Adaptive Control Through Reinforcement Learning

J. Miranda Lemos

INESC-ID
Instituto Superior Técnico/Universidade de Lisboa, Portugal
jlml@inesc-id.pt

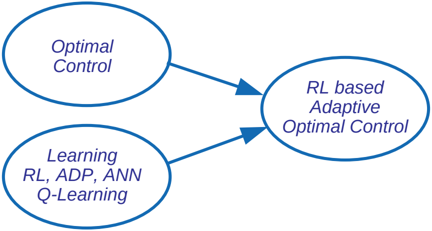
Seminars on Mathematics, Physics and Machine Learning

July 16, 2020

Objective

Explain to a wide audience:

How to design adaptive optimal controllers by combining optimal control with reinforcement learning, approximate dynamic programming, and artificial neural networks?

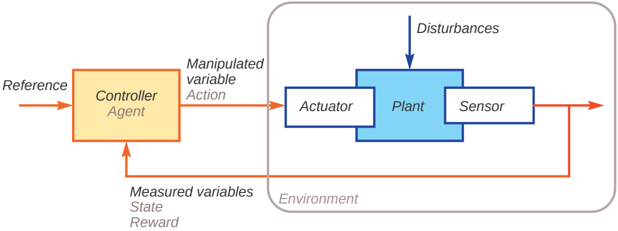


Presentation road map

- What is adaptive control?
- Approaches to adaptive control
- Early Reinforcement Learning based controllers
- RL based linear Model Predictive Control (MPC)
- How to tackle adaptive nonlinear optimal control?
- Approximate Dynamic Programming (ADP)
- Q-Learning
- Conclusions

What is control?

Stimulate a system such that it behaves in a specified way.



Physical system (good old gravity law!)

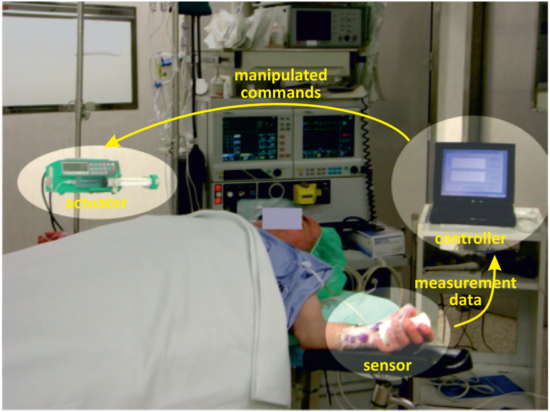
Control modifies dynamic behaviour (Cyber-Physical Systems)

Cyber-physical system

Help of Paula and Francisco kindly acknowledged.

What is control? An example: anesthesia

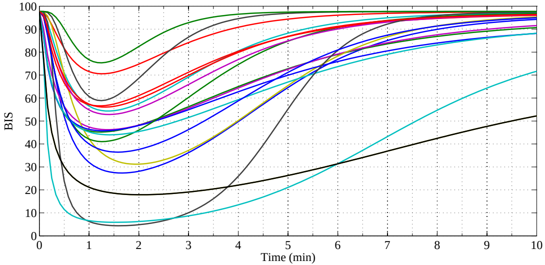
Controlling neuromuscular blockade for a patient subject to general anesthesia



Source: Project GALENO, Photo taken at Hospital de S. António, Porto, Portugal.

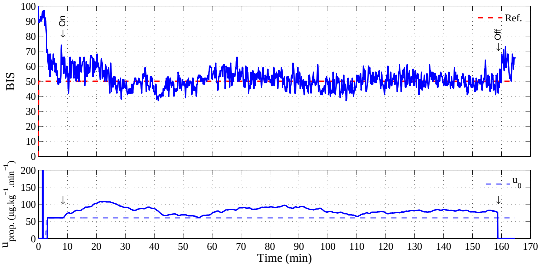
Uncertainty

Uncertainty: Unpredictable variability in plant dynamics.



Robustness

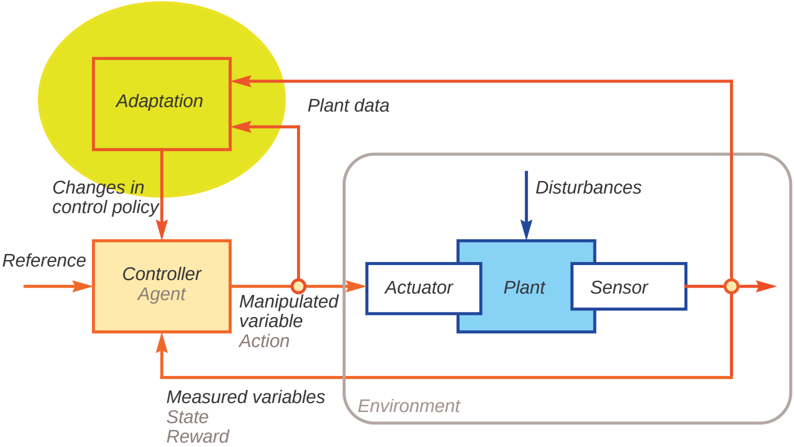
Robustness: Design the controller for a nominal model, but it works with nearby systems (with graceful degradation in performance)
Example: control of the level of self-unconsciousness in patients subject to general anesthesia Clinical results



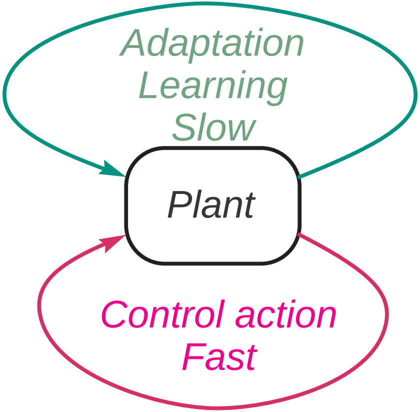
J. M. Lemos, D. V. Caiado, B. A. Costa, L. A. Paz, T. F. Mendonça, R. Rabiço, S. Esteves and M. Seabra (2014). Robust Control of Maintenance Phase Anesthesia. *IEEE Control Systems*, 34(6):24-38.

What is adaptive control?

Modify the control law (= control policy) to make it match the plant.
Learn the "best" control policy. **Not** merely the plant inverse.



Two time-scales system



Why use adaptive control?

Controlling time varying processes.
Controlling processes with big variability.

Source: Hizook, 2012

KIVA robots for automatic warehouses (now Amazon robotics)

Use low cost components causes big variability

Use adaptive control to compensate uncertainty.

Approaches to adaptive control

- Joint parameter and state nonlinear estimation
- Certainty equivalence
- SMMAC - Supervised Multiple Model Adaptive Control
- Model falsification
- Reinforcement Learning (RL)
- Control Lyapunov Functions (CLF)

Joint parameter and state nonlinear control

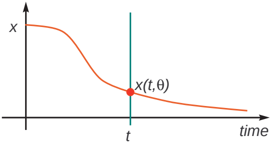
$$\frac{dx}{dt} = f(x, \theta)$$

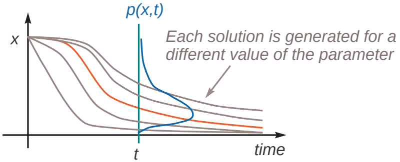
Stochastic control of the hyperstate untractable in computational terms.
Need for approximate solutions.

Augment the state:

$$z(t) = \begin{bmatrix} x(t) \\ \theta \end{bmatrix} \qquad dz = \begin{bmatrix} f(x,\theta) \\ \theta \end{bmatrix} dt + \begin{bmatrix} 0 \\ \sigma \end{bmatrix} dw$$

For a given parameter, the state has a well defined evolution. If the parameter is a r.v. with a known distribution, how can we compute the state pdf?





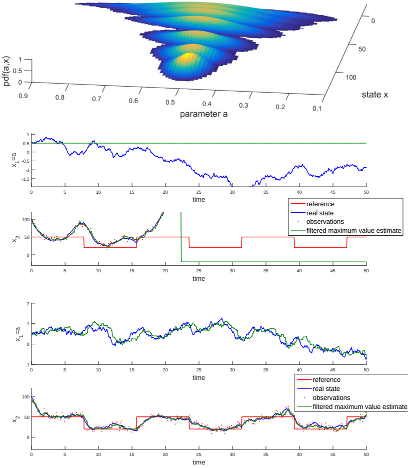
Suboptimal solution: Joint state-parameter estimation

$$dz_t = f(z_t)dt + \sigma dw_t$$

$p(z,t)$ satisfies the Fokker-Planck equation (scalar case for simplicity)

$$\frac{\partial p}{\partial t} = -f_z(z)p - f(z)\frac{\partial p}{\partial z} + \frac{\sigma^2}{2}\frac{\partial^2 p}{\partial z^2}$$

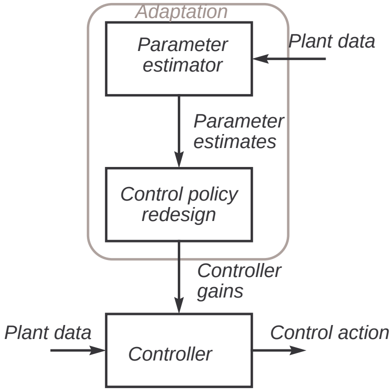
Example with an unknown gain.
Cautious adaptive control.



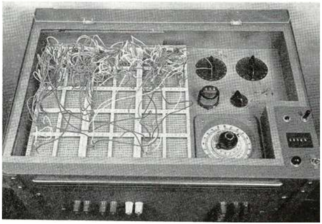
Joint work with António Silva

Certainty equivalence

Assume the estimated model to be the true model



Kalman, 1958 Self-optimizing controller



Åstrom and Wittenmark, 1972
Self-tuning controller

Issues with Certainty equivalence: Complex dynamics (1)

Plant dynamics (linear)

$$y(t) + a_1 y(t-1) + a_2 y(t-2) = Ku(t-1)$$

Controller

$$\theta(t) = \theta(t-1) + py(t-1)[y(t) - \theta(t-1)y(t-1) - \hat{K}u(t-1)],$$

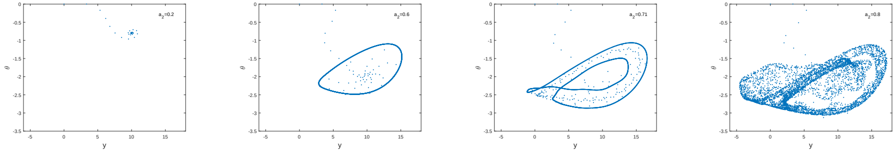
$$u(t) = (r - \theta(t)y(t))/\hat{K}$$

The plant is assumed to be 1^{st} order although it is of 2^{nd} order

Issues with Certainty equivalence: Complex dynamics (2)

With moderate un-modelled dynamics, the output converges to the reference.

Increase the level of un-modelled dynamics causes a sequence of bifurcations that leads to chaos



B. E. Ydstie (1986). Bifurcations and complex dynamics in adaptive control systems. *Proc. 25th CDC*, Athens, 2232 - 2236

B. E. Ydstie and M. P. Golden (1987). Chaos and strange attractors in adaptive control systems. *Proc. 10th IFAC World Congress*, Munich, 10: 127-132.

B. E, Ydstie (1991). Stability of the Direct Self-Tuning Regulator. *in* P. V. Kokotovic (ed.), *Foundations of Adaptive Control*,

Springer, 1992, 201-237.

Issues with Certainty equivalence: Equivocation

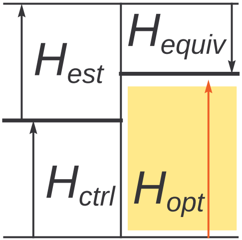
Maximum entropy approach to control, Saridis, 1988

Equivalence between optimal cost and entropy.

Describe the possible controls by a pdf p.

Maximize the entropy subject to

$$\int_{\Omega} p = 1, \quad \mathbb{E}(J(u)) = J(u^*)$$



Linear case: Separation theorem

Non-linear case: There is no separation theorem

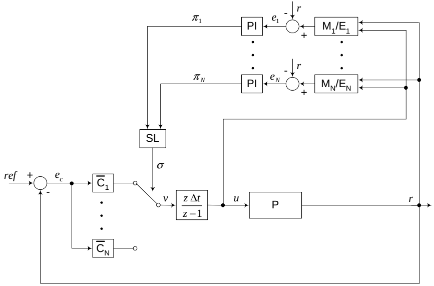
Use good adaptation and good control to reduce equivocation

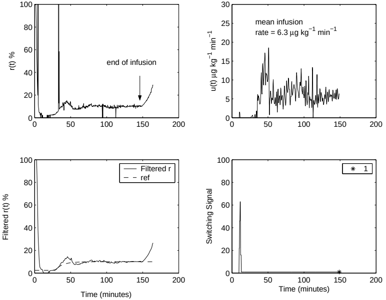
TÉCNICO LISBOA

SMMAC - Supervised Multiple Model Adaptive Control

Lainiotis 1974 Partitioning (lots of critics at the time)
Morse, Hespanha, Mosca, ... (1997 - present)

Clinical results for neuromuscular blockade





T. Mendonça, J. M. Lemos, H. Magalhães, P. Rocha and S. Esteves (2009). Drug delivery for neuromuscular blockade with supervised multimodel adaptive control. *IEEE Trans. Control Systems Technology*, 17(6):1237-1244.

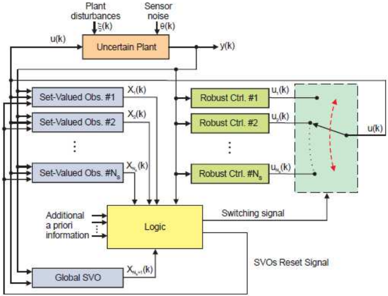
Model falsification

Based on Karl Popper falsification approach to Philosophy.

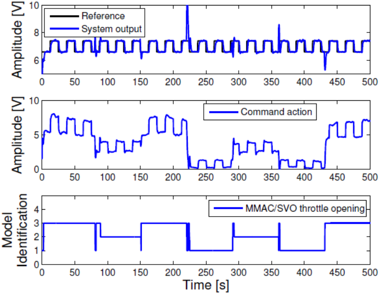
Carve the model bank by eliminating models incompatible with data.

Computationally very heavy.

Architecture based on Set-Value Observers

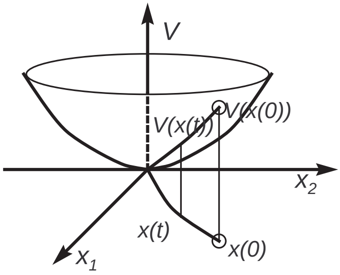


Experimental results – fan with varying flow



P. Rosa, T. Simão, C. Silvestre, J. M. Lemos (2016). Fault tolerant control of an air heating fan using set-value observers: an experimental evaluation. *Int. J. Adaptive Control and Signal Proc.*, 30(2):336-358

Control Lyapunov Functions (CLF)



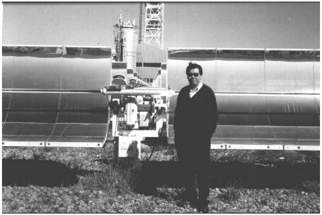
In adaptive control: Postulate a Lyapunov function for the hyperstate.
Choose the adaptation law such as to force the LF time derivative to be negative semi-definite.

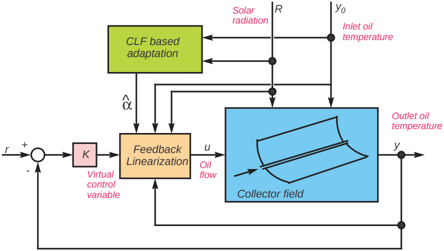
Convergence follows the set-invariant theorem.

Parks, 1966 and many others since then

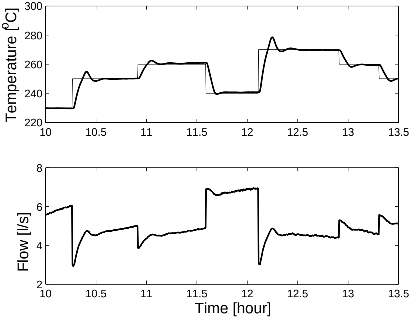
Alexander Lyapunov (1857-1918)
Lyapunov, 1892
Lasalle, 1950

Example: Lyapunov adaptation of a solar field



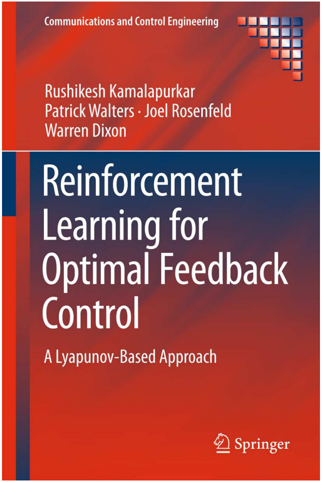


Experimental results



Barão, M., J. M. Lemos e R. N. Silva (2002). Reduced complexity adaptive nonlinear control of a distributed collector solar field. *J. Process Control*,12:131-141

Control Lyapunov Functions and Reinforcement Learning



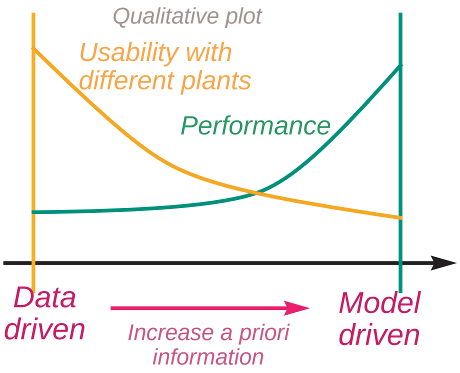
Control Lyapunov functions play a key role in control using reinforcement learning.

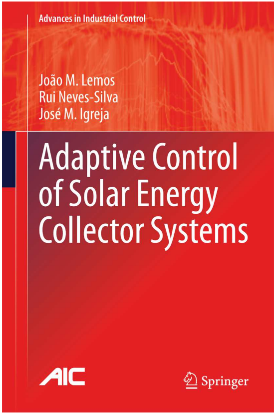
See the recent book (2018) and many papers on the subject.

The long term reward can be used to build Lyapunov functions.

A priori information versus performance

Increasing *a priori* information on plant dynamics increases performance but reduces the range of possible applications





Lemos, Neves Silva, Igreja *Adaptive Control of Solar Energy Collector Systems*, Springer, 2014

Reinforcement Learning

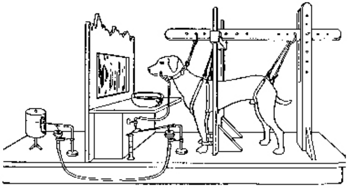
- Perception causes action
- Action influences perception
- Learn the optimal action by trial and error to maximize a reward
- Apply non-optimal actions with a low probability to learn by exploiting different regions of the state space

Exploitation and exploration

What is an adequate reward for control design?

How can exploitation be made in control?

Early roots: Pavlov's (1849-1936) experiments on reflex conditioning



Countless works since then.

Early RL based adaptive controllers

Whitaker, 1958 MIT rule

A gradient rule to maximize the instantaneous squared tracking error e of a Model Reference Adaptive Controller (MRAC) by adjusting a gain:

$$\frac{d\theta}{dt} = -\gamma e \frac{\partial e}{\partial \theta}$$

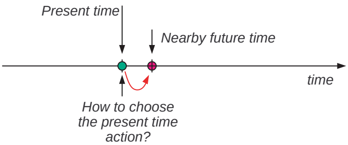
Due to technology limitations they used

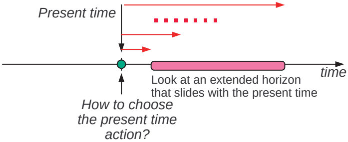
$$\frac{d\theta}{dt} = -\gamma e \, sign \left[\frac{\partial e}{\partial \theta} \right]$$

Crash of the X-15 aircraft in 15 Nov. 1967, that caused the death of the pilot Michael J. Adams.
A lot of enthusiasm, poor technology, and no theory at all.

The road to Predictive Adaptive Control (Adaptive MPC)

- Self-tuning regulator, Åstrom and Wittenmark, 1972, RLS + Minimum variance. Unable to stabilize non-minimum-phase plants
- Detuned Self-tuning regulator, Clarke and Gawthrop, 1974, Include a penalty on the action Unable to stabilize non-minimum-phase plants that are also unstable
- GPC, Clarke, Mohtadi and Tufts, 1980, Stabilizes any linear plant for a sufficiently large horizon

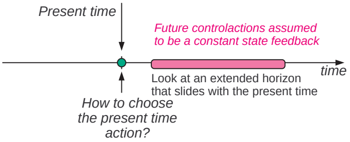


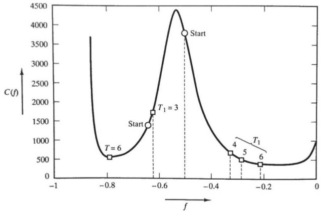


Key ideas

- Enlarge the horizon
- Receding horizon control

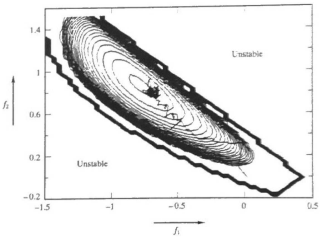
RL based linear adaptive MPC

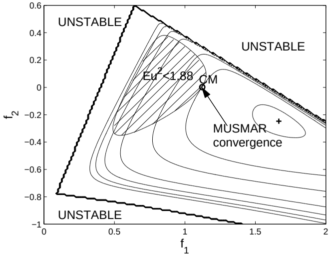




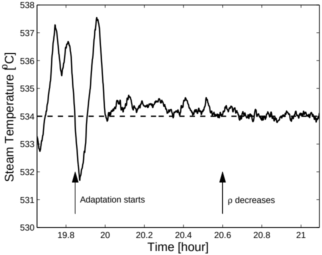
$$F_k = F_{k-1} - \gamma R_s^{-1} \nabla J$$

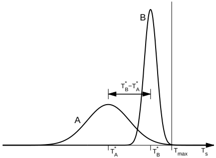
May start from a non-stabilizing gain.





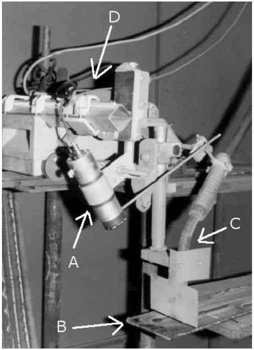
Example 1: Steam temperature control in a boiler

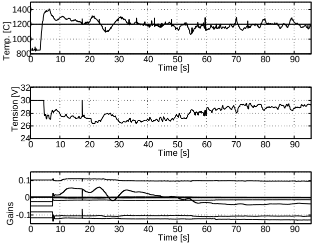




Silva, R. N., P. O. Shirley, J. M. Lemos and A. C. Gonçalves (2000). Adaptive regulation of super-heated steam temperature: a case study in an industrial boile. *Control Engineering Practice*, 8:1405-1415

Example 2: Rate of cooling in arc-welding





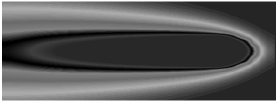
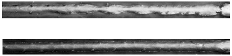


Plate with varying thickness.

Resulting seams, nonadaptive pole placement (above) and

adaptive MPC (below)



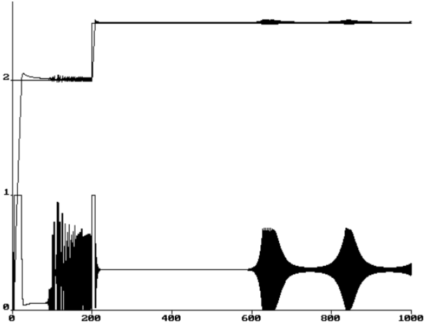
Santos, T. O., R. B. Caetano, J. M. Lemos and F. J. Coito (2000). Multipredictive Adaptive Control of Arc Eelding Trailing Centerline Temperature. *IEEE Trans. Control Systems Technology*, 8(1):159-169

Dual control and persistency of excitation

Duality: Learning implies exploitation and conflicts with optimal control.

Feldbaum, 1961

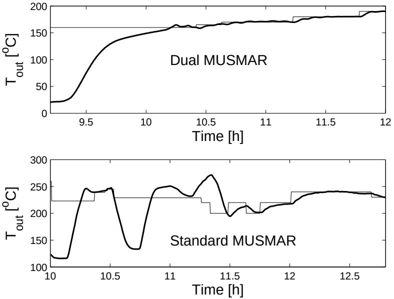
Optimal dual controller impossible to design, except in very simple cases. Need to resort to suboptimal dual strategies.



Dual adaptive MPC

Temperature control of solar field

Experimental results



Use a multicriterion approach to adjust the action, reaching a balance between persistency of excitation and good control performance.

Optimize the exploitation to improve learning.

Silva, R. N., N. Filatov, J. M. Lemos and H. Unbehauen (2005). A dual approach to start-up of an Adaptive Predictive Controller. *IEEE Trans. Control Systems Technology*, 13(6):877-883.

How to tackle adaptive nonlinear optimal control

Approximate Dynamic Programming

Computationally feasible approach to compute the long-term reward

Q-learning

Eliminate model knowledge assumptions

Recursive learning/estimation algorithms

Embed adaptation

Werbos, 1992

Sutton and Barto, 1998

Bertsekas, 1996

But much work and publications before.

See F. Lewis and D. Vrabble (2009), Reinforcement Learning and Adaptive Dynamic Programming for Feedback Control, *IEEE Circuits and Systems Mag.*, 9(3):32-50, for a tutorial on details.

Dynamic Programming

Bellman, 1957 (but actually since Jacob Bernouilli, XVII cent.)

Performance measure (infinite horizon)

$$V(h_h) = \sum_{i=k}^{\infty} \gamma^{i-k} r(x_i, u_i)$$

$$r(x_k, u_k) = Q(x_k) + u_k^T R u_k$$

Plant state model

$$x_{k+1} = f(x_k) + g(x_k) u_k$$

Control policy $u_k = h(x_k)$ Minimize the performance subject to the dynamics

Bellman's optimality principle

Hamilton-Jacobi-Bellman equation

$$V^*(x_k) = \min_{h(\cdot)} (r(x_k, h(x(k)) +$$

$$\gamma V^*(h_k + 1))$$

Optimal policy

$$h^*(x_k) = arg \min_{h(\cdot)} (r(x_k, h(x(k)) +$$

$$\gamma V^*(h_k + 1))$$

Policy iteration (PI)

Requires a stabilizing initial estimate of the control policy
Policy evaluation step

$$V_{j+1}(x_k) = r(x_k, h_j(h_k)) + \gamma V_{j+1}(x_{k+1})$$

Policy improvement step

$$h_{j+1}(x_k) = arg\min(r(x_k, h(x_k)) + \gamma V_{j+1}(x_{k+1}))$$

Corresponds to the difference Riccati equation in the LQ case.
Value iteration
At each time step do just a limited (e. g. 1) number of policy update.

Adaptive Dynamic Programming

Temporal Difference error

$$e_k = r(x_k, h(x_k)) + \gamma V_h(x_{k+1} - V_h(x_k)$$

Approximate the policy by $V_h(x) \approx W^T \phi(x)$ ϕ estimated from data.

On-line Policy iteration algorithm

Policy evaluation step (obtain W from RLS):

$$W_{j+1}^T(\phi(x(_k) - \gamma \phi x(k+1)) = r(x_k, h_j(x_k))$$

Policy improvement step

$$h_{j+1}(x_k) = arg \min_h (r(x_k, h(x_k)) + \gamma W_{j+1}^T \phi(x_{k+1}))$$

May start from a non-stabilizing policy.

Q-Learning

Q (quality) function

$$Q_h(x_k, u_k) = r(x_k, u_k) + \gamma V_h(x_{k+1})$$

u is the control action.
Assume a parametric approximatior of NN of the form

$$Q_h(x, u) = W^T \phi(x, u)$$

The optimal value for the action may be computed from

$$\frac{\partial}{\partial u} Q^*(x_k, u) = 0$$

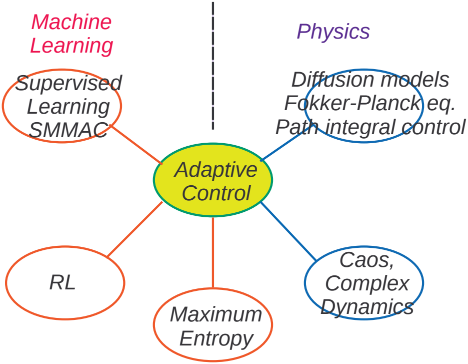
Does not require any derivatives involving model parameters.

Other problems and issues

- Difference and differential adaptive games (Soccer!)
- Distributed adaptive control
- Minimum attention and event-driven adaptive control
- Forgetting and adaptation
- Dynamic weights and robustness

TÉCNICO LISBOA

Conclusions



Adaptive control provides a meeting arena for machine learning and physics (as well as for mathematics!).

The cross breeding between RL, ADP and Q-Learning is boosting algorithms with increased performance for adaptive nonlinear optimal control.

A final word

Guy de Maussant (1850 - 1893): *Il fit une philosophie comme on faitun bon roman: tout parut vraisemblable, et rien ne fut vrais.*

He did a philosophy as one writes a good novel: everything looks plausible, but nothing is true

We can easily develop plausible algorithms for adaptive control based on "intuition", but that they actually do not work.

To avoid this pitfall, use the anchors provided by mathematical theories for stability, robustness, limits of performance.

Combining machine learning and model based methods is a far reaching ship, but the above anchors must be used to avoid shipwrecks.

TÉCNICO LISBOA

It is now time to stop and rest

Thank you for your attention