Optimal Control and Reinforcement Learning

Michele Palladino
University of L'Aquila - DISIM

Machine Learning and PDEs Workshop
Research Center for Mathematics of Data (MoD)
Friedrich-Alexander Universität (FAU)
Erlangen

28/04/2025

Overall goal of reinforcement learning

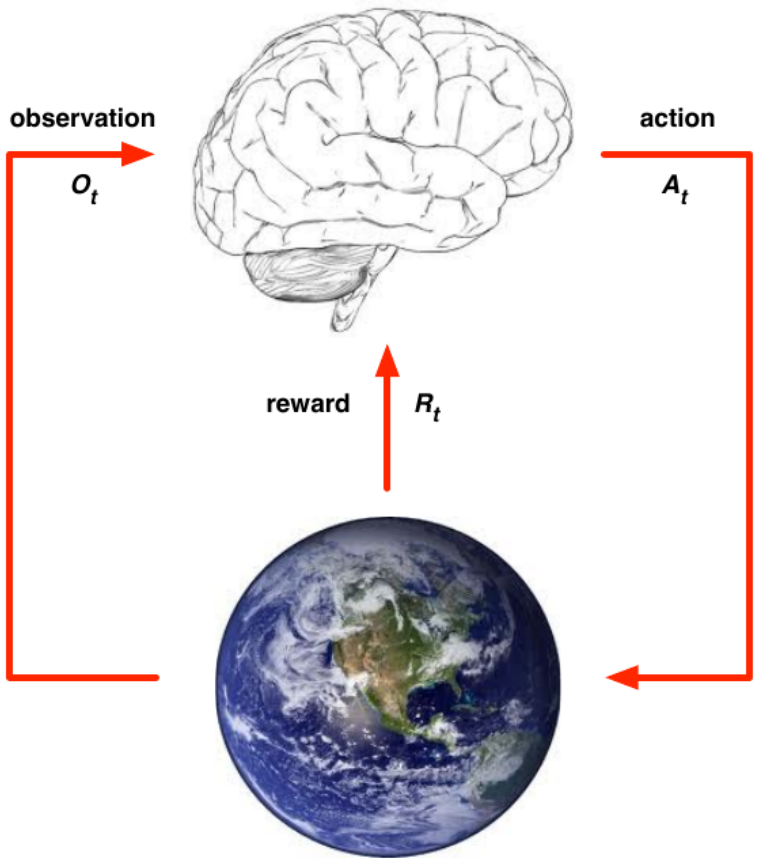


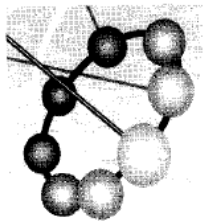
Figure: Framework for reinforcement learning (diagram from David Silver's lectures)

Some notable successes

Go has 10^{170} legal states. Recently *AlphaGo* beat humans (years ahead of "schedule")

RL and Optimal Control

Reinforcement Learning is Direct Adaptive Optimal Control



Richard S. Sutton, Andrew G. Barto, and Ronald J. Williams

Reinforcement learning is one of the major neural-network approaches to learning control. How should it be viewed from a control systems perspective? Control problems can be divided into two classes: 1) regulation and tracking problems, in which the objective is to follow a reference trajectory, and 2) optimal control problems, in which the objective is to extremize a functional of the controlled system's behavior that is not necessarily defined in terms of a reference tlajectory. Adaptive methods for problems of the first kind are well known, and include self-tuning regulators and model-reference methods, whereas adaptive methods for optimal-control problems have received relatively little attention. Moreover, the adaptive optimal-control methods that have been studied are almost all *indirect* methods, in which controls are recomputed from an estimated system model at each step. This computation is inherently complex, making adaptive methods in which the optimal controls are estimated directly more attractive. We view reinforcement learn-

meets a collection of specifications constituting the control objective. In some problems, the control objective is defined in terms of a reference level or reference trajectory that the controlled system's output should match or track as closely as possible. Stability is the key issue in these regulation and tracking problems. In other problems, the control objective is to extremize a functional of the controlled system's behavior that is not necessarily defined in terms of a reference level or trajectory. The key issue in the latter problems is constrained optimization; here optimal-control methods based on the calculus of variations and dynamic programming have been extensively studied. In recent years, optimal control has received less attention than regulation and tracking, which have proven to be more tractable both analytically and computationally, and which produce more reliable controls for many applications.

When a detailed and accurate model of the system to be controlled is not available, adap-

Ideally, one would like to have both the trajectories and the required controls determined so as to extremize the objective function.

For both tracking and optimal control, it is usual to distinguish between *indirect* and *direct* adaptive control methods. An indirect method relies on a system identification procedure to form an explicit model of the controlled system and determines then the control rule from the model. Direct methods determine the control rule without forming such a system model.

In this paper we briefly describe learning methods known as *reinforcement learning methods*, and present them as a direct approach to adaptive optimal control. These methods have their roots in studies of animal learning and in early learning control work (e.g., [22]), and are now an active area of research in neural networks and machine learning (e.g., see [1], [41]). We summarize here an emerging deeper understanding of these methods that is being obtained by viewing

An RL algorithm: Q-learning

Given a reward function $r(x)$, $\alpha_t \in (0, 1]$ for all t, x_0 initial state, an initial action-value function $Q(x, u)$ and a set of actions U:

- Pick u_t using Q and x_t (e.g. $u_t = \arg\max Q(x_t, u)$, or ϵ-greedy);
- Observe the state x_{t+1} and the reward $r(x_{t+1})$
- Update Q as follows

$$Q(x_t, u_t) \leftarrow Q(x_t, u_t) + \alpha_t \big(r(x_{t+1}) + \max_{u \in U} Q(x_{t+1}, u) - Q(x_t, u_t)\big)$$

- Repeat the procedure at x_{t+1} until the end of the episode (e.g. a time T or a recurrent state condition)

RL: Model Free VS Model Based

1) **Model Free Algorithms:** aim to approximate the Value Function, combining the Dynamic Programming Principle and/or the Monte Carlo Method. Examples:
   - SARSA;
   - Q-learning;
   - etc.
2) **Model Based Algorithms:** aim to approximate the control system and to control it *simultaneously*. Examples:
   - PILCO (using GPs)
   - DeepPILCO (using DNN)

PILCO: problem setting

Given x_0 initial condition, consider the deterministic problem

$$x_{t+1} = f(x_t, u_t), \qquad t = 0, 1 \ldots, T,$$

where the function f is **unknown**.

A policy is a deterministic mapping $\pi : \mathbb{R}^n \times \mathbb{R}^k \to \mathcal{A}$, $\pi(x, \theta) \in \mathcal{A}$.
θ is the parameter which one can use to improve the policy.

Goal: Maximize the reward (or minimize the cost)

$$J^\pi(\theta) = \sum_{t=0}^{T} \mathbb{E}_f \left[c(x_t) \right],$$

where

$$p\Big(f(x_t, u_t) \,|(x_t, u_t), \ldots, (x_0, u_0) \Big) \sim \mathcal{N}(\mu_t, \Sigma_t).$$

$(x_t, u_t), \ldots, (x_0, u_0)$ are the training inputs.

Summary of PILCO

For a fixed class of policies $\{\pi(\cdot, \theta) : \theta \in \Theta\}$:

- Given a prior distribution over f, compute the expected value

$$\mathbb{E}_f [c(x_t)]$$

  (policy evaluation);
- Improve the policy by adjusting the parameter θ (policy improvement).
- Given the new experience on the system, compute the posterior distribution on the dynamics f.

Prior/Posterior distributions in GPs

Suppose one has $x_1, \ldots, x_d$ observed inputs and $y_1, \ldots, y_d$ observed outputs. Assume the representation $h(x_i) = y_i + \epsilon_i$ where h is an unknown function and $\epsilon_i \sim \mathcal{N}(0, \sigma_{\epsilon_i})$, for $i = 1, \ldots, d$ independent.
Bayes' Theorem yields:

$$p(h\,|\mathbf{x},\,\mathbf{y},\,\theta) = \frac{p(\mathbf{y}|\,h, \mathbf{x},\,\theta)p(h|\,\theta)}{p(\mathbf{y}\,|\mathbf{x},\,\theta)}$$

where

- θ paramater controlling the Gaussian distribution (*hyper-parameter*);
- $p(h|\,\theta)$ is the *prior* distribution;
- $p(\mathbf{y}|\,h, \mathbf{x},\,\theta)$ is a jointly (finite) Gaussian distribution with mean $h(\mathbf{x}) = (h(x_1), \ldots, h(x_d))$ and covariance matrix $\mathrm{diag}(\sigma^2_{\epsilon_1}, \ldots, \sigma^2_{\epsilon_d})$.

Figure: From YouTube Channel: PilcoLearner (https://www.youtube.com/watch?v=XiigTGKZfks)

References on PILCO

- Deisenroth, Rasmussen, "PILCO: A Model-Based and Data-Efficient Approach to Policy Search",in Proceedings of the 28th International Conference on machine learning (ICML-11), 2011
- Deisenroth, Fox, Rasmussen, "Gaussian processes for data-efficient learning in robotics and control", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.
- Gal, McAllister, Rasmussen "Improving PILCO with Bayesian Neural Network Dynamics Models", (ICML-16), 2016
- Deisenroth. "Efficient Reinforcement Learning using Gaussian Processes", PhD Thesis.

Question

Assume that the underlying (unknown) physical system is represented by a function $\hat{f}$ and that does not drastically change in time (no failure, no fault detections etc).

Does the approximated optimal policy converge to the optimal policy of the physical system?

Our framework

The state equation of the system is

$$(S)\begin{cases} \dot{x}(t) = f(x(t)) + \sum_{i=1}^{I} u^i(t) g_i(x(t)) & t \in [t_0, T] \\ x(t_0) = x_0. \end{cases}$$

We define the cost functional (or payoff functional)

$$J_{t_0}[x(\cdot), u(\cdot)] = \int_{t_0}^{T} G(x(t)) + u^T(t) R u(t) dt + h(x(T)), \quad u(\cdot) \in \mathcal{U}_{t_0},$$

where $\mathcal{U}_{t_0} = \{u : [t_0, T] \to \mathbb{R}^I, \text{ measurable}\}$ and R is a positive definite matrix.

The goal is to minimize J_{t_0} over the trajectory/control pairs of (S).

Our framework

However, we assume as known just the functions $g_1, \dots, g_l$, while the drift f is **unknown**. This leads to the optimal control

$$(OC)_N \begin{cases} \text{minimize} \displaystyle\int_X J_{t_0}[x_f(\cdot), u(\cdot)] p^N(df) \\ \text{over the } (x, u) \text{ s.t.} \\ \dot{x}(t) = f(x(t)) + \displaystyle\sum_{i=1}^{l} u^i(t) g_i(x(t)) \qquad t \in [t_0, T] \\ x(t_0) = x_0. \end{cases}$$

Here, $p^N(df)$ is a probability measure defined on $X \subset C^0(\mathbb{R}^n)$ (constructed at the N-th episode).

Research question: (Falcone - P. - Pesare, '21), (P. - Pesare- Scarinci, '25)

Suppose that the real, underlying, physical system is

$$(\hat{S}) \begin{cases} \dot{x}(t) = \hat{f}(x(t)) + \sum_{i=1}^{I} u^i(t) g_i(x(t)) & t \in [t_0, T] \\ x(t_0) = x_0. \end{cases}$$

where $\hat{f}$ is the real drift. Assume that $\hat{f} \in X$ and that $p^N \rightharpoonup \delta_{\hat{f}}$ (that is, p^N weakly converges to $\delta_{\hat{f}}$).

Then what does it happen to the optimal controls?
Here we will tackle the more general case in which $p^N \rightharpoonup \hat{p}$ for a generic $\hat{p}$.

Standing Assumptions

H1) For all $f \in X \subset C^0(\mathbb{R}^n)$, exist L_f, M_f s.t.

$$|f(x) - f(y)| \leq L_f |x - y|, \qquad \text{for all} \quad x, y \in \mathbb{R}^n$$

and

$$|f(x)| \leq M_f (1 + |x|), \qquad \text{for all} \quad x \in \mathbb{R}^n;$$

H2) For $g : \mathbb{R}^n \to \mathbb{M}^{n \times l}$, exist L_g, M_g s.t.

$$||g(x) - g(y)|| \leq L_g |x - y|, \qquad \text{for all} \quad x, y \in \mathbb{R}^n$$

and

$$||g(x)|| \leq M_g (1 + |x|) \qquad \text{for all} \quad x \in \mathbb{R}^n;$$

H3) $G : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}$ are bounded below and continuous.

H4) $R \in \mathbb{M}^{l \times l}$ is positive definite.

A first convergence result (P. - Pesare- Scarinci, '25)

Theorem (Γ-convergence): Let the "Standing Assumptions" be satisfied. Assume that $p^N \rightharpoonup \hat{p}$. Then, for each $s \in [0, T]$ and $x_0 \in \mathbb{R}^n$, the sequence of functionals $F_N : (L^2([s, T], \mathbb{R}^l), \tau_w) \to \mathbb{R}$ defined as

$$F_N(u) := \int_X \left[\int_s^T G(x_f(t; u))dt \right] dp^N(f) + \int_s^T u(t)^T Ru(t)dt$$

Γ-converges to $\hat{F} : (L^2([s, T], \mathbb{R}^l), \tau_w) \to \mathbb{R}$ defined as

$$\hat{F}(u) := \int_X \left[\int_0^T G(x_f(t; u))dt \right] d\hat{p}(f) + \int_0^T u(t)^T Ru(t)dt.$$

Implications of the Γ-convergence

I1) Take a sequence of minimizers $u_N \in \mathcal{U}$ to the optimal control problem $(OC)_N$. If $\hat{u}$ is a cluster point of $\{u_N\}_{N\in\mathbb{N}}$ (with resepct to the weak convergence in $L^2([s,T],U)$, then $\hat{u}$ is a minimizer of $(\hat{OC})$.

I2)

$$V_{\hat{p}}(s,x_0) = \limsup_{N\to\infty} V_{p^N}(s,x_0), \qquad \text{for all} \quad (s,x_0) \in [0,T]\times\mathbb{R}^n$$

where $V_{\hat{p}}$ is the value function of $(\hat{OC})$ and V_{p^N} is the value function of $(OC)_N$

A stronger result on the value functions' convergence (Pesare-P.-Falcone, '21)

Theorem: Given the sequence $\{p^N\}$, assume that $W_1(p^N, \hat{p}) \to 0$. We use V_{p^N} and $V_{\hat{p}}$ to denote the value function for the average problem with p^N and $\hat{p}$ respectively.

Then, for each $K \subset [0, T] \times \mathbb{R}^n$ compact,

$$V_{p_N} \to V_{\hat{p}} \quad \text{uniformly} \quad \text{in} \quad K \quad \text{for} \quad N \to \infty,$$

and, in particular

$$||V_{p^N} - V_{\hat{p}}||_{\infty,K} \leq C_K W_1(p^N, \hat{p})$$

with C_K positive constant depending just on the set K.

Necessary conditions of optimality for $(\hat{OC})$

- **(D)**: Assume that $X \subset C^1(\mathbb{R}^n;\mathbb{R}^n)$, X complete metric space and that g, G, h are mappings of class C^1.

Theorem: Let $\{(\hat{x}_f, \hat{u})(\cdot) : \ f \in X\}$ be a minimizer for $(\hat{OC})$. Assume the Standing Assumptions and **(D)**. Then, there exists a function $\lambda_f \in W^{1,1}([s,T];\mathbb{R}^n)$ for all $f \in X$ s.t.

$$\hat{u}(t) = R^{-1} \int_X g(\hat{x}_f(t))^T \lambda_f(t)\, d\hat{p}(f) \qquad \text{for all} \quad t \in [s,T]\ ;$$

$$-\dot{\lambda}_f(t) = \left(J_x f(\hat{x}_f(t)) + \sum_{i=1}^{l} \hat{u}^i(t)\, J_x g_i(\hat{x}_f(t)) \right)^T \lambda_f(t) - \nabla_x G(\hat{x}_f(t))$$

$$a.e.\ t \in [s,T], \text{ for all} \quad f \in supp(\hat{p})\,,$$

$$-\lambda_f(T) = \nabla_x h(\hat{x}_f(T)) \qquad \text{for all} \quad f \in supp(\hat{p})\,.$$

Convergence of the multipliers

Theorem: Assume the "Standing Assumptions" and **(D)**.
Consider $\{p^N\}$ sequence of measures such that

$$p^N \rightharpoonup \hat{p}.$$

Suppose that $u^N(\cdot)$ is the optimal control of $(OC)_N$ and is converging weakly in $L^2([s,T];\mathbb{R}^l)$ to $\hat{u}(\cdot)$. Then $\hat{u}(\cdot)$ is optimal for $(\hat{OC})$.

Furthermore a cluster point of any sequence of adjoint variables $\{\lambda_f^N : \ f \in X\}_{N\in\mathbb{N}}$ related to $\left\{(x_f^N, u^N)(\cdot) : \ f \in X\right\}$ is an adjoint arc for $\{(\hat{x}_f, \hat{u})(\cdot) : \ f \in X\}$, implying that $u^N(\cdot) \to \hat{u}(\cdot)$ uniformly in $[s,T]$.

Comments

- The previous theorem requires knowing the minimizers u^N to be applied (not very convenient...);
- This is due to the fact that the problem is nonlinear, hence there are several minimizers as well as extremals that are not minimizers;
- A natural question is then to provide conditions such that, if we take $\{(\hat{x}_f, \hat{u}) : \quad f \in X\}$ satisfying the necessary conditions, then $\{(\hat{x}_f, \hat{u}) : \quad f \in X\}$ is a minimizer for $(\hat{OC})$.
- In this case, the previous theorem provides a way to find sub-optimal controls.

Sufficient Conditions

Call

$$H(x_f, \lambda_f) = \max_{u \in \mathbb{R}^l} \left\{ \int_X (\lambda_f \cdot (f(x_f) + g(x_f) \cdot u) - G(x_f)) \, \hat{p}(df) - u^T R u \right\}$$
$$\left(= \int_X \left(f(x_f)^T \lambda_f - G(x_f) \right) \hat{p}(df) \right).$$

(CC) Assume that $x_f \mapsto H(x_f, \lambda_f)$ is concave for all λ_f, while the map $x \mapsto h(x)$ is convex.

Theorem: Assume the "Standing Assumptions", **(D)** and **(CC)**. Take $\{(\hat{x}_f, \hat{u})(\cdot) : \ f \in X\}$ feasible process such that the Necessary Conditions hold true. Then $\{(\hat{x}_f, \hat{u})(\cdot) : \ f \in X\}$ is a global minimizer for $(\hat{OC})$.

Comments

Suppose that the real, underlying, physical system is

$$(\hat{S}) \quad \begin{cases} \dot{x}(t) = \hat{f}(x(t)) + \sum_{i=1}^{I} u^i(t) g_i(x(t)) & t \in [t_0, T] \\ x(t_0) = x_0. \end{cases}$$

where $\hat{f}$ is the real drift and that $p^N \rightharpoonup \delta_{\hat{f}}$ (that is, p^N weakly converges to $\delta_{\hat{f}}$).
Condition **(CC)** reads as

$$x \mapsto H(x, \lambda) = \hat{f}(x)^T \lambda - G(x)$$

is concave for each λ. Notice that, in RL, G is a design function...

Special case: Linear Quadratic Regulator (Pesare-P.-Falcone '21)

Consider the state dynamics:

$$\dot{x}_A(t) = Ax_A(t) + Bu(t), \qquad x(0) = x_0 \tag{0.1}$$

- State vector: $x_A \in \mathbb{R}^n$ State matrix: $A \in \Omega \subset \mathbb{M}^{n\times n}$ **unknown**
- Control vector: $u \in \mathbb{R}^l$ Control matrix: $B \in \mathbb{M}^{n\times l}$

Quadratic cost:

$$\min_{u\in\mathcal{U}} \left\{ \frac{1}{2} \int_\Omega \left(\int_0^T x_A(s)^T Q x_A(s) + u^T(s) R u(s)\, ds + x_A^T(T) G x_A(T) \right) p(dA) \right\}$$

- $Q, G \in \mathbb{M}^{n\times n}$, symmetric, semipositive definite ($Q,\ G \geq 0$), $R \in \mathbb{M}^{l\times l}$ symmetric, positive definite ($R > 0$).

LQR Framework

- p is a measure over the compact set of matrices Ω;
- The set of possible dynamics is

$$X := \{f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n : \ f(x,u) = Ax + Bu, \ A \in \Omega, \ u \in \mathcal{U}\}.$$

  p can be regarded as a measure over X;
- p tracks the belief that an agent has on the dynamics.

Remark: In the LQR setting, the optimal policy is a linear feedback control!

Conclusions and Open questions

- We presented a mathematical framework able to justify certain model-based RL approaches
- We showed certain convergence properties of the optimal controls (optimal policies, in the RL language)
- Open question: Stability analysis for Hamilton-Jacobi Equations
- Open question: Convergence result for Riccati equations;
- Open question: Convergence results for the control constrained case.
- Open question: Applications to PDEs.

Thanks for your attention!

References

- R.W. Murray, M. P., *"A model for system uncertainties in Reinforcement Learning"*, Systems and Control Letters, 2018.
- A. Pesare, M. P., M. Falcone, *"Convergence of the Value Function in Optimal Control Problems with Unknown Dynamics"*, European Control Conference (ECC) 2021
- A. Pesare, M. P., M. Falcone, *"A convergent approximation of the linear quadratic optimal control problem for Reinforcement Learning"*, Mathematics of Control, Signals and Systems, 2021
- M. P., A. Pesare, T. Scarinci, *"Convergence results for control problems with unknown dynamic and applications to reinforcement learning"*, under revision.
- A. Alla, A. Pacifico, M. P., A. Pesare, *Online identification and control of PDEs via Reinforcement Learning methods*, Advances in Computational Mathematics, 2024.