Neural Fitted Actor-Critic

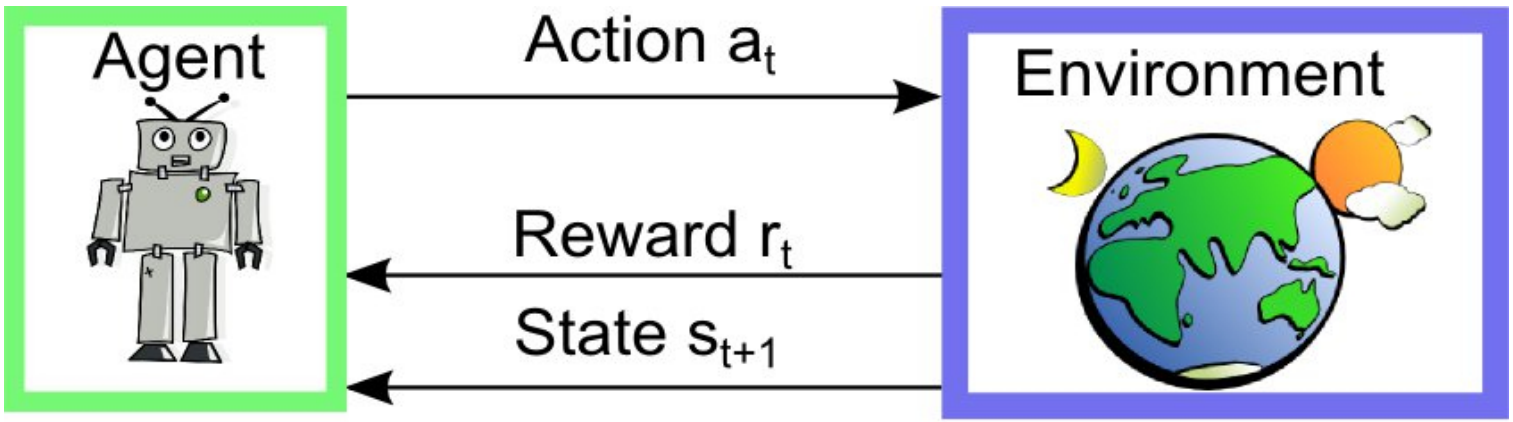
Matthieu Zimmer
Alain Dutech
Yann Boniface

University of Lorraine, LORIA

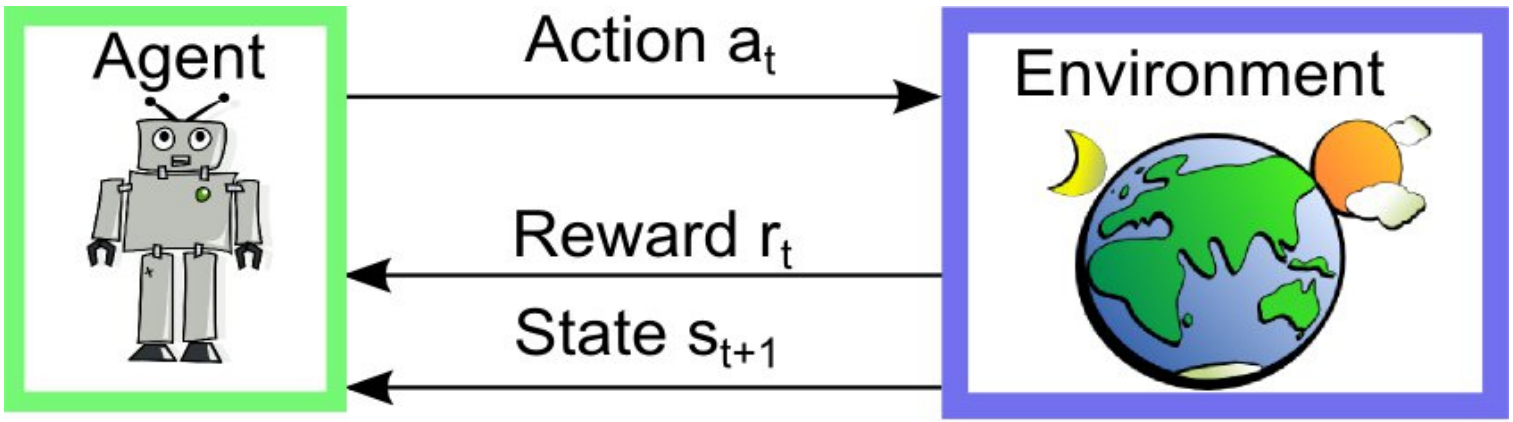
8^{th} July 2016

Outline

Reinforcement Learning



Reinforcement Learning



Optimization problem

Find function $\pi : S \rightarrow A$ that maximize rewards $\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \right]$

Constraints and Motivations

Reinforcement learning + Developmental robotics :

1. Continuous environments
2. No prior models of agent or environment
3. Use non linear approximator (neural networks)
4. No prior goal states or trajectories

How to solve reinforcement learning problems ?

Actor-only $\pi : S \rightarrow A$

$$\pi^k \xrightarrow{\text{play}} \boxed{\sum_{t=0}^{\infty} \gamma^t r_t} \xrightarrow{\text{update}} \pi^{k+1}$$

How to solve reinforcement learning problems ?

Actor-only $\pi : S \to A$

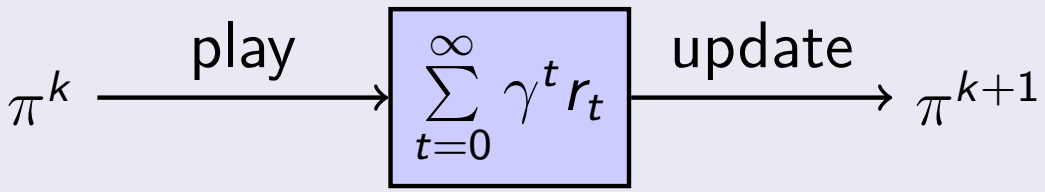
$$\pi^k \xrightarrow{\text{play}} \boxed{\sum_{t=0}^{\infty} \gamma^t r_t} \xrightarrow{\text{update}} \pi^{k+1}$$

Critic-only $Q : S \times A \to \mathbb{R}$

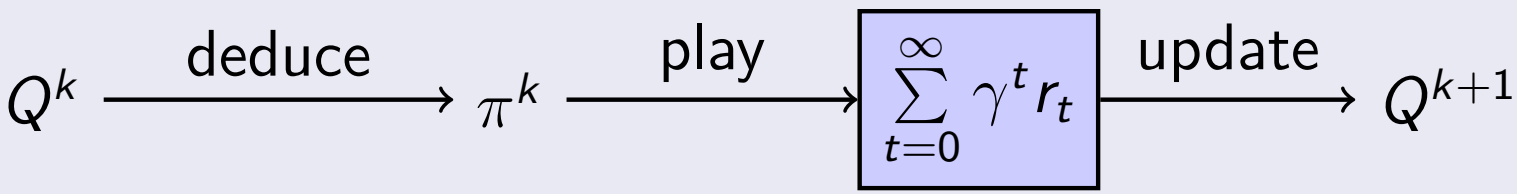
$$Q^k \xrightarrow{\text{deduce}} \pi^k \xrightarrow{\text{play}} \boxed{\sum_{t=0}^{\infty} \gamma^t r_t} \xrightarrow{\text{update}} Q^{k+1}$$

How to solve reinforcement learning problems ?

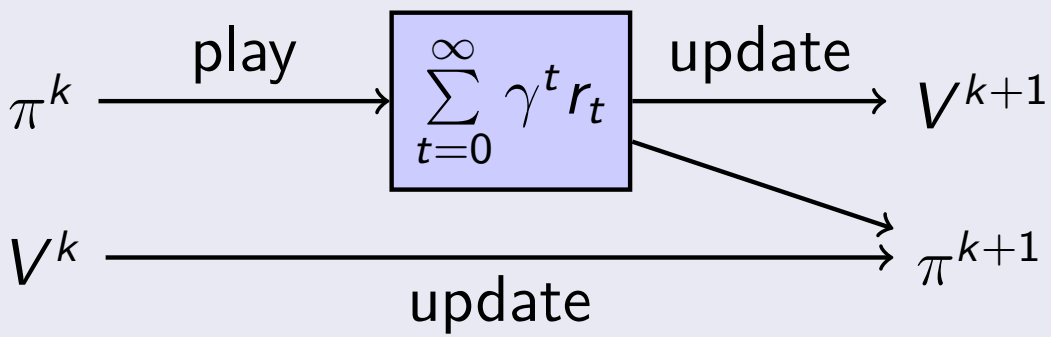
Actor-only $\pi : S \to A$



Critic-only $Q : S \times A \to \mathbb{R}$



Actor-Critic $\pi : S \to A$ $V : S \to \mathbb{R}$



State of the art

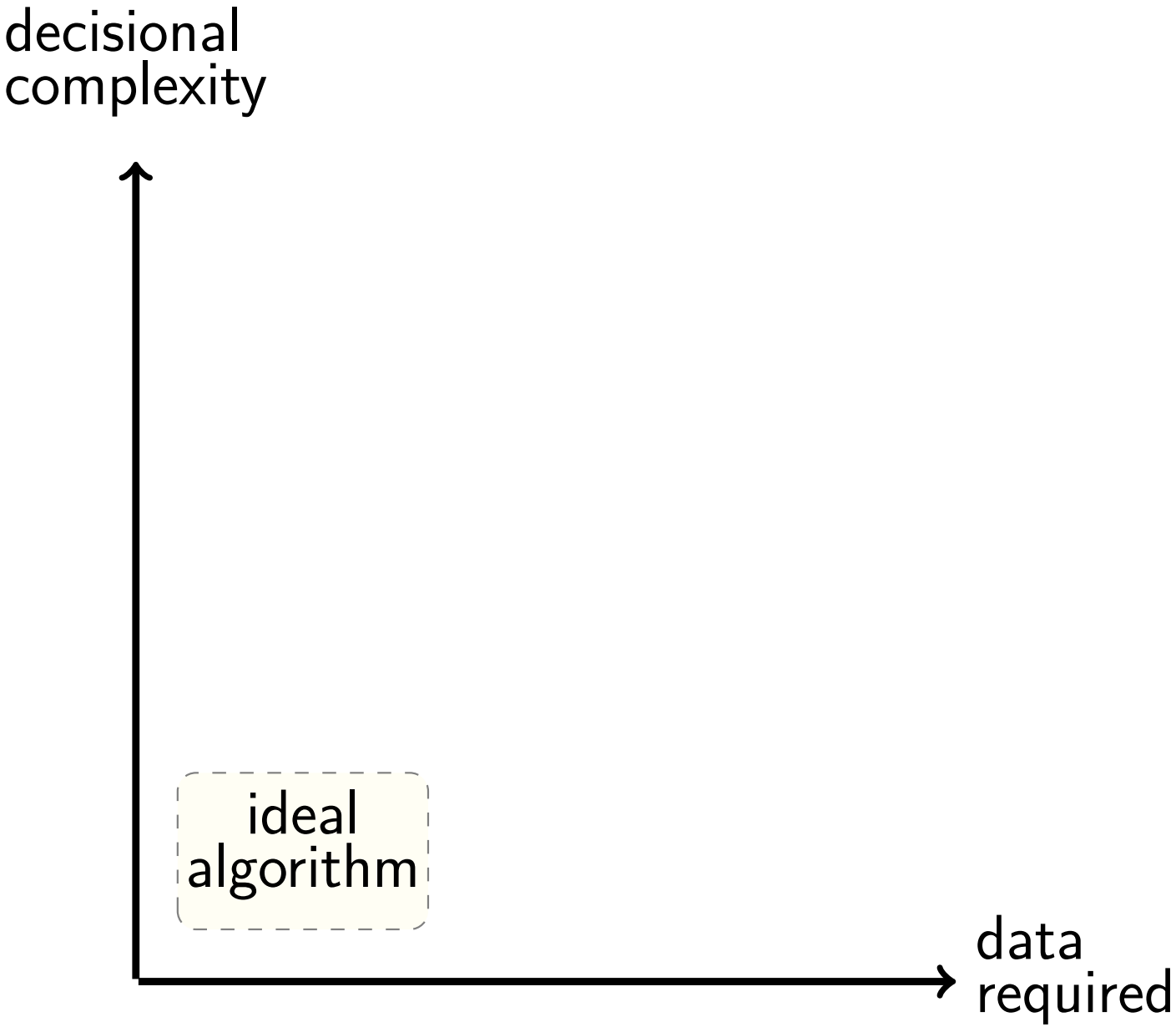
- Critic only
  - Fitted Q Iteration
  - Q Learning, Sarsa
- Actor only
  - Evolutionnary algorithms (CMA-ES, ...)
  - PI^2
- Actor-critic
  - Natural Actor Critic
  - Cacla

State of the art

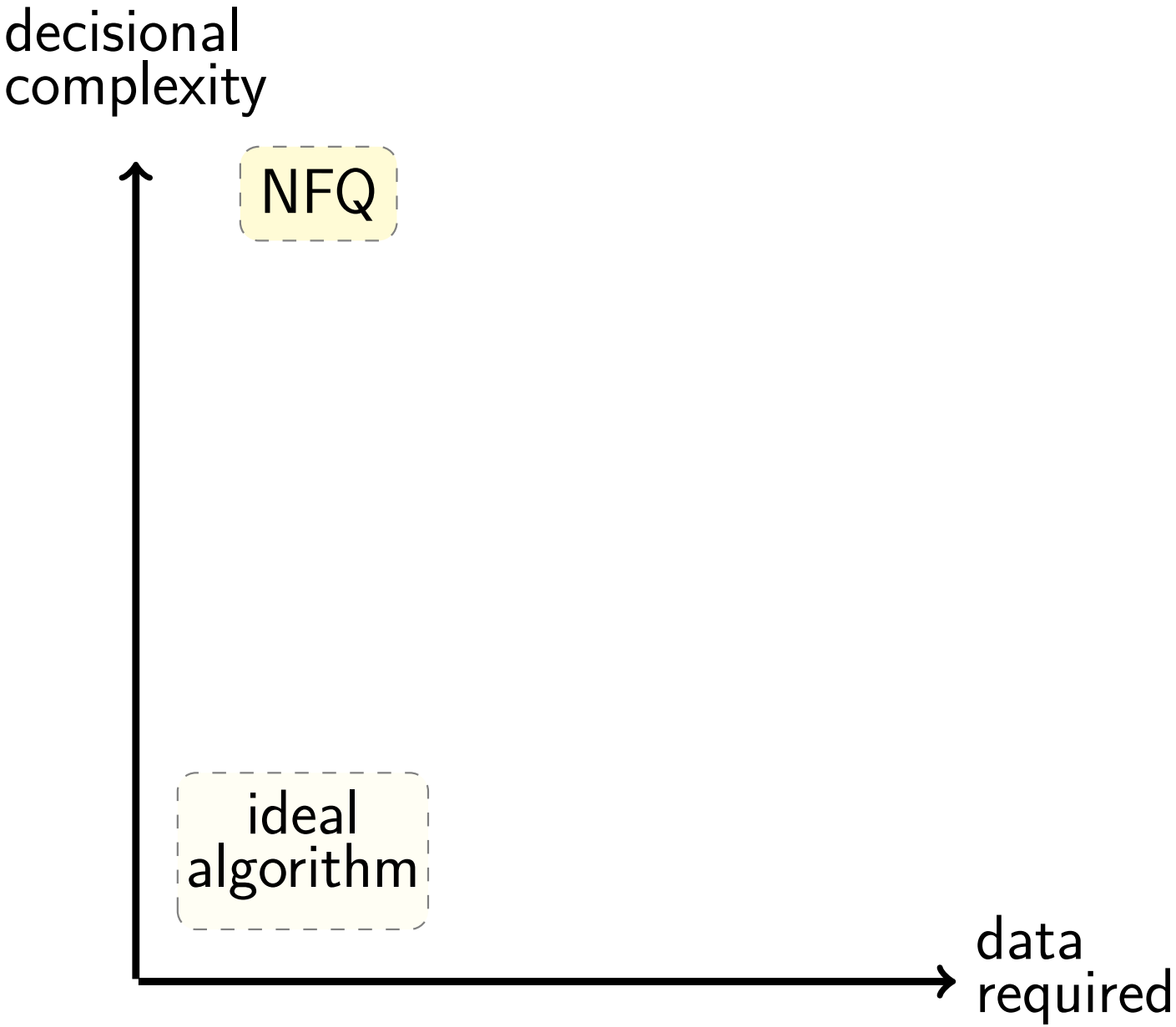
Unsatisfied Constraints :
(1) No Continuous environments
(2) No prior models of agent or environment
(3) Use linear approximator
(4) No prior goal states or trajectories

- Critic only
  - Fitted Q Iteration (1)
  - Q Learning, Sarsa (1)
- Actor only
  - Evolutionnary algorithms (CMA-ES) $\rightarrow$ poor data efficiency
  - PI^2 (3) (4)
- Actor-critic
  - Natural Actor Critic (3) (4)
  - Cacla $\rightarrow$ poor data efficiency, lot of meta-parameters

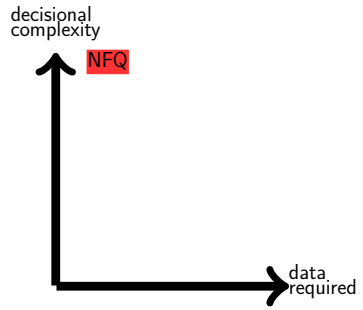
Landscape algorithms



Landscape algorithms

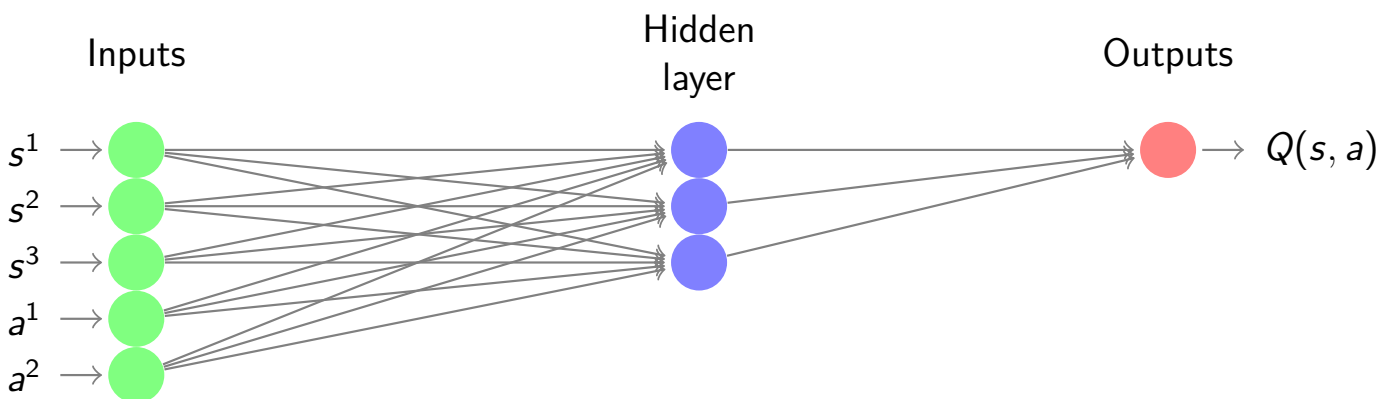


Neural Fitted Q (NFQ)



$$Q_{k+1} = \arg\min_{Q \in \mathcal{F}_c} \sum_{t=1}^{N} \left[Q(s_t, a_t) - \left(r_{t+1} + \gamma \max_{a' \in A} Q_k(s_{t+1}, a') \right) \right]^2$$

$$\pi^*(s) = \arg\max_{a \in A} Q(s, a)$$



CACLA

Temporal Difference Error

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

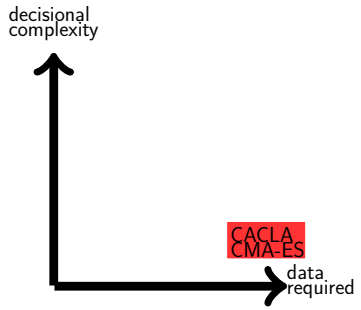
Critic

$$V_{k+1}(s) = V_k(s) + \alpha_v \delta_t$$

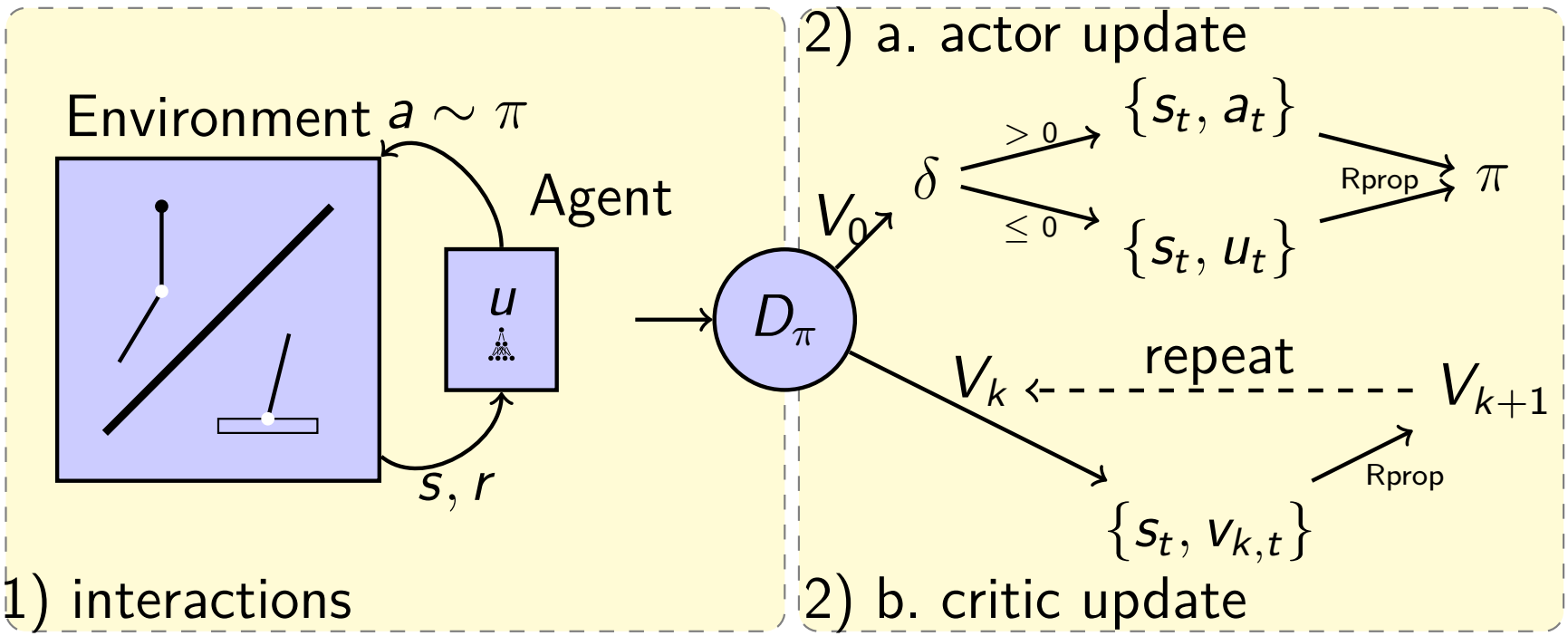
$$\theta^V_{i,k+1} = \theta^V_{i,k+1} + \alpha_v \delta_t \frac{\partial V_t(s_t)}{\partial \theta^V_{i,k+1}}$$

Actor

$$\theta_{t+1} = \theta_t + \begin{cases} \alpha_a (a_t - u_t) \frac{\partial u_t(s_t)}{\partial \theta_t}, & \text{if } \delta > 0 \\ 0, & \text{otherwise} \end{cases}$$

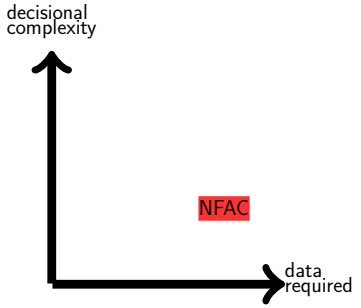


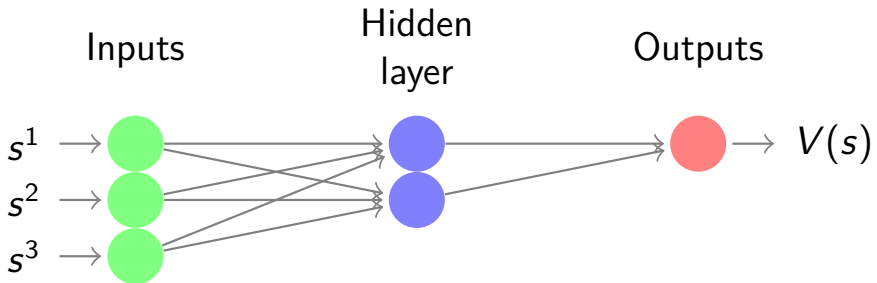
Neural Fitted Actor Critic



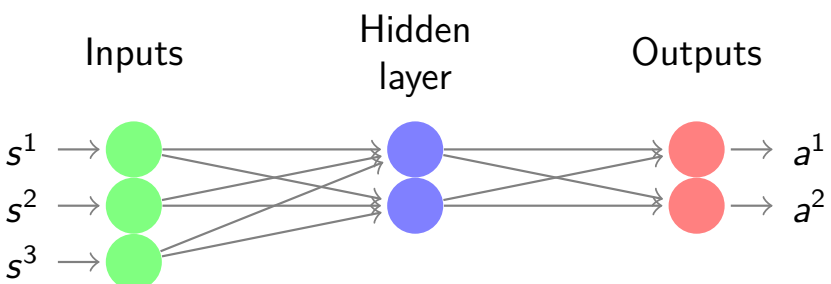
Neural Fitted Actor Critic

$$V_{k+1} \leftarrow \underset{V \in \mathcal{F}_c}{\operatorname{argmin}} \sum_{s_t \in \mathcal{D}_\pi} \left[V(s_t) - r_{t+1} + \gamma V_{k-1}(s_{t+1}) \right]^2$$

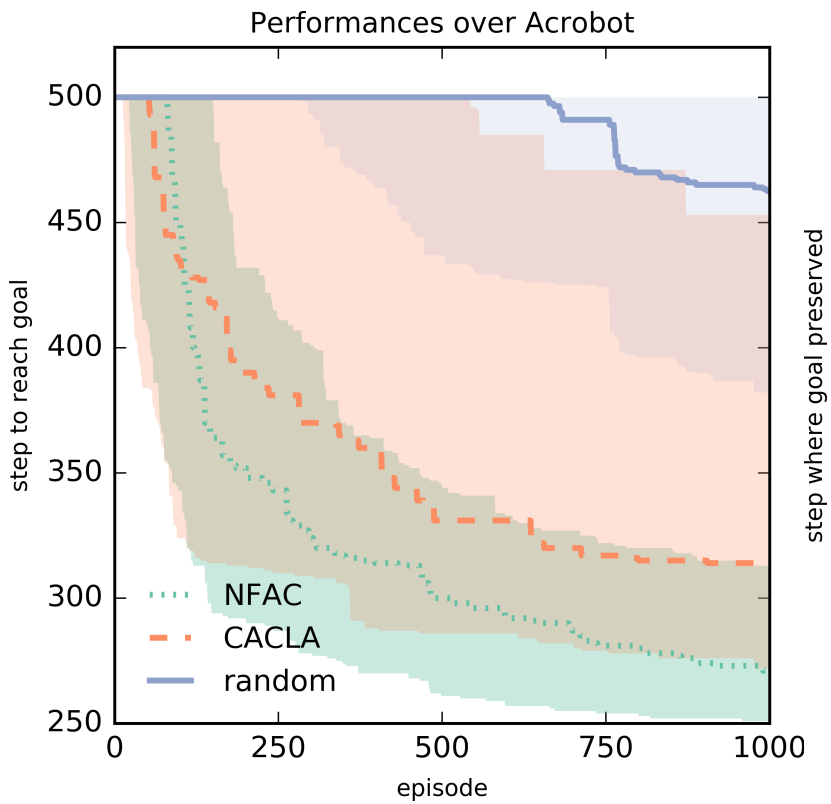


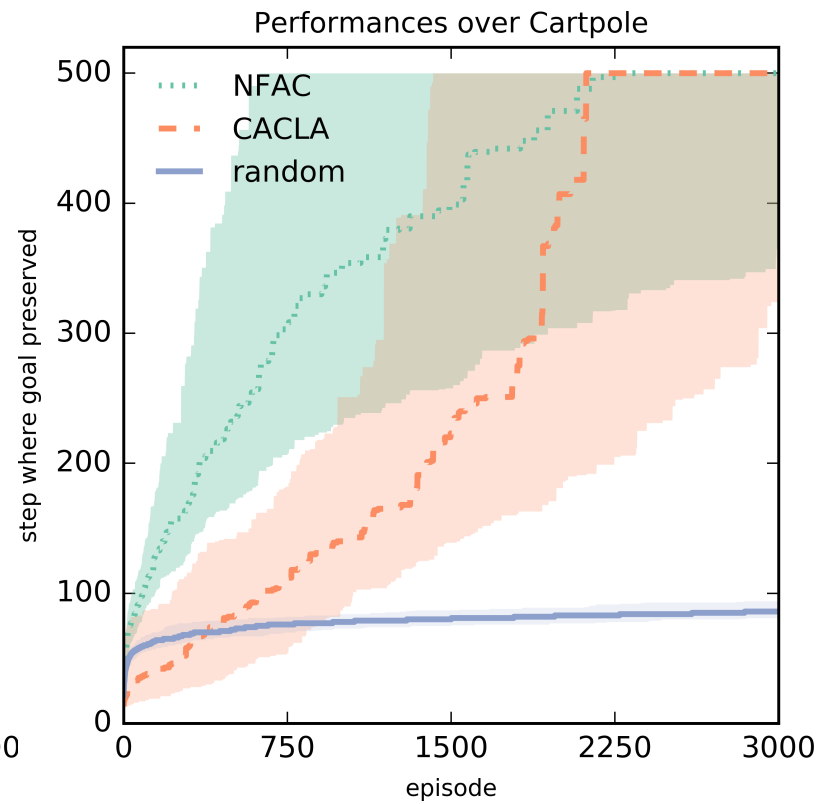


$$\pi_{k+1} \leftarrow \underset{\pi \in \mathcal{F}_a}{\operatorname{argmin}} \sum_{s_t \in \mathcal{D}_\pi} \left[\pi(s_t) - \begin{cases} a_t, & \text{if } \delta_t > 0 \\ u_t, & \text{otherwise} \end{cases} \right]^2$$

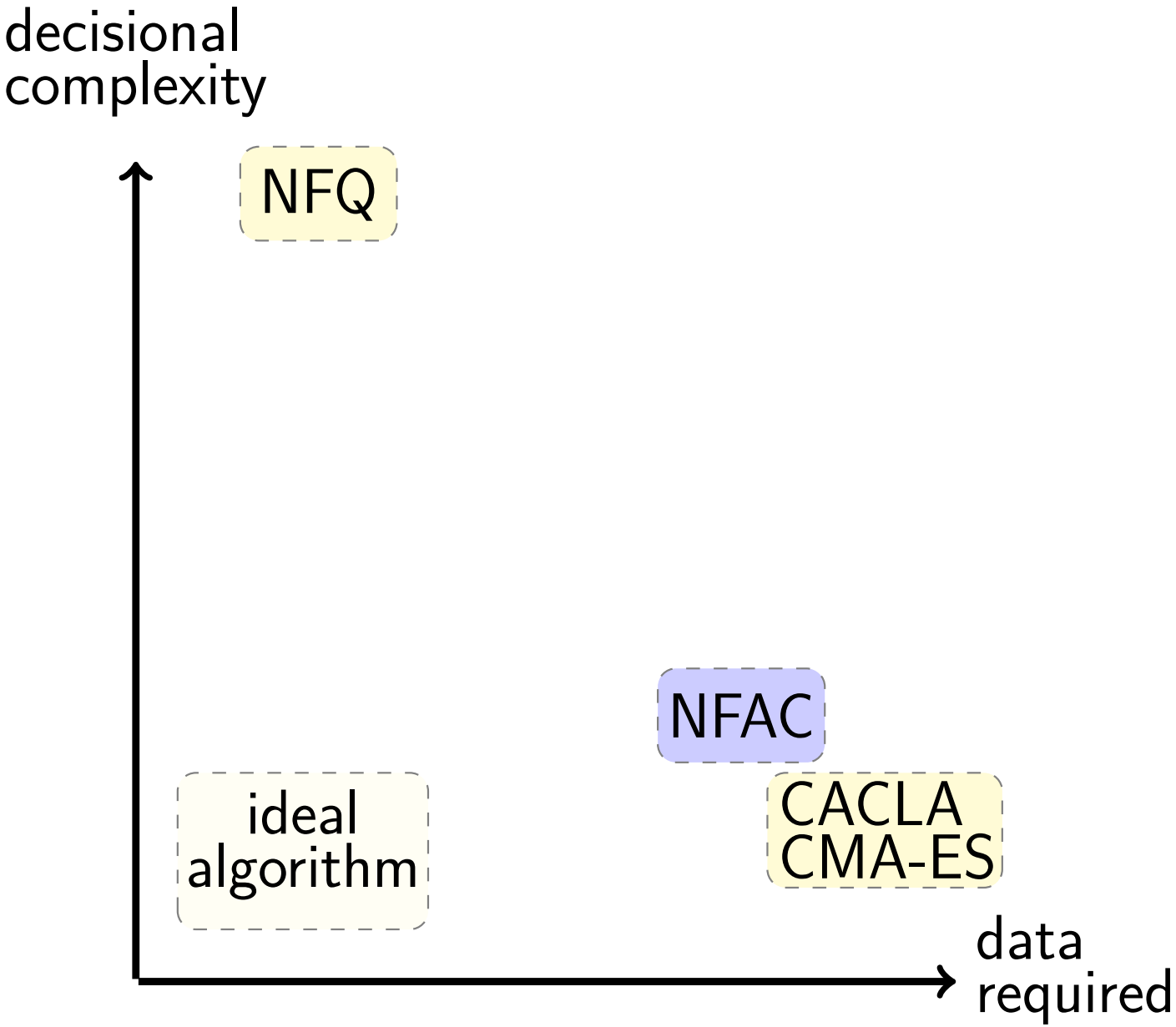


Experimental Results

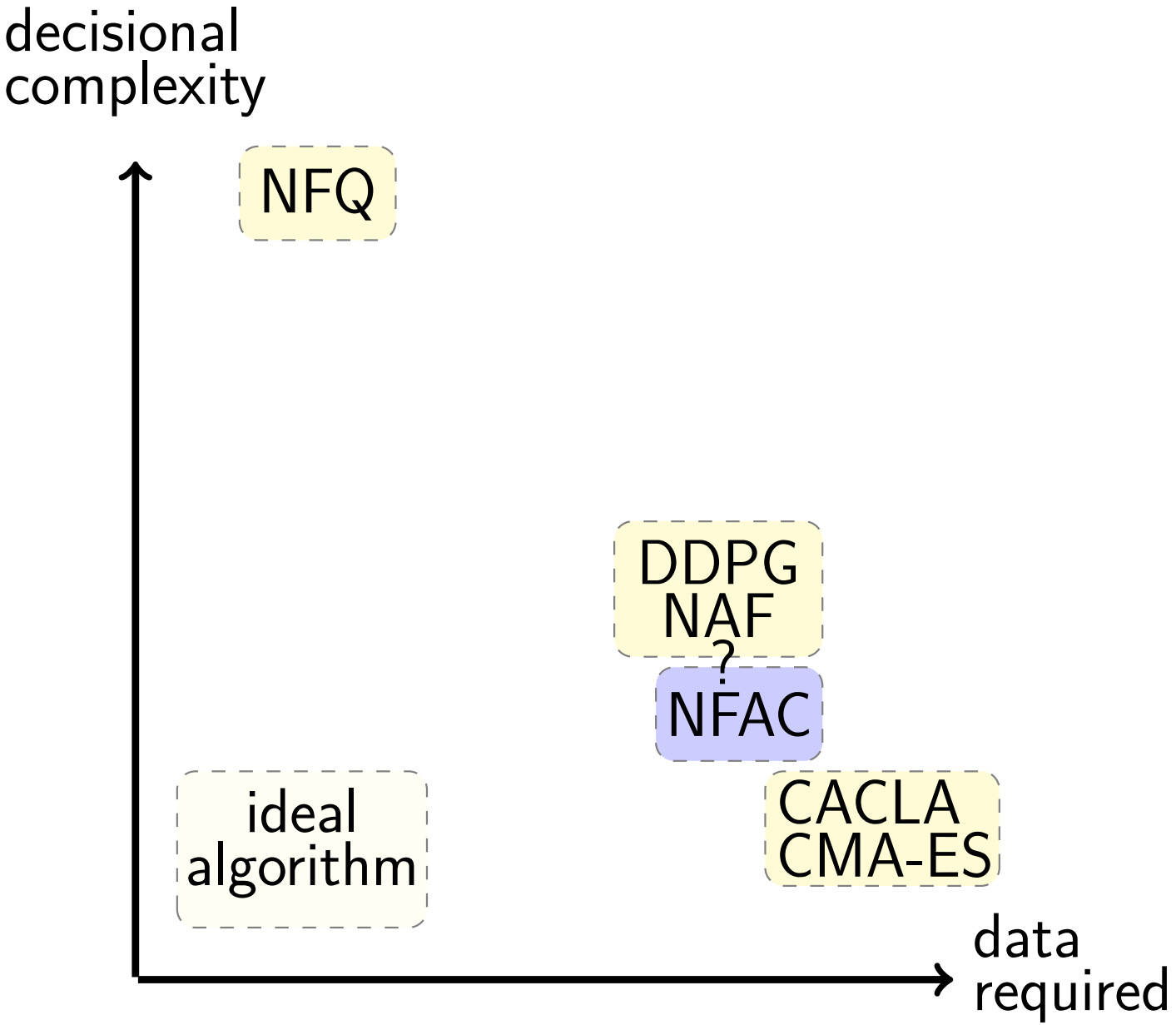




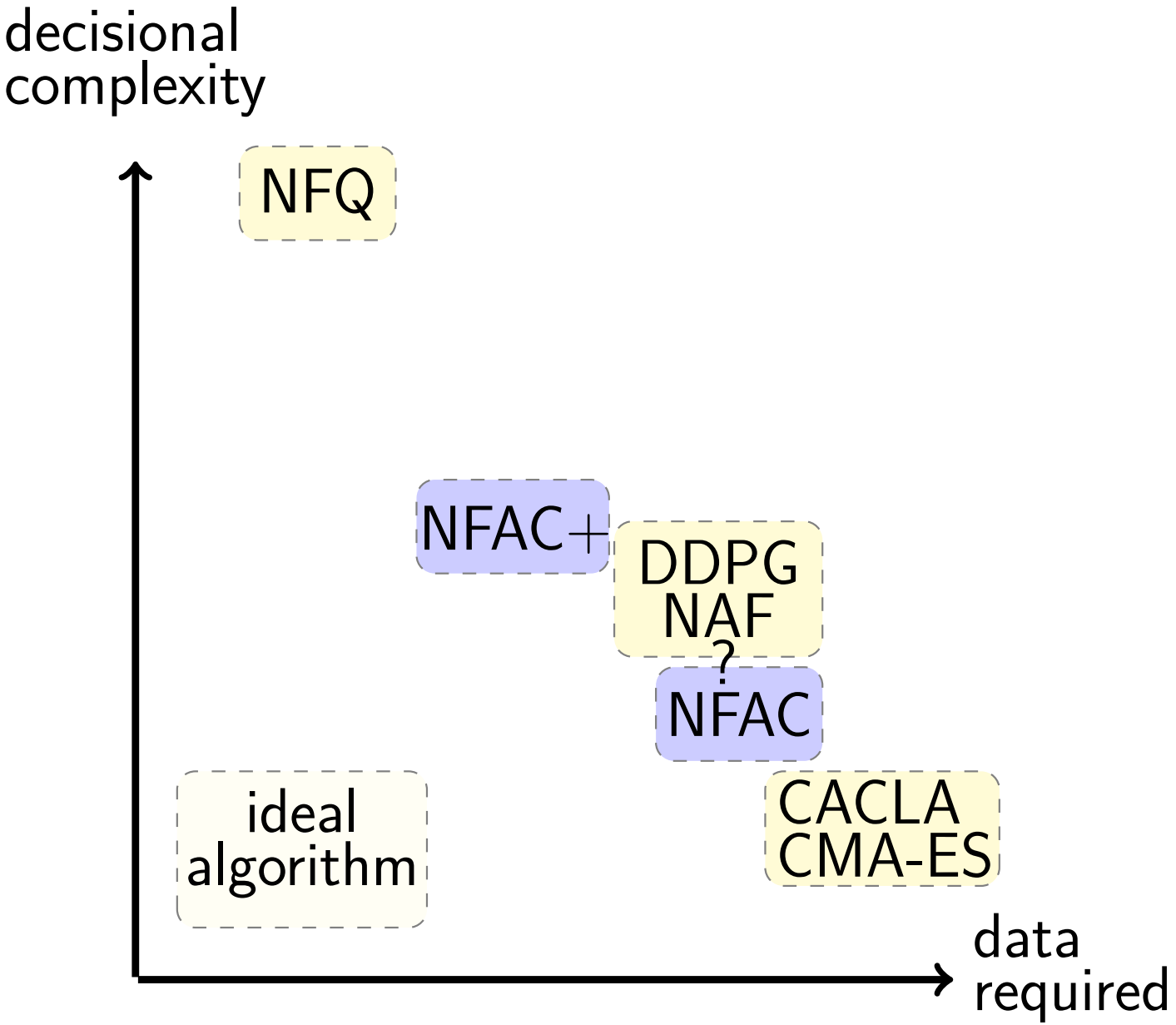
Landscape algorithms



Landscape algorithms



Methods landscape



Toward a better data efficiency

Fitted Actor-Critic

$$Q^{\pi}_{k+1} = \underset{Q \in \mathcal{F}_c}{\operatorname{argmin}} \sum_{t=1}^{N} c(a_t|s_t) \Big[Q(s_t, a_t) - \Big(r_{t+1} + \gamma Q^{\pi}_{k}(s_{t+1}, \pi(s_{t+1})) \Big) \Big]^2$$

$$\pi_{k+1} = \underset{\pi \in \mathcal{F}_a}{\operatorname{argmax}} \sum_{t=1}^{N} Q_{k+1}\Big(s_t, \pi_k(s_t)\Big)$$

$$c(a_t|s_t) = min\left(1, \frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)}\right)$$

Conclusion & Further Works

Neural Fitted Actor Critic

- Compare to DDPG
- Don't forget previous data

Guided exploration of sensorimotor space

- Increase the dimension of states/actions
- Redefine the reward function for the new sub-goal