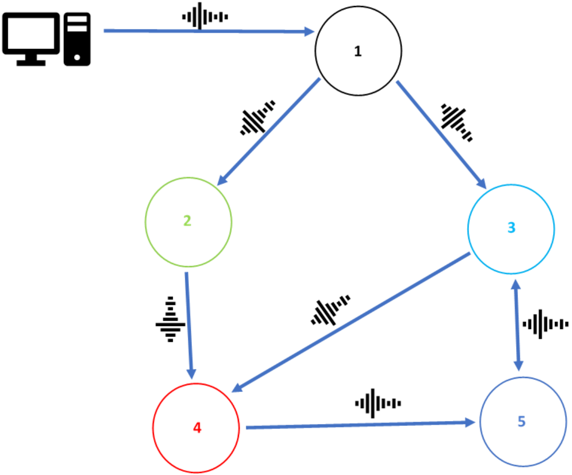
Learning near-optimal Broadcasting Intervals in Decentralized Multi-Agent Systems Using Online Least-Square Policy Iteration

Ivana Palunko

LARIAT - Laboratory for intelligent autonomous systems
University of Dubrovnik

Workshop on Control of Dynamical Systems
15. June, 2021. Dubrovnik



Link

Agent Dynamics

- consider N heterogeneous linear agents given by

$$\begin{aligned} \dot{\xi}_i &= A_i \xi_i + B_i u_i + \omega_i, \\ \zeta_i &= C_i \xi_i, \end{aligned} \tag{1}$$

where $\xi_i \in \mathbb{R}^{n_{\xi_i}}$ is the state, $u_i \in \mathbb{R}^{n_{u_i}}$ is the input, $\zeta_i \in \mathbb{R}^{n_\zeta}$ is the output of the i^{th} agent, $i \in \{1, 2, \ldots, N\}$, and $\omega_i \in \mathbb{R}^{n_{\xi_i}}$ reflects exogenous disturbances and/or modeling uncertainties

Agent Dynamics

- consider N heterogeneous linear agents given by

$$\begin{aligned} \dot{\xi}_i &= A_i \xi_i + B_i u_i + \omega_i, \\ \zeta_i &= C_i \xi_i, \end{aligned} \tag{1}$$

where $\xi_i \in \mathbb{R}^{n_{\xi_i}}$ is the state, $u_i \in \mathbb{R}^{n_{u_i}}$ is the input, $\zeta_i \in \mathbb{R}^{n_\zeta}$ is the output of the i^{th} agent, $i \in \{1, 2, \ldots, N\}$, and $\omega_i \in \mathbb{R}^{n_{\xi_i}}$ reflects exogenous disturbances and/or modeling uncertainties

- a common decentralized policy is

$$u_i(t) = -K_i \sum_{j \in \mathcal{N}_i} (\zeta_i(t) - \zeta_j(t)), \tag{2}$$

where K_i is an $n_{u_i} \times n_\zeta$ gain matrix

Closed-Loop Dynamics

- define $\xi := (\xi_1, \ldots, \xi_N)$, $\zeta := (\zeta_1, \ldots, \zeta_N)$ and $\omega := (\omega_1, \ldots, \omega_N)$
- utilizing the Laplacian matrix L of the communication graph $\mathcal{G}$, we reach

$$\begin{aligned}\dot{\xi}(t) &= A^{\mathrm{cl}}\xi(t) + A^{\mathrm{cld}}\xi(t-d) + \omega(t),\\ \zeta &= C^{\mathrm{cl}}\xi,\end{aligned}$$

with

$$\begin{aligned}A^{\mathrm{cl}} &= \mathrm{diag}(A_1, \ldots, A_N), & A^{\mathrm{cld}} &= [A_{ij}^{\mathrm{cld}}],\\ A_{ij}^{\mathrm{cld}} &= -l_{ij}B_iK_iC_j, & C^{\mathrm{cl}} &= \mathrm{diag}(C_1, \ldots, C_N),\end{aligned}$$

Optimal Intermittent Feedback

- $t_i^j \in \mathcal{T}, i \in \mathbb{N}$ – broadcasting instants of the j^{th} agent
- asynchronous communication
- $x_i := (\ldots, \zeta_i - \zeta_j, \ldots)$, where $i \in \{1, \ldots, N\}$ and $j \in \mathcal{N}_i$

Optimal Intermittent Feedback

- $t_i^j \in \mathcal{T}$, $i \in \mathbb{N}$ – broadcasting instants of the j^{th} agent
- asynchronous communication
- $x_i := (\ldots, \zeta_i - \zeta_j, \ldots)$, where $i \in \{1, \ldots, N\}$ and $j \in \mathcal{N}_i$

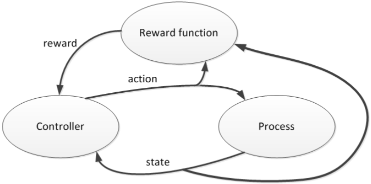
Problem

For each $j \in \{1, \ldots, N\}$, minimize the following cost function that captures performance vs. energy trade-offs

$$\mathop{\mathbb{E}}_{\omega}\left\{ \sum_{i=1}^{\infty} (\gamma_j)^i \underbrace{\left[\int_{t_{i-1}^j}^{t_i^j} (x_j^\top P_j x_j + u_j^\top R_j u_j)\mathrm{d}t + S_j \right]}_{r_j(x_j, u_j, \tau_i^j)} \right\} \tag{3}$$

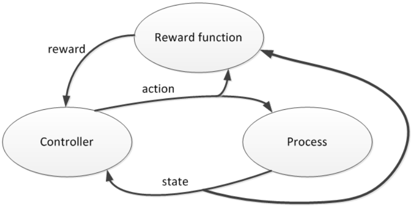
for the j^{th} agent of MAS (1)-(2) over all sampling policies τ_i^j and for all initial conditions $x_j(t_0) \in \mathbb{R}^{n_{x_j}}$.

The goal of RL is to solve a stochastic discrete-time optimal control problem



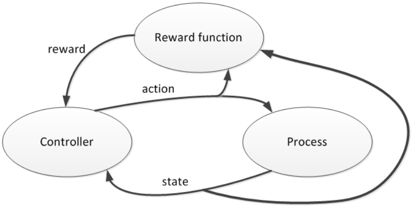
Markov decision process (MDP) $(\mathcal{X}, \mathcal{A}, f, \rho)$,

- $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ is the state space of the process,
- $\mathcal{A} \subseteq \mathbb{R}^{n_a}$ is the action space,
- $f : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, \infty)$ is the transition probability function of the process,
- $\rho : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ is the reward function



- A deterministic Markov Decision Process (MDP)

$$x_{k+1} = f(x_k, a_k)$$

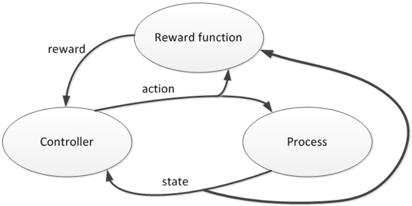


- A deterministic Markov Decision Process (MDP)

$$x_{k+1} = f(x_k, a_k)$$

- Reward function $\rho : X \times U \to \mathbb{R}$

$$r_{k+1} = \rho(x_k, a_k, x_{k+1})$$



- A deterministic Markov Decision Process (MDP)

$$x_{k+1} = f(x_k, a_k)$$

- Reward function $\rho : X \times U \to \mathbb{R}$

$$r_{k+1} = \rho(x_k, a_k, x_{k+1})$$

- The controller chooses actions according to its policy $h : X \to U$

$$a_k = h(x_k)$$

- The return R

$$R^h(x_0) = \mathbb{E}\left\{ \sum_{k=0}^{\infty} \gamma^k \rho(x_k, h(x_k), x_{k+1}) \right\}$$

where $\gamma \in (0,\ 1]$ is the discount factor

Any policy h^* that attains the minima in this equation is optimal

$$V^*(x_0) := \min_h V^h(x_0), \quad \forall x_0.$$

Q-learning

- Q-functions $Q : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$ fix the initial action.
- Once Q^* is available, an optimal (greedy) policy can be computed easily by selecting at each state an action with the smallest optimal Q^* value:

$$h^*(x) \in \arg\min_a Q^*(x, a).$$

- The state value functions can be expressed in terms of Q-functions

$$V^h(x) = Q^h(x, h(x)),$$
$$V^*(x) = \min_a Q^*(x, a) = Q^*(x, h^*(x)).$$

Q-iteration

Bellman equations

$$Q^h(x,a) = \mathbb{E}\Big\{\rho(x,a,x') + \gamma Q^h(x',h(x'))\Big\}, \tag{4}$$

$$Q^*(x,a) = \mathbb{E}\Big\{\rho(x,a,x') + \gamma \min_{a'} Q^*(x',a')\Big\}. \tag{5}$$

Policy

- The optimal policy (greedy policy in Q^*)

$$h(x) \in \arg\max_u Q^*(x, u)$$

Policy

- The optimal policy (greedy policy in Q^*)

$$h(x) \in \arg\max_{u} Q^*(x, u)$$

- Policy evaluation
  - at every iteration l solving the Bellman equation for Q^{h_l} of the current policy h_l

Policy

- The optimal policy (greedy policy in Q^*)

$$h(x) \in \arg\max_u Q^*(x, u)$$

- Policy evaluation
  - at every iteration l solving the Bellman equation for Q^{h_l} of the current policy h_l
- Policy improvement

$$h_{l+1}(x) \in \arg\max_u Q^{h_l}(x, u)$$

Approximation of Q

- In continuous spaces, policy evaluation cannot be solved exactly

Approximation of Q

- In continuous spaces, policy evaluation cannot be solved exactly
- Linearly parametrized Q-function approximator $\hat{Q}$
  - n basis function (BFs) $\phi_1, \dots, \phi_n : X \times U \to \mathbb{R}$
  - n dimensional parameter vector θ

$$\hat{Q} = \sum_{l=1}^{n} \phi_l(x, u)\theta_l = \phi^T(x, u)\theta$$

where $\phi(x, u) = [\phi_1(x, u), \dots, \phi_n(x, u)]^T$.

Approximation of Q

- In continuous spaces, policy evaluation cannot be solved exactly
- Linearly parametrized Q-function approximator $\hat{Q}$
  - n basis function (BFs) $\phi_1, \ldots, \phi_n : X \times U \to \mathbb{R}$
  - n dimensional parameter vector θ

$$\hat{Q} = \sum_{l=1}^{n} \phi_l(x, u)\theta_l = \phi^T(x, u)\theta$$

where $\phi(x, u) = [\phi_1(x, u), \ldots, \phi_n(x, u)]^T$.

- Control action u is scalar which is bounded to an interval $U = [u_L \quad u_H]$.

Approximation of Q

- In continuous spaces, policy evaluation cannot be solved exactly
- Linearly parametrized Q-function approximator $\hat{Q}$
  - n basis function (BFs) $\phi_1, \ldots, \phi_n : X \times U \rightarrow \mathbb{R}$
  - n dimensional parameter vector θ

$$\hat{Q} = \sum_{l=1}^{n} \phi_l(x,u)\theta_l = \phi^T(x,u)\theta$$

  where $\phi(x,u) = [\phi_1(x,u), \ldots, \phi_n(x,u)]^T$.
- Control action u is scalar which is bounded to an interval $U = [u_L \quad u_H]$.
- Chebyshev polynomials of the first kind

$$\begin{aligned} \psi_0(\bar{u}) &= 1, \\ \psi_1(\bar{u}) &= \bar{u}, \\ \psi_{j+1}(\bar{u}) &= 2\bar{u}\psi_j(\bar{u}) - \psi_{j-1}(\bar{u}), \end{aligned}$$

Least Square Policy Iteration (LSPI)

- define $\tau(t_i) := t_{i+1} - t_i$
- decision $\tau(t_i) \in \mathcal{A}$ is given by

$$\tau(t_i) = h_\kappa\big(x(t_i)\big),$$

where

$$h_\kappa\big(x(t_i)\big) = \begin{cases} \text{u.r.a.} \in \mathcal{A} & \text{every } \varepsilon \text{ iterations,} \\ h_\kappa\big(x(t_i)\big) & \text{otherwise,} \end{cases}$$

Least Square Policy Iteration (LSPI)

- define $\tau(t_i) := t_{i+1} - t_i$
- decision $\tau(t_i) \in \mathcal{A}$ is given by

$$\tau(t_i) = h_\kappa(x(t_i)),$$

where

$$h_\kappa(x(t_i)) = \begin{cases} \text{u.r.a.} \in \mathcal{A} & \text{every } \varepsilon \text{ iterations,} \\ h_\kappa(x(t_i)) & \text{otherwise,} \end{cases}$$

where "u.r.a." stands for "uniformly chosen random action" and yields exploration every ε steps while $h_\kappa(x(t_i))$ is the policy obtained according to

$$h_\kappa(x(t_i)) \in \arg\max_u \hat{Q}(x(t_i), \tau(t_i)) \tag{6}$$

Least Square Policy Iteration (LSPI)

- α_κ is updated every $\kappa \geq 1$ steps from the projected Bellman equation for model-free policy iteration

$$\Gamma_i \alpha_\kappa = \gamma \Lambda_i \alpha_\kappa + z_i,$$

where γ is from (3) and

$$\begin{aligned}
\Gamma_0 &= \beta_\Gamma I, \quad \Lambda_0 = \mathbf{0}, \quad z_0 = \mathbf{0}, \\
\Gamma_i &= \Gamma_{i-1} + \phi\big(x(t_i), \tau(t_i)\big)\phi\big(x(t_{i-1}), \tau(t_{i-1})\big)^\top, \\
\Lambda_i &= \Lambda_{i-1} + \phi\big(x(t_i), \tau(t_i)\big)\phi\big(x(t_i), h(x(t_{i+1}))\big)^\top, \\
z_i &= z_{i-1} + \phi\big(x(t_i), \tau(t_i)\big) r(t_i),
\end{aligned}$$

where Γ_i, Λ_i and z_i are updated at every iteration step i

Least Square Policy Iteration (LSPI)

- α_κ is updated every $\kappa \geq 1$ steps from the projected Bellman equation for model-free policy iteration

$$\Gamma_i \alpha_\kappa = \gamma \Lambda_i \alpha_\kappa + z_i,$$

where γ is from (3) and

$$\Gamma_0 = \beta_\Gamma I, \quad \Lambda_0 = \mathbf{0}, \quad z_0 = \mathbf{0},$$
$$\Gamma_i = \Gamma_{i-1} + \phi\big(x(t_i), \tau(t_i)\big)\phi\big(x(t_{i-1}), \tau(t_{i-1})\big)^\top,$$
$$\Lambda_i = \Lambda_{i-1} + \phi\big(x(t_i), \tau(t_i)\big)\phi\big(x(t_i), h(x(t_{i+1}))\big)^\top,$$
$$z_i = z_{i-1} + \phi\big(x(t_i), \tau(t_i)\big)r(t_i),$$

where Γ_i, Λ_i and z_i are updated at every iteration step i

- new α_κ improves the Q-function
- improved policies (in the sense of Problem) are obtained from (6)

Bellman equations (4) and (5) can be written as

$$Q^h = T^h(Q^h), \qquad Q^* = T(Q^*).$$

Contraction

Mapping T, as well as T^h, is a contraction with factor $\gamma < 1$ in L_∞-norm

$$\|T(Q) - T(Q')\|_\infty \leq \gamma \|Q - Q'\|_\infty .$$

T has the unique fixed point Q^*.

Q-iteration

An arbitrary initial Q-function Q_0 can be iterated to reach Q^*:

$$Q_{l+1} = T(Q_l),$$

which is known as the Q-iteration.

Contraction

$$\|Q_{l+1} - Q^*\|_\infty \leq \gamma \|Q_l - Q^*\|_\infty$$

Estimate of state-action value function in policy iteration

$$Q^{h_k}(x, a) = \mathbb{E}\left\{\rho(x, a, x') + \gamma Q^{h_k}(x', h_k(x'))\right\}.$$

Approximated policy iteration converts to

$$\alpha_{l+1} = (P \circ T \circ F)(\alpha_l),$$

where $F(\alpha)$ equals the right-hand side of

$$\hat{Q}(x(t_i), \tau(t_i)) = \Phi^\top\big(x(t_i), \tau(t_i)\big)\alpha_\kappa,$$

while the projection $P(Q)$ equals ΦQ when orthonormal bases (e.g., Chebyshev polynomials) are employed.

The expansiveness coefficient of $P \circ T \circ F$ in

$$\alpha_{l+1} = (P \circ T \circ F)(\alpha_l),$$

is upper bounded by

$$\mathrm{E} := \gamma \sqrt{2}^{n_x + n_a} M_b^{n_x} N_b^{n_a}.$$

Theorem

If $\mathrm{E} < 1$, than the composite mapping $P \circ T \circ F$ built upon Chebyshev polynomials is a contraction, that is,

$$\alpha_{l+1} = (P \circ T \circ F)(\alpha_l),$$

converges to a unique fixed point.

Near-Optimality Bounds

The approximate policy evaluation is accurate to within δ in the $\mathcal{L}_\infty$ sense, that is, if

$$\|\hat{Q}^{h_k} - Q^{h_k}\|_\infty \leq \delta, \qquad \forall k \in \{1, 2, \ldots\},$$

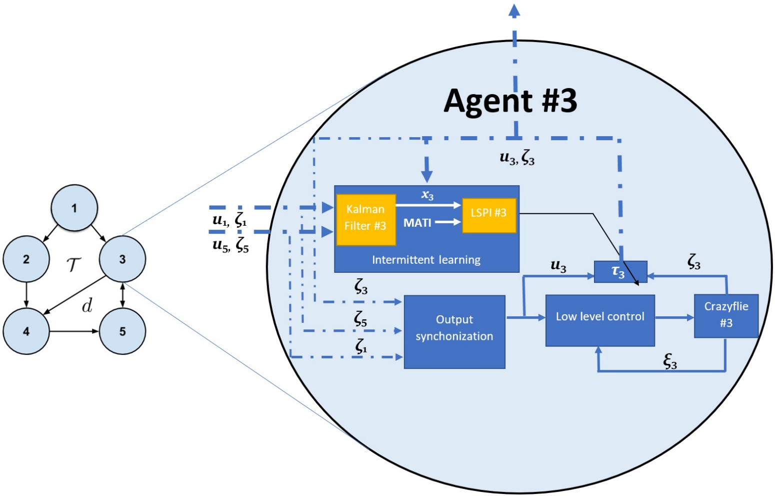
then in the limit as $k \to \infty$ the following near-optimality bound holds

$$\limsup_{k\to\infty} \|\hat{Q}^{h_k} - Q^*\|_\infty \leq \frac{2\gamma\delta}{(1-\gamma)^2}.$$

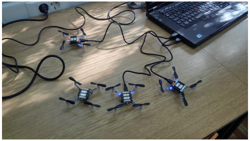
Moreover, if the sequence of obtained policies converges to some $\tilde{h}$, then the following tighter bound holds:

$$\|\hat{Q}^{\tilde{h}} - Q^*\|_\infty \leq \frac{2\gamma\delta}{1-\gamma}.$$

Agent Interconnections



Crazyflie model identification

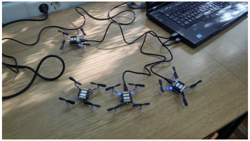


- Transfer function form

$$\frac{X(s)}{\Phi(s)} = \frac{K_T}{s(T_T s + 1)} e^{-T_d s}$$

where $K_T = 0.944$, $T_T = 0.297$ and $T_d = 0.45$

Crazyflie model identification



- Transfer function form

$$\frac{X(s)}{\Phi(s)} = \frac{K_T}{s(T_T s + 1)} e^{-T_d s}$$

where $K_T = 0.944$, $T_T = 0.297$ and $T_d = 0.45$

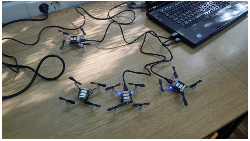
- State-space form

$$\dot{\xi}(t) = A\xi(t) + Bu(t) + \omega$$

$$\begin{bmatrix} \dot{x}(t) \\ \ddot{x}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -T_s \end{bmatrix} \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ K_s \end{bmatrix} \phi(t) + \omega,$$

where $K_s = 3.17$ and $T_s = 3.37$

Crazyflie model identification



- Transfer function form

$$\frac{X(s)}{\Phi(s)} = \frac{K_T}{s(T_T s + 1)} e^{-T_d s}$$

where $K_T = 0.944$, $T_T = 0.297$ and $T_d = 0.45$

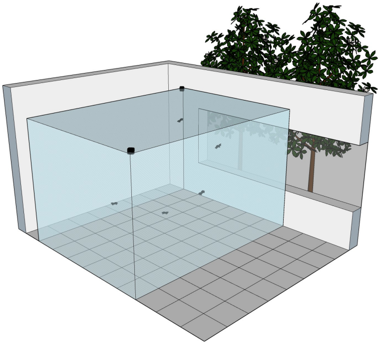
- State-space form

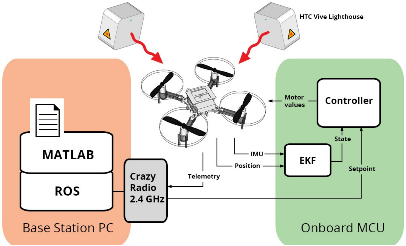
$$\dot{\xi}(t) = A\xi(t) + Bu(t) + \omega$$

$$\begin{bmatrix} \dot{x}(t) \\ \ddot{x}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -T_s \end{bmatrix} \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ K_s \end{bmatrix} \phi(t) + \omega,$$

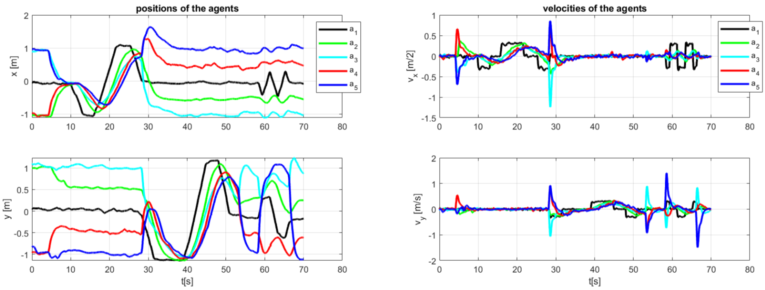
where $K_s = 3.17$ and $T_s = 3.37$

- Communication delay is $d = 0.45$ s

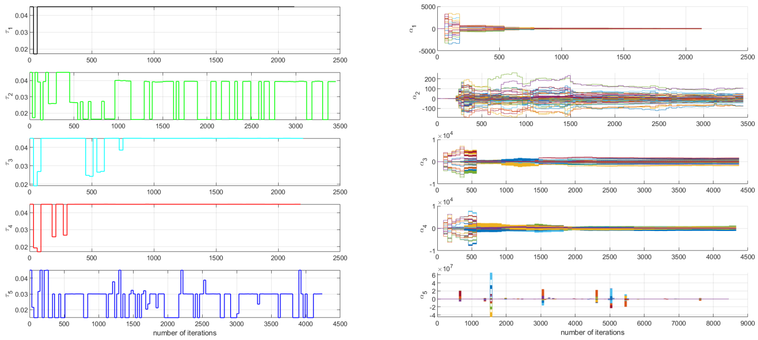



HTC Vive Lighthouse
Motor values
Controller
State
IMU
EKF
Position
Setpoint
Telemetry
MATLAB
ROS
Crazy Radio 2.4 GHz
Base Station PC
Onboard MCU

MAS with Crazyflie - Experimental validation



MAS with Crazyflie - Experimental validation



- *Domagoj Tolić, Ivana Palunko, Learning Suboptimal Broadcasting Intervals in Multi-Agent Systems, IFAC-PapersOnLine, Volume 50, Issue 1, 2017, Pages 4144-4149*
- *Lucian Busoniu, Tim de Bruin, Domagoj Tolić, Jens Kober, Ivana Palunko, Reinforcement learning for control: Performance, stability, and deep approximators, Annual Reviews in Control, Volume 46, 2018, Pages 8-28*
- *Palunko, I, Tolić, D, Prkačin, V. Learning near-optimal broadcasting intervals in decentralized multi-agent systems using online least-square policy iteration. IET Control Theory Appl. 2021; 15: 1054– 1067*

Thank you for your attention!
Questions?!

ConDyS

Croatian Science Foundation

DATACROSS
Advanced methods and technologies for data science and cooperative systems

ACROSS
Centre of Research Excellence for Data Science and Cooperative Systems
Research unit for Advanced Cooperative systems (ACROSS)