Zertifizierte Benchmarks für Reinforcement Learning mittels stochastischer konverser Optimalität

P. Osinenko[1], **S. Ibrahim**[2], **G. Ouerdane**[2], **H. Ouerdane**[2], **S. Streif**[4]

[1] Central University (Moscow); CESS (Moscow); Sirius University, E-Mail: p.osinenko@yandex.ru

[2] CESS (Moscow); [3] Innopolis University / Moscow Center for Advanced Studies; [4] Technische Universität Chemnitz

Das Benchmarking von Reinforcement-Learning-Verfahren für Regelungsaufgaben ist oft nur relativ: man vergleicht erzielte Kosten oder Belohnungen, ohne die optimale Politik des Testproblems zu kennen. Dadurch bleibt unklar, ob ein Verfahren nahe am Optimum liegt oder lediglich besser als eine schwache Referenz ist. Wir stellen einen konstruktiven Ansatz vor, der dieses Problem umkehrt: Statt für ein gegebenes System die optimale Politik zu suchen, wird ein stochastisches nichtlineares System so konstruiert, dass eine vorgegebene Wertfunktion und die zugehörige Politik nachweislich optimal sind.

Ausgangspunkt ist ein diskretes stochastisches System

$$s_{k+1} = f(s_k) + g(s_k)a_k + w_k, \quad c(s,a) = q(s) + \rho(a).$$

Unter Regularitäts- und Integrierbarkeitsannahmen liefert eine stochastische konverse Optimalitätsaussage notwendige und hinreichende Bedingungen dafür, dass eine gewählte Funktion V^* die Bellman-Gleichung erfüllt. Im quadratischen Fall

$$V^*(s) = s^\top P s + b, \quad c(s,a) = s^\top Q s + a^\top R a,$$
$$a^*(s) = -\gamma B(s)^{-1} g(s)^\top P f(s), \quad B(s) = R + \gamma\, g(s)^\top P g(s)$$

ergibt sich die optimale Rückführung. Die verbleibende Freiheit in der Dynamik wird durch eine Energiegleichung parametrisiert. Damit lassen sich Familien nichtlinearer, verrauschter und skalierbarer Testsysteme erzeugen, deren Optimum analytisch bekannt ist und deren Bellman-Residual vor der Verwendung numerisch zertifiziert wird.

Die resultierenden Benchmarks erlauben absolute statt nur relative Bewertung: Jede gelernte Politik wird gegen das bekannte Optimum auf identischen Anfangszuständen und Rauschrealisationen ausgewertet. Die Metriken sind insbesondere die normierte Optimalitätslücke und der gepaarte Regret. Neben klassischem Deep RL kann derselbe Aufbau auch für interaktive, durch Large Language Models vorgeschlagene Regler verwendet werden: Das Modell sieht nur die black-box-Interaktion mit dem System, während die Auswertung verborgen gegen das zertifizierte Optimum erfolgt. So entsteht ein reproduzierbarer Prüfstand für lernende Regelungsverfahren, bei dem Nichtlinearität, Rauschstärke, Aktuatorgeometrie und Anfangszustandsbereich kontrolliert variiert werden.

Literatur.

[1] Yaremenko, G.; Ibrahim, S.; Moreno-Mora, F.; Osinenko, P.; Streif, S.: Generating Informative Benchmarks for Reinforcement Learning. *IEEE Control Systems Letters* 9, 480–485, 2025.

[2] Ibrahim, S.; Ouerdane, G.; Salloum, H.; Ouerdane, H.; Streif, S.; Osinenko, P.: Benchmarking Reinforcement Learning via Stochastic Converse Optimality. *IEEE Control Systems Letters* 12, 1321–1326, 2026.