetanto D. Nonlinear path-following control of an AUV. *Ocean Eng*, 2007, 34: 1734–1744

- 28 Elmokadem T, Zribi M, Youcef-Toumi K. Trajectory tracking sliding mode control of underactuated AUVs. *Nonlinear Dyn*, 2016, 84: 1079–1091
- 29 Antonelli G. *Underwater Robots*. 4th ed. Berlin: Springer, 2018
- 30 Kober J, Bagnell J A, Peters J. Reinforcement learning in robotics: a survey. *Int J Robot Res*, 2013, 32: 1238–1274
- 31 El-Fakdi A, Carreras M. Two-step gradient-based reinforcement learning for underwater robotics behavior learning. *Robot Auton Syst*, 2013, 61: 271–282
- 32 Ferreira F, Machado D, Ferri G, et al. Underwater optical and acoustic imaging: a time for fusion? A brief overview of the state-of-the-art. In: *Proceedings of Oceans 2016 MTS/IEEE*, Monterey, 2016
- 33 Lu H M, Li Y J, Zhang L F, et al. Contrast enhancement for images in turbid water. *J Opt Soc Am A*, 2015, 32: 886–893
- 34 Heidemann J, Stojanovic M, Zorzi M. Underwater sensor networks: applications, advances and challenges. *Proc R Soc A*, 2012, 370: 158–175
- 35 Yoo B, Kim J. Path optimization for marine vehicles in ocean currents using reinforcement learning. *J Mar Sci Technol*, 2016, 21: 334–343
- 36 Wang C F, Wei L, Wang Z H, et al. Reinforcement learning-based multi-AUV adaptive trajectory planning for under-ice field estimation. *Sensors*, 2018, 18: 3859
- 37 Hu H K, Song S J, Chen C L P. Plume tracing via model-free reinforcement learning method. *IEEE Trans Neural Netw Learn Syst*, 2019, 30: 2515–2527
- 38 Wu H, Song S J, You K Y, et al. Depth control of model-free AUVs via reinforcement learning. *IEEE Trans Syst Man Cybern Syst*, 2019, 49: 2499–2510
- 39 Carlucho I, de Paula M, Wang S, et al. Adaptive low-level control of autonomous underwater vehicles using deep reinforcement learning. *Robot Auton Syst*, 2018, 107: 71–86
- 40 Carlucho I, de Paula M, Wang S, et al. AUV position tracking control using end-to-end deep reinforcement learning. In: *Proceedings of Oceans 2018 MTS/IEEE*, Charleston, 2018
- 41 Ahmadzadeh S R, Kormushev P, Caldwell D G. Multi-objective reinforcement learning for AUV thruster failure recovery. In: *Proceedings of IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, Orlando, 2014
- 42 Stokey R P, Roup A, von Alt C, et al. Development of the remus 600 autonomous underwater vehicle. In: *Proceedings of Oceans 2005 MTS/IEEE*, Washington, 2005. 1301–1304
- 43 Fernandez-Gauna B, Osa J L, Graña M. Effect of initial conditioning of reinforcement learning agents on feedback control tasks over continuous state and action spaces. In: *Proceedings of International Joint Conference SOCO'14-CISIS'14-ICEUTE'14*, Bilbao, 2014. 125–133
- 44 Walters P, Kamalapurkar R, Voight F, et al. Online approximate optimal station keeping of a marine craft in the presence of an irrotational current. *IEEE Trans Robot*, 2018, 34: 486–496
- 45 Knudsen K B, Nielsen M C, Schjølberg I. Deep learning for station keeping of AUVs. In: *Proceedings of Oceans 2019 MTS/IEEE*, Seattle, 2019
- 46 Frost G, Lane D M. Evaluation of Q-learning for search and inspect missions using underwater vehicles. In: *Proceedings of Oceans 2014 MTS/IEEE*, St. John's, 2014
- 47 Jamali N, Kormushev P, Ahmadzadeh S R, et al. Covariance analysis as a measure of policy robustness. In: *Proceedings of Oceans 2014 MTS/IEEE*, Taipei, 2014
- 48 Leonetti M, Ahmadzadeh S R, Kormushev P. On-line learning to recover from thruster failures on autonomous underwater vehicles. In: *Proceedings of Oceans 2013 MTS/IEEE*, San Diego, 2013
- 49 Ahmadzadeh S R, Leonetti M, Carrera A, et al. Online discovery of AUV control policies to overcome thruster failures. In: *Proceedings of IEEE International Conference on Robotics and Automation*, HongKong, 2014. 6522–6528
- 50 Palomeras N, El-Fakdi A, Carreras M, et al. COLA2: a control architecture for AUVs. *IEEE J Ocean Eng*, 2012, 37: 695–716
- 51 Sun T T, He B, Nian R, et al. Target following for an autonomous underwater vehicle using regularized ELM-based reinforcement learning. In: *Proceedings of Oceans 2015 MTS/IEEE*, Washington, 2015
- 52 Shi W J, Song S J, Wu C. High-level tracking of autonomous underwater vehicles based on pseudo averaged Q-learning. In: *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, Miyazaki, 2018. 4138–4143

- 53 Shi W J, Song S J, Wu C, et al. Multi pseudo Q-learning-based deterministic policy gradient for tracking control of autonomous underwater vehicles. *IEEE Trans Neural Netw Learn Syst*, 2019, 30: 3534–3546
- 54 Zhang Q L, Lin J Y, Sha Q X, et al. Deep interactive reinforcement learning for path following of autonomous underwater vehicle. *IEEE Access*, 2020, 8: 24258–24268
- 55 Yu R S, Shi Z Y, Huang C X, et al. Deep reinforcement learning based optimal trajectory tracking control of autonomous underwater vehicle. In: *Proceedings of the 36th Chinese Control Conference*, Dalian, 2017. 4958–4965
- 56 Cui R X, Yang C G, Li Y, et al. Neural network based reinforcement learning control of autonomous underwater vehicles with control input saturation. In: *Proceedings of UKACC International Conference on Control*, Loughborough, 2014. 50–55
- 57 Cui R X, Yang C G, Li Y, et al. Adaptive neural network control of AUVs with control input nonlinearities using reinforcement learning. *IEEE Trans Syst Man Cybern Syst*, 2017, 47: 1019–1029
- 58 Guo X X, Yan W S, Cui R X. Integral reinforcement learning-based adaptive NN control for continuous-time nonlinear MIMO systems with unknown control directions. *IEEE Trans Syst Man Cybern Syst*, 2019. doi: 10.1109/TSMC.2019.2897221
- 59 Guo X X, Yan W S, Cui R X. Event-triggered reinforcement learning-based adaptive tracking control for completely unknown continuous-time nonlinear systems. *IEEE Trans Cybern*, 2020, 50: 3231–3242
- 60 Lin L X, Xie H B, Shen L C. Application of reinforcement learning to autonomous heading control for bionic underwater robots. In: *Proceedings of IEEE International Conference on Robotics and Biomimetics*, Bangkok, 2009. 2486–2490
- 61 Lin L X, Xie H B, Zhang D B, et al. Supervised neural Q-learning based motion control for bionic underwater robots. *J Bionic Eng*, 2010, 7: 177–184
- 62 Wang J, Kim J. Optimization of fish-like locomotion using hierarchical reinforcement learning. In: *Proceedings of the 12th International Conference on Ubiquitous Robots and Ambient Intelligence*, Goyang, 2015. 465–469
- 63 Yan S Z, Wang J, Wu Z X, et al. Motion optimization for a robotic fish based on adversarial structured control. In: *Proceedings of IEEE International Conference on Robotics and Biomimetics*, Dali, 2019. 346–351
- 64 Prahacs C, Saudnars A, Smith M K, et al. Towards legged amphibious mobile robotics. In: *Proceedings of Canadian Design Engineering Network Conference*, Montreal, 2004
- 65 Meger D, Higuera J C G, Xu A, et al. Learning legged swimming gaits from experience. In: *Proceedings of IEEE International Conference on Robotics and Automation*, Seattle, 2015. 2332–2338
- 66 Higuera J C G, Meger D, Dudek G. Synthesizing neural network controllers with probabilistic model-based reinforcement learning. In: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, 2018. 2538–2544
- 67 Zhang X L, Li B, Chang J, et al. Gliding control of underwater gliding snake-like robot based on reinforcement learning. In: *Proceedings of the 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems*, Tianjin, 2018. 323–328
- 68 Zang W C, Nie Y L, Song D L, et al. Research on constraining strategies of underwater glider's trajectory under the influence of ocean currents based on DQN algorithm. In: *Proceedings of Oceans 2019 MTS/IEEE*, Seattle, 2019
- 69 Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay. 2015. ArXiv:1511.05952
- 70 Wu H, Song S J, Hsu Y C, et al. End-to-end sensorimotor control problems of AUVs with deep reinforcement learning. In: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Macau, 2019. 5869–5874

A selected review of reinforcement learning-based control for autonomous underwater vehicles

Yachu HSU, Hui WU, Keyou YOU* & Shiji SONG

Department of Automation and Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

* Corresponding author. E-mail: youky@tsinghua.edu.cn

Abstract Recently, reinforcement learning (RL) has been studied extensively and achieved promising results in a wide range of control tasks. In addition, autonomous underwater vehicles (AUVs) are important tools for executing complex and challenging tasks in marine environments. The advances in RL offers ample opportunities to develop intelligent AUVs. This paper provides a selected review of RL-based control for AUVs with a focus on applications of RL in low-level control tasks for underwater regulation and tracking. We first present a concise introduction to the RL-based control framework. We then provide an overview of RL methods for AUV control problems, where primary challenges and recent progress are discussed. Finally, two representative RL-based controller cases are presented in detail for model-free RL methods for AUVs.

Keywords reinforcement learning, autonomous underwater vehicle, low-level control, autonomous control, model-free

Yachu HSU received her B.S. degree from the Department of Mechanical Engineering, Tsinghua University, Beijing, China in 2017. She is currently pursuing a master's degree at the Department of Automation, Institute of System Integration, Tsinghua University, Beijing, China. Her current research interests include reinforcement learning and robot control.

Hui WU received his B.S. degree from the Department of Automation, Tsinghua University, Beijing, China in 2014. He is currently pursuing a Ph.D. degree at the Department of Automation, Institute of System Integration, Tsinghua University, Beijing, China. His current research interests include reinforcement learning and robot control, specifically continuous control for underwater vehicles.

Keyou YOU received his B.S. degree in statistical science from Sun Yat-sen University, Guangzhou, China in 2007 and his Ph.D. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore in 2012. After briefly working as a research fellow at NTU, he joined Tsinghua University in Beijing, China, where he is currently an associate professor in the Department of Automation. He has held visiting positions at

Politecnico di Torino, the Hong Kong University of Science and Technology, and the University of Melbourne. His current research interests include networked control systems, distributed algorithms, and their applications.

Shiji SONG received his Ph.D. degree from the Department of Mathematics, Harbin Institute of Technology, Harbin, China in 1996. He is currently a professor at the Department of Automation, Tsinghua University, Beijing, China. His current research interests include system modeling, optimization and control, computational intelligence, and pattern recognition.