Empirical study of the sensitivity of CACLA to sub-optimal parameter setting in learning feedback controllers

Borja Fernandez-Gauna, Igor Ansoategui, Ismael Etxeberria-Agiriano, Manuel Graña

Computational Intelligence Group
University of the Basque Country (UPV/EHU)

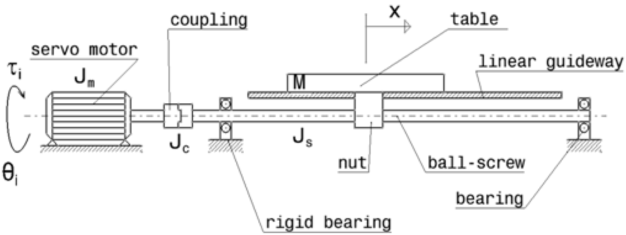
SOCO 2013, Salamanca

Outline

Problem statement

- Goal: design a feedback controller with minimal input from the designer
  - Typically, manufacturers employ some kind of Proportional Integrative Derivative (PID) controller
    - require manual tuning of parameters
  - Researchers have started using Reinforcement Learning (RL) as an alternative
    - require little input from the designer
    - CACLA is considered the state of the art

The ball screw feed drive model



$$\ddot{x} = \frac{\tau}{M \cdot \frac{p}{2 \cdot \Pi} + (J_c + J_s + J_m)\left(\frac{2\pi}{P}\right)}$$

Control goal

- The goal of the controller is to minimize the error $e_x(t)$ between the position of the table (x) and the setpoint $(w(t))$

$$e_x(t) = |x(t) - w(t)|$$

Research question

- How robust is CACLA to suboptimally learning tuned parameters?

Markov Decision Process

- General RL methods model environments as MDPs
  - S: set of states (discrete / continuous)
  - A: set of actions (discrete / continuous)
  - P: transition function defined by the model
  - R: reward signal to be maximized, defined by the system designer

Actor-Critic methods

- Two separate learning components are defined:
  - The actor: learns a policy $\pi_a(s)$
  - The critic: estimates the value $\hat{V}_t(s)$ of each state s:

$$\hat{V}_t(s) \simeq E^{\pi}\left\{\sum_{k=1}^{\infty} r_{t+k}\gamma^{k-1} \,|s_t = s\right\}$$

Actor-Critic methods

- Each time step
  - The actor observes the state s and selects an action following its policy $\pi_a(s)$
  - The critic observes the new state s', receives the reward r_t and updates its value estimate of s
  - The critic sends a critique δ_t to the actor, and the actor updates accordingly its policy $\pi_a(s)$

CACLA actor

- Instead of directly using the output of the policy $\pi_a(s)$, some disturbance signal $\eta(t)$ is added in order to explore unknown policies: $a_t = \pi_a(s) + \eta(t)$
- The update rule used by the actor is:

$$if\, \delta_t > 0: \qquad \pi_t^a(s_t) \leftarrow \pi_t^a(s_t) + \alpha_t \cdot (a_t - \pi_a(s_t))$$

- This means
  - the policy is only updated if an improvement is observed
  - the update is proportional to the distance in action space from the actually taken action a_t to the output of the policy $\pi_a(s)$

Critic

- We have used a standard $TD(\lambda)$ critic, which is similar to $TD(0)$:

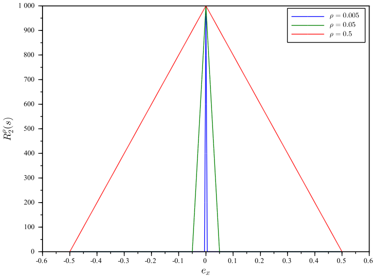
$$\hat{V}_t(s_t) \leftarrow \hat{V}_{t-1}(s_t) + \alpha_t \left(r_t + \gamma * \hat{V}_t(s_t) - \hat{V}_t(s_{t-1}) \right)$$

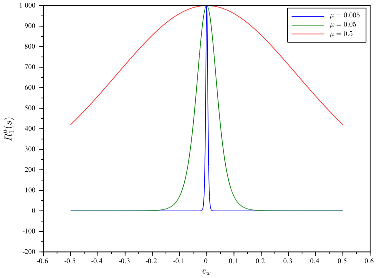
Experiments

- One experiment with each of the design parameters:
  - Experiment A: the reward signal
  - Experiment B: the number of features used to approximate the value function and policy (Gaussian RBF)
  - Experiment C: the learning gain α
- Performance measurement
  - Average absolute off-set error:

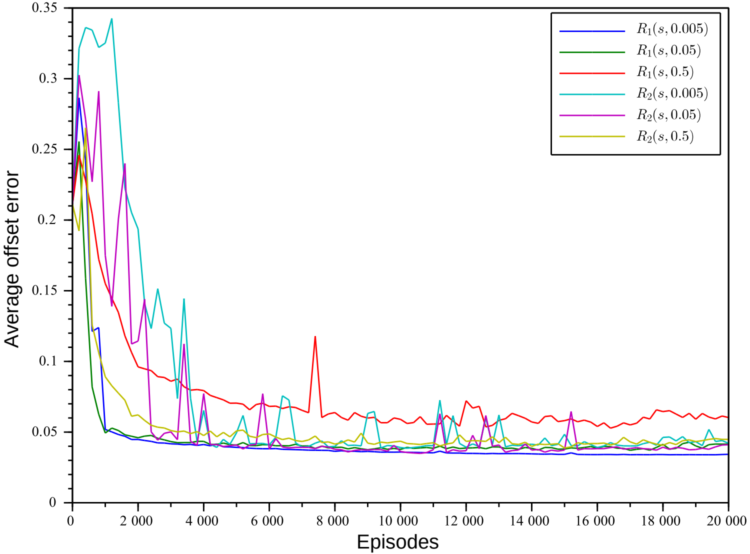
$$e_T(t) = \frac{1}{T}\sum_{t=0}^{T} e_x(t).$$

Experiment A: reward signals



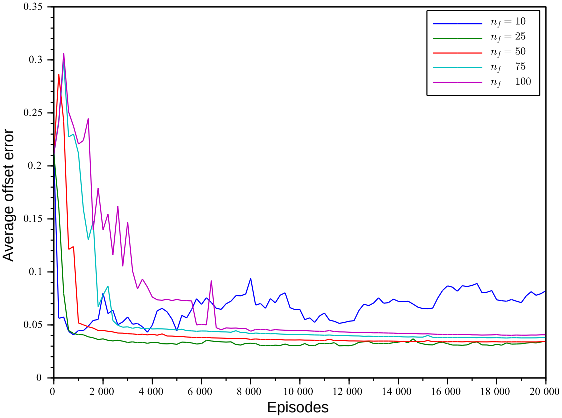


Experiment A: results



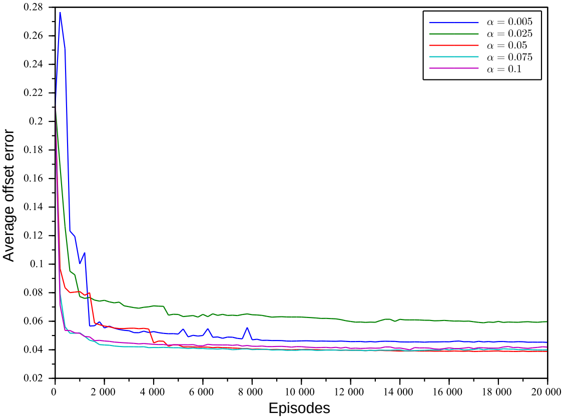
Experiment B: number of features

- Different number of features n_f to represent both the policy and the value function: $n_f = \{ 10, 25, 50, 75, 100 \}$



Experiment C: learning gain

- Different gains were tested: $\alpha = \{0.005, 0.025, 0.05, 0.075, 0.1\}$



Conclusions

- CACLA offers an interesting alternative to classic PID controllers in feedback control processes
  - minimal input required from the designer
  - robust behavior to suboptimal parameters

Thanks

Thank you very much for your attention.

- Contact:
  - Borja Fernández Gauna.
  - Computational Intelligence Group.
  - University of the Basque Country (UPV/EHU).
  - E-mail: **borja.fernandez@ehu.es**
  - Web page: **http://www.ehu.es/computationalintelligence**