Zurich University
of Applied Sciences

School of
Engineering

IAMP Institute of Applied
Mathematics and Physics

Artificial Intelligence for HVAC Systems

V. Ziebart, ACSS IAMP ZHAW

Motivation

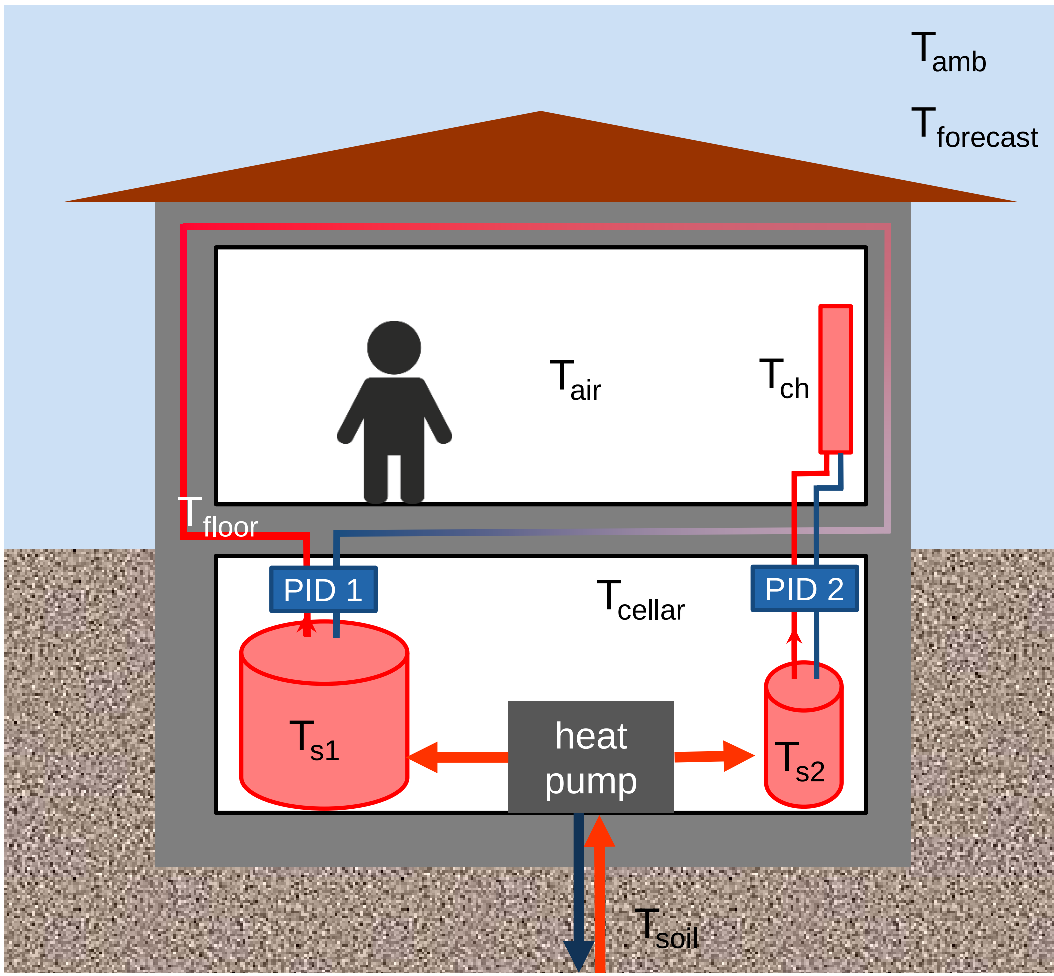
Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics

- More than 80% of the energy consumed in Swiss households is used for room heating & cooling and domestic hot water.*
- The increasing combined use of various energy conversion and storage technologies (PT, solar thermal collectors, heat pumps, combustion, batteries, hot water storage, ice storage systems) requires intelligent and optimal control systems.
- Reinforcement Learning (RL) showed promising performance in different fields (AlphaZero, computer games).
- RL is a data-driven approach.
- Can RL be used as control method for HVAC**-Systems? (optimality, learning behavior, robustness,…

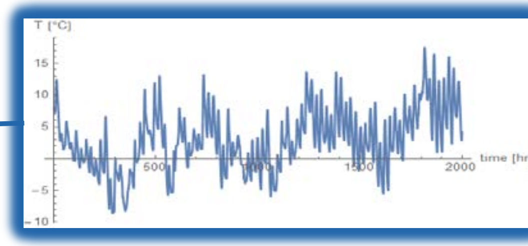
* Energieverbrauch in der Schweiz und weltweit, EnergieSchweiz, Bundesamt für Energie BFE Dienst Aus-und Weiterbildung, Juli 2015
** HVAC: heating, ventilation, air conditioning

Model of Building and Heating System

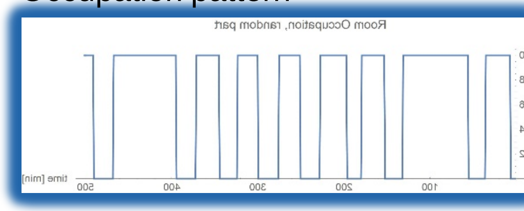
Zurich University of Applied Sciences
School of Engineering
IAMP Institute of Applied Mathematics and Physics



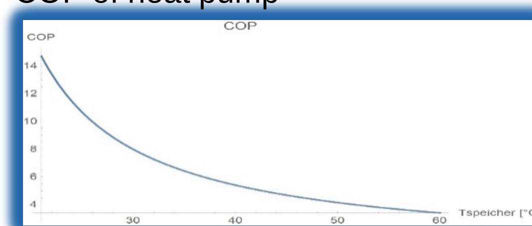
Weather data of 11 Swiss cities



Occupation pattern



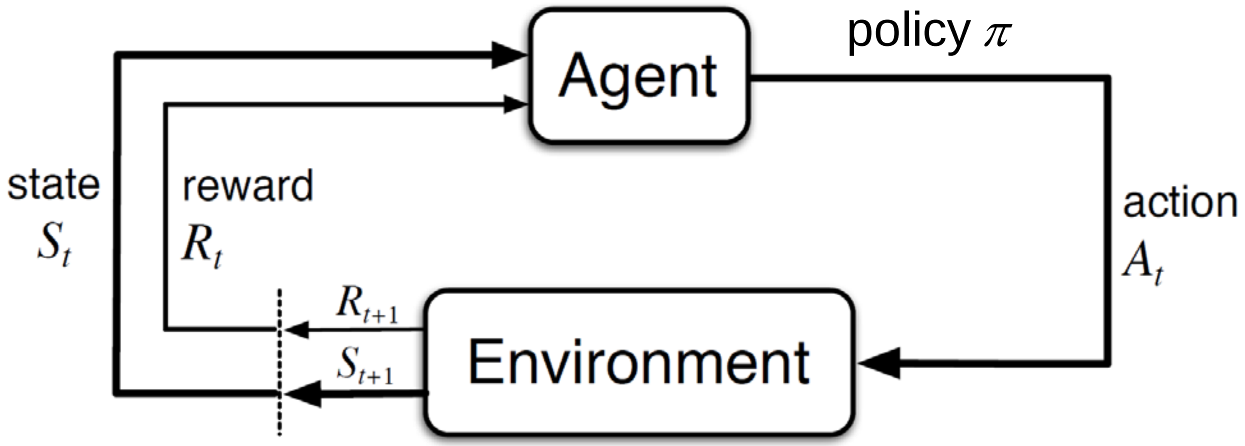
COP of heat pump



Enery price

night rate	0.5
date rate	1

Reinforcement Learning Control



$$\textit{value function}: \quad q_\pi(s,a) = \mathrm{E}_\pi\left[G_t \middle| S_t = s, A_t = a\right]$$

$$\textit{return}: \quad G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$$\textit{discount rate}: \quad \gamma \in [0,1]$$

Schematic from: Richard S. Sutton and Andrew G. Barto, An introduction to reinforcement learning, 2018

Q-Learning and SARSA

Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics

Q-learning:

$$q(s_t,a_t) \leftarrow q(s_t,a_t) + \underset{\text{learning rate}}{\alpha}\left(\overbrace{R_t + \gamma \max_a q(s_{t+1},a)}^{\text{new estimate of return G}} - q(s_t,a_t)\right)$$

SARSA:

$$q(s_t,a_t) \leftarrow q(s_t,a_t) + \alpha\left(R_t + \underbrace{\gamma q(s_{t+1},a_{t+1}) - q(s_t,a_t)}_{\text{difference of «large» numbers!}}\right)$$

These *interacting, interdependent, interative* jobs must be done:

1. learn *q* for actions performed in state visited
2. approximate a generalized *q-function* on state and action space (*SxA*) (typically with gradient descent)
3. improve control performance by e.g. (ε-)greedy policy

→ **potentially unstable**

Improvement of Stability

Zurich University
of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied
Mathematics and Physics

- Ad 1: More efficient learning of *q* through n-step SARSA

$$G_{t,n} = \underbrace{\sum_{k=0}^{n-1} \gamma^k R_{t+k}}_{\text{truncated sum of } n \text{ rewards}} + \gamma^n q(s_{t+n}, a_{t+n})$$

instead of $G_{t,1} = R_t + \gamma\, q(s_{t+1}, a_{t+1})$.

Average fraction return: $$\frac{\sum_{k=0}^{n-1} \gamma^k \overline{R}}{\sum_{k=0}^{\infty} \gamma^k \overline{R}} = \frac{\frac{1-\gamma^n}{1-\gamma}}{\frac{1}{1-\gamma}} = 1-\gamma^n$$

- Ad 2: Least-Squares fit to $\{s_t, a_t, G_{t,n}\}$ with polynomials and trigonometric functions to find $q(s,a)$ on SxA instead of iterative gradient descent methods.

State, Action, Reward

- **State** variables for RLC (continous and discrete)
  - **6 temperatures**, $\mathbb{R}^6$ (T_{amb}, T_{air}, T_{floor}, $T_{storage1}$, $T_{storage2}$, $T_{forecast}$)
  - **time**, real interval [0,24[
  - **room occupancy**, boolean
- **Action** variables (discrete, 12 combinations)
  - **heat pump off/loading storage 1 or 2**, $\in \{0,1,2\}$
  - **PID floor heating on/off**, boolean
  - **PID convection heater on/off**, boolean
- **Reward** (≤ 0):
  - **energy costs** (negative, night and day rate)
  - **temperature deviation** from setpoint, (negative, proportional to ΔT, only if house is occupied)

Approximation of State-Action Value Function q(s,a)

Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics

- Partitioning of state-action space in continous (c) and discrete (d) subspaces:

$$S \times A = (S_c \times S_d) \times (A_c \times A_d) = \underbrace{S_c \times A_c}_{\text{continous}} \times \underbrace{S_d \times A_d}_{\text{discrete}}$$

- For each set of discrete variables in $S_d \times A_d$ (24 *configurations*) $q(s,a)$ is approximated in the continous subspace $S_c \times A_c$ by a linear combination of polynomials (temperatures) and trigonometric functions (time):

$$q_k(s,a) = \sum_{j,l} \tilde{c}_{kjl} \prod_{i=1}^{6} T_i^{n(k,i,j)}\, trig_l(t) \qquad k = 1,\ldots,24$$

rewritten as flattened vector product

$$q_k(s,a) = \sum_j c_{kj}\, p_j(T_1,\ldots,T_6,t) = \boldsymbol{c}_k \cdot \boldsymbol{p} \qquad k = 1,\ldots,24$$

typically: $j = 252 \;\rightarrow\; k{\cdot}j = 6'048$ coefficients c_{kj}

Analytical Solution

Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics

older experience is discounted by γ_Q

$$\boldsymbol{c}_k = \arg\min \sum_{i=1}^{t} \gamma_Q^{t-i} \left\{ \left(\sum_{j=0}^{n-1} \gamma^j R_{s(i,k)+j} + \gamma^n q\left(s_{s(i,k)+n}, a_{s(i,k)+n}\right) \right) - \boldsymbol{c}_k \boldsymbol{p}_{s(i,k)} \right\}^2, \quad k = 1,\ldots,24$$

s(i,k) is the index when $(s_d,a_d)_i=(s_d,a_d)_k$ the *i*-th time

rearranging yields:

$$\boldsymbol{Q}_k^{(i)} = \gamma_Q \boldsymbol{Q}_k^{(i-1)} + \left\{ \boldsymbol{p}_{s(i,k),\,r}\ \boldsymbol{p}_{s(i,k),\,l} \right\}_{rl}$$

$$\boldsymbol{b}_k^{(i)} = \gamma_Q \boldsymbol{b}_k^{(i-1)} - 2\boldsymbol{p}_{s(i,k)} \left(\sum_{j=0}^{n-1} \gamma^j R_{s(i,k)+j} + \gamma^n q\left(s_{s(i,k)+n}, a_{s(i,k)+n}\right) \right)$$

$$\Rightarrow \quad \boldsymbol{c}_k = 0.5 \cdot \left(\boldsymbol{Q}_k + \varepsilon_k \boldsymbol{I}\right)^{-1} \boldsymbol{b}_k$$

for nummerical reasons

Decision Making, Reward Normalization

Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics

- probability of choosing a_i:

$$\pi\left(a_i \middle| s\right) = \frac{e^{q(s,a_i)/\tau}}{\sum_{j=1}^{n} e^{q(s,a_j)/\tau}}$$

$$\tau \to \infty : \pi\left(a_i \middle| s\right) = \pi\left(a_j \middle| s\right)$$
$$\tau \to 0 : \pi\left(\arg\max q(s,a_i) \middle| s\right) = 1$$
$$q(s,a_i) \le 0!$$

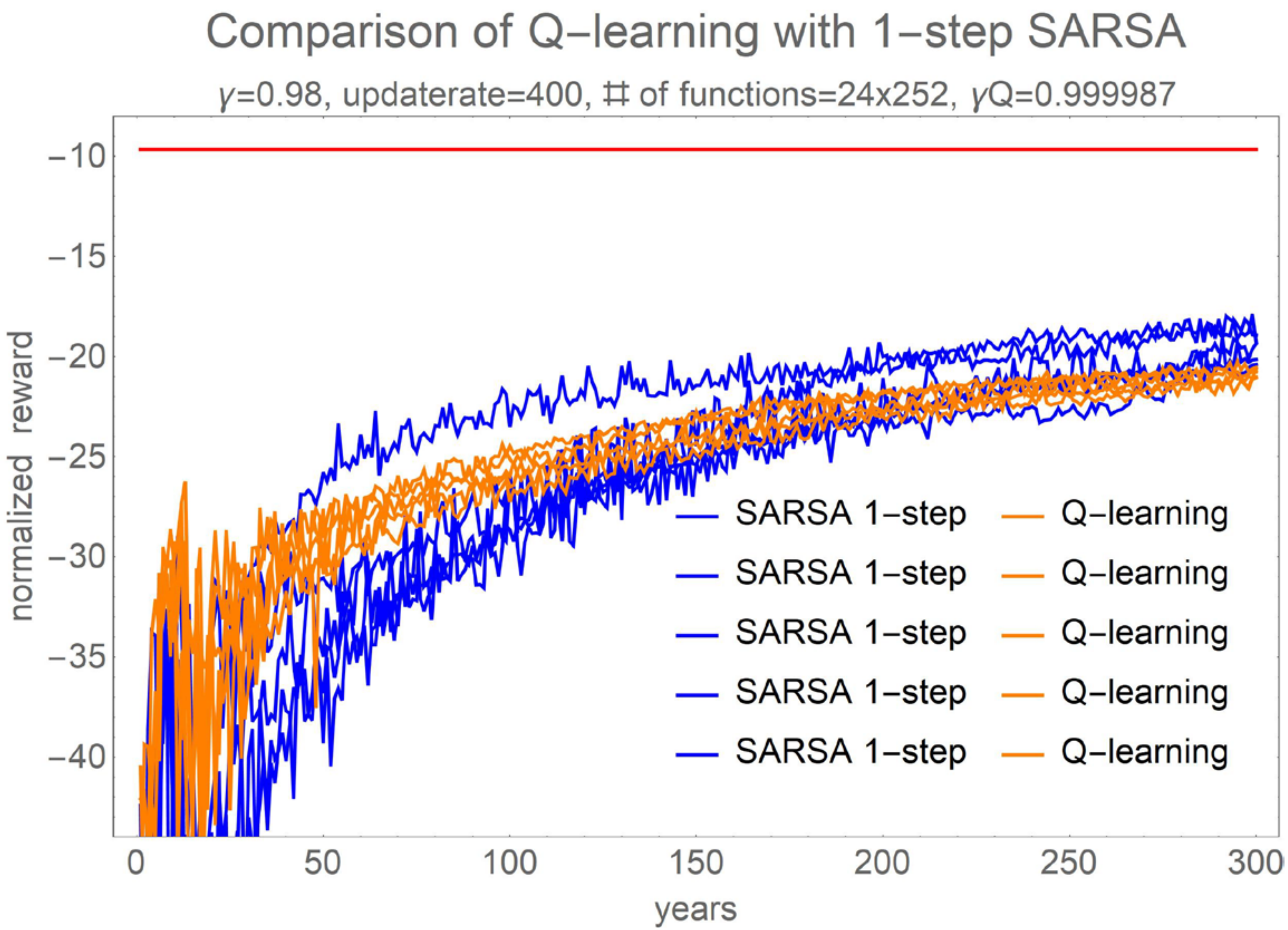
- normalization of reward per year (plots only!) by difference of average outside temperature and room set point temperature

$$R_{year,norm} = \frac{R_{year}}{\Delta T + 0.005 \Delta T^2} \quad with\, \Delta T = T_{sp} - \bar{T}_{amb}$$

Hardware & Software

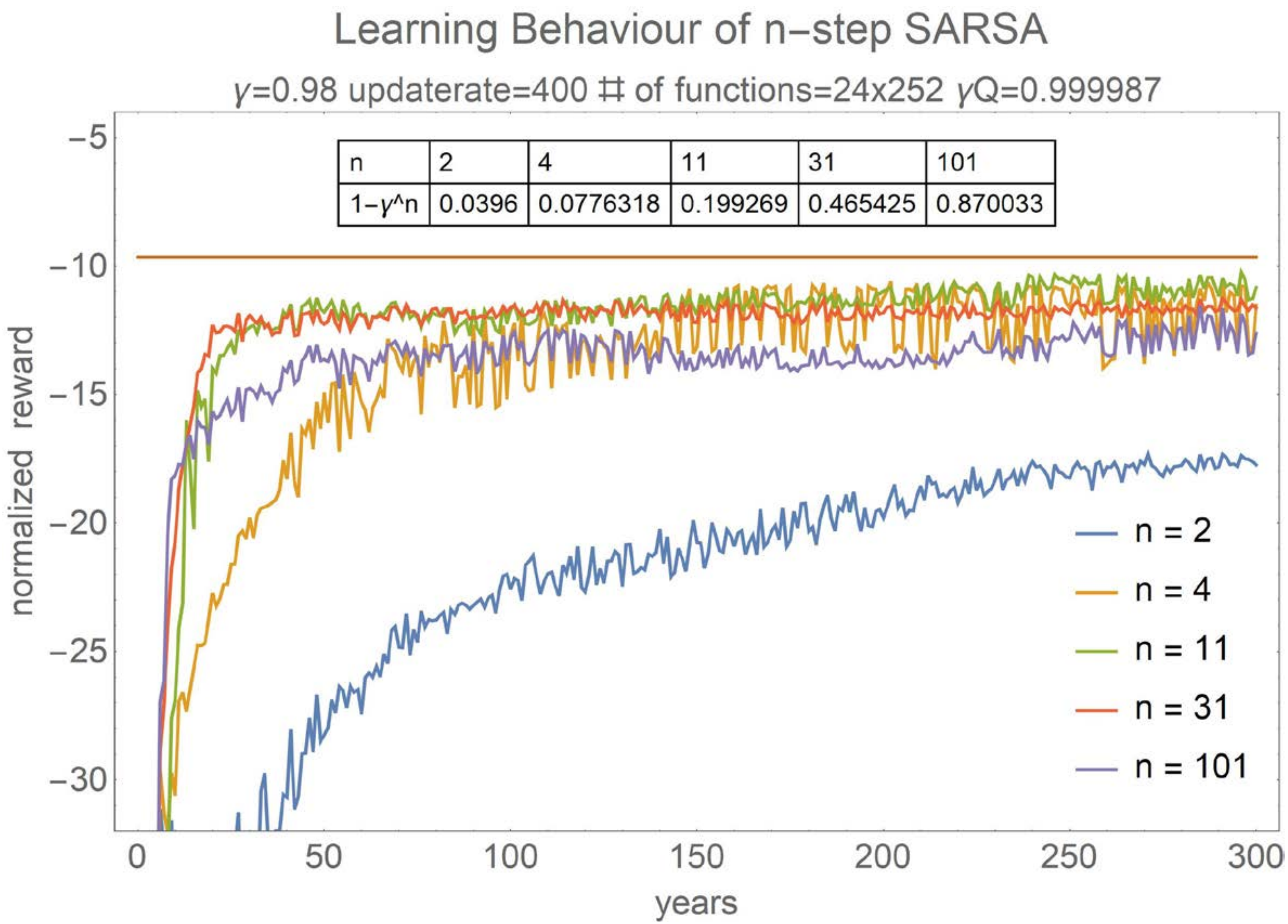
- 2 Servers with 2 CPUs Intel Xeon Platinum 8164 each
- Each CPU with 26 cores
- 768 GB RAM
- Code written in Mathematica
- Computation time for 1 year simulation: 1min

SARSA vs Q-Learning

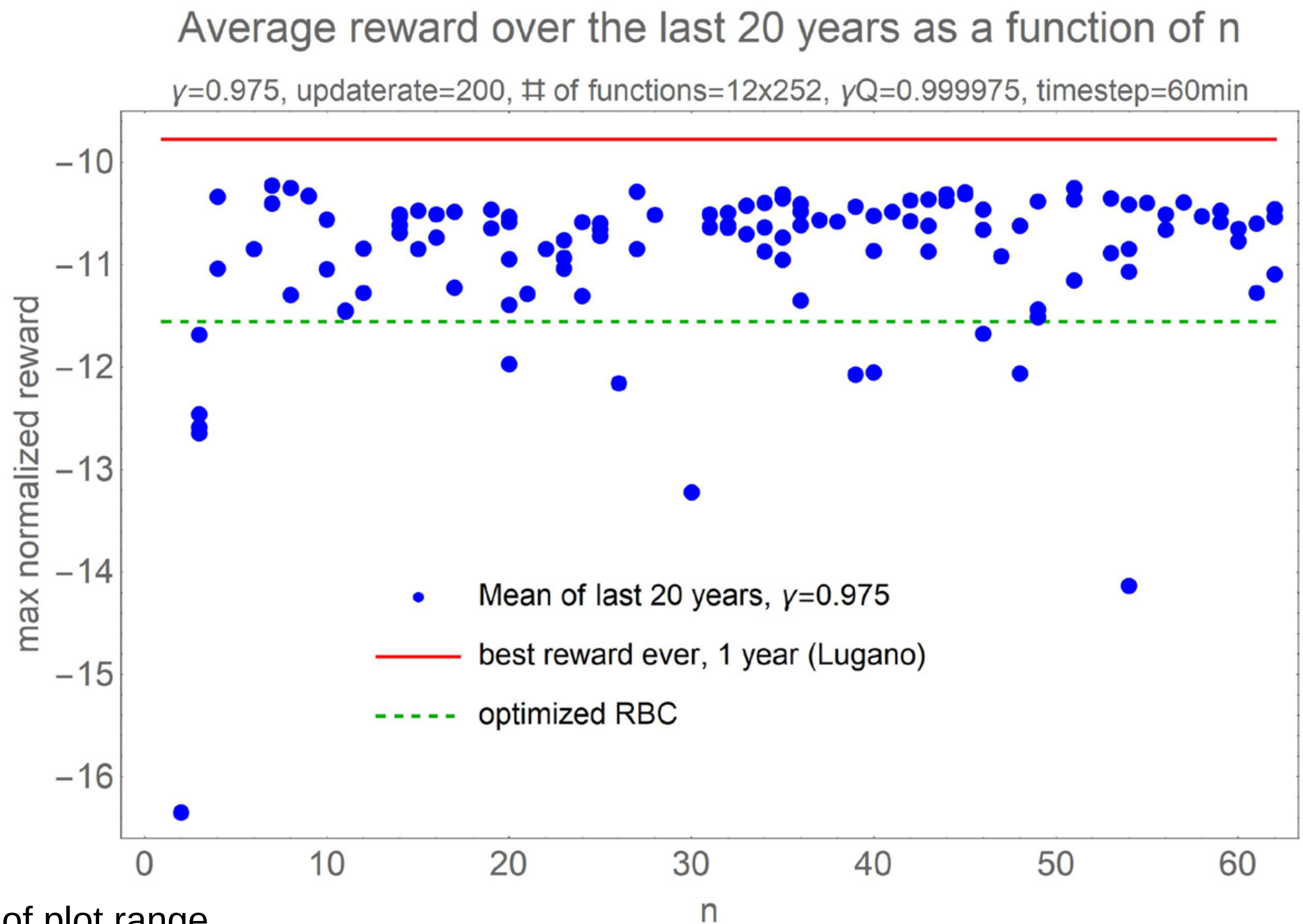


Effect of n in n-step SARSA

Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics



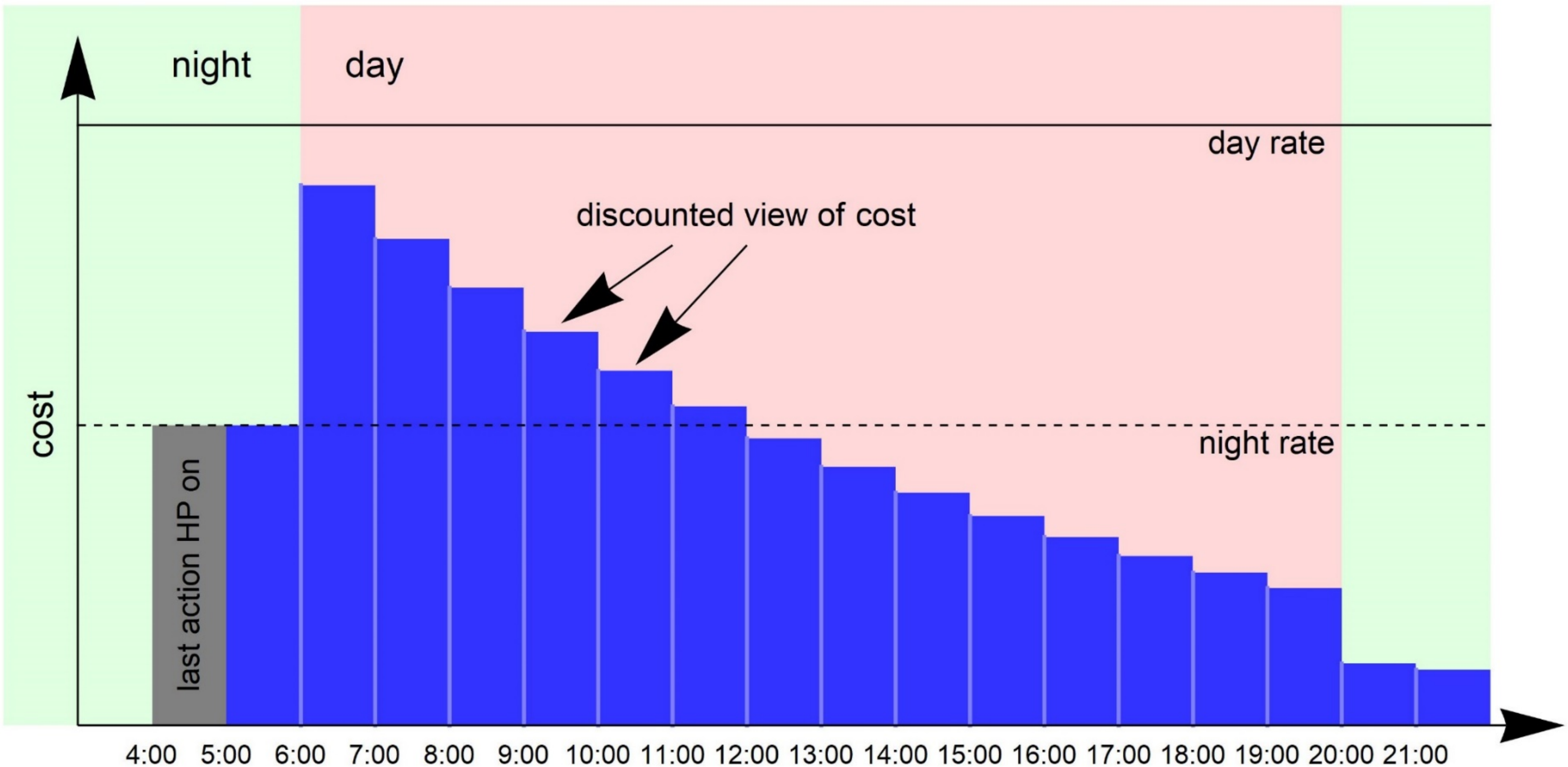
Effect of *n* in *n*-step SARSA



n=1 out of plot range

Influence of Discount Factor γ

Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics



Actual time 5:00. One more heating hour before 20:00 is needed. What hour seems to be the best choice? (If γ>0.952: heating at 5:00, otherwise heating at 19:00)

$\rightarrow$ especially too low γ's lead to suboptimal decisions and unrealistic *q*-values

n-step SARSA without Discount

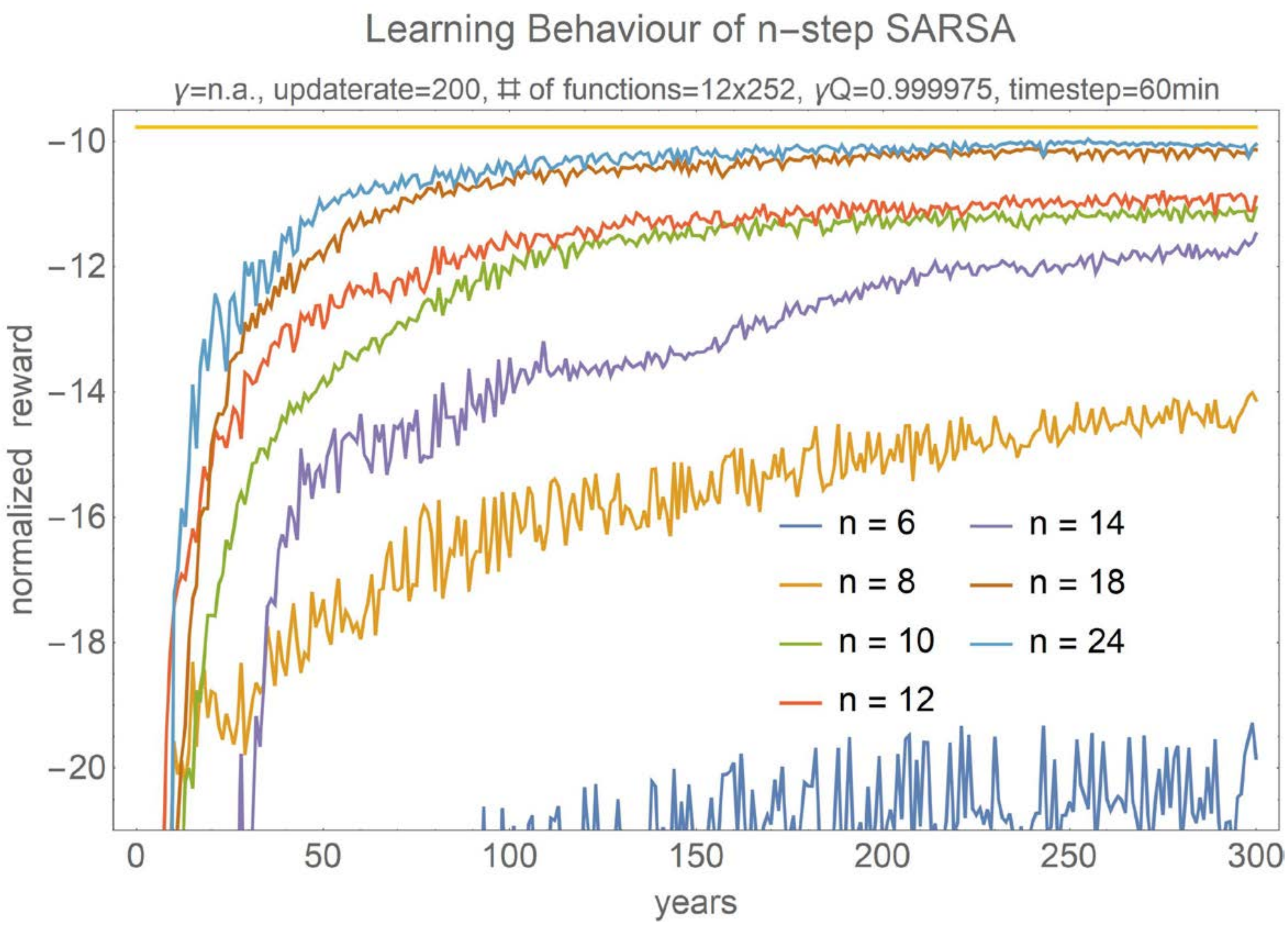
Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics

- The discount of future rewards disturbs the optimal scheduling of actions with fixed and known costs.
- Alternative approach: n-step SARSA without discount

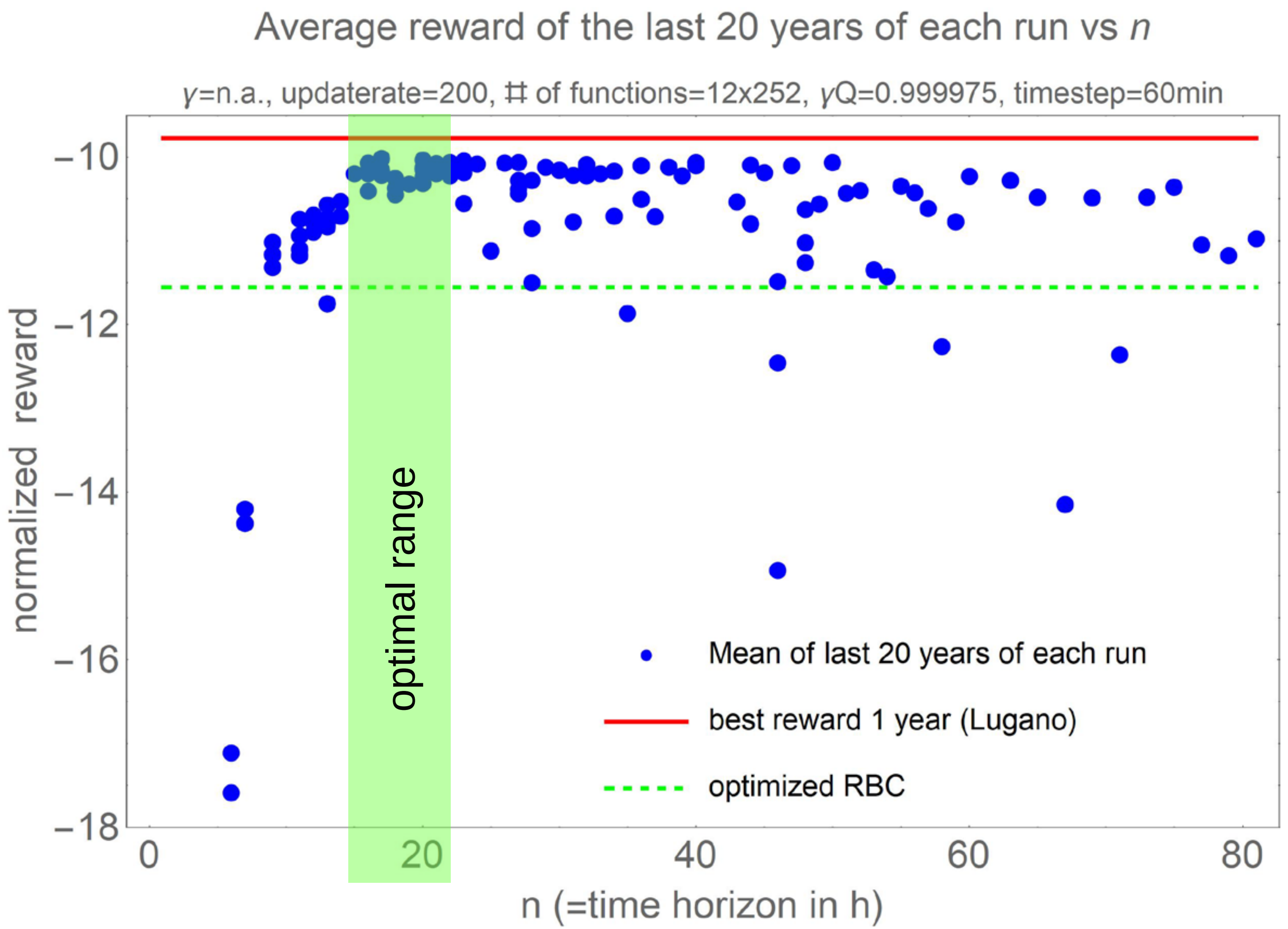
$$G_{t,n} = \sum_{k=0}^{n-1} \cancel{\gamma^k} R_{t+k} + \cancel{\gamma^n q(s_{t+n}, a_{t+n})}$$

- The time horizon is defined by n instead of γ!

n-step SARSA without Discount

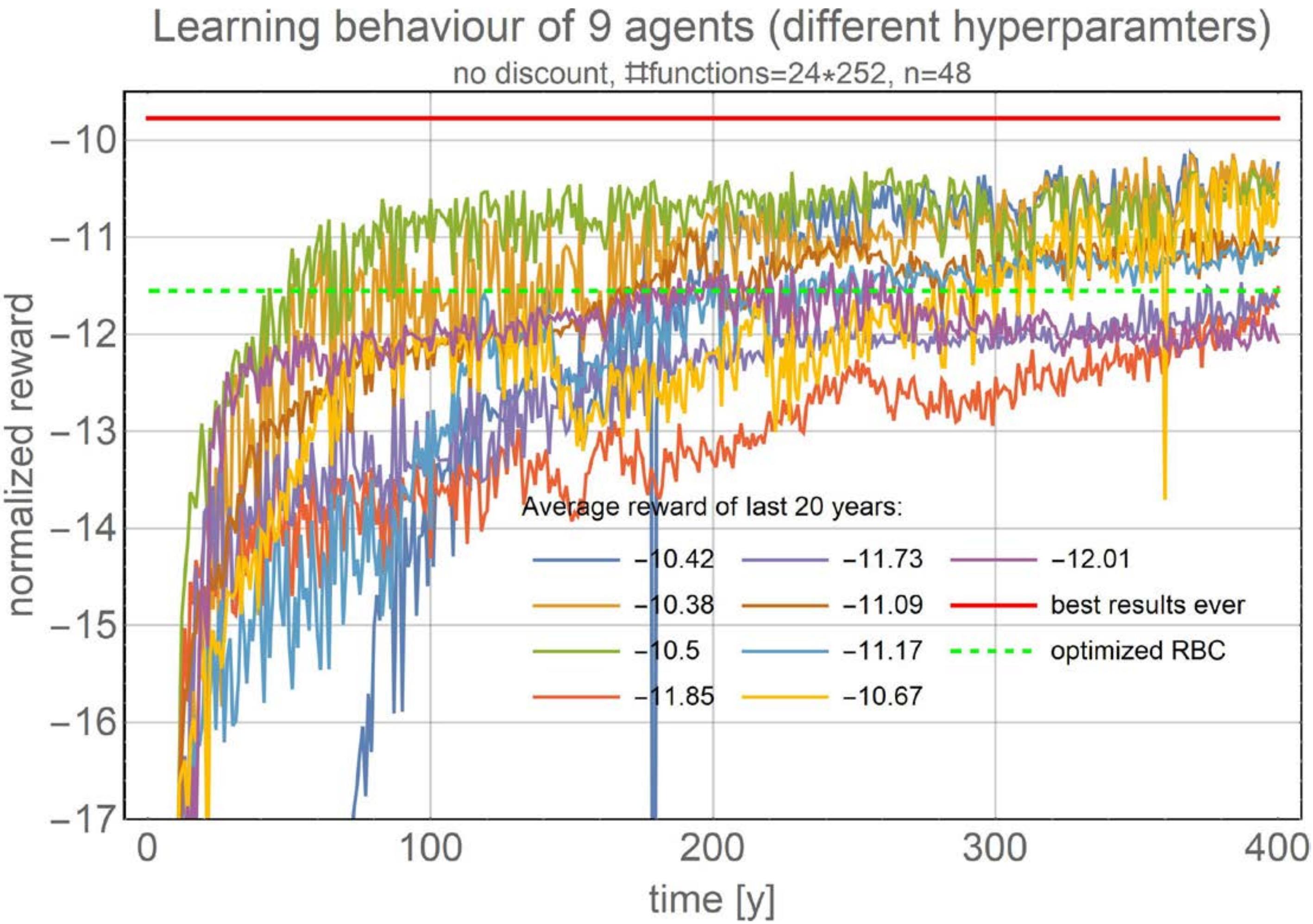


Performance of *n*-step SARSA without Discount



Parallel «Supporting» Agents, Motivation

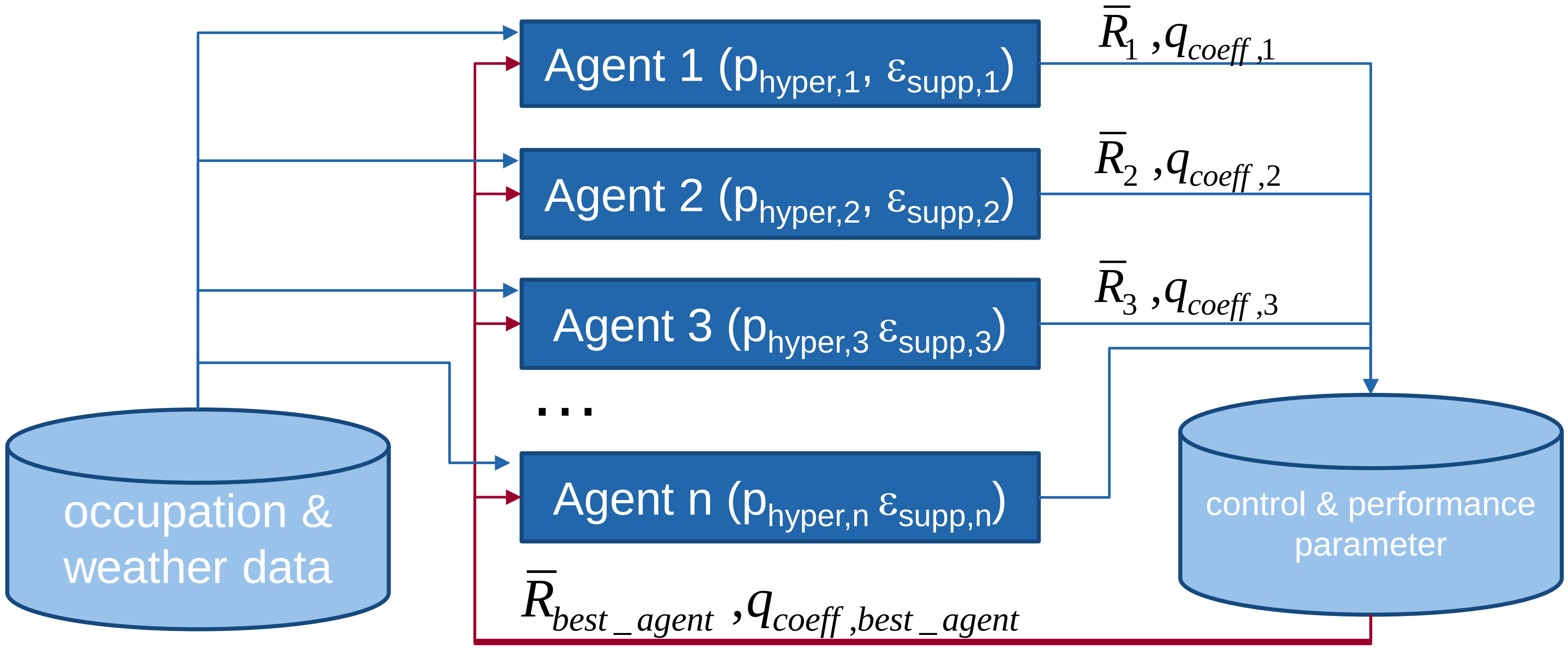
Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics



Could larger scatter be exploited by helping worse agents become better and then by chance even be better than best agent?

Parallel «Supporting» Agents, Structure

Zurich University of Applied Sciences
School of Engineering
IAMP Institute of Applied Mathematics and Physics



$$\varepsilon_{supp,i} = \begin{cases} \min\left[c_{supp} \dfrac{\bar{R}_i}{\left(\bar{R}_i - \bar{R}_{best\_agent}\right)} \left(1 - e^{\#year/\tau_{supp}}\right), 1 \right] & \text{if } \#year \text{ even} \\ 0 & \text{if } \#year \text{ odd} \end{cases}$$

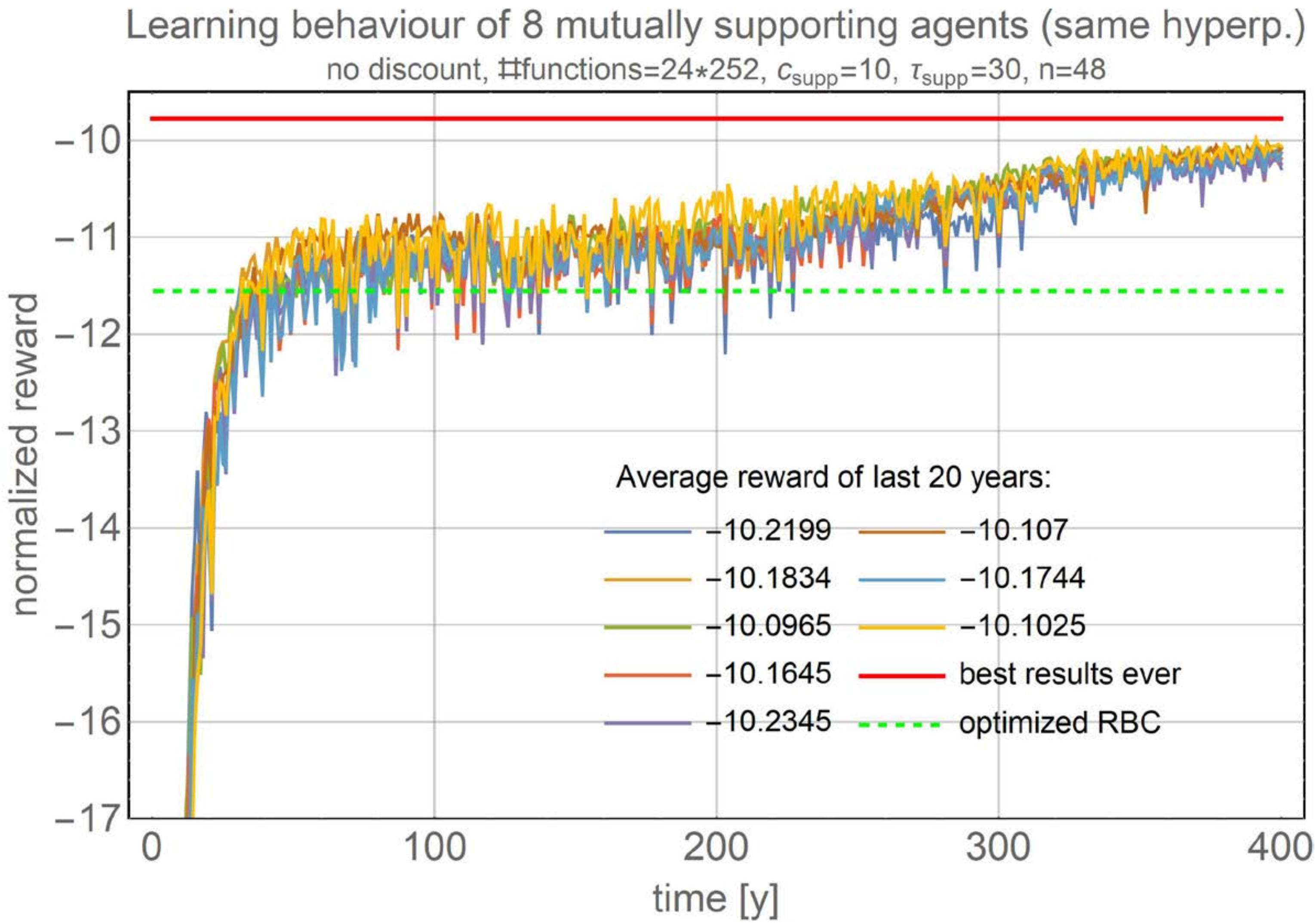
needed to evalute performance

$$\varepsilon > \varepsilon_{supp} : \text{decision based on own parameters}$$

$$\varepsilon < \varepsilon_{supp} : \text{decision based on best agents parameters}$$

Parallel «Supporting» Agents, Results

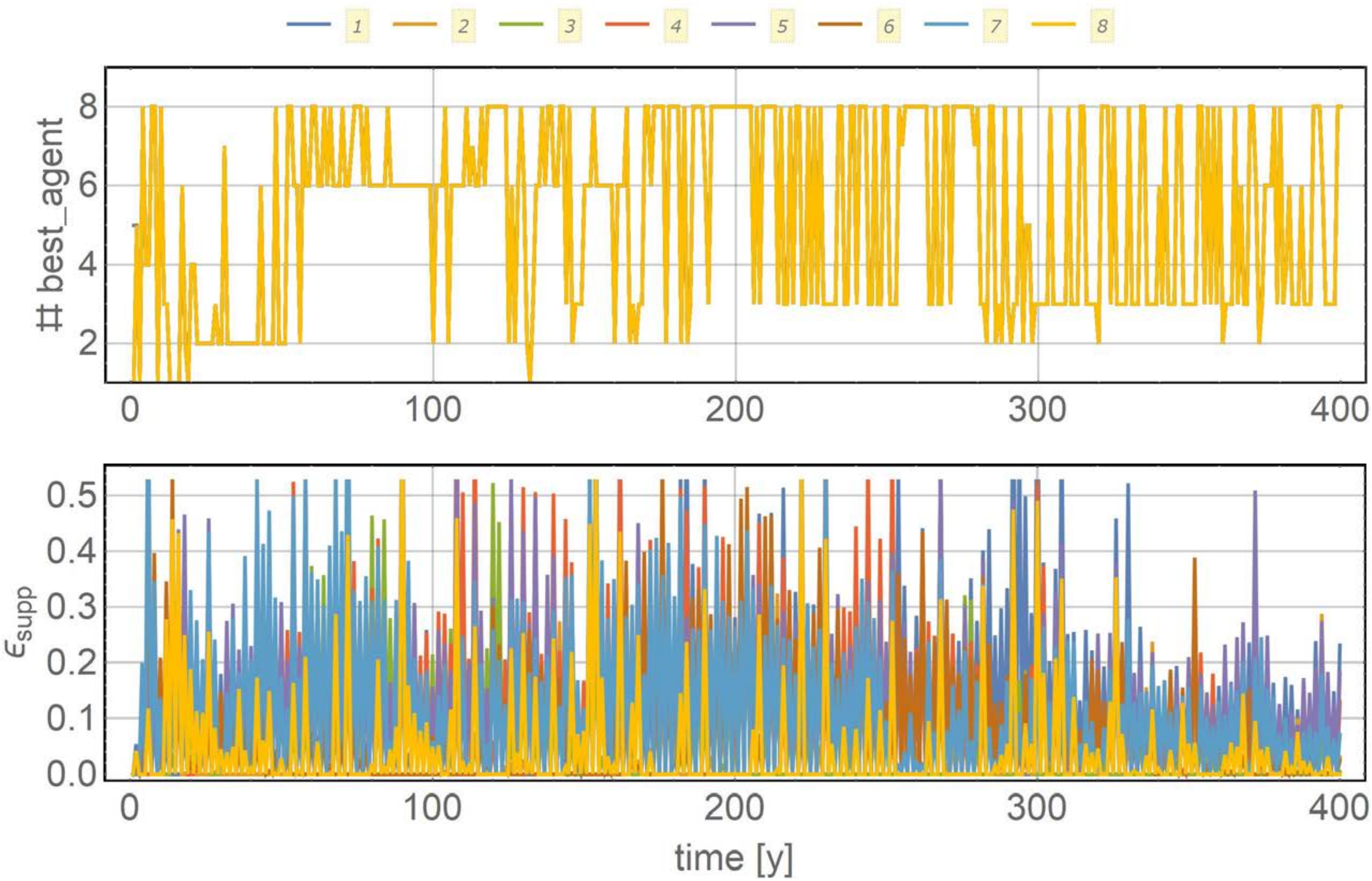
Zurich University
of Applied Sciences
zh
aw
School of Engineering
IAMP Institute of Applied Mathematics and Physics



→ all agents show better performance than best agent without support

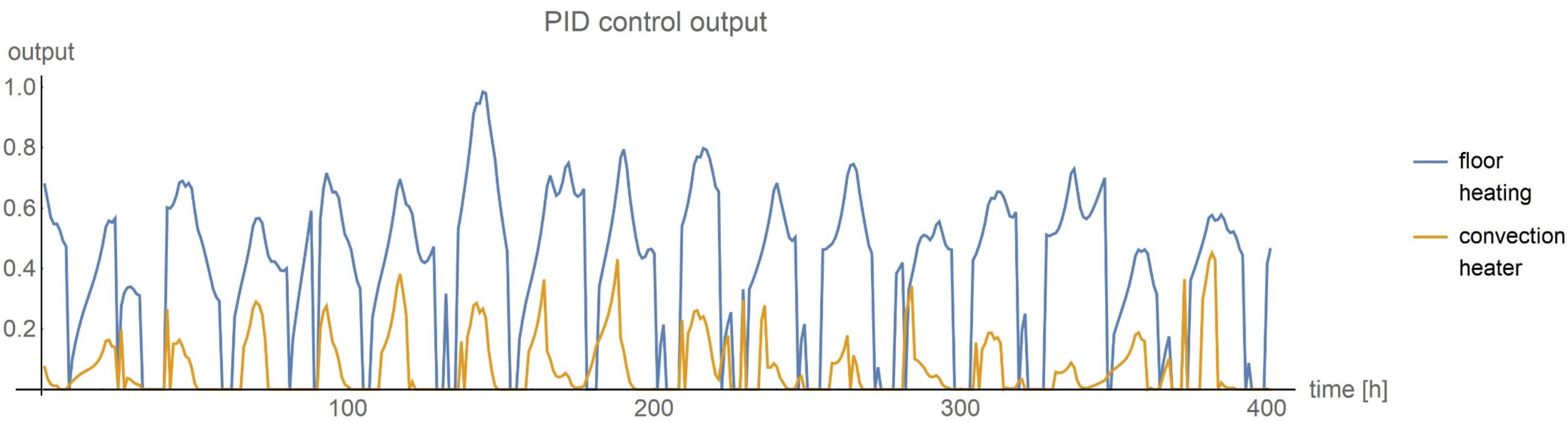
Parallel «Supporting» Agents, Results

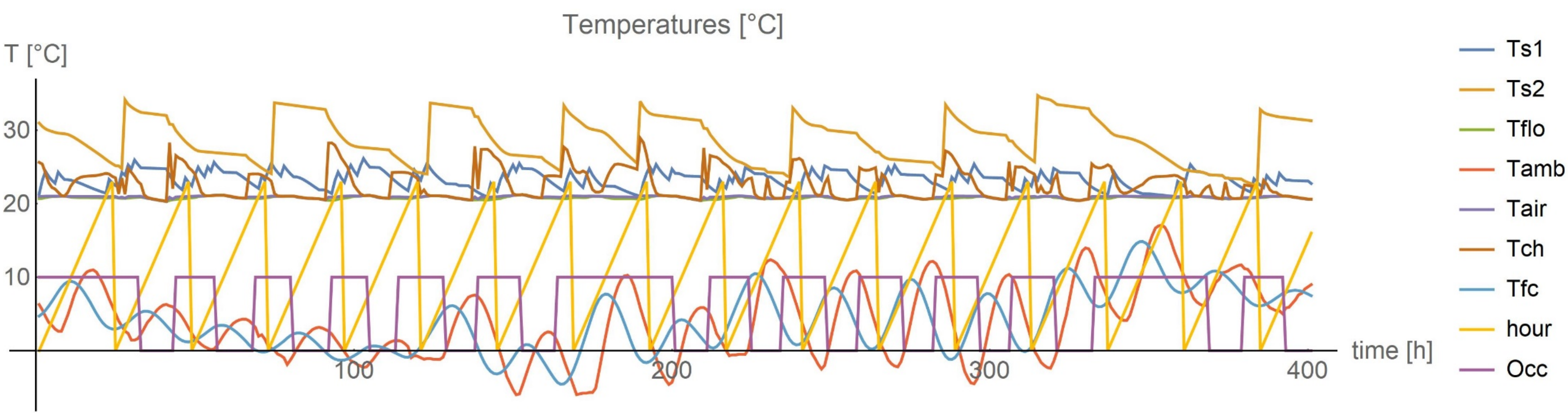
Zurich University of Applied Sciences
zh aw
School of Engineering
IAMP Institute of Applied Mathematics and Physics



RL Control Example (1)

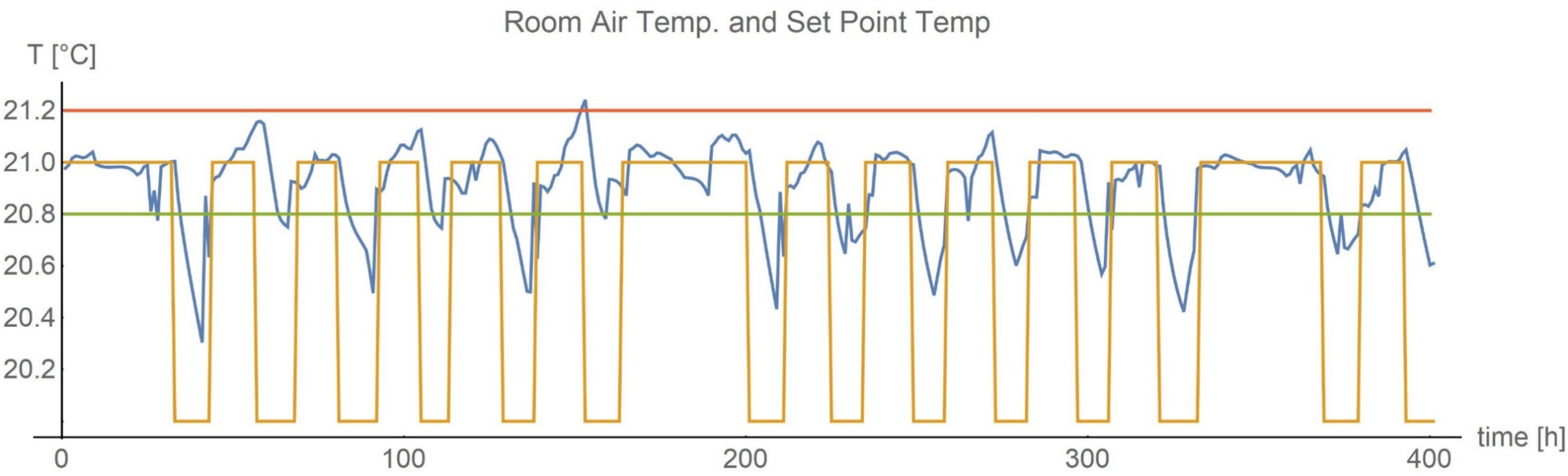
Zurich University of Applied Sciences
School of Engineering
IAMP Institute of Applied Mathematics and Physics

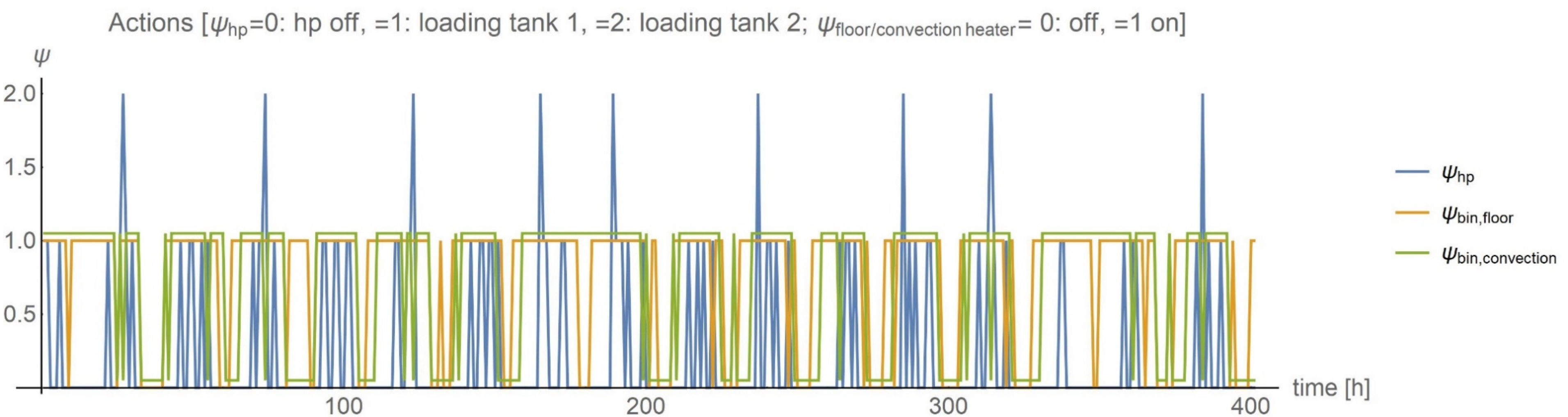




RL Control Example (2)

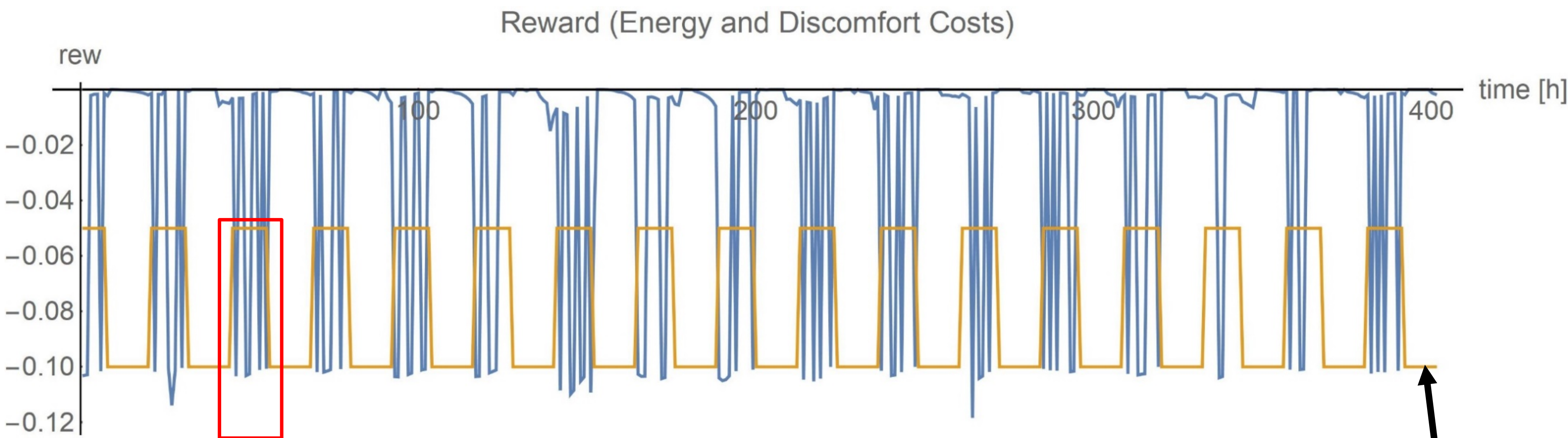
Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics





RL Control Example (3)

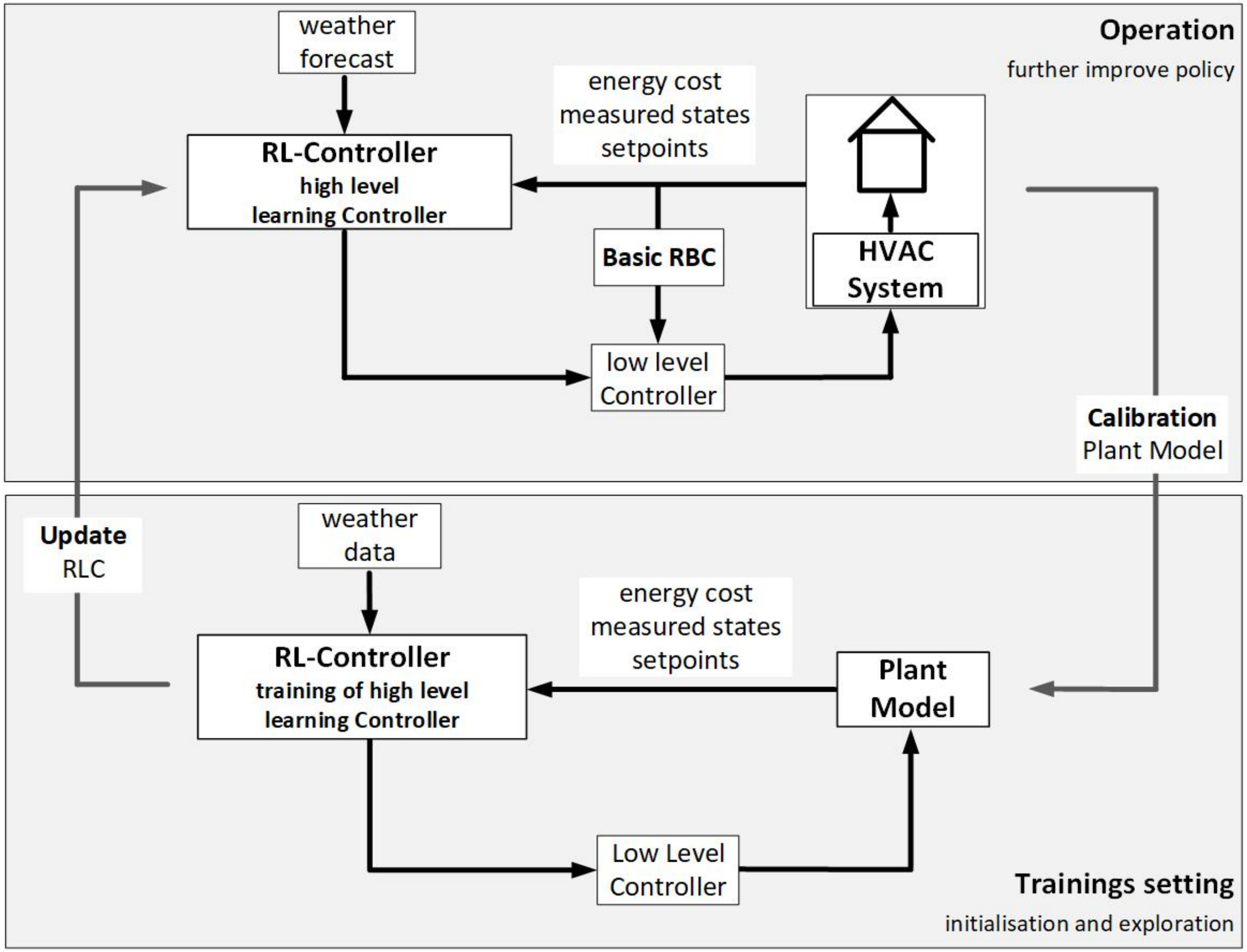
Zurich University of Applied Sciences
School of Engineering
IAMP Institute of Applied Mathematics and Physics



-0.1 *(energy rate)

agent heats at cheaper night rate only

Simulated Reinforcement Learning



Schematic: Peter Bolt, ACSS IMAP ZHAW

RBC with Parameters Optimized by RL

Zurich University of Applied Sciences
zh aw School of Engineering
IAMP Institute of Applied Mathematics and Physics

- Some simple fixed rules
- Setpoints for water storage heating 1 & 2 are parametrized, k_1 and k_2 learned

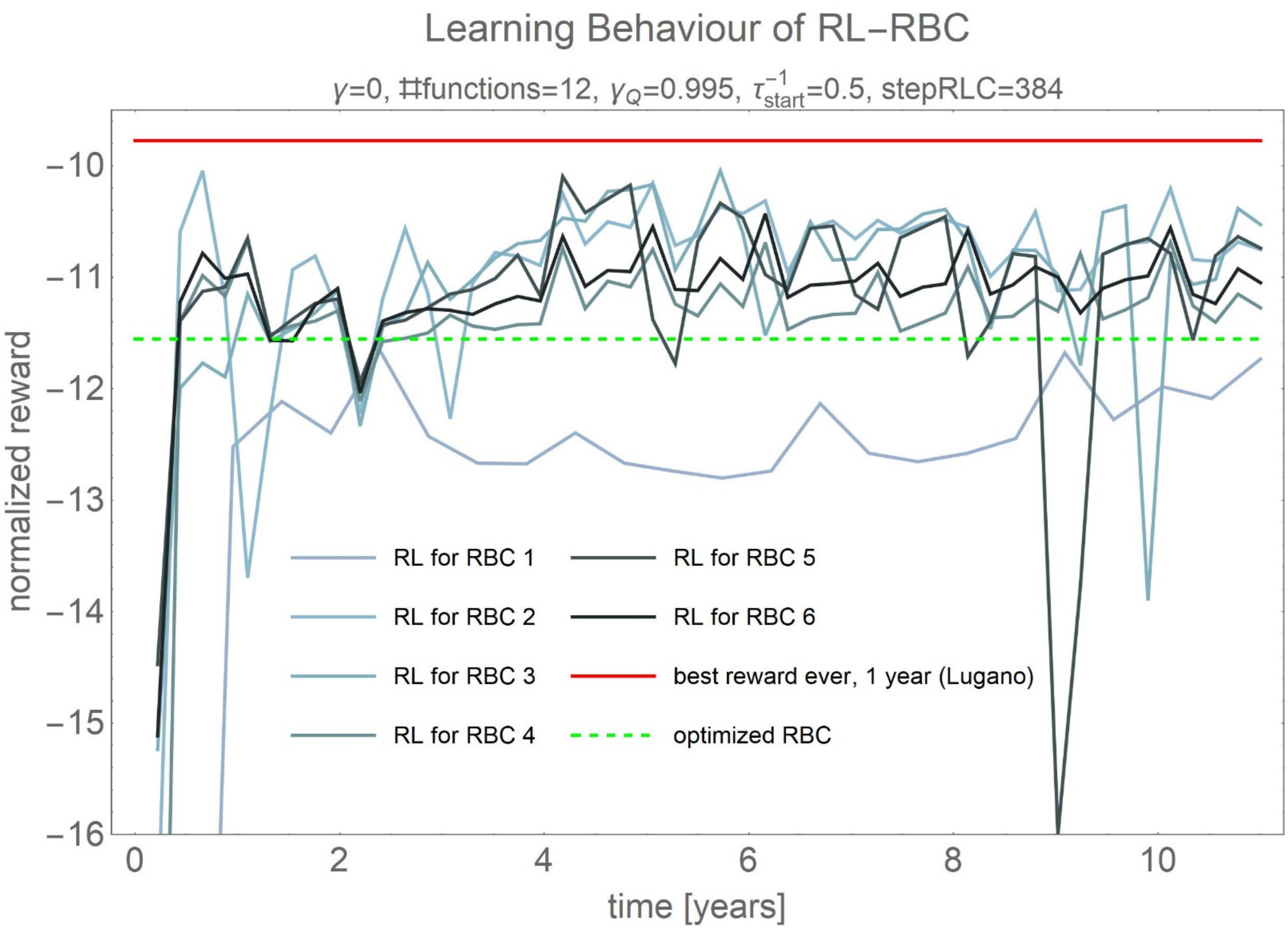
$$T_{sp,heating\ st1} = T_{sp,room} + \left(T_{sp,room} - \frac{T_{amb} + T_{forecast}}{2} \right) \cdot k_1$$

$$T_{sp,heating\ st2} = T_{sp,room} + \left(T_{sp,room} - \frac{T_{amb} + T_{forecast}}{2} \right) \cdot k_2$$

- Reward (energy consumption & set point violation) is normalized and thus almost independent of outside temperature:

$$R_{norm} = \frac{R}{\Delta T + 0.005 \Delta T^2} \quad with\, \Delta T = T_{sp} - T_{amb}$$

RL with Parametrized RBC



→ acceptable performance in less than 1 year!

Summary

- RLC for a heating system
  - converges to nearly optimal trajectories
  - needs order of 100 simulated years for convergence ($\rightarrow$ simulated RL)
- Improvements to RL
  - least-squares fit to get q(s,a) in one step
  - *n*-step SARSA
  - truncated reward sum without discount
  - parallel, mutually supporting agents
- Alternative approach with parametrized RBC
  - much shorter learning time due to less coefficients
  - good performance