

强化学习闭环中的非正规性： 从动作平滑到源-放大器诊断

独立双语论文包草稿

摘要

连续控制强化学习通常把动作抖动视为平滑性问题。本文将其重构为源-放大器问题：残差动作、探索噪声、执行噪声和执行器扰动本身不是最终风险；它们在进入一个非正规闭环 surrogate 后，可能被方向性放大为状态协方差、约束裕度损失、投影修正负担和后验 Bellman-Lyapunov 残差。本文给出一个最小局部模型、一组 source/amplifier/consequence 诊断量，以及基于 $M = \Phi_u^\top P \Phi_u$ 的证书加权残差协方差响应。当前证据支持的是局部 surrogate 分析与数值仿真，而不是全局非线性安全定理。已有 controlled intervention 显示：在固定特征值和 source covariance 时，增大特征向量非正交性可使状态协方差 trace 从约 0.37 增至 15.66；在固定非正规放大器时，源侧滤波使动作方差从 0.03992 降至 0.00317，状态协方差从 15.849 降至 12.393。二维受约束强化学习消融进一步说明，普通 $\text{Var}(u)$ 和动作差分惩罚本身不足以控制下游风险。

关键词：强化学习；非正规性；瞬态放大；动作平滑；Koopman surrogate；Lyapunov 证书；安全控制

1 引言

连续控制策略即使取得较高回报，也可能生成抖动动作、密集约束裕度损失，并频繁触发安全层修正。常见处理方式是惩罚动作方差、动作差分或执行器粗糙度。这些惩罚有用，但它们不能解释为什么同样大小的残差动作噪声在一个闭环中无害、在另一个闭环中却显著放大。

本文的核心判断是：关键问题不是动作在欧氏范数下是否平滑，而是残差动作协方差是否被注入了一个非正规闭环 surrogate 的高放大方向。残差动作是 source，闭环动力学是 amplifier，状态协方差、约束裕度损失、后验 Bellman-Lyapunov 残差和投影修正负担是 downstream consequence。

因此，本文不是宽泛的 safe RL benchmark，也不是全局非线性安全证明，而是一个问题优先的诊断/控制框架。CG-RVC 在这里仅作为该 source-amplifier 视角的一

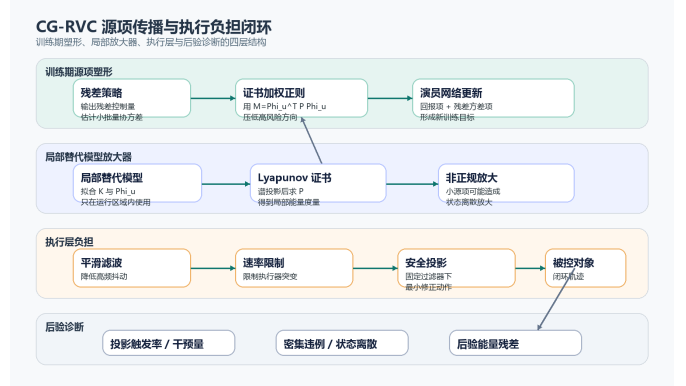


图 1: Source-amplifier 视角。残差动作、探索噪声、执行噪声和扰动通过输入通道 Φ_u 进入非正规闭环放大器 K ，并在下游表现为状态协方差、安全裕度损失、后验 Bellman-Lyapunov 残差和投影修正负担。

种响应方案：它结合局部 surrogate、Lyapunov 加权、残差协方差控制、平滑、rate limiting 和安全投影。

2 非正规闭环放大器

考虑局部 lifted surrogate

$$z_{t+1} = Kz_t + \Phi_u r_t + w_t, \quad (1)$$

其中 $z_t = \psi(x_t)$ 是 lifted state， r_t 是残差动作或命令扰动， Φ_u 是输入通道， w_t 表示扰动或模型误差。若 source covariance 为 Σ_r ，扰动协方差为 W ，则局部平稳协方差满足

$$\Sigma_z = \sum_{i=0}^{\infty} K^i (\Phi_u \Sigma_r \Phi_u^\top + W) (K^i)^\top. \quad (2)$$

式 (2) 表明：动作方差不是状态方差本身；动作方差只有经过闭环放大器传播后才成为状态协方差。

谱半径稳定 $\rho(K) < 1$ 并不能排除短时能量增长。如果 $K = V\Lambda V^{-1}$ 高度非正规，即使特征值位于单位圆内，特征向量也可能病态。因此， $\kappa(V)$ 、峰值瞬态增益 $G_{\text{peak}} = \max_{t \leq H} \|K^t\|_2$ 、resolvent norm 和 Kreiss-like proxy 都应进入强化学习闭环评估。

表 1: source-amplifier 论文应报告的诊断量。

层次	诊断量	含义
Source	$\text{tr}(\Sigma_r)$, applied variance, diff MSE	注入了多少残差或执行动作抖动。
Amplifier	$\rho(K)$, $\kappa(V)$, G_{peak} , resolvent norm	稳定外观下是否仍能短时放大。
Certificate	P , γ_P , $M = \Phi_u^\top P \Phi_u$	哪些动作方向在局部证书下能量更高。
Consequence	状态协方差 trace、dense violation、B-L residual	source 抖动是否变成状态扩散或安全裕度损失。
Execution	projection trigger rate, normalized correction	安全层需要多少修正。

3 为什么普通动作平滑不够

标量动作方差惩罚

$$R_{\text{var}} = \text{tr}(\Sigma_u) \quad (3)$$

只度量 source 总能量，不知道这些能量进入了哪些输入方向。时间粗糙度惩罚

$$R_{\Delta u} = \mathbb{E}\|u_t - u_{t-1}\|^2 \quad (4)$$

主要抑制快速变化，但不判断剩余残差协方差是否对齐了高放大方向。

若存在局部 Lyapunov 度量 P 和测得的收缩系数 γ_P ,

$$K^\top P K \preceq \gamma_P^2 P, \quad (5)$$

则残差输入方向可以被

$$M = \Phi_u^\top P \Phi_u \quad (6)$$

加权。相应的 P-weighted action Lyapunov energy 为

$$R_{\text{P-ALE}} = \frac{\text{tr}(M \Sigma_r)}{1 - \gamma_P^2}. \quad (7)$$

动作平滑是时间正则性；P-ALE 是放大器感知的残差协方差控制。二者相关，但并不等价。

4 最小证书引导响应

响应框架包含五步。第一，使用局部线性化、EDMDc 回归或 replay-buffer 局部模型拟合 $z_{t+1} = K_0 z_t + \Phi_u u_t + w_t$ 。第二，构造名义或投影闭环矩阵，例如 $K_{\text{raw}} = K_0 -$

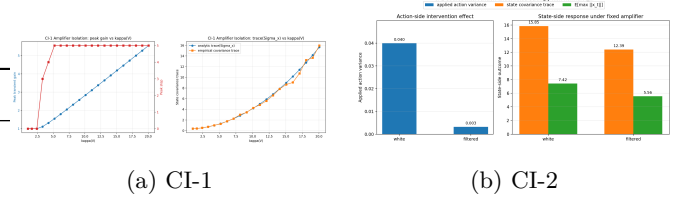


图 2: Controlled interventions. CI-1 操纵非正规 amplifier geometry, 同时固定 source statistics; CI-2 操纵 source statistics, 同时固定 amplifier。

$\Phi_u L$, 再得到 $K_{\text{proj}} = \Pi_\beta(K_{\text{raw}})$ 。第三，求解 Lyapunov 方程或不等式得到 P 。第四，在 actor objective 中加入式 (7) 的 P-ALE 惩罚。第五，在执行层加入平滑、rate limiting 和安全投影。

这里必须区分 β 与 γ_P 。 β 是谱半径目标，不是 P-metric contraction guarantee。理论界限和经验诊断应使用测得的 γ_P 。此外，平滑与投影使执行层变成非线性操作，因此证书应被理解为局部设计和诊断对象，而不是全局非线性 plant 证书。

如果残差策略依赖状态，则残差不再是纯外生 source。局部诊断应估计

$$K_\theta = K + \Phi_u J_r, \quad (8)$$

其中 J_r 是有限差分 actor-Jacobian proxy。当 J_r 很大时，feedback component 必须被纳入 amplifier。

5 证据

实验设计应分离 source、amplifier 和 safety repair。图 2 汇总了两个 controlled interventions。CI-1 固定特征值、谱半径、输入通道和 source covariance，只改变特征向量几何。在已归档结果中， $\kappa(V)$ 从约 1.0 增至 20.05，状态协方差 trace 从约 0.37 增至 15.66，峰值瞬态增益从约 1.0 增至 5.55。这说明在受控矩阵族中，非正规几何可以独立放大固定 source。

CI-2 固定强非正规 amplifier，只改变 source。在已归档结果中，filtered source 使动作方差从 0.03992 降至 0.00317，状态协方差 trace 从 15.849 降至 12.393，而结构峰值增益保持在约 5.55。这是源侧证据，但 filtered input 是 colored input，必须通过增广 colored-input dynamics 解释。

表 2 说明，在二维高噪声受约束任务中，普通 smoothing 并不完整。标量动作方差惩罚和动作差分惩罚没有消除 dense violation 或后验残差。Projection only 能降低 vi-

表 2: 二维受约束 RL 高噪声消融摘要。数值来自已归档 matched ablation 的均值。

方法	RMSE	Var.	Viol.	B-L max
Direct TD3	0.537	0.1138	0.9389	92.4
TD3 + Var(u)	0.612	0.1026	0.9278	161.9
TD3 + action diff	0.565	0.1118	0.9389	109.1
Projection only	0.505	0.1033	0.1222	82.3
Full response	0.284	0.0060	0.0000	5.24

olation, 但仍保留较高 applied variance 和 posterior residual。Full response 同时组合 source shaping 与 execution repair。

Projection burden 是一个操作层面的下游诊断。在已归档二维高噪声 run 中, normalized correction 约为: projection only 0.09362, P-ALE + projection 0.05426, full method 0.01093。Residual-feedback diagnostic 则给出另一条边界: 有限差分 actor-Jacobian proxy 在 Direct TD3 中约为 16.23, 在 full method 中约为 0.274。这支持 source-side 解释, 但不是证明。

6 局限性

本文 claim level 有意保持克制。第一, 所有证书声明都是局部 surrogate 层面的, 不是全局非线性安全保证。第二, P-ALE 降低的是进入 amplifier 的 certificate-weighted residual covariance, 并不直接降低 K 的非正规性。第三, 如果 residual feedback 很强, 则 source-only interpretation 只能作为诊断, 除非将 feedback 纳入式 (8)。第四, projection 依赖局部可行性, 不能凭空创造 feasibility。第五, Full3D quadrotor 和 ControlGym 结果是 simulation-level relevance check, 不是硬件验证或 broad safe-RL benchmark dominance。

7 结论

低方差安全强化学习应同时报告 source 与 amplifier。关键对象不仅是 $\text{Var}(u)$ 或 $\|u_t - u_{t-1}\|^2$, 而是从 Σ_r 经过 Φ_u 、 K 、 P 和 M , 最终进入状态协方差、安全裕度损失、后验 Bellman-Lyapunov 残差和 projection burden 的完整链条。该 source-amplifier 视角解释了为什么普通 action smoothing 在强非正规闭环中可能不足, 也解释了为什么证书加权残差协方差是一个自然的诊断和控制响应。