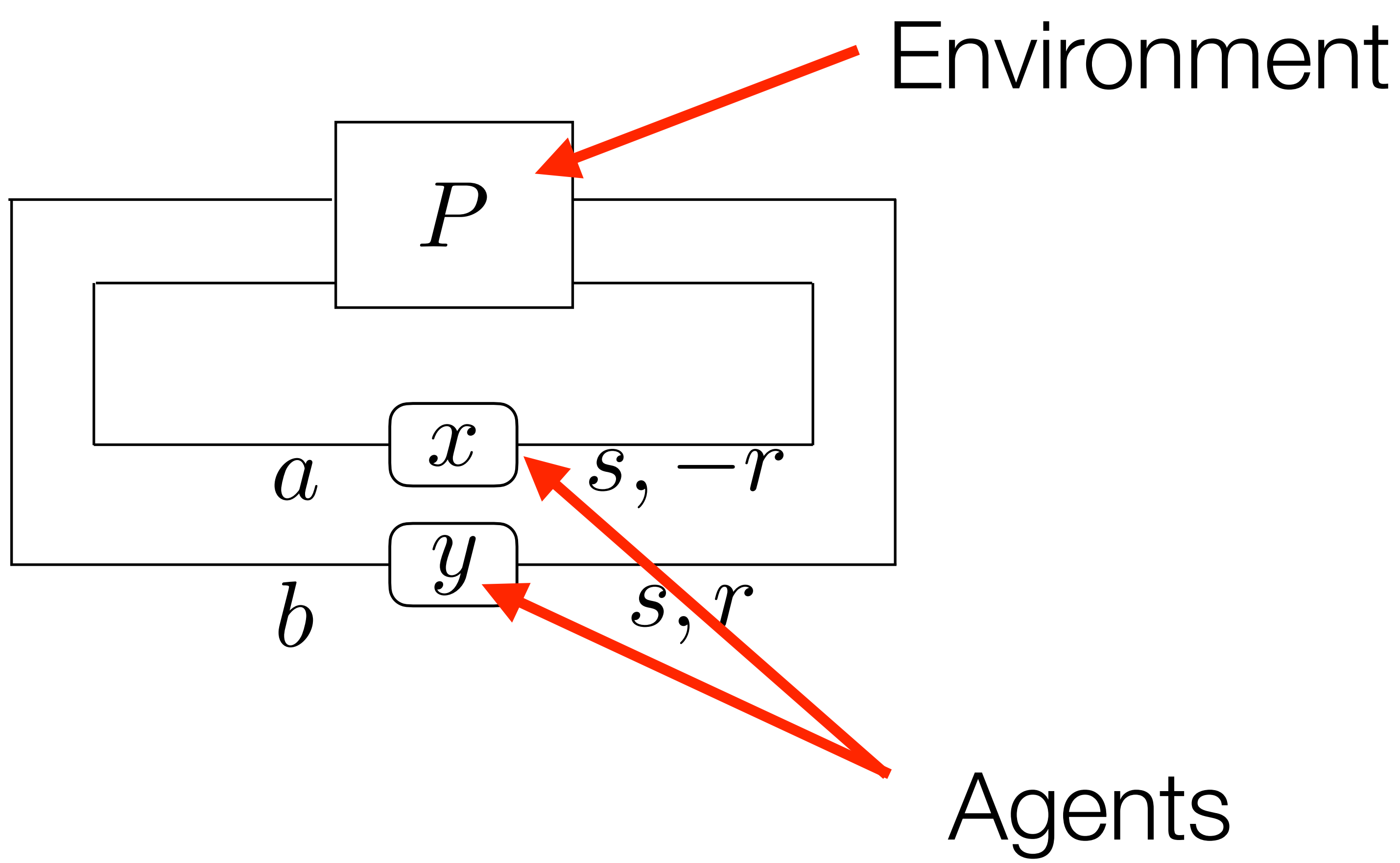
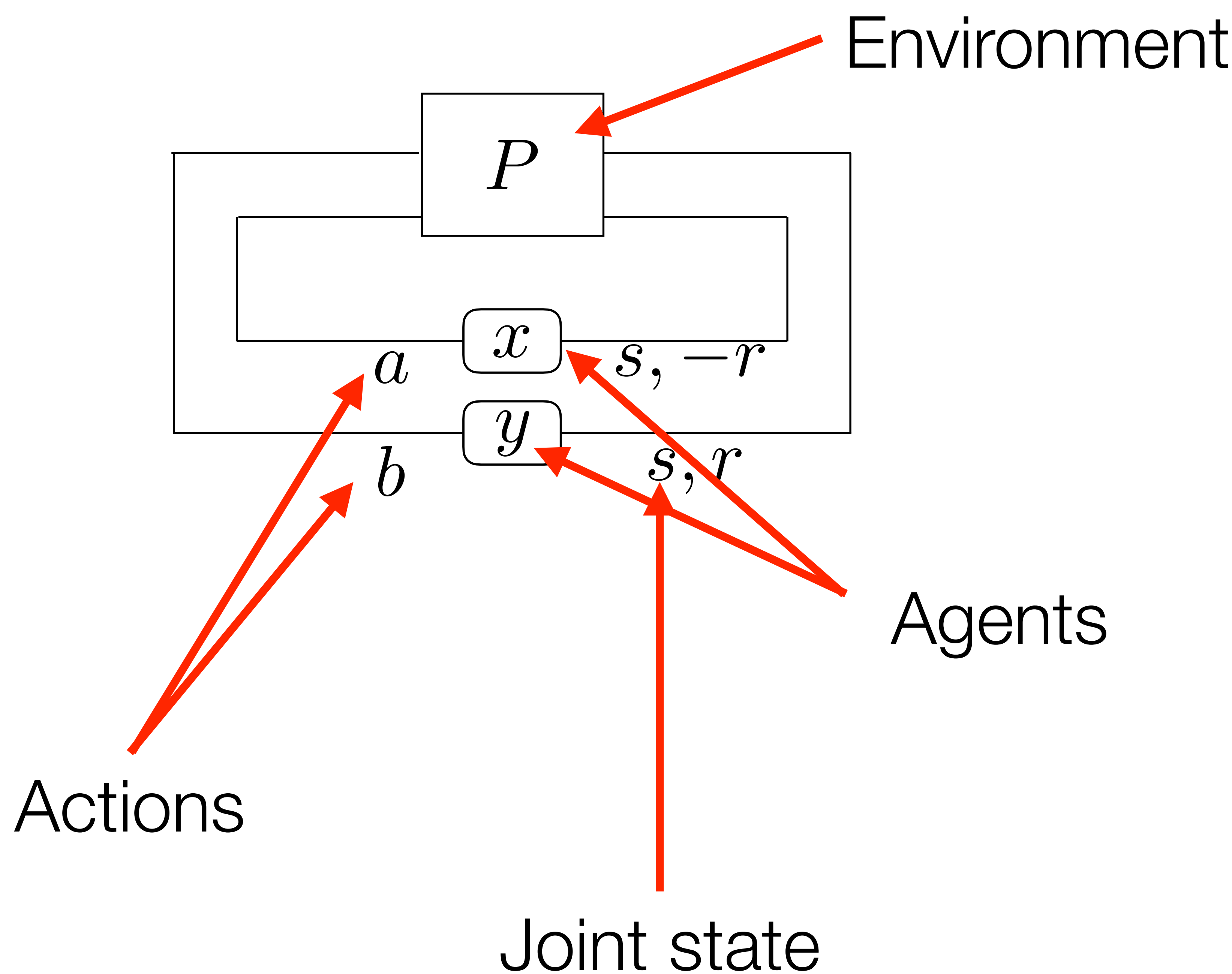
A Natural Actor-Critic Framework for Zero-Sum Markov Games

Ahmet Alacaoglu, UW-Madison
Luca Viano, EPFL
Niao He, ETH Zurich
Volkan Cevher, EPFL

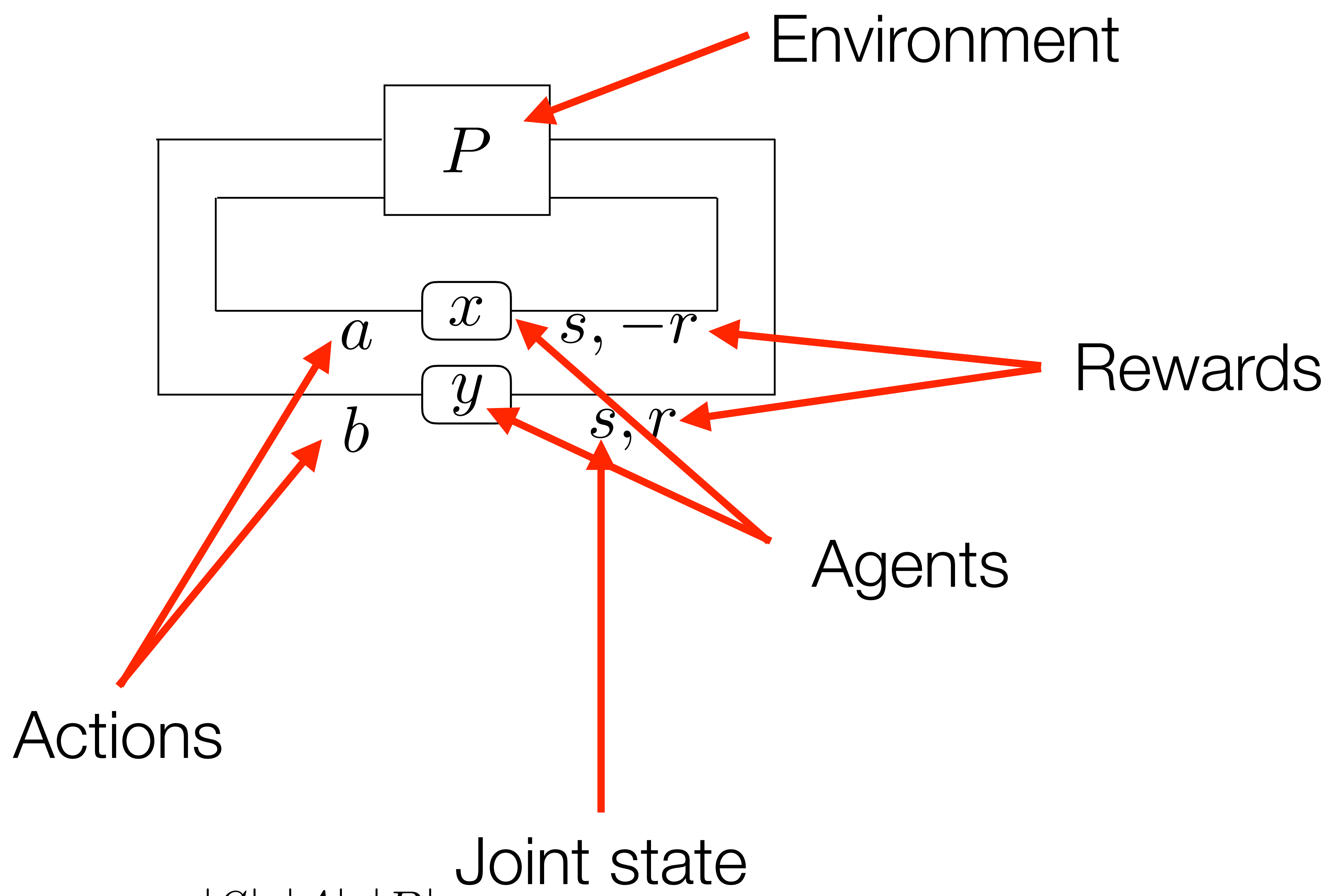
Setup: Markov Games



Setup: Markov Games



Setup: Markov Games



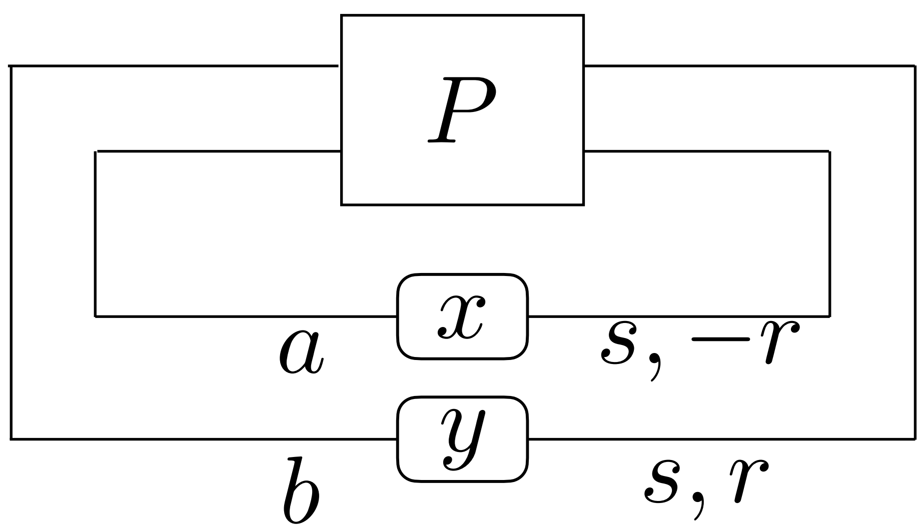
finite state/action spaces: $|S|, |A|, |B|$
"tabular case"

Setup: Markov Games

Markov games: $$\min_{x \in \Delta} \max_{y \in \Delta} \mathbb{E}_{s \sim \rho_0} V^{x,y}(s)$$

$$V^{x,y}(s) = \mathbb{E}_{x,y}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) | s_0 = s\right],$$

with $a_t \sim x(\cdot|s_t), b_t \sim y(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t, b_t)$

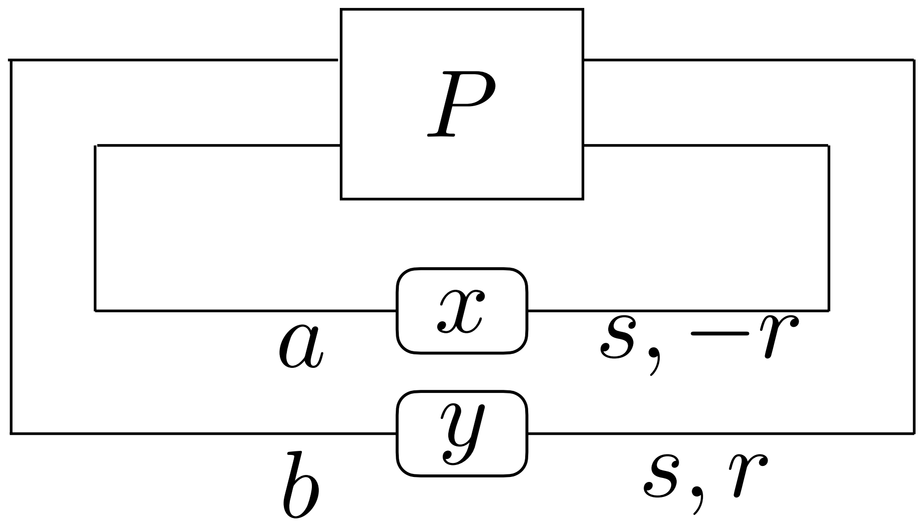


Setup: Markov Games

Markov games: $$\min_{x\in\Delta}\max_{y\in\Delta}\mathbb{E}_{s\sim\rho_0}V^{x,y}(s)$$

$$V^{x,y}(s)=\mathbb{E}_{x,y}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,b_t)|s_0=s\right],$$

with $a_t\sim x(\cdot|s_t), b_t\sim y(\cdot|s_t), s_{t+1}\sim P(\cdot|s_t,a_t,b_t)$



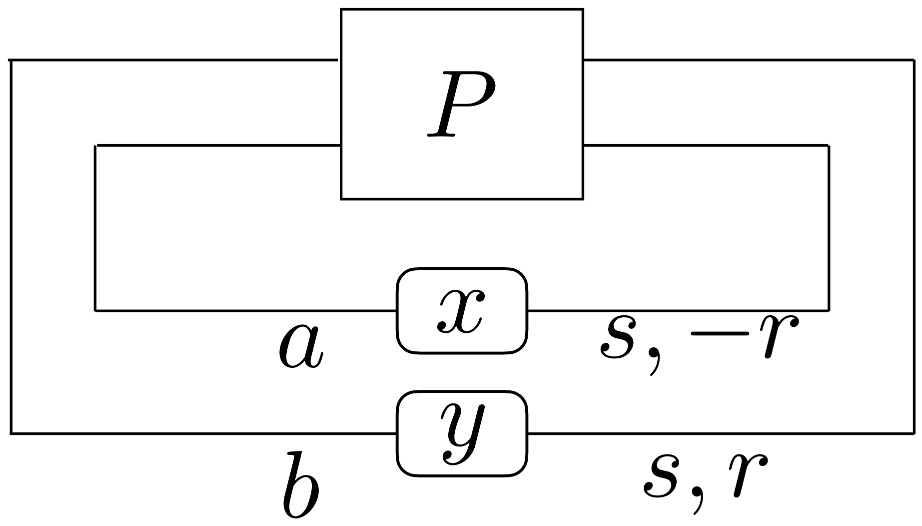
Setup: Markov Games

Markov games: $$\min_{x \in \Delta} \max_{y \in \Delta} \mathbb{E}_{s \sim \rho_0} V^{x,y}(s)$$

Nonconvex-nonconcave

$$V^{x,y}(s) = \mathbb{E}_{x,y}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) | s_0 = s\right],$$

with $a_t \sim x(\cdot|s_t), b_t \sim y(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t, b_t)$



Setup: Markov Games

Markov games: $$\min_{x \in \Delta} \max_{y \in \Delta} \mathbb{E}_{s \sim \rho_0} V^{x,y}(s)$$

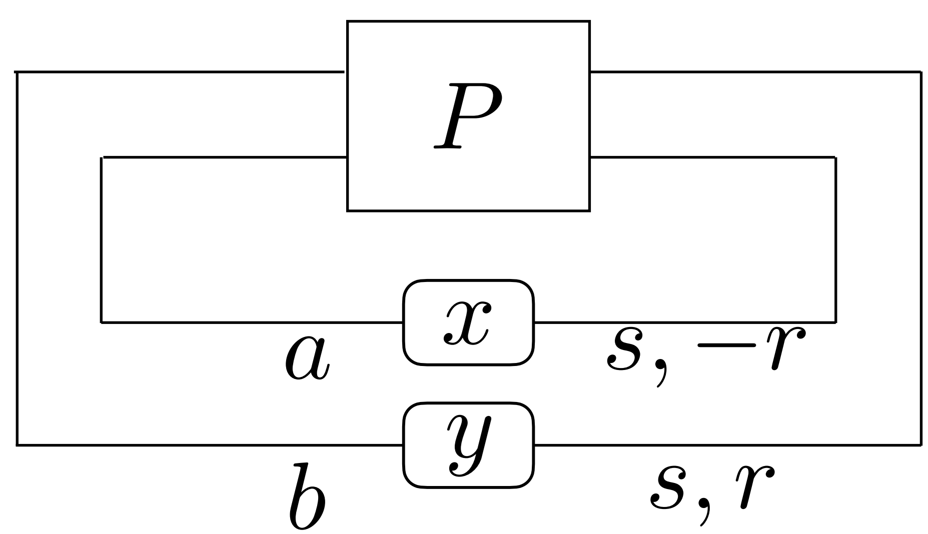
Nonconvex-nonconcave

$$V^{x,y}(s) = \mathbb{E}_{x,y}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) | s_0 = s\right],$$

with $a_t \sim x(\cdot|s_t), b_t \sim y(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t, b_t)$

still tractable

ε-Nash eq.



Agents only access their own actions

Setup: Markov Games

Markov games: $$\min_{x\in\Delta}\max_{y\in\Delta}\mathbb{E}_{s\sim\rho_0}V^{x,y}(s)$$

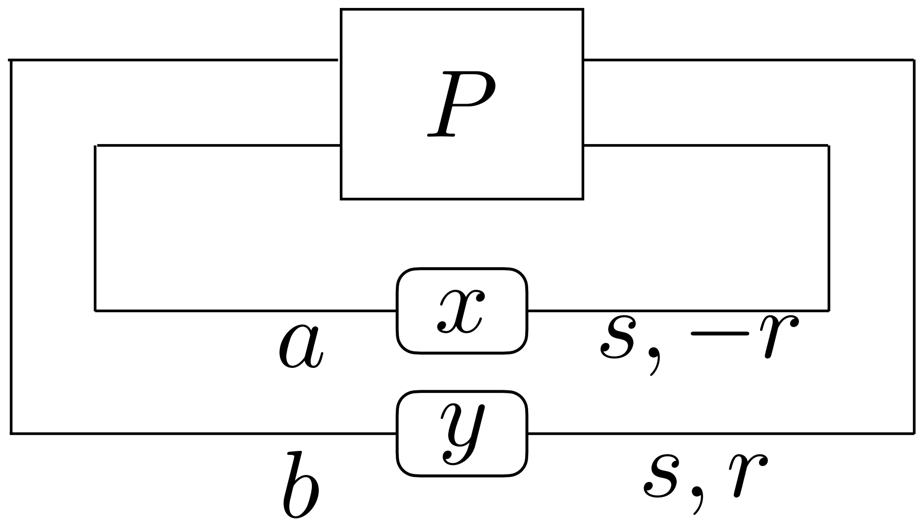
Nonconvex-nonconcave

$$V^{x,y}(s)=\mathbb{E}_{x,y}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,b_t)|s_0=s\right],$$

with $a_t\sim x(\cdot|s_t), b_t\sim y(\cdot|s_t), s_{t+1}\sim P(\cdot|s_t,a_t,b_t)$

still tractable

ε-Nash eq.



Sample Complexity Results

Solve as a sequence of single-agent MDP and zero-sum matrix game

Policy mirror descent + stochastic FoRB

$\mathcal{O}(\varepsilon^{-2})$ ← ε-Nash eq.

matching the sample complexity
of solving single-agent MDP

Sample Complexity Results

Solve as a sequence of single-agent MDP and zero-sum matrix game

Policy mirror descent + stochastic FoRB

$\mathcal{O}(\varepsilon^{-2})$ ← ε-Nash eq.

matching the sample complexity
of solving single-agent MDP

Careful analysis of the bias
between the two stages

Sample Complexity Results

Solve as a sequence of single-agent MDP and zero-sum matrix game

Policy mirror descent + stochastic FoRB

$\mathcal{O}(\varepsilon^{-2})$ ← ε-Nash eq.

matching the sample complexity
of solving single-agent MDP

Careful analysis of the bias
between the two stages

Assumption: lower bounded policies (same as single agent)

Sample Complexity Results

Solve as a sequence of single-agent MDP and zero-sum matrix game

Policy mirror descent + stochastic FoRB

$\mathcal{O}(\varepsilon^{-4})$ ← ε-Nash eq.

~~Assumption: lower bounded policies (same as single agent)~~

use greedy exploration

Sample Complexity Results

Solve as a sequence of single-agent MDP and zero-sum matrix game

Policy mirror descent + stochastic FoRB

$\mathcal{O}(\varepsilon^{-4})$ ← ε-Nash eq.

Better dependence on $|S|, |A|, |B|, \gamma$ than SOTA

~~Assumption: lower bounded policies (same as single agent)~~

use greedy exploration