中国科学: 信息科学 2023 年 第 53 卷 第 6 期: 1027–1049
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
评述

数据驱动的策略优化控制律设计最新研究综述

赵斐然, 游科友*

清华大学自动化系, 北京 100084
* 通信作者. E-mail: youky@tsinghua.edu.cn

收稿日期: 2022–12–16; 修回日期: 2023–01–13; 接受日期: 2023–01–31; 网络出版日期: 2023–06–13

国家自然科学基金 (批准号: 62033006) 和清华大学自主科研计划资助项目

摘要 随着通信技术和新一代人工智能的迅速发展, 强化学习这一数据驱动的控制方法引起了极大的关注. 本文回顾了强化学习中的一类典型方法 —— 策略优化法 —— 在系统控制律设计上的最新研究进展. 主要讨论了其在各种重要线性最优控制问题上的收敛性及样本复杂度, 例如线性二次控制、输出反馈控制、$\mathcal{H}_\infty$ 控制、分布式控制等. 此外, 对策略优化法在网络化系统控制中的应用作了展望.

关键词 线性系统, 最优控制, 策略梯度法, 强化学习, 数据驱动控制

1 引言

随着人类科技的进步与发展, 现代自动控制系统的物理结构、运行环境以及传感模式变得越来越复杂. 以系统机理模型为基础的传统控制理论难以应对这种变化: 一方面, 以物理学定律为基础的建模方式难以准确地建立控制系统的机理模型; 另一方面, 即使能够建立, 复杂的数学模型也可能给系统的分析与设计带来困难. 而随着计算、存储以及通信技术的快速发展, 现代控制系统能够更容易地获取及实时处理大量运行数据. 同时, 以机器学习为核心的新一代人工智能的兴起带来了统计和计算方法层面的突破, 大大提高了数据的挖掘和利用效率. 近年来直接使用数据的端到端控制设计方法引起了学者的极大兴趣, 并在应用中展现出了优势 [1~5]. 由我国著名学者侯忠生教授发起的数据驱动控制与学习系统会议 (DDCLS) 已成功举办了 10 余次, 吸引了大量国内外知名学者积极参与. 在美国国家科学基金 (NSF)、空军 AFOSR、陆军 ARO、海军 ONR 等 4 家权威机构资助下, 来自自动控制、人工智能、运筹优化等多学科领域的 19 位国际权威学者于 2019 年发起了 Learning for Dynamics and Control (L4DC) 会议, 分别得到了 MIT, Berkeley, ETH, Stanford 等世界顶尖大学的积极主办. 此外, 数据驱动控制在机器学习三大主流学术会议 (ICML/ICLR/NeurIPS) 上的关注度连年攀升. 这引发了人们对控制理论根本立足点的思考: 能否从以 “模型” 为基础的控制, 转变成以 “数据” 为基础的控制?

引用格式: 赵斐然, 游科友. 数据驱动的策略优化控制律设计最新研究综述. 中国科学: 信息科学, 2023, 53: 1027–1049, doi: 10.1360/SSI-2022-0455
Zhao F R, You K Y. Survey of recent progress in data-driven policy optimization for controller design (in Chinese). Sci Sin Inform, 2023, 53: 1027–1049, doi: 10.1360/SSI-2022-0455

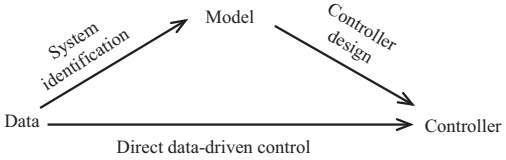

图 1　策略优化法能够利用数据直接学习控制律, 而不需要系统辨识过程

Figure 1　The policy gradient method aims to directly learn a controller from data without identification of an internal model

自卡尔曼 (Kalman) [6] 在 1960 年前后将状态空间表示法引入到控制理论以来, 现代控制理论牢牢植根于系统的显式动力学模型, 如卡尔曼滤波器、线性二次调节器 (linear quadratic regulator, LQR) 等. 因此, 动力学建模及其模型参数辨识几乎成为了现代控制理论的默认前提. 对于复杂的动力学系统, 用物理学定律进行显式动力学建模需要该领域的专家知识, 且通常费时费力. 而系统辨识 [7] 通常不需要人为干预: 通过权衡近似精度与模型复杂度, 其能够将复杂的物理过程用简化的模型近似表示. 因此, 系统辨识更适用于复杂系统的建模. 不过, 由于我们的最终目标是设计一个 "最优" 的控制律, 而系统辨识模型参数存在估计误差, 便引发了一个重要问题: "从控制律的最优性上来说, 什么样的辨识模型是最好的?" 本质上, 传统控制理论将控制律的设计问题 (1) 建模成如下双层优化 (bi-level optimization) 问题:

$$\begin{aligned} &\text{maximize}\quad \text{闭环性能指标}(\omega)\\ &\text{subject to}\ \ \omega \text{ 是模型 } \mathcal{M} \text{ 的一条轨迹},\\ &\text{其中}\qquad \mathcal{M} \text{ 从给定数据中辨识}. \end{aligned} \tag{1}$$

其包含两个耦合的子问题: 系统辨识和基于模型的控制律设计. 系统辨识的目的是找到一组模型参数, 使得数据的拟合误差最小化. 基于模型的控制律设计的目的是在给定辨识模型的情况下, 求解最大化闭环性能指标的控制器. 因为两个子问题之间的分离原理通常不成立, 所以拟合精度高的辨识模型未必能保证控制律是最优的. 从这一角度来看, 传统基于模型的控制方法求得的控制律可能是次优的.

近年来, 强化学习 (reinforcement learning, RL) 这一数据驱动的控制方法在动力学系统的控制上取得了诸多成功, 从机器人控制, 如控制机器人完成复杂的操作 [8~10], 到游戏领域的序列决策问题, 如 AlphaGo [11]、Atari 游戏 [12] 等. 以上许多成就都有赖于数据驱动的强化学习算法, 如策略优化法. 与传统控制中的先 "辨识" 再 "控制" 的两段式思路不同, 策略优化法通过估计梯度, 直接在策略空间中搜索能够最优化闭环系统性能指标的策略 (控制律), 而不需要系统辨识过程. 常用的性能指标包括系统的 $\mathcal{H}_2$ 及 $\mathcal{H}_\infty$ 范数等. 因此, 策略优化法属于以 "数据" 为基础的控制, 其与传统控制方法的概念区别如图 1 所示. 这样一种 "端到端" 的方法不仅在概念上简单, 而且在实际中易于实施, 即使在系统数学模型未知时也是如此. 从控制律的最优性角度来说, 由于避免了辨识过程, 策略优化法的效果可能优于以 "模型" 为基础的控制. 有关强化学习与基于模型的控制方法的讨论和对比, 读者可参考文献 [13~15].

尽管策略优化法在实际决策问题上取得了巨大的成功, 但是从统计和计算的角度看, 人们对这类算法的性能与效率在理论层面的理解还不成熟. 即使对于最简单的线性二次控制, 由于策略优化法涉及非凸优化问题的求解, 也难以分析其收敛性以及所需要的样本数目. 直到 2018 年, Fazel 等 [16] 才首次将其应用在线性系统上, 利用其线性结构探索其在 LQR 问题中的收敛性. 尽管 LQR 问题的损失函数是策略的非凸函数, 但其满足梯度支配 (gradient dominance) 这一重要性质 (优化理论中又被称为 Polyak-Lojasiewicz 条件), 进而策略优化法能够收敛到全局最优解. 这一发现引发了许多学者的研究

兴趣, 并对其他的经典线性二次控制问题展开探索 [17~22], 主要集中于诸如输出反馈控制、$\mathcal{H}_\infty$ 控制、分布式控制等. 此外, 也有文献研究策略优化法求解系统镇定问题以及约束 LQR 问题等. 本综述对上述问题逐一展开讨论.

2 综述概要

本文将主要讨论策略优化法在控制律设计上的理论问题.

第 3 节将讨论 LQR 问题以及策略优化法的样本复杂度. 然后, 为去除初始控制律可镇定这一假设, 提出折扣策略优化法解决系统镇定问题.

第 4 节将讨论输出反馈下的线性二次控制问题, 包括线性二次高斯 (linear quadratic Gaussian, LQG) 问题和静态输出反馈控制. 由于状态不可测量, 该设定下的参数化方法也有所不同, 因此策略优化法的分析有本质不同.

第 5 节将讨论鲁棒控制问题, 如线性二次博弈 (linear quadratic game), 线性指数高斯控制 (linear exponential quadratic Gaussian, LEQG) 等. 在这类问题下, 策略的可行域通常不是可镇定策略的集合, 而需同时考虑鲁棒性约束, 如 $\mathcal{H}_\infty$ 范数约束等. 因此, 优化图景也与 LQR 问题存在本质区别.

第 6 节将讨论当控制律带有结构约束时的控制问题, 包括分布式、子空间约束以及风险约束控制, 其目的是探究控制律结构约束对策略优化法收敛性及算法复杂度的影响.

第 7 节总结了全文, 并讨论了几个值得思考的课题.

3 状态反馈的线性二次控制

3.1 LQR 问题

考虑离散时间线性时不变系统

$$x_{t+1} = Ax_t + Bu_t, \quad x_0 \sim \mathcal{D}, \tag{2}$$

其中, $x_t \in \mathbb{R}^n$ 是系统的状态, $u_k \in \mathbb{R}^m$ 是控制输入. A 和 B 是系统矩阵, 且 (A, B) 是可镇定的. x_0 是初始状态, 服从随机分布 $\mathcal{D}$.

策略优化法用一个反馈增益矩阵 $K \in \mathbb{R}^{n\times m}$ 参数化控制策略, 并在增益矩阵空间中搜索下述 LQR 问题的最优解:

$$\min\ J(K) := \mathbb{E}\sum_{t=0}^{\infty}(x_t^{\mathrm{T}}Qx_t + u_t^{\mathrm{T}}Ru_t) \quad \text{s.t. 式 (2)},\ \ u_t = -Kx_t,\ \text{and } K \in \mathcal{S}, \tag{3}$$

其中, $J(K)$ 表示损失函数, Q 和 R 是惩罚矩阵, $\mathcal{S} = \{K \in \mathbb{R}^{n\times m} | \rho(A - BK) < 1\}$ 表示可镇定策略集合, $\rho(\cdot)$ 表示方阵的谱半径. 在初始策略 K^0 可以镇定系统 (2) 的假设下, 策略优化法按以下方式更新 K:

$$K^+ = K - \eta\nabla J(K), \tag{4}$$

其中, $\nabla J(K)$ 为 $J(K)$ 关于 K 的梯度, η 表示更新步长. 当 (A, B) 未知时, 式 (4) 可以用零阶优化的方式实施, 其中梯度可以根据系统的损失函数数据估计.

关于策略优化法的一个最基本问题是: 当 $\nabla J(K)$ 准确时, 式 (4) 能否收敛到全局最优解? 为此, Fazel 等 [16] 首先分析了 LQR 问题 (3) 的优化图景. 当 $n \geqslant 3$ 时, 式 (8) 一般是非凸的: 目标函数 $J(K)$

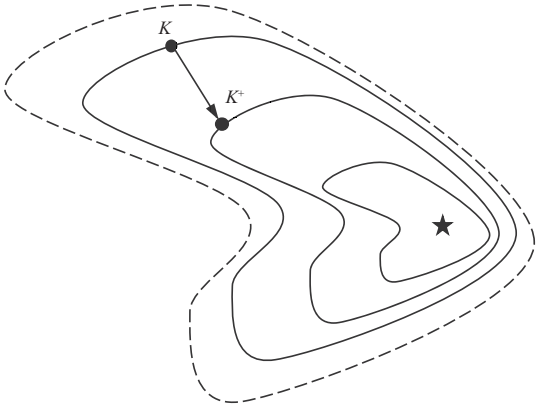

图 2　LQR 优化问题 (3) 的优化图景

Figure 2　The optimization landscape of the LQR problem (3)

和可行域 $\mathcal{S}$ 均为非凸. 图 2 展示了 $J(K)$ 的等值线, 其中虚线为 $\mathcal{S}$ 的边界. 第一个性质是 $J(K)$ 是强制 (coercive) 的, 即当 K 从 $\mathcal{S}$ 的内部趋近于 $\mathcal{S}$ 的边界时, $J(K)$ 会趋于无穷. 这意味着在合适的步长下, 梯度算法会自动地远离 $\mathcal{S}$ 的边界. 第二个重要性质是 $J(K)$ 的驻点唯一, 且海森 (Hessian) 矩阵是正定的. 这两个性质说明了式 (8) 的唯一驻点就是全局最优点. 不过, 仅依靠这两条性质无法刻画梯度算法的收敛速率. Fazel 等 [16] 通过分析最优策略 K^* 相对于 K 的 "优势函数", 证明了 LQR 问题 (3) 还满足梯度支配条件:

$$J(K)-J(K^*)\leqslant\lambda\|\nabla J(K)\|_F^2,$$

其中, $\lambda>0$ 被称为梯度支配常数. 不难看出, 梯度支配条件意味着所有驻点均为全局最优点. 事实上, 在优化理论中通常认为该条件是凸性的一种弱化, 几乎能起到和凸性一样的作用 [23]. 因此, 再结合 $J(K)$ 在 $\mathcal{S}$ 内部的光滑性, 可以证明梯度下降能够以线性速度收敛到全局最优点. 文中还证明了另外两种梯度算法的收敛性, 分别是自然梯度法和高斯 – 牛顿法. 其中, 自然梯度法等价于在一个黎曼流形 (Riemann manifold) 上做梯度下降; 而高斯 – 牛顿法是一类伪牛顿更新法, 详见文献 [24]. 特别地, 文献 [24] 进一步指出, 高斯 – 牛顿法的更新过程等价于策略迭代 (policy iteration), 并具有超线性收敛速度. 受制于篇幅, 在此不再赘述.

以上考虑的是 $\nabla J(K)$ 准确, 即 (A,B,Q,R) 完全已知的情况. 当 (A,B,Q,R) 未知时, 我们假定只能获得模型的仿真数据. 在此设定下, Fazel 等 [16] 提出用如下方式获得梯度的估计 $\widehat{\nabla J(K)}$: 单位超球面上 $\mathbb{S}_n$ 均匀随机选取 N 个点 $U_i, i\in\{1,\dots,N\}$ 得到策略 $\widehat{K}_i=K+rU_i$,; 然后, 仿真 N 条轨迹并获得损失函数的估计 $\widehat{J}_i, i\in\{1,\dots,N\}$, 则梯度的估计为

$$\widehat{\nabla J(K)}=\frac{1}{N}\sum_{i=1}^{N}\frac{n}{r}\widehat{J}_iU_i. \tag{5}$$

值得注意的是, 这里的 $\widehat{J}_i$ 为有限时域 τ 内的估计,

$$\widehat{J}_i=\sum_{t=0}^{\tau}(x_t^{\mathrm{T}}Qx_t+(K_ix_t)^{\mathrm{T}}RK_ix_t).$$

此设定下策略的更新方式为

$$K^+=K-\eta\widehat{\nabla J(K)}. \tag{6}$$

Fazel 等 [16] 证明了当 N 足够大时, $\widehat{\nabla J(K)}$ 足够接近真实梯度. 进而, 通过离散时间代数里卡提方程 (algebraic Riccati equation, ARE) 的扰动性分析, 证明了式 (6) 能够收敛到全局最优解.

文献 [25] 讨论了 (Q,R) 是不定矩阵的 LQR 策略优化问题. 这种情况下, 损失函数不是强制的, 而且策略梯度法也无法保证策略序列是稳定的. 尽管如此, 策略梯度法仍然有全局的收敛性保证. 读者可参考文献 [25] 中的详细讨论.

3.2 样本复杂度

在强化学习中, 样本复杂度指的是找到最优策略所需要的采样系统轨迹的条数. 通过研究其与系统参数的关系, 包括系统阶数 n, 控制输入个数 m, 以及系统矩阵的范数 $\|A\|$ 和 $\|B\|$ 等, 能够进一步理解策略优化法的性能. 文献 [16] 指出, 求解离散时间 LQR 问题的样本复杂度为参数 n, m, $\|A\|$, $\|B\|$, $\|Q\|$, $\|R\|$, $J(K^0)$ 的多项式. 同时, 其关于损失函数求解精度 ϵ 的关系为 $\mathcal{O}(1/\epsilon^4)$. 文献 [26] 借鉴零阶优化的方法, 提出单点和两点估计法, 即每次梯度估计只需要一条或两条系统轨迹, 使得复杂度能减少到 $\mathcal{O}(1/\epsilon^2)$ (单点) 或 $\mathcal{O}(1/\epsilon)$ (两点). 定义 U 为单位超球面 $\mathbb{S}_n$ 上一点, 则在策略 K 处的单点和两点梯度估计分别为

$$\begin{aligned}&\widehat{\nabla J(K)} = J(K+rU)\frac{n}{r}U,\\ &\widehat{\nabla J(K)} = (J(K+rU)-J(K-rU))\frac{n}{2r}U.\end{aligned}\tag{7}$$

在任意包含最优点的下水平集 $\mathcal{G}$ 上, 可以通过设计梯度更新的步长和光滑半径 r 来调整梯度估计的范数及其方差

$$\sup_{x\in\mathcal{G}}\|\widehat{\nabla J(K)}\|_2,\quad \sup_{x\in\mathcal{G}}\mathbb{E}[\|\widehat{\nabla J(K)}-\mathbb{E}[\widehat{\nabla J(K)}]\|_2^2]$$

的上界. 对比式 (5), 尽管文献 [26] 可能需要更多次的迭代以弥补更大的梯度估计误差, 但每次梯度估计 (7) 所需样本远小于文献 [16], 故有更低的样本复杂度. 文献 [27,28] 分别考虑离散和连续时间系统, 在 LQR 的非凸优化问题与其标准凸参数化下问题之间建立了联系, 并基于此进行了随机梯度算法的收敛性分析. 文献提出, 对于梯度的估计只要求方向足够接近即可, 而无需要求数值上的准确性. 这进一步降低了样本量, 达到了 $\log(\epsilon)$ 的样本复杂度. 这也是目前关于收敛精度的样本复杂度最低的结论.

3.3 系统镇定问题

策略优化法要求初始的策略 K^0 能够镇定系统, 即 $J(K^0)<\infty$. 然而, 在系统完全未知的情况下, 如何获取 K^0 本身就是一个难题. 另外, 从系统的角度看, 系统稳定是所有最优控制问题的基础, 也是一切数据驱动控制方法的首要前提. 因此, 如何除去策略优化法对初始策略的这一假设成为了最近研究的重要问题.

考虑线性状态反馈 $u=-Kx$, 则线性系统 (2) 的镇定问题可以描述为: 找到一个策略 K 使得 $\rho(A-BK)<1$. 策略优化法求解此问题的难点在于, 局部迭代搜索要求目标函数始终是有限值, 而初始策略不满足该条件. 因此, 从有噪的损失函数观测中学习一个镇定策略不是一个简单的任务, 而且应该和求解 LQR 问题同等重要 [29]. 事实上, 这也是 Fazel 等 [16] 提出的一个公开问题.

文献 [30] 借助于被称为 “折扣法” [21] 的策略设计方法, 首次提出了 “折扣退火” 算法, 利用策略优化法求解稳定控制策略. 折扣法 [21,30~34] 指的是通过求解一系列折扣 LQR 问题获得所期望策略的一类方法, 其中折扣因子 (discount factor) 通常是不断增大的. 折扣法最初被用于多智能体控制系统中以避免策略陷入局部最优 [32], 近来也被用于学习线性及非线性系统的镇定策略 [21,30,34]. 定义折

扣 LQR 问题:

$$\min\ J_\gamma(K) := \mathbb{E}\sum_{t=0}^{\infty}\gamma^t(x_t^{\mathrm{T}}Qx_t + u_t^{\mathrm{T}}Ru_t) \quad \text{s.t. 式 (2), } u_t = -Kx_t, \text{ and } K \in \mathcal{S}_\gamma, \tag{8}$$

其中 $\mathcal{S}_\gamma = \{K|\sqrt{\gamma}\rho(A-BK) < 1\}$. 折扣法的机制是: 通过选取一个足够小的折扣因子 γ^0, 使得零策略的损失函数是有限值, 即 $J_{\gamma^0}(0) < \infty$, 进而满足策略优化法的要求; 然后, 交替进行折扣 LQR 问题 (8) 的求解和折扣因子的更新, 最终获得镇定策略. 折扣法的框架可用如下迭代过程描述:

$$K^{k+1} \in \underset{K\in\mathcal{S}}{\operatorname{argmin}}\ J_{\gamma^k}(K), \tag{9}$$

$$\gamma^{k+1} = (1+\alpha^k)\gamma^k, \tag{10}$$

其中式 (9) 通过策略优化法近似求解, 式 (10) 中的 α^k 为更新速率.

文献 [30] 提出用随机搜索方法在每一步搜索满足如下条件的折扣因子 $\gamma^{k+1} \in [\gamma^k, 1]$:

$$2.5J_{\gamma^k}(K^{k+1}) \leqslant J_{\gamma^{k+1}}(K^{k+1}) \leqslant 8J_{\gamma^k}(K^{k+1}), \tag{11}$$

并证明了如果式 (11) 有解, 那么算法会在有限步内求解出一个可镇定策略. 进一步地, 将上述结论推广至一类光滑的非线性系统. 不失一般性, 假定非线性系统的平衡点在原点, 并且非线性动力学 $x_{t+1} = f(x_t, u_t)$ 满足局部李普希兹光滑性 (Lipschitz smoothness), 即对于充分小的 $\|x\|$ 和 $\|u\|$, 有

$$\|\nabla_{x,u}f(x,u) - \nabla_{x,u}f(0,0)\| \leqslant \beta(\|x\| + \|u\|).$$

依据这一条件, 就可以将 $f(x,u)$ 在原点泰勒 (Taylor) 展开, 近似为线性系统和余项, 再利用之前的结论进行分析.

然而, 上述折扣因子的更新方式 (11) 存在明显缺陷. 首先, 满足式 (11) 的 γ^{k+1} 存在性无法保证. 这是因为当 γ^k 接近于 1 时, 再增加折扣因子可能无法使损失函数增大 2.5 倍. 其次, 这种更新方式依赖于随机搜索方法, 增加了采样数量. 而且, 该搜索算法的样本复杂度是不可计算的, 这使得求解可镇定策略的样本复杂度无法找到上界.

事实上, 折扣因子的选取与控制策略的稳定裕度有关. 定义如下衰减系统 (damped system):

$$x_{t+1} = \sqrt{\gamma}(A-BK)x_t, \quad x_0 \sim \mathcal{D}. \tag{12}$$

根据式 (8), 容易看出 $J_\gamma(K)$ 实际上等于

$$J_\gamma(K) = \mathbb{E}\sum_{t=0}^{\infty}(x_t^{\mathrm{T}}Qx_t + u_t^{\mathrm{T}}Ru_t) \quad \text{s.t. 式 (12), } u_t = -Kx_t, \text{ and } K \in \mathcal{S}_\gamma. \tag{13}$$

因此, $J_\gamma(K) < 0$ 等价于衰减系统 (12) 稳定. 于是, 折扣因子更新规则设计问题可以归结为如下问题: 在给定一个策略 K 满足 $\sqrt{\gamma}\rho(A-BK) < 1$ 后, 我们能否找到一个更大的折扣因子 $\gamma' > \gamma$, 使得 $\sqrt{\gamma'}\rho(A-BK) < 1$? 如果能的话, 我们就可以在新折扣因子 γ' 下以 K 为初始策略进行策略优化法的更新. 这个问题本质上是关于任意策略 K 的稳定裕度的问题. 文献 [21] 借助李雅普诺夫 (Lyapunov) 方法率先回答了这一问题: 当新折扣因子满足 $\gamma' \leqslant (1+\alpha)\gamma$ 时, 其中

$$\alpha = \frac{\underline{\sigma}(Q + K^{\mathrm{T}}RK)}{J_\gamma(K) - \underline{\sigma}(Q + K^{\mathrm{T}}RK)}, \tag{14}$$

其对应的衰减系统稳定. 这一结论揭示了策略 K 的稳定裕度及它所对应的 LQR 损失函数 $J_\gamma(K)$ 之间的关系. 更具体地说, 降低 $J_\gamma(K)$ 能够提高策略 K 稳定裕度的上界. 在这种意义上, 该结论是对著名的 LQR 稳定裕度 [35] 的泛化, 而后者只关注最优策略的鲁棒性. 更重要的是, 条件 (14) 可以用数据直接计算: 惩罚矩阵 Q 和 R 是自定义的, $J_\gamma(K)$ 可以由系统采样轨迹估计. 基于这一结论, 文献 [21] 改进了文献 [30] 的方法, 提出 "精确折扣退火" 算法, 其主要特点在于折扣因子的更新不需要搜索过程, 进而为算法样本复杂度的分析带来了便利. 文献证明了式 (10) 中的更新速率 α^k 存在下界, 进而可以证明算法在有限步即可收敛. 利用策略优化法求解 LQR 问题的样本复杂度 [27] 结论, 首次指出策略优化法用于系统镇定的样本复杂度正比于 $\rho(A)$ 乘以求解 LQR 问题的复杂度.

折扣法仍有许多值得继续探究的问题. 例如, 文献 [21] 中关于稳定裕度的结论能否推广至非线性系统? 如何求解非线性系统下的样本复杂度? 折扣法的基础在于式 (13) 成立, 而其依赖于二次型的损失函数. 当损失函数非二次型时, 如何改进折扣法以解决系统镇定问题? 另外, 强化学习中的另一经典算法 Q-learning 也可以应用到折扣法以最小化损失函数, 探究其样本复杂度是一个有趣的研究方向.

4 输出反馈的线性二次控制

与 LQR 问题最接近的是输出反馈的线性二次控制. 这种情况下状态信号不可直接测量, 而只能获取输出信号 (状态的线性函数). 考虑如下的连续线性时间时不变系统:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + w(t), \\ y(t) &= Cx(t) + v(t), \end{aligned} \tag{15}$$

其中, $x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m, y(t) \in \mathbb{R}^p$, C 为观测矩阵, $w(t) \in \mathbb{R}^n$ 和 $v(t) \in \mathbb{R}^p$ 分别是系统在 t 时刻过程和测量噪声. 控制目标为求得最优的控制输入以优化 LQG 的损失函数:

$$J = \lim_{T\to\infty} \frac{1}{T} \mathbb{E} \int_{t=0}^{T} \left(x^{\mathrm{T}} Q x + u^{\mathrm{T}} R u \right) \mathrm{d}t. \tag{16}$$

输出反馈控制的策略参数化方式通常有两种, 静态输出反馈和基于卡尔曼滤波的动态控制律, 这两种情况下策略梯度法都只有局部的收敛性保证. 最近, 文献 [20] 提出了输入输出反馈, 在新的参数化形式下, 能够得到全局收敛结果.

4.1 动态输出反馈

对于 $w(t)$ 和 $v(t)$ 为高斯白噪声的情况, 根据著名的分离原理 [36], 系统 (15) 的最优策略是动态的, 且具有如下形式:

$$\begin{aligned} \dot{\xi}(t) &= (A - BK)\xi(t) + L(y(t) - C\xi(t)), \\ u(t) &= -K\xi(t), \end{aligned} \tag{17}$$

其中, $\xi(t) \in \mathbb{R}^q$ 为策略的中间状态, LQR 增益 K 及卡尔曼增益 L 分别满足两个代数里卡提方程 [36]. 文献 [37] 首先研究了动态控制律的优化图景. 定义 $A_K = A - BK - LC$, $B_K = L$, $C_K = -K$, 则 LQG 的控制律可以写成

$$\begin{aligned} \dot{\xi} &= A_K \xi + B_K y, \\ u &= C_K \xi. \end{aligned} \tag{18}$$

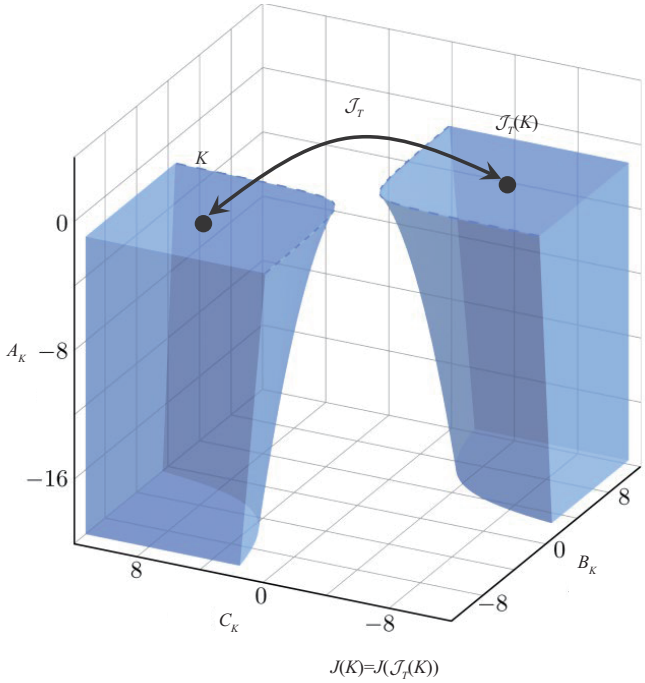

图 3 (网络版彩图) LQG 优化问题 (3) 的可行域至多有两个非连通部分 [37, 例2]

Figure 3 (Color online) There are at most two path-connected components for the stabilizing controller of LQG problem (3) [37, Example 2]

因此, 可以用 (A_K, B_K, C_K) 参数化策略. 定义 q 阶的镇定策略集合为

$$\mathcal{C}_q = \left\{ K = \begin{bmatrix} 0 & C_K \\ B_K & A_K \end{bmatrix} \middle| \begin{bmatrix} A & BC_K \\ B_K C & A_K \end{bmatrix} \text{稳定} \right\}.$$

若考虑 n 阶的动态控制律, 则 LQG 问题可以建模为如下的约束优化问题:

$$\min\ J(K) \quad \text{s.t. } K \in \mathcal{C}_n. \tag{19}$$

为了研究策略优化法在 LQG 问题上的性能, 文献 [37] 主要分析了集合 $\mathcal{C}_n$ 的拓扑特征以及损失函数 $J(K)$ 的分析性质. 一个比较直接的结论是 $\mathcal{C}_n$ 是非凸的, 而损失函数 $J(K)$ 是实解析函数. 对于输出反馈控制来说, 镇定策略集合通常是不连通的. 例如, 上文提到的静态输出反馈的情况, 可能存在指数级数量的不连通区域. 但对于动态控制律 (18), 尽管 $\mathcal{C}_n$ 可能是不连通的, 但至多只有两个不连通区域. 定义动力学控制律 K 的相似性变换 $\mathscr{T}_T(K)$ 为

$$\mathscr{T}_T(K) = \begin{bmatrix} I_m & 0 \\ 0 & T \end{bmatrix} K \begin{bmatrix} I_d & 0 \\ 0 & T \end{bmatrix}^{-1},$$

其中非奇异矩阵 T 为变换矩阵. 可以证明, 两区域之间总存在一个相似性变换定义的双射, 并且其有相等的损失函数, 如图 3 所示.

这个结论意味着可以在任一区域内使用梯度算法进行局部搜索, 而不影响结果的最优性. 文献 [37] 同时还研究了 LQG 问题驻点的性质. 最小策略 (minimal controllers) 指的是能控能观策略, 其在刻画

驻点时起着重要作用. 通过研究相似性变换带来的对称性, 可以证明 LQG 问题 (19) 可能存在很多严格次优的驻点, 并且都是非最小策略. 相对应地, 所有的最小驻点都是 LQG 问题的全局最优解, 并且构成了维数为 n^2 的子流形, 其有两个连通的部分. 这些最小驻点对应于不同的相似性变换. 这个结论意味着, 如果梯度算法收敛到了驻点, 并且其对应一个能控能观策略, 那么算法就找到了一个全局最优解. 文献 [38] 进一步将 LQG 问题的驻点分为两类, 严格驻点和高阶驻点. 严格驻点的海森矩阵具有至少一个负特征根, 因此标准的扰动策略优化法就可以逃离严格驻点. 然而, LQG 有许多高阶驻点, 其海森矩阵为 0, 标准的扰动梯度法因而可能会陷于高阶驻点. 事实上, 所有的驻点都是由于能控或能观性的缺失 [38] 造成的. 确实, 任何一个 n 阶最小策略都是全局最优点. 进一步地, 文献 [38] 证明了所有驻点在模型降阶为能控能观之后仍然是驻点, 并且是严格的. 因此, 我们可以依据此结论先扰动驻点, 将其变为严格驻点, 然后用标准的扰动梯度算法寻优 [38].

文献 [39] 研究了离散时间的动力学输出反馈 LQR (dynamic output-feedback linear quadratic regulation, dLQR) 问题, 即假设 $w(t) = v(t) = 0$ 并用式 (18) 参数化策略. 与文献 [37] 中的时间平均损失函数不同, 由于没有噪声, 这里考虑的是累加线性二次损失函数

$$J = \mathbb{E}_{x_0\sim\mathcal{D}} \sum_{t=0}^{\infty} (x_t^{\mathrm{T}} Q x_t + u_t^{\mathrm{T}} R u_t),$$

其中 $\mathcal{D}$ 为已知概率分布. 因此, 系统初始状态带来的影响不可忽略, 而且在 LQG 中相似性变换带来的对称性 [37] 在 dLQR 问题中也未必成立. 文献 [39] 首先证明了, 在不同的相似性变换下, dLQR 的损失函数是不同的. 这是因为初始状态分布是固定的, 而相似性变换又对应了控制律状态的坐标变换. 因此, 一个自然的问题是什么样的相似性变换是最优的, 即何时有 T^* 使得

$$J(\mathscr{T}_{T^\star}(K)) \leqslant J(\mathscr{T}_T(K)), \ \ \forall T \text{ 可逆}.$$

文献 [39] 指出, 如果 K 是可观且可镇定的且 T^* 存在, 那么 T^* 是唯一的并具有显式表达式. 进一步地, 如果一个能观的驻点存在, 那么它是 dLQR 问题唯一的驻点并且具有如下特殊形式:

$$K^* = \mathscr{T}_{T^*} \left(\begin{bmatrix} 0 & -K^* \\ L^* & A - BK^* - L^*C \end{bmatrix} \right),$$

其中 K^* 和 L^* 分别为 LQR 增益及卡尔曼增益. 也就是说, 若 dLQR 问题的最优策略是能观的, 则其为 LQG 问题 (19) 的解再进行最优相似性变换. 这个结论可以用于判断策略优化法收敛的点是否为最优解. 最后, 当初始状态的估计满足一个特定的结构约束时, dLQR 问题与 LQG 是等价的, 即它们的解相同. 这一发现建立了输出反馈控制在确定系统与随机系统的关系.

4.2 静态输出反馈

假设 $w(t) = v(t) = 0$. 在控制策略 $u(t) = -Ky(t)$ 下, 闭环系统动力学为 $\dot{x}(t) = (A - BKC)x(t)$. 由于观测矩阵 C 通常是秩亏的, 损失函数 $J(K)$ 与 LQR 问题具有截然不同的性质. 第一, 稳定区域 $\mathcal{S} = \{K|\rho(A - BKC) < 1\}$ 可能是非连接的, 并且每个区域内可能存在多个局部最优策略; 第二, 梯度支配这一关键性质在静态输出反馈中并不成立. 尽管文献 [40] 证明了 $J(K)$ 在 $\mathcal{S}$ 内部是光滑的, 但由于以上不利条件, 策略优化法在合适的步长下只能保证收敛到局部最优点. 对于离散的线性时不变系统, 文献 [41] 也得出了类似的结论.

4.3 输入输出反馈

区别于以上两种控制律的参数化形式, 文献 [20] 提出了输入输出反馈控制, 其假设 $w(t)=v(t)=0$ 并考虑如下参数化方式:

$$u_t=-Kz_{t,p},$$

其中 $z_{t,p}=[u_{t,p}^{\mathrm{T}},y_{t,p}^{\mathrm{T}}]^{\mathrm{T}}$, $u_{t,p}=[u_{t-1}^{\mathrm{T}},\dots,u_{t-p}^{\mathrm{T}}]^{\mathrm{T}}$, $y_{t,p}=[y_{t-1}^{\mathrm{T}},\dots,y_{t-p}^{\mathrm{T}}]^{\mathrm{T}}$ 是历史的输入输出数据, p 是系统能控性指数 c 和能观性指数 o 的最大值, 即 $p=\max\{o,c\}$. 策略梯度法的目的是求得最优的增益矩阵 $K\in\mathbb{R}^{m\times q}$, 其中 $q=p(m+d)$.

通过系统的动力学方程, 我们可以将状态用一段固定长度的历史输入输出轨迹表示出来, 即

$$x_t=(\mathcal{C}-A^p\mathcal{O}^{\dagger}\mathcal{T})u_{t,p}+A^p\mathcal{O}^{\dagger}y_{t,p}:=Sz_{t,p},$$

其中

$$\mathcal{O}_p=\begin{bmatrix}CA^{p-1}\\ \vdots\\ CA\\ C\end{bmatrix},\quad \mathcal{C}_p=[B\ \ AB\ \ \cdots\ \ A^{p-1}B],\quad \mathcal{T}_p=\begin{bmatrix}0 & CB & CAB & \cdots & CA^{p-2}B\\ 0 & 0 & CB & \cdots & CA^{p-3}B\\ \vdots & \vdots & \ddots & \ddots & \vdots\\ 0 & \cdots & & 0 & CB\\ 0 & 0 & 0 & 0 & 0\end{bmatrix},$$

以及 $S=[\mathcal{C}-A^pO^{\dagger}\mathcal{T},A^pO^{\dagger}]$.

进而, 我们可以写出损失函数的表达式. 定义可行域

$$\mathcal{S}=\{K\in\mathbb{R}^{m\times q}|\rho(A-BKS^{\dagger})<1\},$$

则对于 $K\in\mathcal{S}$, 损失函数可以写为

$$J(K)=\mathrm{Tr}(P_K\Sigma_0),$$

其中 Σ_0 为初始状态分布的协方差矩阵, $P_K\geqslant 0$ 是如下李雅普诺夫方程的解:

$$P_K=Q+(S^{\dagger})^{\mathrm{T}}K^{\mathrm{T}}RKS^{\dagger}+(A-BKS^{\dagger})^{\mathrm{T}}P_K(A-BKS^{\dagger}).\tag{20}$$

显然, 输入输出反馈的一个最优策略为 $u_t=-K^*z_{t,p}$, 其中 $K^*=(R+B^{\mathrm{T}}P^*B)^{-1}B^{\mathrm{T}}P^*AS$ 满足如下李雅普诺夫方程:

$$P^*=Q+(S^{\dagger})^{\mathrm{T}}(K^*)^{\mathrm{T}}RK^*S^{\dagger}+(A-BK^*S^{\dagger})^{\mathrm{T}}P^*(A-BK^*S^{\dagger}).$$

然而, 因为 $S^{\dagger}$ 并非行满秩, 最优解并不唯一. 定义矩阵空间

$$\mathcal{F}=\{\Delta\in\mathbb{R}^{m\times q}|\Delta\cdot S^{\dagger}=0\}.$$

我们证明了最优解集合是一个与 $\mathcal{F}$ 平行的矩阵空间 $\mathcal{K}=\{K\in\mathbb{R}^{m\times q}|K=K^*+\Delta,\Delta\in\mathcal{F}\}$. 这似乎是一个负面的结果, 因为已有文献里能够证明策略梯度法具有全局收敛性的问题其最优解都是唯一的. 不过, 式 (20) 与标准 LQR 的李雅普诺夫方程具有相似的结构, 使得我们的问题具有更好的性质. 例如, 解空间 $\mathcal{K}$ 对相似性变换具有不变性, 这意味着我们只需要考虑系统的一个最小实现. 而且, 即使解是一个空间, 基于式 (20), 我们仍然能证明梯度支配的性质, 进而得到全局收敛性.

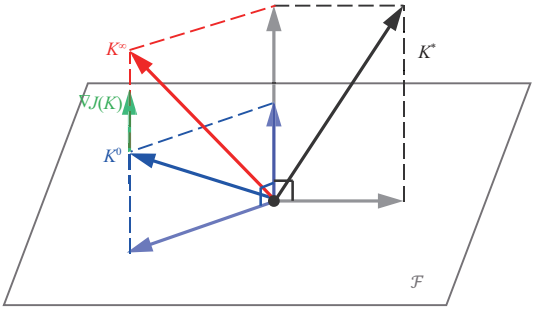

图 4 (网络版彩图) K^0, K^∞, K^* **以及** $\nabla J(K)$ **的示意图**

Figure 4 (Color online) An illustration of K^0, K^∞, K^* and $\nabla J(K)$

此外, 我们观察到一个有趣的性质: 损失函数的梯度 $\nabla J(K)$ 与 $\mathcal{K}$ 是垂直的. 这意味着梯度算法的解 K^∞ 与初始点 K^0 和最优解 K^* 满足如下关系:

$$K^\infty = \Pi_{\mathcal{F}}(K^0) + \Pi_{\mathcal{F}}^{\perp}(K^*).$$

示意图如图 4 所示.

输出反馈控制仍有许多难题亟待解决, 例如对于连续时间的 LQG 问题 [37], 其性质是否在离散时间下也成立? 是否存在一种梯度算法能够达到全局收敛, 收敛速度如何? dLQR 问题里不能观的驻点, 具有怎么样的性质? 能否类比 LQG 问题 [38], 进行模型降解, 转变为严格驻点? 对于输入输出反馈 [20], 在有过程噪声和测量噪声的情况下, 是否也有全局收敛的性质? 得到的控制律与 LQG 相比如何?

5 鲁棒控制

鲁棒控制的目的是设计能够处理系统不确定性的策略 [36]. 在基于机理模型的方法中, 最优控制理论能够提供很多高效方法和数学工具处理有界的模型误差, 例如小增益定理、$\mathcal{H}_\infty$ 控制、μ 综合等. 对于策略优化法, 重要的问题包括: 如何从理论的角度理解鲁棒性 [16]? 如何设计具有鲁棒性的策略? 本节主要讨论线性二次博弈 [42, 43] 以及 $\mathcal{H}_2/\mathcal{H}_\infty$ 混合控制 [44, 45].

5.1 线性二次博弈

文献 [42] 考虑如下离散时间线性系统:

$$x_{t+1} = Ax_t + Bu_t + Cv_t, \quad x_0 \sim \mathcal{D},$$

其中 $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^{m_1}$ 和 $v_t \in \mathbb{R}^{m_2}$ 分别为玩家 1 和 2 的控制输入. 玩家 1 (玩家 2) 的目标是最小化 (最大化) 值函数,

$$\inf_{\{u_t\}} \sup_{\{v_t\}} \mathbb{E} \sum_{t=0}^{\infty} (x_t^{\mathrm{T}} Q x_t + u_t^{\mathrm{T}} R^u u_t - v_t^{\mathrm{T}} R^v v_t), \tag{21}$$

其中 $Q, R^u, R^v > 0$. 问题 (21) 被称为线性二次博弈问题. 特别地, 当 $R^v = \gamma^2 \cdot I$ 时, 为 $\mathcal{H}_\infty$ 次优控制问题. 若式 (21) 有解, 可以看出所求的的控制策略是对噪声 v_t 具有鲁棒性的.

在一些标准假设下 [46, Theorem 3.2], 式 (21) 存在纳什平衡点, 且玩家 1 和 2 的最优策略分别为如下线性状态反馈:

$$u_t^* = -K^* x_t, \quad v_t^* = -L^* x_t,$$

其中增益矩阵 K^* 和 L^* 可以由一个泛化代数里卡提方程 (generalized algebraic Riccati equation, GARE) 解出. 因此, 我们用两个矩阵 K 和 L 参数化策略, 并用策略优化法来搜寻最优解. 根据式 (21) 定义损失函数:

$$J(K,L)=\mathbb{E}\sum_{t=0}^{\infty}(x_t^{\mathrm{T}}Qx_t+(Kx_t)^{\mathrm{T}}R^u(Kx_t)-(Lx_t)^{\mathrm{T}}R^v(Lx_t)).$$

我们的目标是找到纳什平衡点 (K^*,L^*) 求解以下最小最大化问题:

$$\min_K\max_L J(K,L), \tag{22}$$

使得对于任意 K 和 L 有 $J(K^*,L)\leqslant J(K^*,L^*)\leqslant J(K,L^*)$. 为了保证平衡点的存在, 做如下假设:

$$R^v-C^{\mathrm{T}}P^*C>0,\quad Q-(L^*)^{\mathrm{T}}R^vL^*>0, \tag{23}$$

其中 P^* 为 GARE [42, 式(2.2)] 的解.

显然, 问题 (22) 为非凸 – 非凹优化问题. 文献 [42] 提出将式 (22) 分为内外环, 用嵌套梯度算法循环求解. 由于将最小最大换序后, 内环为标准的 LQR 问题, 利用此特点可以在不影响最优解的情况下将式 (22) 的求解转化为最大最小化问题:

$$\max_L\min_K J(K,L).$$

这样, 在外环 L 固定时, 内环可以用文献 [16] 的结果对策略优化法进行分析. 定义 $K(L)$ 为外环策略为 L 时内环的最优策略. 可以证明, 梯度下降算法、自然梯度法以及高斯 – 牛顿算法都可以以线性速率收敛到最优 $K(L)$. 对于外环, 由于要保证假设 (23) 的满足, 用投影梯度法进行更新. 定义集合 $\Omega=\{L|Q-L^{\mathrm{T}}R^vL>0\}$, 则投影梯度法更新 L 为

$$L^+=\mathbb{P}_\Omega[L+\eta\nabla_L J(K(L),L)].$$

这样一来, 由于集合 Ω 包含了纳什平衡点 (K^*,L^*), 因此投影算子不影响最优解. 尽管为非凹优化问题, 可以证明外环的投影梯度算法具有全局的次线性收敛速度, 同时在纳什平衡点附近有局部线性收敛速度. 这与非凸优化的最新收敛性结果 [47, 48] 相契合.

5.2 混合控制

本小节以著名的风险敏感控制 (risk-sensitive control) [49] 问题 (又被称为 LEQG 问题) 为例展开讨论.

文献 [44] 考虑如下离散时间线性系统:

$$x_{t+1}=Ax_t+Bu_t+w_t, \tag{24}$$

其中 $x_t\in\mathbb{R}^n,u_t\in\mathbb{R}^m$, x_0 和 w_t 是独立的零均值高斯随机变量, 协方差矩阵分别为 X_0 和 W. 定义单步损失函数 $c_t=x_t^{\mathrm{T}}Qx_t+u_t^{\mathrm{T}}Ru_t$. LEQG 的目标是最小化长期损失函数:

$$J=\lim_{T\to\infty}\sup\frac{1}{T}\frac{2}{\beta}\mathbb{E}\exp\left[\frac{\beta}{2}\sum_{t=0}^{T-1}c_t\right]. \tag{25}$$

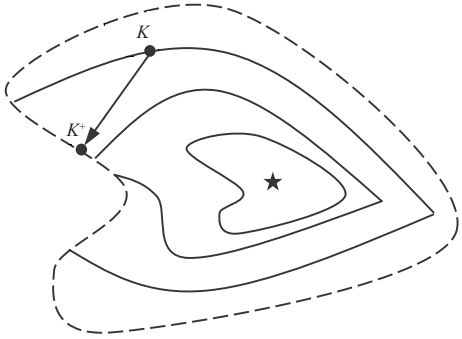

图 5 LEQG 问题 (26) 的优化图景

Figure 5 The optimization landscape of the LEQG problem (26)

该损失函数的含义可以从其在 $\beta = 0$ 处的泰勒展开直观看出, 即

$$J \approx \lim_{T\to\infty} \sup \frac{1}{T}\left\{\mathbb{E}\sum_{t=0}^{T-1} c_t + \frac{\beta}{2}\mathrm{Var}\sum_{t=0}^{T-1} c_t\right\} + O(\beta^2).$$

因此, 若 $\beta > 0$, LEQG 同时惩罚了损失的期望和方差, 达到了风险控制的目的. 对于 LEQG, 最优策略是线性状态反馈的形式 [49], 即 $u_t = -Kx_t$. 因此只需要搜索最优的控制增益. 假设式 (25) 的最优策略存在, 我们的目标是求解如下优化问题:

$$\min J(K) := \lim_{T\to\infty} \frac{1}{T}\frac{2}{\beta}\mathbb{E}\exp\left[\frac{\beta}{2}\sum_{t=0}^{T-1} c_t\right] \quad \text{s.t. 式 (24) and } u_t = -Kx_t. \tag{26}$$

有趣的是, 尽管式 (26) 看似只要求 K 可镇定, 但实际上隐含了一个 $\mathcal{H}_\infty$ 范数的隐式约束 [50]. 定义由噪声 w_t 到 x_t 的传递函数 $\mathcal{T}(K)$ 为

$$\mathcal{T}(K) = \left[\begin{array}{c|c} A - BK & W^{1/2} \\ \hline (Q + K^{\mathrm{T}}RK)^{1/2} & 0 \end{array}\right].$$

文献 [50] 指出, $J(K)$ 的可行域实际上是可镇定集合与 $\mathcal{T}(K)$ 的 $\mathcal{H}_\infty$ 范数约束的交集, 即

$$\mathcal{K} = \{K | \rho(A - BK) < 1,\ \|\mathcal{T}(K)\|_\infty < 1/\sqrt{\beta}\}.$$

事实上, LEQG 问题 (26) 是一个约束非凸优化问题, 其属于一类更宽泛的问题, 即 $\mathcal{H}_2/\mathcal{H}_\infty$ 混合控制.

图 5 展示了 LEQG 问题 (26) 的优化图景, 其中虚线表示 $\mathcal{K}$ 的边界. 与 LQR 问题不同, LEQG 的损失函数不是强制的. 特别地, 当 K 趋近于 $\mathcal{K}$ 的边界时, $J(K)$ 并不一定趋于无穷. 这个特殊性质给 $\mathcal{H}_2/\mathcal{H}_\infty$ 混合控制问题策略优化法的分析带来了挑战. 由于 $J(K)$ 不是强制的, 策略在 $\mathcal{K}$ 的边界附近可能是有限值, 因此固定步长的梯度下降法无法自动保证策略始终在可行域内. 一种解决办法是投影梯度法, 即每次梯度更新后将策略投影到可行域 $\mathcal{K}$ 中. 然而, 集合 $\mathcal{K}$ 是非凸的, 故不太可能进行高效的投影.

为解决此问题, 首先定义隐式正则 (implicit regularization) 的概念: 对于问题 (26), 假定一种迭代算法能够在不投影到 $\mathcal{K}$ 的情况下产生一系列策略 $K_n \in \mathcal{K}, n \in \{0, 1, \dots\}$, 则这个算法被称为隐式正则的. 事实上, 隐式正则的概念已经在许多非凸优化领域被提出, 包括神经网络的训练 [51]、相位复

原 [52,53]、矩阵完备 [54] 等. 对于 $\mathcal{H}_2/\mathcal{H}_\infty$ 混合控制问题, 可以证明自然梯度法和高斯 – 牛顿法实际上是满足隐式正则这一重要性质的. 这意味着上述两种算法不需要经过投影过程, 而只需要设定一个固定步长, 产生的策略就能自动满足约束集 $\mathcal{K}$. 结论的证明分为两部分. 第一步, 直接用 K 策略对应的李雅普诺夫方程的解 P_K [44, (2.13)] 来构造 K^+ 的李雅普诺夫方程, 并保证 $\|\mathcal{T}(K^+)\|_\infty \leqslant 1/\sqrt{\beta}$; 第二步, 受 KYP 引理 [55] 的启发, 进一步扰动 P_K 来证明严格不等式 $\|\mathcal{T}(K^+)\|_\infty < 1/\sqrt{\beta}$. 由于 $J(K)$ 不具有全局的梯度支配性质, 以上两种梯度算法只具有次线性的收敛速度. 不过, 在最优策略附近, 可以证明其具有局部的梯度支配性质. 因此, 自然梯度 (高斯 – 牛顿法) 在最优策略附近有线性 (超线性) 收敛速度.

除以上讨论的线性二次博弈问题以及混合控制问题外, 文献 [56] 讨论了系统矩阵 (A, B) 带有随机乘性噪声下的最优控制. 对于此问题, 全局的梯度支配性质仍然成立, 由此可得到策略优化法的全局收敛性. 由于空间限制, 本文不作深入讨论.

6 结构约束的线性二次控制

6.1 分布式控制

在多智能体控制系统中, 许多智能体通过通信网络连接, 因此控制律通常是分布式的: 由于传感能力、地理距离或隐私上的限制, 每个智能体只能依靠部分传感信息设计控制策略. 目前已知与之相关的优化问题一般是 NP- 难的 [57~59]. 为此, 一般引入凸松弛 [60] 或者某些假设 [61] 来近似求解问题. 这促使学者们研究能否用更直接的策略优化法解决分布式控制问题. 与中心化 (centralized) LQR 问题不同, 分布式控制的最优策略通常不是线性的 [59]. 即使在能够计算显式解时, 最优策略也要求已知某些中间状态, 其形式也可能相当复杂 [62]. 而且, 对于无限时域的静态分布式策略设计问题, 可镇定的策略集合通常是非连通的 [63], 这使得策略优化法难以获得全局收敛的保证. 本小节讨论两种分布式设定下的控制问题, 以揭示网络结构给策略优化法带来的本质影响.

文献 [64] 考虑的是状态反馈的 N 个智能体分布式控制问题, 系统动力学模型为

$$x(t+1) = Ax(t) + Bu(t) + w(t), \tag{27}$$

其中, $x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m$, 噪声 $w(t) \in \mathbb{R}^n$ 服从独立高斯分布 $\mathcal{N}(0, \Sigma_w)$. 假设每个智能体 i 对应一个局部控制输入 $u_i(t) \in \mathbb{R}^{m_i}$, 其构成了全局控制输入 $u(t) = [u_1(t)^{\mathrm{T}}, \dots, u_N(t)^{\mathrm{T}}]^{\mathrm{T}}$. 每个智能体 i 只能观测到部分状态, 其用 $x(t)$ 的子向量 $x_{\mathcal{I}_i}(t) \in \mathbb{R}^{n_i}$ 来表示, 其中 $\mathcal{I}_i$ 是 $\{1, \dots, n\}$ 的子集. 只考虑使用当前观测的局部静态线性策略, 即 $u_i(t) = K_i x_{\mathcal{I}_i}(t)$. 定义 $K = \mathrm{vec}\left((K_i)_{i=1}^N\right) \in \mathbb{R}^{n_K}$, 其中 $n_K = \sum_{i=1}^N n_i m_i$. 不难看出, 全局策略也是静态线性状态反馈, 即 $u(t) = \mathcal{M}(K)x(t)$, 其中 $\mathcal{M}(K)$ 是稀疏增益矩阵. 分布式控制律的示意图请参考文献 [64] 中的图 1.

假设在每个时间点 t, 智能体 i 的局部损失函数为 $c_i(t) = x(t)^{\mathrm{T}} Q_i x(t) + u(t)^{\mathrm{T}} R_i u(t)$, 其中 $Q_i > 0, R_i > 0$, 其依赖于全局的状态和控制信息. 我们的目标是找到一个控制策略, 使得智能体之间的平均损失函数最小化, 即

$$\min_K J(\mathrm{K}) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} c_i(t)\right] \quad \text{s.t.} \ \text{式 (27)}, \ u_i(t) = K_i x_{\mathcal{I}_i}(t), \ i = 1, \dots, N. \tag{28}$$

当模型参数已知时, 式 (28) 可以视作一个去中心化 (decentralized) 的线性二次控制问题. 我们研究智能体 i 如何仅利用局部状态信息 $x_{\mathcal{I}_i}(t)$ 和局部损失函数 $c_i(t)$ 来学习局部策略 K_i.

分布式控制能够通过一个通信网络交换有限的信息. 考虑连通无向通信网络 $\mathcal{G}=(\{1,\dots,N\},\mathcal{E})$, 每个节点代表一个智能体, $\mathcal{E}$ 表示边的集合. 每个时间点 t, 智能体 i 和 j 能够通信当且仅当 $(i,j)\in\mathcal{E}$. 定义与图 $\mathcal{G}$ 相对应的双随机通信矩阵为 $W\in\mathbb{R}^{N\times N}$. 为保证式 (28) 是有解的, 我们需要假设存在控制策略 $K\in\mathbb{R}^{n_K}$ 使得 $x(t+1)=(A+B\mathcal{M}(K))x(t)$ 是渐进稳定的, 并定义可行域 $\mathcal{K}=\{K\in\mathbb{R}^{n_K}|\rho(A+B\mathcal{M}(K))<1\}$.

对于中心化 LQR 问题, 根据式 (7), 策略优化法通过如下方式估计梯度:

$$\widehat{\nabla J(K)}=\frac{n_K}{r}J(K+rU)U,$$

其中, r 是平滑半径, $U\in\mathbb{R}^{n_K}$ 随机从 n_K 维的单位球 $\mathbb{S}_{n_K}$ 中均匀采样. 基于该梯度估计的策略优化法如下:

$$K^+=K-\eta\cdot\frac{n_K}{r}J(K+rU)U. \tag{29}$$

对于去中心化 LQR 问题 (28), 基于类似的梯度估计法, 式 (29) 可以等价地解耦到每个智能体 i 上:

$$K_i^+=K_i-\eta\cdot\frac{n_K}{r}J(K+rU)D_i, \tag{30}$$

其中 $U=\mathrm{vec}((U_i)_{i=1}^N)\sim\mathbb{S}_{n_K}$. 也就是说, 如果每个智能体 i 能够采样 U_i 并得到全局损失函数的值 $J(K+rU)$, 那么式 (29) 就可以以去中心化的方式让每个智能体 i 并行地更新自己的策略 K_i.

基于这个思想, 文献 [64] 提出的分布式零阶策略优化算法由以下三步构成. 第一步, 每个智能体 i 产生一个随机矩阵 $U_i\in\mathbb{R}^{m_i\times n_i}$ 使得叠加的矩阵 U 能够近似地服从 $\mathbb{S}_{n_K}$ 上的均匀分布. 第二步, 每个智能体 i 实施局部策略 K_i+rU_i, 并通过通信网络估计全局损失函数 $J(K+rU)$. 具体来说, 每个智能体 i 用 $\mu_i(t)$ 记录 t 时刻对 $J(K)$ 的估计, 并基于邻居上一时刻的估计 $\mu_j(t)$ 和局部损失函数 $c_i(t)$ 更新:

$$\mu_i(t)=\frac{t-1}{t}\sum_{j=1}^N W_{ij}\mu_j(t-1)+\frac{1}{t}c_i(t).$$

更新规则可以看作是一致过程和在线计算 $\frac{1}{t}\sum_{\tau=1}^t c_i(\tau)$ 的结合. 理论分析可以证明 t 充分大时, 局部估计 $\mu_i(t)$ 可以足够接近真实损失 $J(K)$. 并令 T_G 时刻后的损失函数局部估计为 $\tilde{J}_i$. 第三步, 每个智能体 i 用策略梯度更新

$$K_i^+=K_i-\eta\cdot\frac{n_K}{r}\tilde{J}_iU_i.$$

文献首先证明了 $J(K)$ 是连续可微的, 其非空下水平集都是紧集, 且 $J(K)$ 在任意下水平集上都是光滑的. 基于这些性质, 文献证明了算法以大概率收敛到驻点且策略能镇定系统, 即

$$\frac{1}{T_G}\sum_{s=1}^{T_G}\|\nabla J(K)\|^2\leqslant\epsilon.$$

所需要的样本数为

$$T_GT_J=\Theta\left(\frac{n_K^3}{\epsilon^4}\max\left\{n\beta_0^2,\frac{N}{1-\rho_W}\right\}\right), \tag{31}$$

其中 $\rho_W=\|W-\mathbb{1}\mathbb{1}^{\mathrm{T}}/N\|$ 刻画了一致收敛的速度, β_0 是与 A,B,Σ_w,K_0,Q_i,R_i 相关的常数. 注意到样本复杂度是关于精度 ϵ^{-1}、策略参数数目 n_K 以及智能体数量 N 的多项式, 说明算法规模是可拓展的. 特别地, 式 (31) 与 n_K^3 成正比, 包括单点梯度估计引入的 n_K^2 的复杂度以及损失函数估计非零偏差引入的 n_K 复杂度. 同时, 复杂度跟网络的参数 $N/(1-\rho_W)$ 相关, ρ_W 越大一致速率越小.

6.2 子空间约束

带有结构约束的控制问题中的一类重要情形是控制策略有稀疏性. 文献 [65] 考虑控制律有子空间约束情况下的输出反馈线性二次控制问题. 考虑线性时变的离散时间线性系统:

$$x_{t+1} = A_t x_t + B_t u_t + w_t, \quad y_t = C_t x_t + v_t, \tag{32}$$

其中 $x_t \in \mathbb{R}^n, u_t \in \mathbb{R}^m, y_t \in \mathbb{R}^p$, 随机变量 $x_0, w_t, v_t, t \geqslant 0$ 服从独立的零均值分布, 同时是有界的. 与文献 [64] 不同, 这里考虑有限时域 $t = 0, \ldots, N$ 的控制以及线性输出反馈策略:

$$u_t = K_{t,0} y_0 + K_{t,1} y_1 + \cdots + K_{t,t} y_t,$$

其中, 增益矩阵 $K_{t,0}, K_{t,1}, \ldots, K_{t,t}$ 可能是稀疏的. 令 $u = [u_0^{\mathrm{T}} \ \cdots \ u_{N-1}^{\mathrm{T}}]^{\mathrm{T}}$ 和 $y = [y_0^{\mathrm{T}} \ \cdots \ y_{N-1}^{\mathrm{T}}]^{\mathrm{T}}$, 则控制策略可更紧凑地参数化为

$$u = Ky, \quad K \in \mathcal{K},$$

其中 $\mathcal{K}$ 是 $\mathbb{R}^{mN \times p(N+1)}$ 中的一个子空间并满足 (1) 因果性; (2) 稀疏性约束. 则分布式线性二次最优控制问题为

$$\min_{K \in \mathcal{K}} J(K) := \mathbb{E}\left[\sum_{t=0}^{N-1} (y_t^{\mathrm{T}} M_t y_t + u_t^{\mathrm{T}} R_t u_t) + y_N^{\mathrm{T}} M_N y_N\right], \tag{33}$$

其中惩罚矩阵 $M_t \geqslant 0, R_t \geqslant 0$. 注意到 $J(K)$ 是一个非凸约束优化问题, 而且由于 $\mathcal{K}$ 的稀疏性, 直接求解式 (33) 比较困难 [66]. 为此, 首先将式 (33) 转化为一个无约束优化问题. 假设 $\mathcal{K}$ 的维数为 d, 我们通过定义以子空间 $\mathcal{K}$ 的基为列构成的矩阵 P, 以及函数 $f : \mathbb{R}^d \to \mathbb{R}, f(z) = J(\mathrm{vec}^{-1}(Pz))$, 可以很自然地将式 (33) 转化为无约束优化问题:

$$\min\nolimits_{z \in \mathbb{R}^d} \ f(z). \tag{34}$$

函数 $f(z)$ 通常是非凸的多元多项式, 可能包含多个局部最优. 而且, 与无约束的 LQR 问题不同, 一般无法用里卡提方程求解. 受 Fazel 等 [16] 的启发, 文献 [65] 研究了在什么条件下问题 (34) 也具有梯度支配的性质. 为此, 定义分块矩阵 A, B, C 以及时移矩阵 Z:

$$A = \mathrm{blkdg}(A_0, \ldots, A_N), \quad B = \begin{bmatrix} \mathrm{blkdg}(B_0, \ldots, B_{N-1}) & \\ & 0_{n \times mN} \end{bmatrix},$$

$$C = \mathrm{blkdg}(C_0, \ldots, C_N), \quad Z = \begin{bmatrix} 0_{1 \times N} & 0 \\ I_N & 0_{N \times 1} \end{bmatrix} \otimes I_n,$$

并令 $P = (I - ZA)^{-1} ZB$. 根据分布式控制中的著名结论 [67], 问题 (34) 能够转化为一个无约束的强凸优化问题, 当且仅当如下二次不变性 (quadratic invariance, QI) 条件成立, 即

$$\mathrm{KCPK} \in \mathcal{K}, \quad \forall K \in \mathcal{K}.$$

文献 [65] 利用这一特点, 将强凸性与问题 (34) 建立联系, 证明了 $f(z)$ 具有局部梯度支配的性质. 具体来说, QI 条件意味着存在双射 $h(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ 以及函数 $g : \mathbb{R}^d \to \mathbb{R}$, $g(q) = f(h(q))$, 使得问题 (33) 等价于如下无约束强凸优化:

$$\min\nolimits_{q \in \mathbb{R}^d} \ g(q). \tag{35}$$

由于 $g(q)$ 是强凸的, 故存在常数 $\mu>0$ 使得

$$2\mu(g(q)-J^*)\leqslant\|\nabla g(q)\|_2^2,\quad\forall q\in\mathbb{R}^d.$$

令 $z=h(q)$, 则由于 $f(z)=g(q)$, 我们有

$$2\mu\left(f(z)-J^\star\right)\leqslant\left\|\nabla g\left(h^{-1}(z)\right)\right\|_2^2=\left\|J_h\left(h^{-1}(z)\right)\nabla f\left(h\left(h^{-1}(z)\right)\right)\right\|_2^2\leqslant\left\|J_h\left(h^{-1}(z)\right)\right\|_F^2\|\nabla f(z)\|_2^2,$$

其中, 雅可比矩阵 (Jacobian matrix) $J_h(h^{-1}(z))$ 在紧集上是有界的. 因此, 梯度支配性质在任意紧集上成立. 这一关键性质使得策略优化法能够收敛. 实际上, 许多线性二次控制的优化问题都可以通过变量的代换转化为凸问题, 进而利用类似的分析推导出梯度支配条件, 这正是文献 [68] 的主要发现.

然而, 对于不满足 QI 条件的控制问题, 一般无法转化为凸优化问题, 因此梯度支配的性质也无法保证. 而且文献 [65] 考虑的是有限时域控制问题, 控制策略的可行域实际上是连通的, 而无限时域的分布式控制则通常非连通, 因此降低了优化的难度. 文献 [69] 研究了当 QI 条件不成立时的分布式控制问题, 其定义了一个唯一驻点 (uniquely stationary, US) 性质, 即问题 (33) 是 US 的当且仅当

$$\nabla J(\overline{K})\in\mathcal{K}^\perp\Longrightarrow\overline{K}\in\arg\min_{K\in\mathcal{K}}J(K).$$

US 条件意味着, 稀疏优化问题 (33) 的驻点都是全局最优点. 因此, 投影策略优化法更新 K 在一定步长下可以证明收敛到最优. 尽管利用 US 条件克服了通常的 QI 条件假设, 但文献提出的 US 条件检验需要已知系统数学模型 (32), 因此在模型未知的情况下仍然难以应用.

6.3 风险约束

由于标准的 LQR 问题只关注二次调节性能, 闭环系统可能会容易被极端噪声所影响, 特别是当噪声是重尾分布 (heavy-tailed distribution) 时. 这样一种风险中立 (risk-neutral) 的机制难以应用在一些安全性要求高的物理系统上. 经典文献中通常考虑风险敏感控制, 如在目标函数中引入 LQR 损失函数的方差来补偿风险 [70~73], 或是优化某个风险指标 [74, 75]. 然而, 风险敏感控制只适用于噪声为高斯的情况, 而其他多数方法也都需要已知系统数学模型. 为此, 可以考虑在 LQR 问题中引入风险约束, 同时兼顾二次调节性能和风险. 文献 [76] 首次提出有限时域的风险约束 LQR (risk-constrained LQR), 其风险指标可以化简为一个二次损失函数, 具有和目标函数类似的形式. 进而, 由于有限时域问题可以转化为一个二次约束的二次规划问题 (quadratically constrained quadratic programming, QCQP), 可以证明最优策略是时变的仿射状态反馈. 文献 [18] 将其拓展到了无限时域控制, 并证明最优策略是时不变的仿射反馈. 本小节讨论策略优化法用于无限时域的风险约束 LQR 问题的求解 [17, 19].

文献 [17, 19] 考虑有噪声输入的离散时间线性随机系统:

$$x_{t+1}=Ax_t+Bu_t+w_t,\tag{36}$$

其中, $x_t\in\mathbb{R}^n$, $u_t\in\mathbb{R}^m$. 无限时域的风险约束 LQR 问题为

$$\min\ \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}(x_t^{\mathrm{T}}Qx_t+u_t^{\mathrm{T}}Ru_t)\quad\text{s.t. 式 (36) and }\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}(x_t^{\mathrm{T}}Qx_t-\mathbb{E}[x_t^{\mathrm{T}}Qx_t|h_t])^2\leqslant\rho,\tag{37}$$

其中 ρ 为与风险相关的常数. 假设噪声 w_t 的分布具有有限的四阶矩 [76], 则可将约束简化, 并证明问题 (37) 的最优策略是仿射的 [18], 即 $u^*(x)=-K^*x+l^*$, 且其同时能够镇定系统 (36). 因此, 我们只需

要在策略对 (K,l) 的空间中搜索最优策略. 令 $J(K,l)$ 和 $J_c(K,l)$ 分别为式 (37) 中的目标函数和约束左式, 则我们的目标是求解如下约束非凸优化问题:

$$\min\ J(K,l),\ \ \text{s.t.}\ \ J_c(K,l)\leqslant\rho, \tag{38}$$

其可行域为 $\mathcal{S}=\{[K\in\mathbb{R}^{m\times n},\ l\in\mathbb{R}^m]|\rho(A-BK)<1\}$. 相比于 LQR, 由于非凸约束的存在, 问题 (38) 更加复杂. 特别地, 并不能够证明全局梯度支配性质 [16, 24] 成立. 根据文献 [18, 76], 约束 $J_c(K,l)$ 也具有二次形式. 因此, 我们提出了策略梯度原对偶方法 (policy gradient primal-dual method) 来解决风险约束. 首先, 令 $X=[K\ l]$ 并定义拉格朗日 (Lagrange) 函数:

$$\mathcal{L}(X,\mu)=J(X)+\mu(J_c(X)-\rho),$$

其中 $\mu\geqslant 0$ 是拉格朗日乘子. 定义对偶函数为 $D(\mu)=\min_{X\in\mathcal{S}}\mathcal{L}(X,\mu)$. 以下, 我们将式 (38) 称为主问题, 并将如下:

$$\max_{\mu\geqslant 0}\ D(\mu) \tag{39}$$

称为对偶问题. 我们的原对偶方法按如下方式迭代:

$$X^k\in\operatorname*{argmin}_{X\in\mathcal{S}}\ \mathcal{L}(X,\mu^k), \tag{40}$$

$$\mu^{k+1}=[\mu^k+\zeta^k\cdot d^k]_+, \tag{41}$$

其中, ζ^k 是步长, d^k 是 $D(\mu)$ 在 μ^k 的次梯度, $[x]_+=\max\{0,x\},\forall x\in\mathbb{R}$ 为投影算子.

对于原对偶优化方法来说, 强对偶性的建立是至关重要的. 若强对偶性成立, 那么通过解凹对偶问题 (39), 就能够求得原非凸问题 (38) 的最优解. 为此, 我们首先研究拉格朗日函数的性质. 由于在给定 μ 下, $\mathcal{L}(X,\mu)$ 同样是二次损失函数, 故推测具有与 LQR 损失函数相似的性质. 确实, 我们证明了 $\mathcal{L}(X,\mu)$ 是强制的, 进而其 α- 下水平集 $\mathcal{S}_\alpha=\{X\in\mathbb{R}^{m\times(n+1)}|\mathcal{L}(X,\mu)\leqslant\alpha\}$ 是紧集. 然而, 对于梯度支配性质, 我们只能证明它只在局部成立, 即在任意下水平集 $\mathcal{S}_\alpha$ 内是满足梯度支配的. 不过, 再借助强制性, 我们可以自然地证明, 驻点也是唯一的全局最优点. 如此一来, 考虑到 $\mathcal{L}(X,\mu)$ 是光滑的, 就可以证明式 (40) 在梯度下降、自然梯度法、高斯 – 牛顿法 3 种策略梯度算法下能够全局收敛, 并在固定步长下达到线性收敛速度.

为证明风险约束 LQR 问题的强对偶性, 我们首先证明了问题的一个特殊性质: 给定乘子下的全局最优点 $X^*(\mu)$ 和风险约束 $J_c(X^*(\mu))$ 关于乘子都是连续的. 利用这个性质, 在 Slater 条件下, 就可以通过优化理论中的最优性判据证明强对偶性. 对于对偶迭代 (41) 的求解, 我们可以通过如下方式计算次梯度:

$$d^k=J_c(X^k)-\rho.$$

当动力学模型已知时, $J_c(X)$ 具有显式表达式可以直接计算, 对偶迭代 (41) 可以用标准的投影次梯度算法分析. 当模型未知时, 不能显式地计算次梯度, 而只能通过仿真器估算风险指标 $J_c(X^k)$. 因此, 考虑到估计带来的误差, 可以用随机次梯度算法求解式 (41). 两种情况下都可以得到次线性的收敛速度.

值得一提的是, 风险约束 LQR 实际上是有名的约束马尔可夫过程 (constrained MDP, CMDP) 的一个特例. 已有的文献只证明了当 CMDP 的状态 – 动作空间有限 [77] 或损失函数有界 [78] 时, CMDP 的强对偶性是成立的. 因此, 在这种意义上来说, 文献 [17] 首次证明了二次损失函数下的强对偶性. 不过, 该方向上仍然有许多问题没有解决. 例如, 当约束有多个时, 强对偶性是否成立? 利用扰动里卡提方程理论 [79], 能否挖掘对偶函数的更多性质, 能否加速原对偶算法的收敛? 通过回答这些问题, 能够加深我们对策略优化法在 CMDP 问题上的理解.

7 结语

本文回顾了策略优化法在控制领域当前的发展状况, 对一些重要问题提出了见解. 读者可以从文中所提及的相应文献中找到更深层次的讨论. 我们寄希望于读者能从中发现新的问题, 并在此领域继续努力. 需要指出的是, 因篇幅和作者能力有限, 本文的讨论未能涵盖该领域的所有研究.

基于作者的理解, 本文最后提出一些未来值得研究的方向.

(1) 分布式系统. 目前, 策略优化法在分布式控制领域的研究还存在许多不足. 如仅关注有限时域控制、单个智能体能够获取全局状态及控制输入信息等. 考虑有限时域控制的一个好处在于可行域是连通的, 因此在 QI 条件下策略优化法可以收敛到全局最优点. 但是对于无限时域的静态反馈控制, 通常有指数数量的多个连通区域, 全局保证难以获得. 为此, 可以考虑动态控制策略, 研究其优化图景以及收敛性保证. 此外, 可行域非连通带来的另外一个影响就是策略优化法的效果与初始策略的选取有关. 因此, 如何更高效地初始化, 以及对可行域几何特征的刻画是一个重要的研究方向.

(2) 非线性系统. 策略优化法的研究目前主要集中在线性系统领域, 而实际的物理系统通常是非线性的, 这可能使得本文讨论的方法具有局限性. 尽管如此, 对于一类具有光滑动力学的非线性系统, 仍然可以通过线性化, 来获得平衡点附近的收敛性等理论保证, 这正是文献 [80] 所完成的. 未来可以研究的工作包括如何弱化光滑性条件, 能否扩展到非线性策略如神经网络的分析等. 另外, 策略优化法能否处理非凸损失函数或约束也是一个值得思考的问题.

(3) 鲁棒控制. 目前策略优化法对鲁棒控制的研究主要体现在 $\mathcal{H}_\infty$ 约束上. 一个值得思考的方向是能否直接对闭环系统的 $\mathcal{H}_\infty$ 指标进行优化. 文献 [44] 提到, $\mathcal{H}_\infty$ 范数通常是不光滑的. 因此, 对 $\mathcal{H}_\infty$ 范数性质的研究是一个重要问题. 另外, 目前文献所考虑的都是单个系统的鲁棒控制问题, 因此一个有意义的方向是如何将其扩展到多智能体系统中, 并探索其收敛性.

(4) 正则化方法. 机器学习领域中正则化是一个重要的概念, 能够避免过拟合、加速算法收敛等. 文献 [81] 考虑了卡尔曼滤波器的策略优化问题, 其引入了 “信息性” 的概念, 即需要在损失函数中引入一个正则项来保证策略的信息性, 才能得到全局收敛. 如何理解正则化方法在策略梯度法中的作用, 能否利用正则化方法提高策略优化过程的稳定性或策略的安全性是一个值得探究的方向.

参考文献

1 Hou Z S, Wang Z. From model-based control to data-driven control: survey, classification and perspective. Inf Sci, 2013, 235: 3–35

2 Wang X, Sun J, Berberich J, et al. Data-driven control of dynamic event-triggered systems with delays. 2021. ArXiv:2110.12768

3 Liu W, Sun J, Wang G, et al. Data-driven self-triggered control via trajectory prediction. 2022. ArXiv:2207.08596

4 Kang S, You K. Minimum input design for direct data-driven property identification of unknown linear systems. 2023. ArXiv:2208.13454

5 Zhao F, Li X, You K. Data-driven control of unknown linear systems via quantized feedback. In: Proceedings of Learning for Dynamics and Control Conference, 2022. 467–479

6 Kalman R E. A new approach to linear filtering and prediction problems. J Basic Eng, 1960, 82D: 35–45

7 Ljung L. System identification. In: Proceedings of Signal Analysis and Prediction. Berlin: Springer, 1998. 163–173

8 Kumar V, Todorov E, Levine S. Optimal control with learned local models: application to dexterous manipulation. In: Proceedings of 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016. 378–383

9 Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies. J Machine Learning Res, 2016, 17: 1334–1373

10 Tobin J, Fong R, Ray A, et al. Domain randomization for transferring deep neural networks from simulation to the

real world. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017. 23–30
11 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489
12 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533
13 Recht B. A tour of reinforcement learning: the view from continuous control. Annu Rev Control Robot Auton Syst, 2019, 2: 253–279
14 Tu S, Recht B. The gap between model-based and model-free methods on the linear quadratic regulator: an asymptotic viewpoint. In: Proceedings of Conference on Learning Theory, 2019. 3036–3083
15 Matni N, Proutiere A, Rantzer A, et al. From self-tuning regulators to reinforcement learning and back again. In: Proceedings of the 58th Conference on Decision and Control (CDC), 2019. 3724–3740
16 Fazel M, Ge R, Kakade S, et al. Global convergence of policy gradient methods for the linear quadratic regulator. In: Proceedings of International Conference on Machine Learning, 2018. 1467–1476
17 Zhao F, You K, Başar T. Global convergence of policy gradient primal-dual methods for risk-constrained LQRs. IEEE Trans Automat Contr, 2023, 68: 2934–2949
18 Zhao F, You K, Başar T. Infinite-horizon risk-constrained linear quadratic regulator with average cost. In: Proceedings of the 60th IEEE Conference on Decision and Control (CDC), 2021. 390–395
19 Zhao F, You K. Primal-dual learning for the model-free risk-constrained linear quadratic regulator. In: Proceedings of Learning for Dynamics and Control, 2021. 702–714
20 Zhao F, Fu X, You K. Global convergence of policy gradient methods for output feedback linear quadratic control. 2022. ArXiv:2211.04051
21 Zhao F, Fu X, You K. On the sample complexity of stabilizing linear systems via policy gradient methods. 2022. ArXiv:2205.14335
22 Hu B, Zhang K, Li N, et al. Towards a theoretical foundation of policy optimization for learning control policies. 2022. ArXiv:2210.04810
23 Karimi H, Nutini J, Schmidt M. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In: Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2016. 795–811
24 Bu J, Mesbahi A, Fazel M, et al. LQR through the lens of first order methods: Discrete-time case. 2019. ArXiv:1907.08921
25 Bu J, Mesbahi M. Global convergence of policy gradient algorithms for indefinite least squares stationary optimal control. IEEE Control Syst Lett, 2020, 4: 638–643
26 Malik D, Pananjady A, Bhatia K, et al. Derivative-free methods for policy optimization: guarantees for linear quadratic systems. In: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, 2019. 2916–2925
27 Mohammadi H, Soltanolkotabi M, Jovanovic M R. On the linear convergence of random search for discrete-time LQR. IEEE Control Syst Lett, 2020, 5: 989–994
28 Mohammadi H, Zare A, Soltanolkotabi M, et al. Convergence and sample complexity of gradient methods for the model-free linear-quadratic regulator problem. IEEE Trans Automat Contr, 2022, 67: 2435–2450
29 Hu Y, Wierman A, Qu G. On the sample complexity of stabilizing LTI systems on a single trajectory. 2022. ArXiv:2202.07187
30 Perdomo J C, Umenberger J, Simchowitz M. Stabilizing dynamical systems via policy gradient methods. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34
31 Jing G, Bai H, George J, et al. Learning distributed stabilizing controllers for multi-agent systems. IEEE Control Syst Lett, 2022, 6: 301–306
32 Feng H, Lavaei J. Escaping locally optimal decentralized control polices via damping. In: Proceedings of 2020 American Control Conference (ACC), 2020. 50–57
33 Feng H, Lavaei J. Damping with varying regularization in optimal decentralized control. IEEE Trans Control Netw Syst, 2022, 9: 344–355
34 Lamperski A. Computing stabilizing linear controllers via policy iteration. In: Proceedings of the 59th IEEE Conference

on Decision and Control (CDC), 2020. 1902–1907
35 Safonov M, Athans M. Gain and phase margin for multiloop LQG regulators. IEEE Trans Automat Contr, 1977, 22: 173–179
36 Zhou K, Doyle J C, Glover K. Robust and Optimal Control. Upper Saddle River: Prentice-Hall, Inc., 1996
37 Zheng Y, Tang Y, Li N. Analysis of the optimization landscape of linear quadratic gaussian (LQG) control. 2021. ArXiv:2102.04393
38 Zheng Y, Sun Y, Fazel M, et al. Escaping high-order saddles in policy optimization for linear quadratic Gaussian (LQG) control. 2022. ArXiv:2204.00912
39 Duan J, Cao W, Zheng Y, et al. On the optimization landscape of dynamical output feedback linear quadratic control. 2022. ArXiv:2201.09598
40 Fatkhullin I, Polyak B. Optimizing static linear feedback: gradient method. SIAM J Control Optim, 2021, 59: 3887–3911
41 Duan J, Li J, Li S E, et al. Optimization landscape of gradient descent for discrete-time static output feedback. In: Proceedings of 2022 American Control Conference (ACC), 2022. 2932–2937
42 Zhang K, Yang Z, Başar T. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 11598–11610
43 Zhang K, Hu B, Başar T. On the stability and convergence of robust adversarial reinforcement learning: a case study on linear quadratic systems. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33
44 Zhang K, Hu B, Başar T. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: implicit regularization and global convergence. SIAM J Control Optim, 2021, 59: 4081–4109
45 Zhang K, Zhang X, Hu B, et al. Derivative-free policy optimization for linear risk-sensitive and robust control design: implicit regularization and sample complexity. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 2949–2964
46 Başar T, Bernhard P. H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach. Boston: Birkhäuser, 2008
47 Sanjabi M, Razaviyayn M, Lee J D. Solving non-convex non-concave min-max games under Polyak-Łojasiewicz condition. 2018. ArXiv:1812.02878
48 Nouiehed M, Sanjabi M, Huang T, et al. Solving a class of non-convex min-max games using iterative first order methods. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 32
49 Jacobson D. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. IEEE Trans Automat Contr, 1973, 18: 124–131
50 Glover K, Doyle J C. State-space formulae for all stabilizing controllers that satisfy an H^∞-norm bound and relations to relations to risk sensitivity. Syst Control Lett, 1988, 11: 167–172
51 Allen-Zhu Z, Li Y, Liang Y. Learning and generalization in overparameterized neural networks, going beyond two layers. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 32
52 Chen Y, Candes E. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In: Proceedings of Advances in Neural Information Processing Systems, 2015. 28
53 Ma C, Wang K, Chi Y, et al. Implicit regularization in nonconvex statistical estimation: gradient descent converges linearly for phase retrieval and matrix completion. In: Proceedings of International Conference on Machine Learning, 2018. 3345–3354
54 Zheng Q, Lafferty J. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. 2016. ArXiv:1605.07051
55 Dullerud G E, Paganini F. A Course in Robust Control Theory: A Convex Approach. New York: Springer, 2013
56 Gravell B, Esfahani P M, Summers T. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. IEEE Trans Automat Contr, 2021, 66: 5283–5298
57 Papadimitriou C H, Tsitsiklis J. Intractable problems in control theory. SIAM J Control Optim, 1986, 24: 639–654
58 Blondel V D, Tsitsiklis J N. A survey of computational complexity results in systems and control. Automatica, 2000, 36: 1249–1274
59 Witsenhausen H S. A counterexample in stochastic optimum control. SIAM J Control, 1968, 6: 131–147
60 Fazelnia G, Madani R, Kalbat A, et al. Convex relaxation for optimal distributed control problems. IEEE Trans

Automat Contr, 2016, 62: 206–221
61 Furieri L, Zheng Y, Papachristodoulou A, et al. Sparsity invariance for convex design of distributed controllers. IEEE Trans Control Netw Syst, 2020, 7: 1836–1847
62 Lamperski A, Doyle J C. The $\mathcal{H}_2$ control problem for quadratically invariant systems with delays. IEEE Trans Automat Contr, 2015, 60: 1945–1950
63 Feng H, Lavaei J. On the exponential number of connected components for the feasible set of optimal decentralized control problems. In: Proceedings of 2019 American Control Conference (ACC), 2019. 1430–1437
64 Li Y, Tang Y, Zhang R, et al. Distributed reinforcement learning for decentralized linear quadratic control: a derivative-free policy optimization approach. IEEE Trans Automat Contr, 2022, 67: 6429–6444
65 Furieri L, Zheng Y, Kamgarpour M. Learning the globally optimal distributed LQ regulator. In: Proceedings of Learning for Dynamics and Control, 2020. 287–297
66 Furieri L, Kamgarpour M. Unified approach to convex robust distributed control given arbitrary information structures. IEEE Trans Automat Contr, 2019, 64: 5199–5206
67 Rotkowitz M, Lall S. A characterization of convex problems in decentralized control. IEEE Trans Automat Contr, 2005, 50: 1984–1996
68 Sun Y, Fazel M. Learning optimal controllers by policy gradient: global optimality via convex parameterization. In: Proceedings of the 60th IEEE Conference on Decision and Control (CDC), 2021. 4576–4581
69 Furieri L, Kamgarpour M. First order methods for globally optimal distributed controllers beyond quadratic invariance. In: Proceedings of 2020 American Control Conference (ACC), 2020. 4588–4593
70 Moore J B, Elliott R J, Dey S. Risk-sensitive generalizations of minimum variance estimation and control. J Math Syst Estimation Control, 1997, 7: 123–126
71 Ito Y, Fujimoto K, Tadokoro Y, et al. Risk-sensitive linear control for systems with stochastic parameters. IEEE Trans Automat Contr, 2018, 64: 1328–1343
72 Speyer J L, Fan C H, Banavar R N. Optimal stochastic estimation with exponential cost criteria. In: Proceedings of the 31st IEEE Conference on Decision and Control, 1992. 2293–2299
73 Pan Z, Başar T. Model simplification and optimal control of stochastic singularly perturbed systems under exponentiated quadratic cost. SIAM J Control Optim, 1996, 34: 1734–1766
74 Borkar V, Jain R. Risk-constrained Markov decision processes. IEEE Trans Automat Contr, 2014, 59: 2574–2579
75 Chapman M P, Lacotte J, Tamar A, et al. A risk-sensitive finite-time reachability approach for safety of stochastic dynamic systems. In: Proceedings of American Control Conference, 2019. 2958–2963
76 Tsiamis A, Kalogerias D S, Chamon L F O, et al. Risk-constrained linear-quadratic regulators. In: Proceedings of the 59th IEEE Conference on Decision and Control (CDC), 2020. 3040-3047
77 Altman E. Constrained Markov Decision Processes. Boca Raton: CRC Press, 1999
78 Prashanth L A, Fu M. Risk-sensitive reinforcement learning: a constrained optimization viewpoint. 2018. ArXiv:1810.09126
79 Sun J G. Perturbation theory for algebraic Riccati equations. SIAM J Matrix Anal Appl, 1998, 19: 39–65
80 Qu G, Yu C, Low S, et al. Exploiting linear models for model-free nonlinear control: a provably convergent policy gradient approach. In: Proceedings of the 60th IEEE Conference on Decision and Control (CDC), 2021. 6539–6546
81 Umenberger J, Simchowitz M, Perdomo J C, et al. Globally convergent policy search over dynamic filters for output estimation. 2022. ArXiv:2202.11659

Survey of recent progress in data-driven policy optimization for controller design

Feiran ZHAO & Keyou YOU*

Department of Automation, Tsinghua University, Beijing 100084*, China*
* Corresponding author. E-mail: youky@tsinghua.edu.cn

Abstract With the development of communication technology and artificial intelligence, reinforcement learning (RL), as a data-driven control method, has received tremendous attention. The purpose of this survey is to provide an overview of the state-of-the-art policy optimization method for controller design, which is a typical RL method. In particular, we discuss its convergence and sample complexity in certain fundamental optimal control problems in linear systems, such as linear quadratic regulators, output feedback, $\mathcal{H}_\infty$ control, and distributed control. Additionally, we discuss some future work on the policy optimization for control systems.

Keywords linear system, optimal control, policy gradient method, reinforcement learning, data-driven control