Safety-Aware Algorithms for Adversarial Contextual Bandits

Wen Sun
Carnegie Mellon University, USA

Debadeepta Dey, and Ashish Kapoor
Microsoft Research, Redmond, USA

The Robotics Institute
Carnegie Mellon University

Microsoft Research

Control & Sequential Decision Making

[Georgia Tech Autorally Project, autorally.github.io]

Control & Sequential Decision Making

[Georgia Tech Autorally Project, autorally.github.io]

[Silver, et.al, 2016, Nature]

Control & Sequential Decision Making

[Georgia Tech Autorally Project, autorally.github.io]

[Silver, et.al, 2016, Nature]

Control & Sequential Decision Making

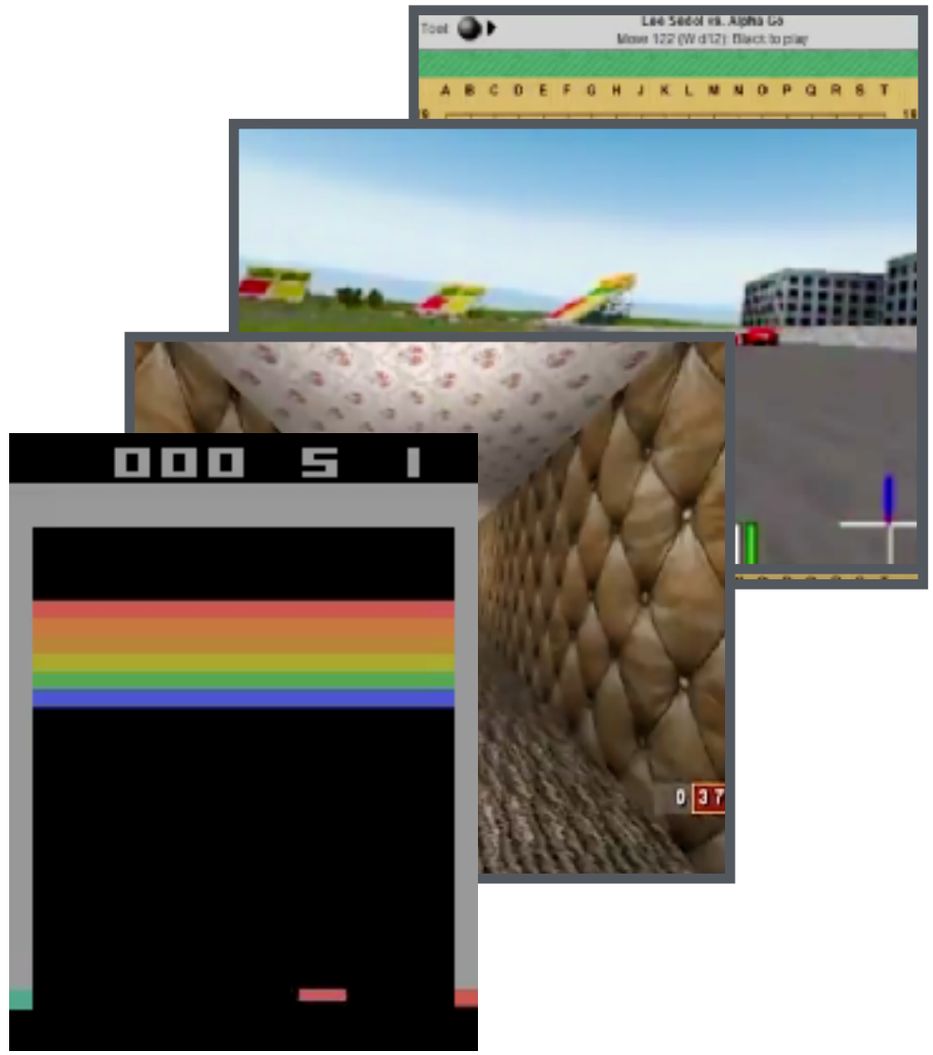
[Georgia Tech Autorally Project, autorally.github.io]

[Silver, et.al, 2016, Nature]

Control & Sequential Decision Making

[Georgia Tech Autorally Project, autorally.github.io]

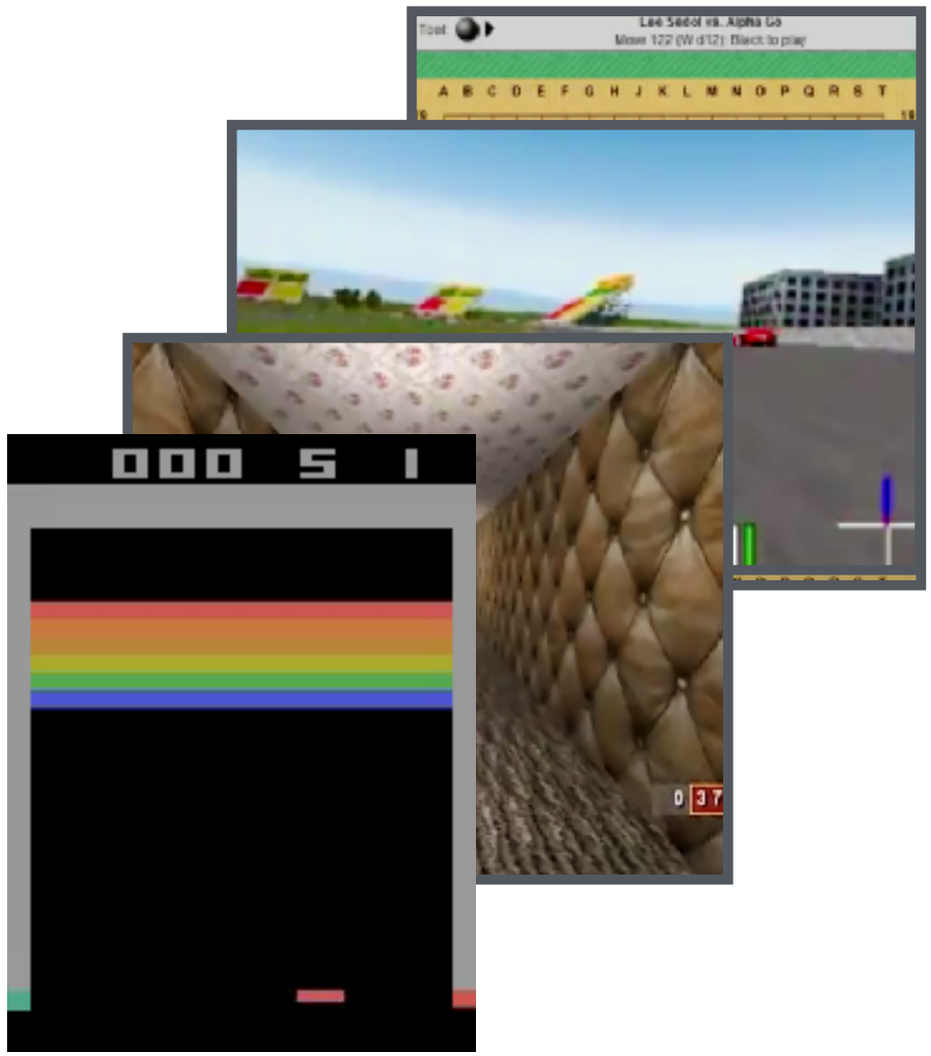
[Silver, et.al, 2016, Nature]



Control & Sequential Decision Making

[Georgia Tech Autorally Project, autorally.github.io]

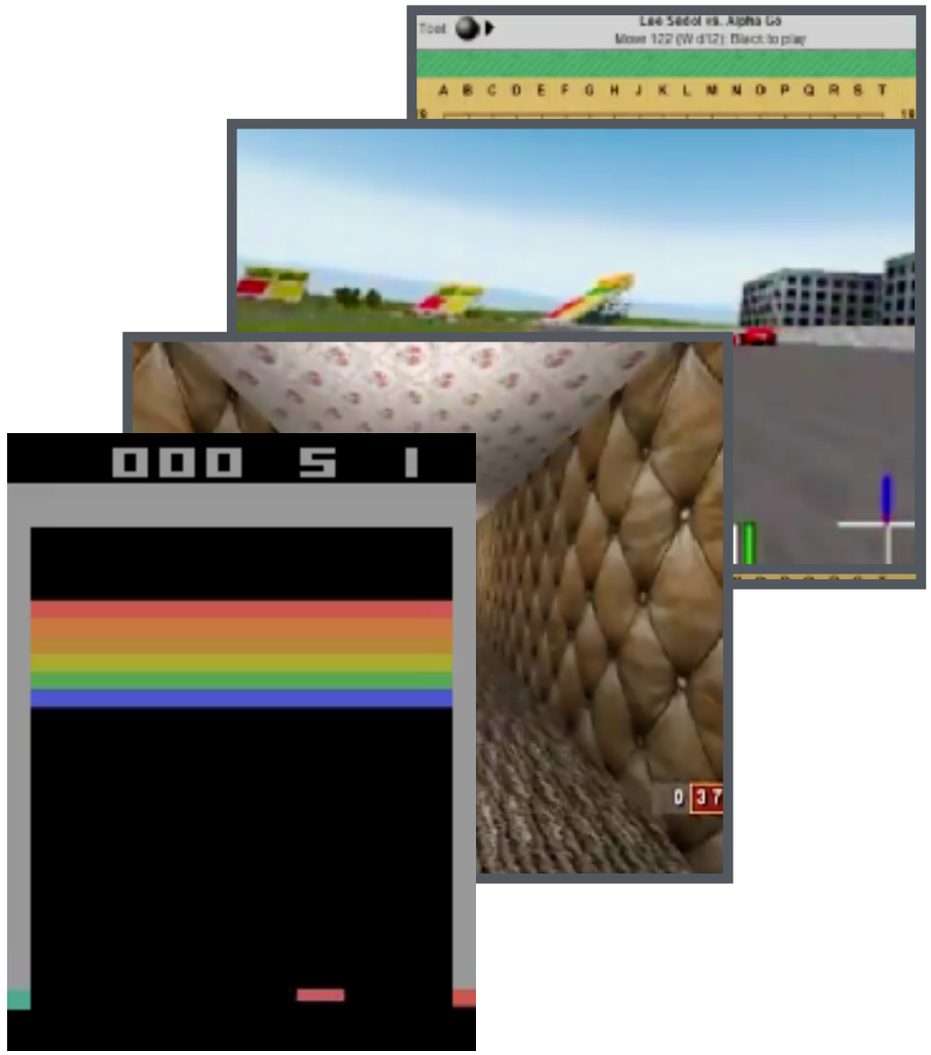
[Silver, et.al, 2016, Nature]



Control & Sequential Decision Making

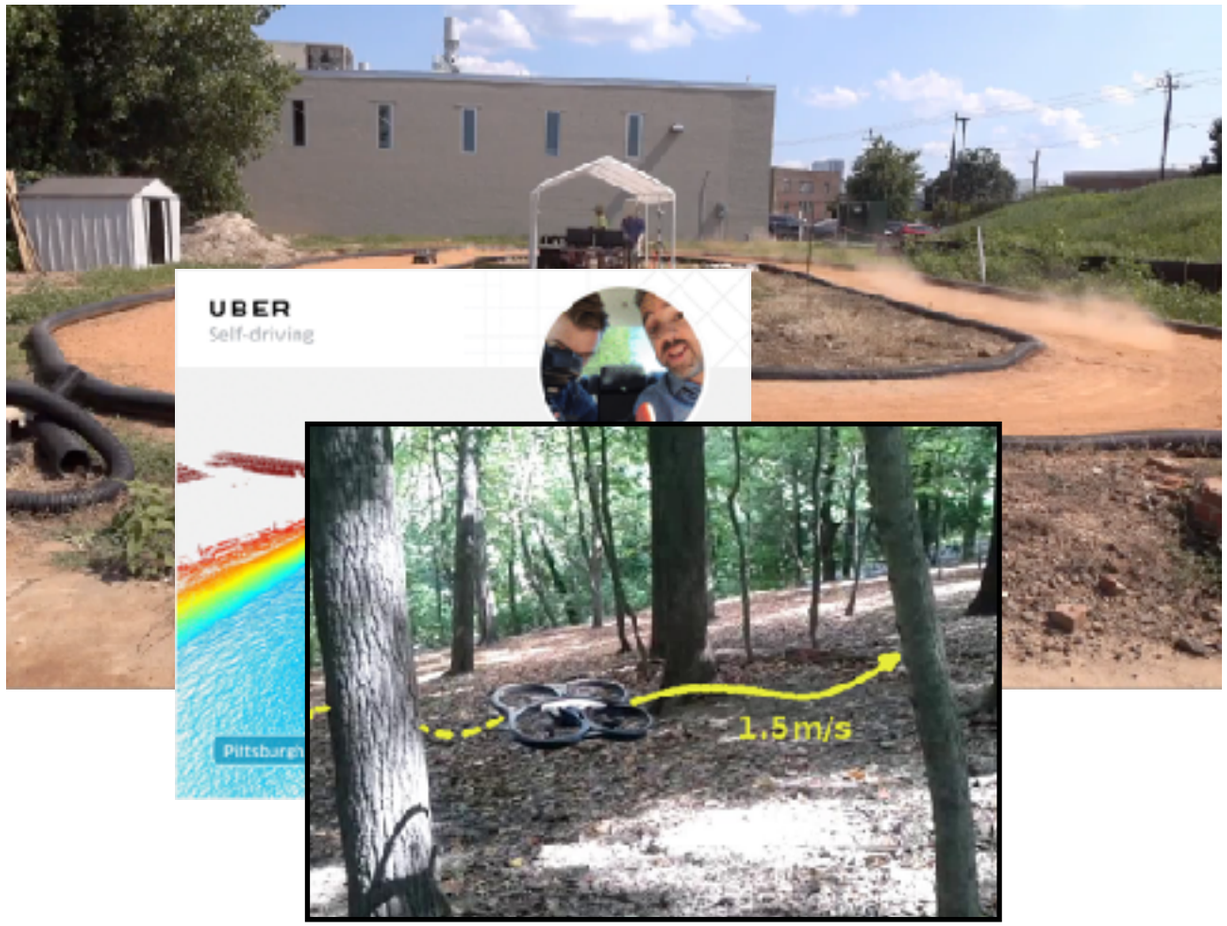
[Georgia Tech Autorally Project, autorally.github.io]

[Silver, et.al, 2016, Nature]

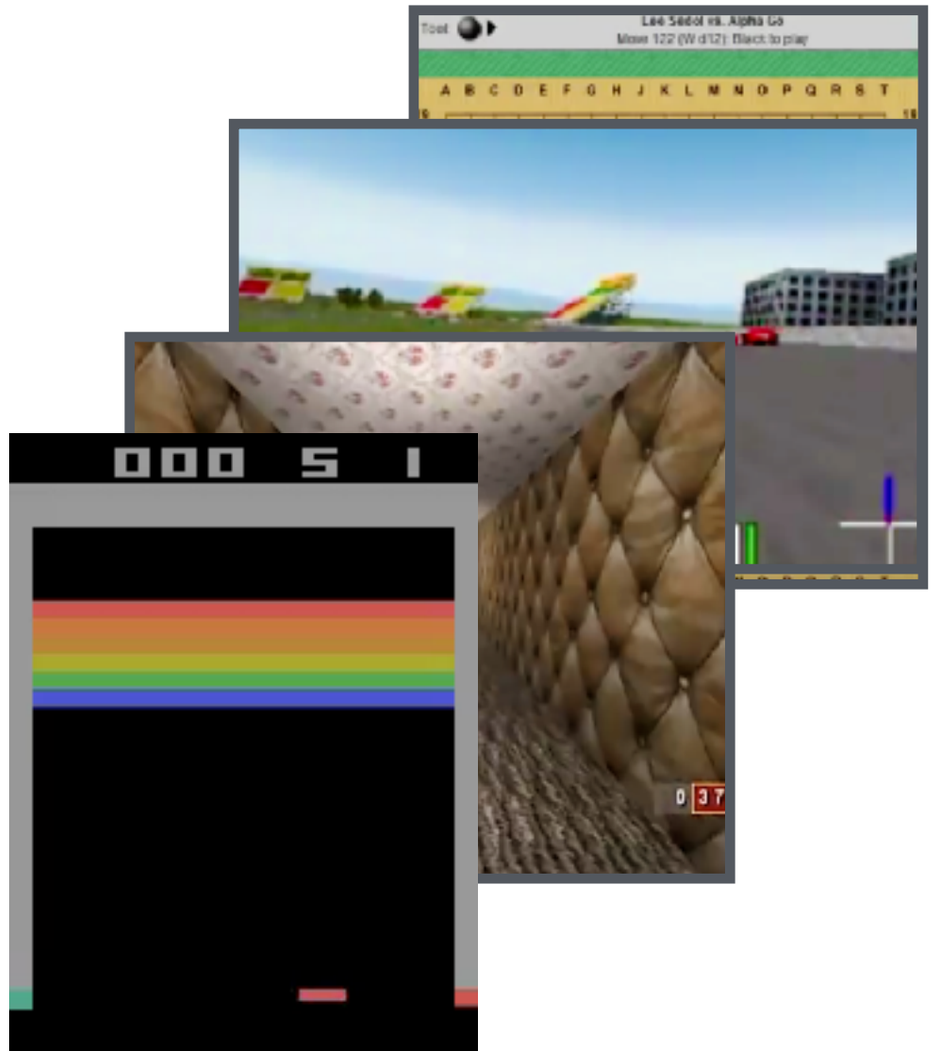


Control & Sequential Decision Making

[Georgia Tech Autorally Project, autorally.github.io]



[Silver, et.al, 2016, Nature]

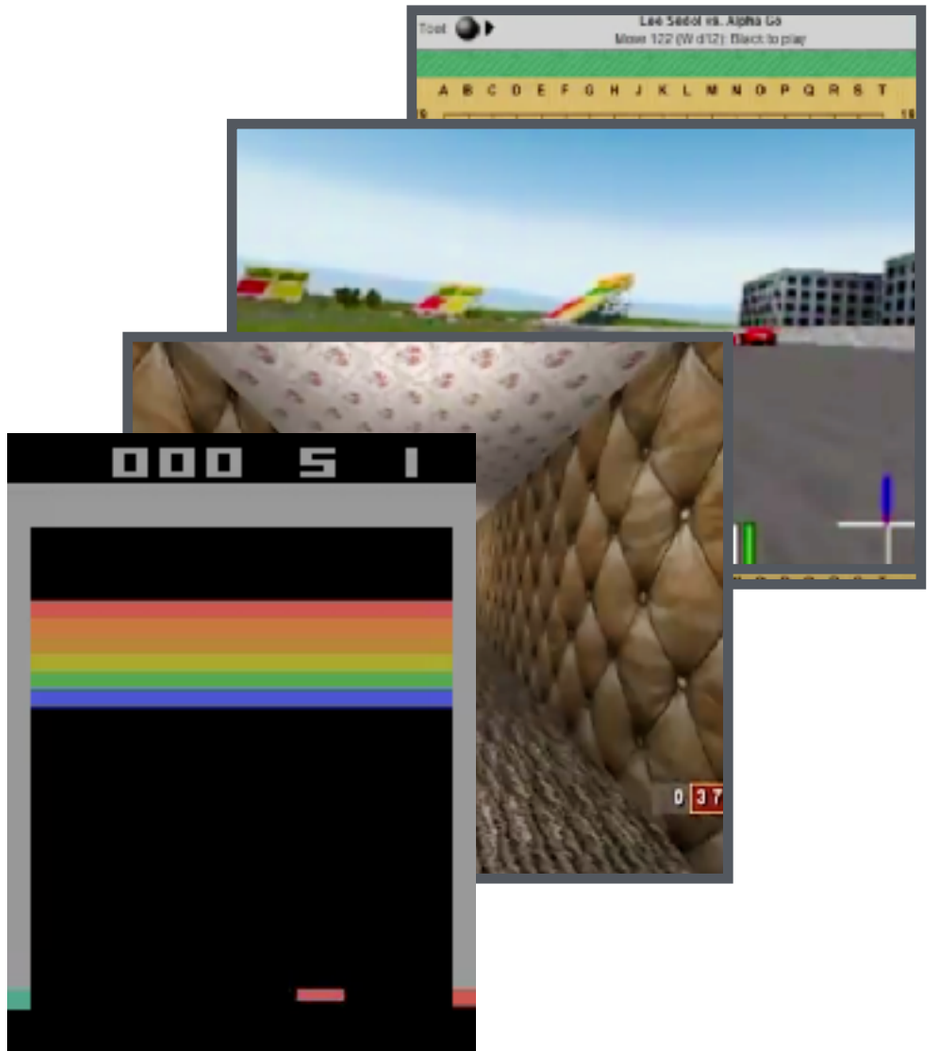


Control & Sequential Decision Making

[Georgia Tech Autorally Project, autorally.github.io]

[Coates et.al,08,ICML]

[Silver, et.al, 2016, Nature]

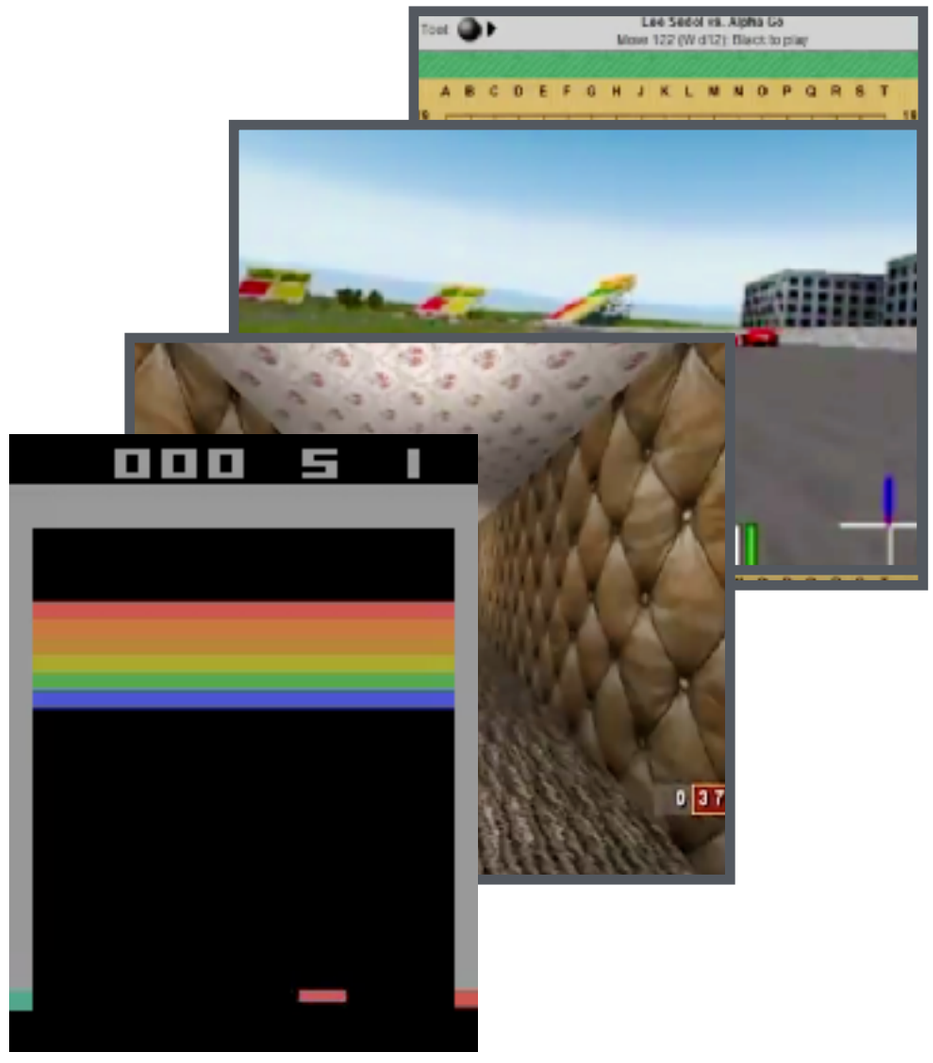


Control & Sequential Decision Making

[Georgia Tech Autorally Project, autorally.github.io]

[Coates et.al,08,ICML]

[Silver, et.al, 2016, Nature]



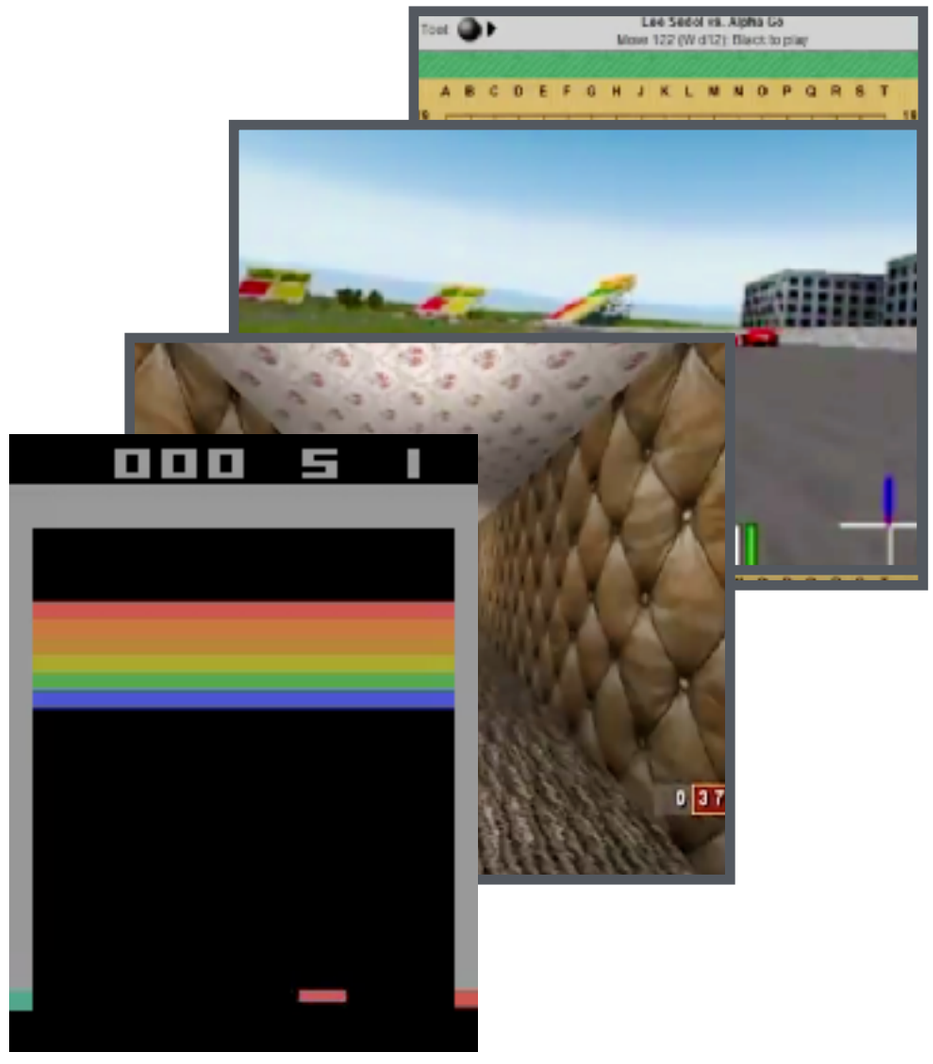
Safety

Control & Sequential Decision Making

[Georgia Tech Autorally Project, autorally.github.io]

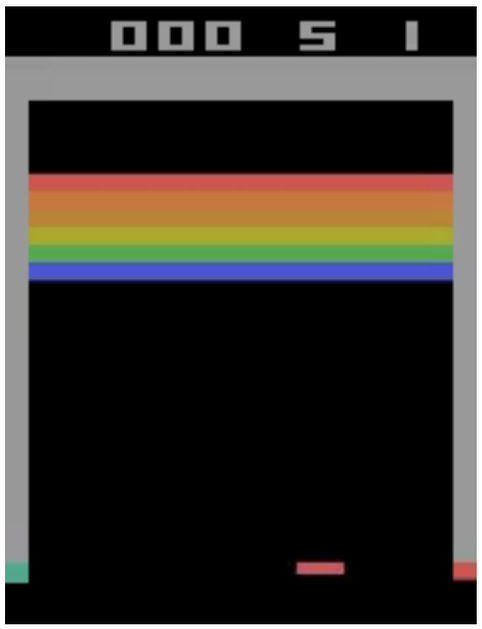
[Coates et.al,08,ICML]

[Silver, et.al, 2016, Nature]

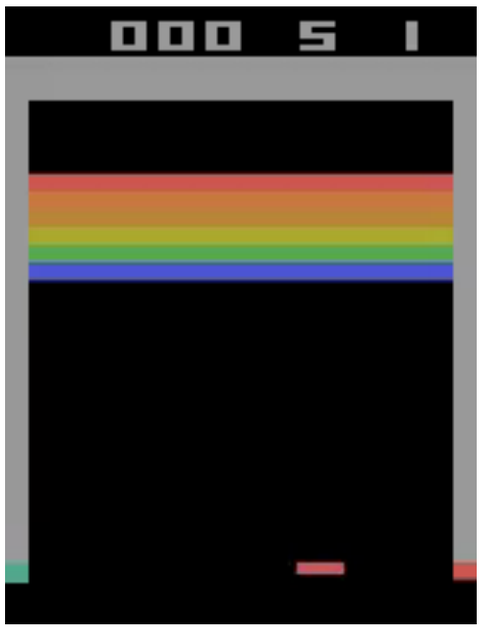


Safety

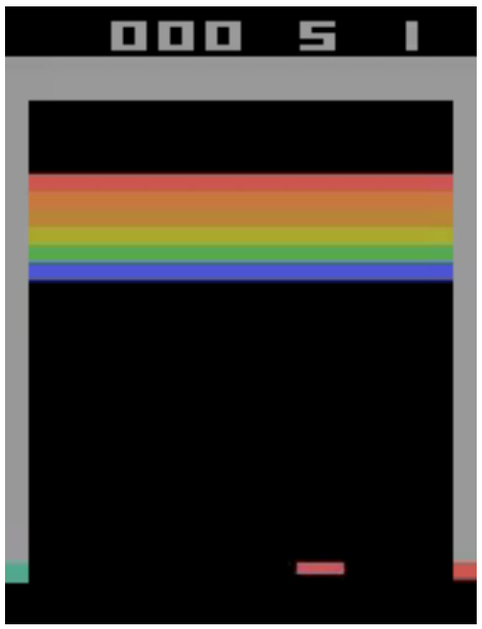
Control & Sequential Decision Making



Control & Sequential Decision Making

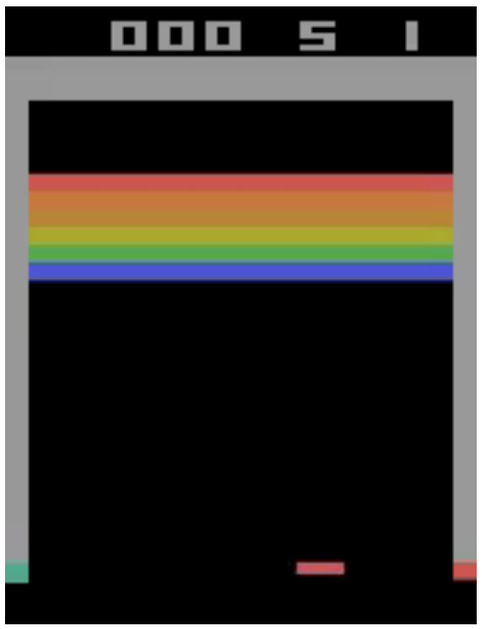


Control & Sequential Decision Making



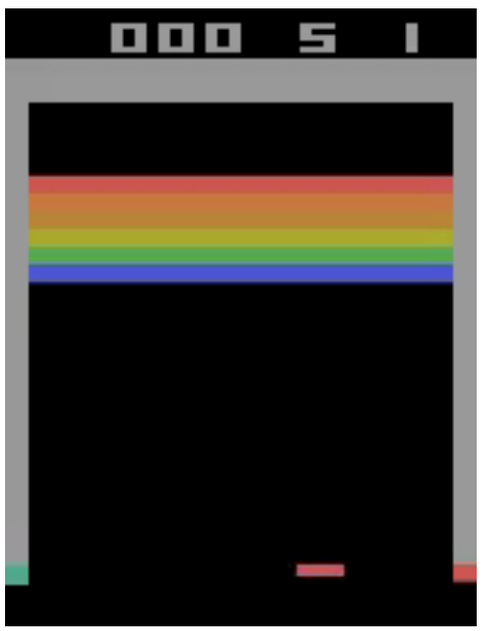
Classic Exploration & Exploitation
(e.g., epsilon greedy, upper confidence bound)
Not enough!

Control & Sequential Decision Making



Classic Exploration & Exploitation
(e.g., epsilon greedy, upper confidence bound)
Not enough!

Control & Sequential Decision Making



Classic Exploration & Exploitation
(e.g., epsilon greedy, upper confidence bound)
Not enough!

Explore more carefully…
Risk of failures
Energy exhaustion
Side effect of a treatment in clinical trial
…

This Work:

We attempt to model this problem in the contextual bandit setting

We introduce extra risk associated with each action

The goal is to maintain small regret for reward while ensuring the cumulative risk is small

Setting:

Context (i.e., features): $s_t \in \mathcal{S}$

Actions (finitely many): $a_t \in [K]$

Setting:

Context (i.e., features): $s_t \in \mathcal{S}$

Actions (finitely many): $a_t \in [K]$

Cost vector: $c_t \in [0,1]^K$

Setting:

Context (i.e., features): $s_t \in \mathcal{S}$

Actions (finitely many): $a_t \in [K]$

Cost vector: $c_t \in [0,1]^K$ and *Risk vector*: $r_t \in [0,1]^K$

Setting:

Context (i.e., features): $s_t \in \mathcal{S}$

Actions (finitely many): $a_t \in [K]$

Cost vector: $c_t \in [0,1]^K$ and *Risk vector*: $r_t \in [0,1]^K$

Pre-Defined Risk Threshold: $\beta \in [0,1]$

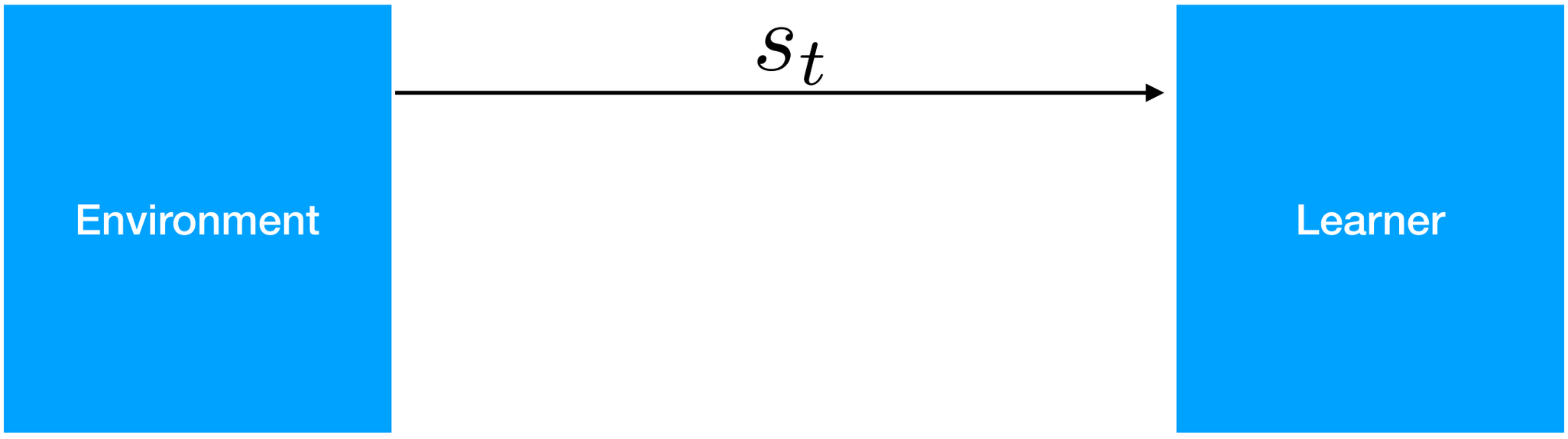
Setting:

Context (i.e., features): $s_t \in \mathcal{S}$

Actions (finitely many): $a_t \in [K]$

Cost vector: $c_t \in [0,1]^K$ and *Risk vector*: $r_t \in [0,1]^K$

Pre-Defined Risk Threshold: $\beta \in [0,1]$



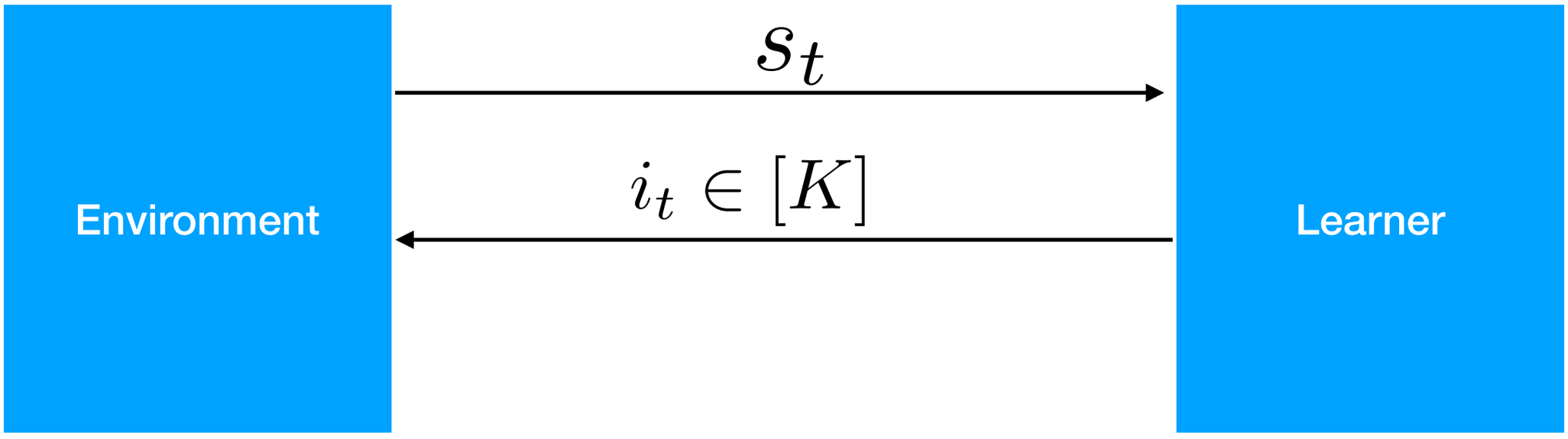
Setting:

Context (i.e., features): $s_t \in \mathcal{S}$

Actions (finitely many): $a_t \in [K]$

Cost vector: $c_t \in [0,1]^K$ and *Risk vector*: $r_t \in [0,1]^K$

Pre-Defined Risk Threshold: $\beta \in [0,1]$



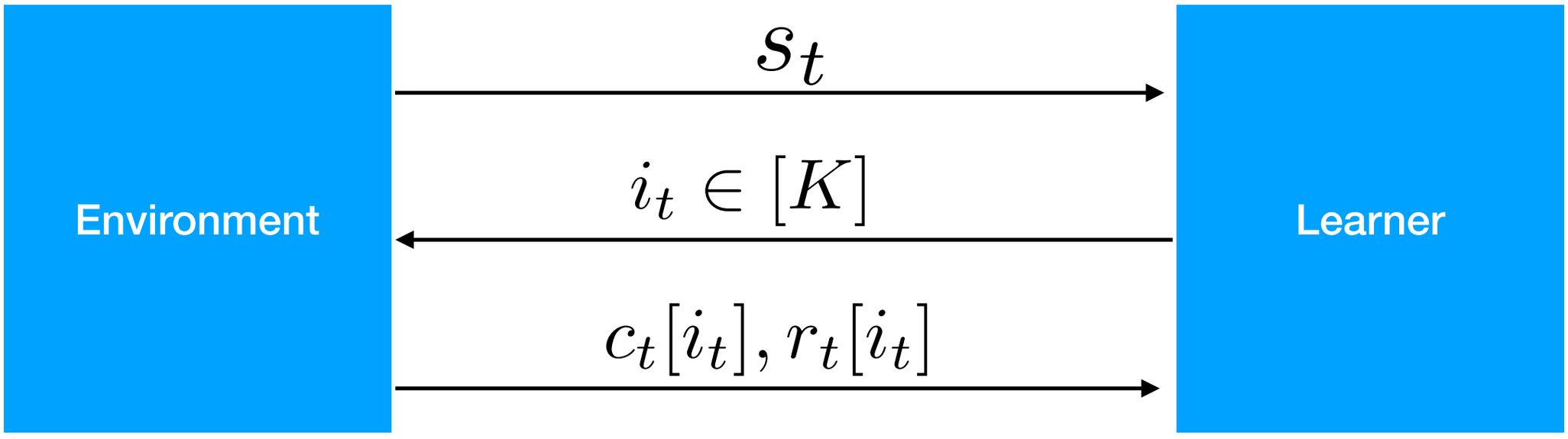
Setting:

Context (i.e., features): $s_t \in \mathcal{S}$

Actions (finitely many): $a_t \in [K]$

Cost vector: $c_t \in [0,1]^K$ and *Risk vector*: $r_t \in [0,1]^K$

Pre-Defined Risk Threshold: $\beta \in [0,1]$



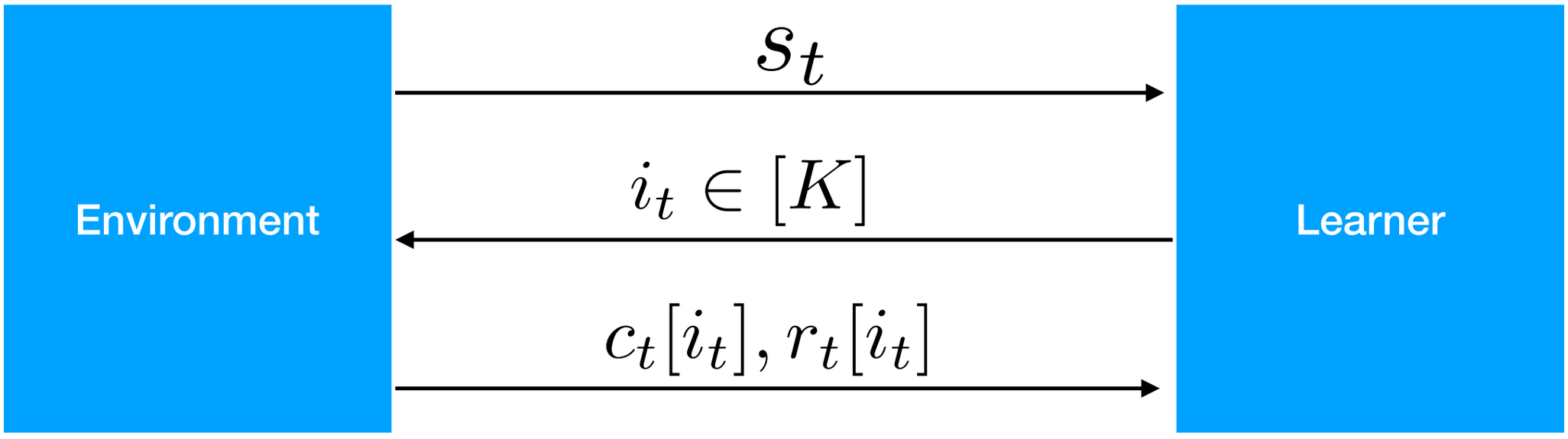
Setting:

Context (i.e., features): $s_t \in \mathcal{S}$

Actions (finitely many): $a_t \in [K]$

Cost vector: $c_t \in [0,1]^K$ and *Risk vector*: $r_t \in [0,1]^K$

Pre-Defined Risk Threshold: $\beta \in [0,1]$



No statistical assumptions on the generation of context, cost or risk….

Goal:

Ideally:

$$\min_{i_1,\ldots,i_T} \sum_{t=1}^{T} c_t[i_t] - \min_{\pi^*} \sum_{t=1}^{T} c_t[\pi^*(s_t)], \quad s.t., r_t[i_t] \leq \beta, \forall t$$

Goal:

Ideally:

$$\min_{i_1,\ldots,i_T} \sum_{t=1}^{T} c_t[i_t] - \min_{\pi^*} \sum_{t=1}^{T} c_t[\pi^*(s_t)], \quad s.t., r_t[i_t] \leq \beta, \forall t$$

More realistically:

$$\min_{i_1,\ldots,i_T} \sum_{t=1}^{T} c_t[i_t] - \min_{\pi^*} \sum_{t=1}^{T} c_t[\pi^*(s_t)], \quad s.t., \sum_{t=1}^{T} (r_t[i_t] - \beta) \leq o(T)$$

Goal:

Ideally:

$$\min_{i_1,\ldots,i_T} \sum_{t=1}^{T} c_t[i_t] - \min_{\pi^*} \sum_{t=1}^{T} c_t[\pi^*(s_t)], \quad s.t., r_t[i_t] \leq \beta, \forall t$$

More realistically:

$$\min_{i_1,\ldots,i_T} \sum_{t=1}^{T} c_t[i_t] - \min_{\pi^*} \sum_{t=1}^{T} c_t[\pi^*(s_t)], \quad s.t., \sum_{t=1}^{T} (r_t[i_t] - \beta) \leq o(T)$$

No constraint violation in a long-term perspective!

Full Information OCP Setting

Decision: $x \in \mathcal{X}$

Convex Loss: $\ell_t(x)$

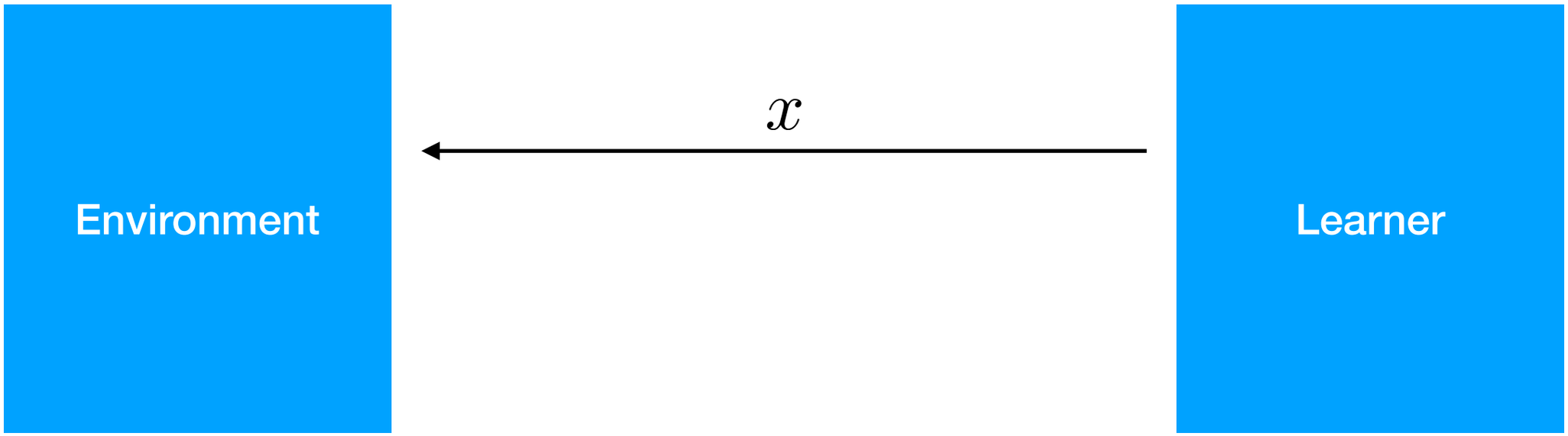
Convex Constraint: $f_t(x) \leq 0$

Full Information OCP Setting

Decision: $x \in \mathcal{X}$

Convex Loss: $\ell_t(x)$

Convex Constraint: $f_t(x) \leq 0$

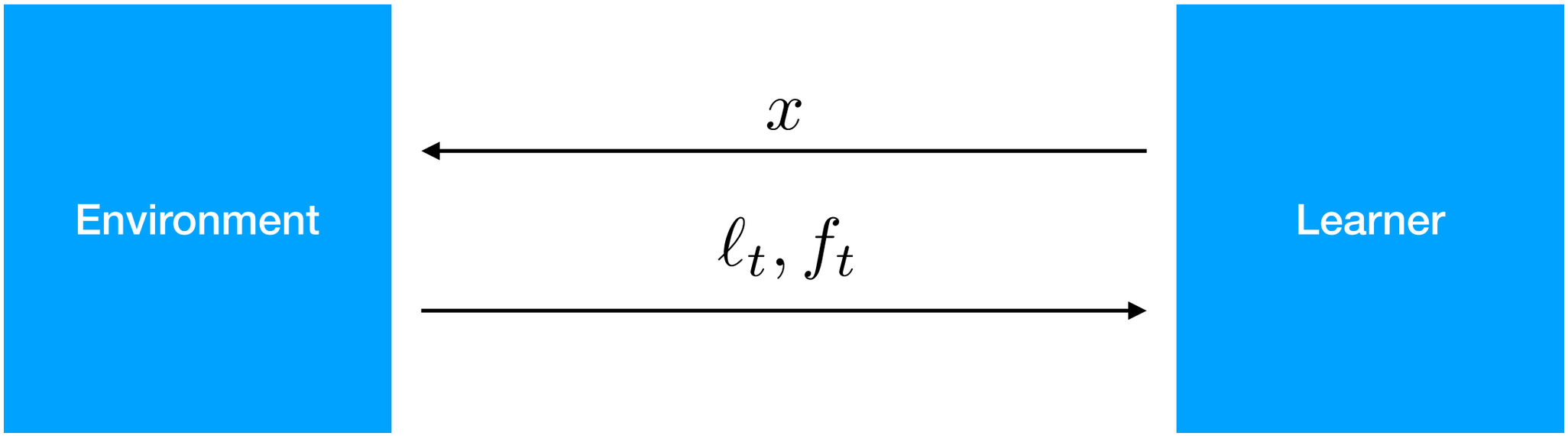


Full Information OCP Setting

Decision: $x \in \mathcal{X}$

Convex Loss: $\ell_t(x)$

Convex Constraint: $f_t(x) \leq 0$

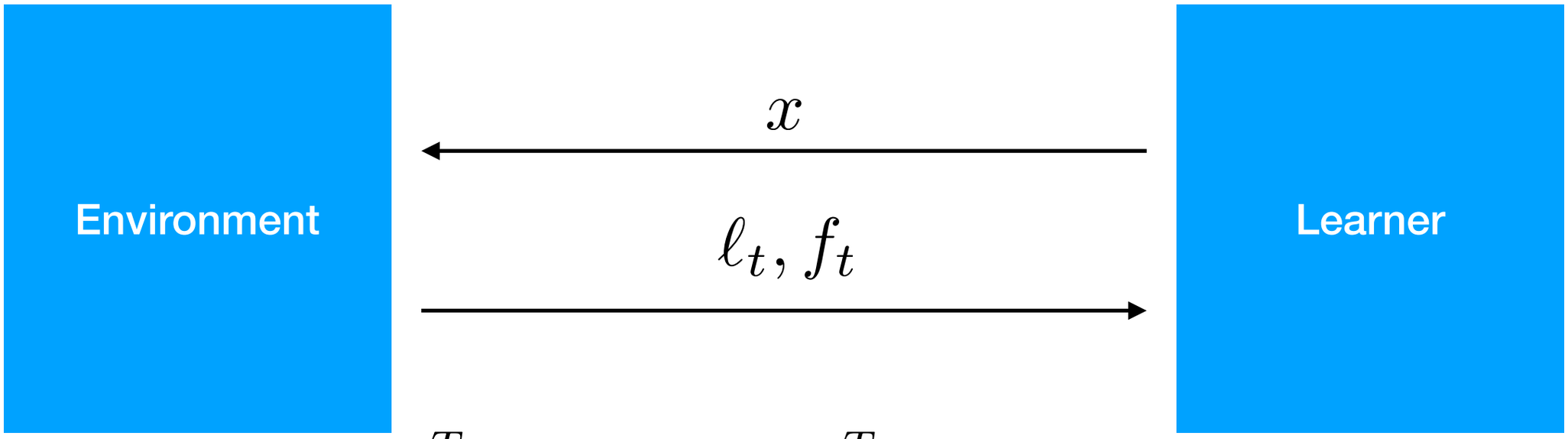


Full Information OCP Setting

Decision: $x \in \mathcal{X}$

Convex Loss: $\ell_t(x)$

Convex Constraint: $f_t(x) \leq 0$



$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^*} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

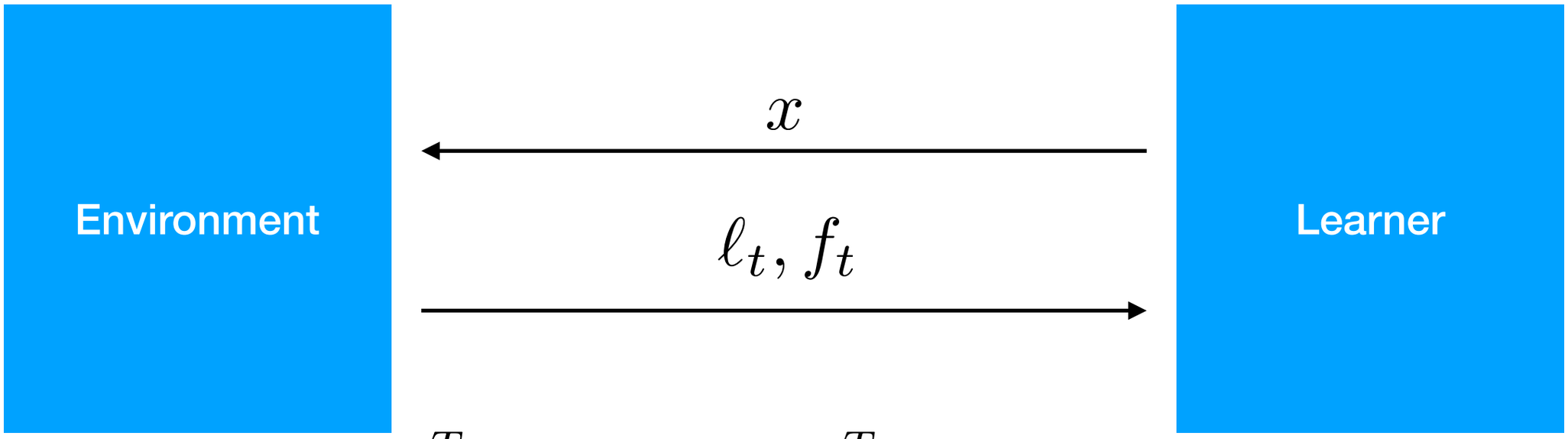
$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

Full Information OCP Setting

Decision: $x \in \mathcal{X}$

Convex Loss: $\ell_t(x)$

Convex Constraint: $f_t(x) \leq 0$



$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^*} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

No constraint violation in a long-term perspective!

Competing Against…

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^*} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

Competing Against…

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^*} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

Competing Against…

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^*} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

Option 1: $\mathcal{O}' = \{x \in \mathcal{X} : \sum_{t=1}^{T} f_t(x) \leq 0\}$

Competing Against…

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^*} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

Option 1: $\mathcal{O}' = \{x \in \mathcal{X} : \sum_{t=1}^{T} f_t(x) \leq 0\}$ Decisions satisfy average constraint

Competing Against…

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^*} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

Option 1: $\mathcal{O}' = \{x \in \mathcal{X} : \sum_{t=1}^{T} f_t(x) \leq 0\}$ Decisions satisfy average constraint

Option 2: $\mathcal{O} = \{x \in \mathcal{X} : f_t(x) \leq 0, \forall t\}$

Competing Against…

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^*} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

Option 1: $\mathcal{O}' = \{x \in \mathcal{X} : \sum_{t=1}^{T} f_t(x) \leq 0\}$ Decisions satisfy average constraint

Option 2: $\mathcal{O} = \{x \in \mathcal{X} : f_t(x) \leq 0, \forall t\}$ Decisions satisfy every constraint

Competing Against…

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^*} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

Option 1: $\mathcal{O}' = \{x \in \mathcal{X} : \sum_{t=1}^{T} f_t(x) \leq 0\}$ Decisions satisfy average constraint

Option 2: $\mathcal{O} = \{x \in \mathcal{X} : f_t(x) \leq 0, \forall t\}$ Decisions satisfy every constraint

Option 1 seems quite natural…..

Competing Against…

Competing Against…

Option 1: $\mathcal{O}' = \{x \in \mathcal{X} : \sum_{t=1}^{T} f_t(x) \leq 0\}$

Competing Against…

Option 1: $\mathcal{O}' = \{x \in \mathcal{X} : \sum_{t=1}^{T} f_t(x) \leq 0\}$

Claim: there exist a sequence of loss and constraints such that for any sequence of decisions that satisfies the average constraint:

$$\limsup_{t\to\infty} \sum_{i=1}^{t} f_i(x_i)/t \leq 0,$$

then, the regret grows linearly when competing against $\mathcal{O}'$*:*

$$\limsup_{t\to\infty} (\sum_{i=1}^{t} \ell_i(x_i) - \min_{x^* \in \mathcal{O}'} \sum_{i=1}^{t} \ell_i(x^*)) = \Omega(t)$$

Competing Against…

Option 1: $\mathcal{O}' = \{x \in \mathcal{X} : \sum_{t=1}^{T} f_t(x) \leq 0\}$

Claim: there exist a sequence of loss and constraints such that for any sequence of decisions that satisfies the average constraint:

$$\limsup_{t\to\infty} \sum_{i=1}^{t} f_i(x_i)/t \leq 0,$$

then, the regret grows linearly when competing against $\mathcal{O}'$*:*

$$\limsup_{t\to\infty} (\sum_{i=1}^{t} \ell_i(x_i) - \min_{x^* \in \mathcal{O}'} \sum_{i=1}^{t} \ell_i(x^*)) = \Omega(t)$$

(Construction adapts a discrete two-player game in [Mannor, et.al, 09])

Algorithm

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^* \in \mathcal{O}} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

where $\mathcal{O} = \{x \in \mathcal{X} : f_t(x) \leq 0, \forall t\}$

Algorithm

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^* \in \mathcal{O}} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

where $\mathcal{O} = \{x \in \mathcal{X} : f_t(x) \leq 0, \forall t\}$

Convex-Concave formulation [Mahdavi et al.,2012]

$$\mathcal{L}_t(x, \lambda) = \ell_t(x) + \lambda f_t(x) - \frac{\delta \mu}{2} \lambda^2, \quad \delta \in \mathbb{R}^+$$

Algorithm

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^* \in \mathcal{O}} \sum_{t=1}^{T} \ell_t(x^*) \leq o(T)$$

$$s.t., \sum_{t=1}^{T} f_t(x_t) \leq o(T)$$

where $\mathcal{O} = \{x \in \mathcal{X} : f_t(x) \leq 0, \forall t\}$

Convex-Concave formulation [Mahdavi et al.,2012]

$$\mathcal{L}_t(x, \lambda) = \ell_t(x) + \lambda f_t(x) - \frac{\delta \mu}{2} \lambda^2, \quad \delta \in \mathbb{R}^+$$

$\lambda \in \mathbb{R}^+$ **dual variable**

$\lambda \in [0, \infty)$

Algorithm

Strongly convex regularizer $R(x)$

Reduction to Online Mirror Descent + Online Gradient Ascent

Algorithm

Strongly convex regularizer $R(x)$

Reduction to Online Mirror Descent + Online Gradient Ascent

At iteration t, x_t, λ_t

Algorithm

Strongly convex regularizer $R(x)$

Reduction to Online Mirror Descent + Online Gradient Ascent

At iteration t, x_t, λ_t

$$\mathcal{L}_t(x, \lambda) = \ell_t(x) + \lambda f_t(x) - \frac{\delta\mu}{2}\lambda^2, \quad \delta \in \mathbb{R}^+$$

Algorithm

Strongly convex regularizer $R(x)$

Reduction to Online Mirror Descent + Online Gradient Ascent

At iteration t, x_t, λ_t

$$\mathcal{L}_t(x, \lambda) = \ell_t(x) + \lambda f_t(x) - \frac{\delta\mu}{2}\lambda^2, \quad \delta \in \mathbb{R}^+$$

$$\mathcal{L}_t(x, \boxed{\lambda_t})$$

Algorithm

Strongly convex regularizer $R(x)$

Reduction to Online Mirror Descent + Online Gradient Ascent

At iteration t, x_t, λ_t

$$\mathcal{L}_t(x, \lambda) = \ell_t(x) + \lambda f_t(x) - \frac{\delta\mu}{2}\lambda^2, \quad \delta \in \mathbb{R}^+$$

$\mathcal{L}_t(x, \lambda_t)$

OMD

Algorithm

Strongly convex regularizer $R(x)$

Reduction to Online Mirror Descent + Online Gradient Ascent

At iteration t, x_t, λ_t

$$\mathcal{L}_t(x, \lambda) = \ell_t(x) + \lambda f_t(x) - \frac{\delta\mu}{2}\lambda^2, \quad \delta \in \mathbb{R}^+$$

$\mathcal{L}_t(x, \lambda_t)$

$\mathcal{L}_t(x_t, \lambda)$

OMD

Algorithm

Strongly convex regularizer $R(x)$

Reduction to Online Mirror Descent + Online Gradient Ascent

At iteration t, x_t, λ_t

$$\mathcal{L}_t(x, \lambda) = \ell_t(x) + \lambda f_t(x) - \frac{\delta\mu}{2}\lambda^2, \quad \delta \in \mathbb{R}^+$$

$\mathcal{L}_t(x, \lambda_t)$ → OMD

$\mathcal{L}_t(x_t, \lambda)$ → OGA

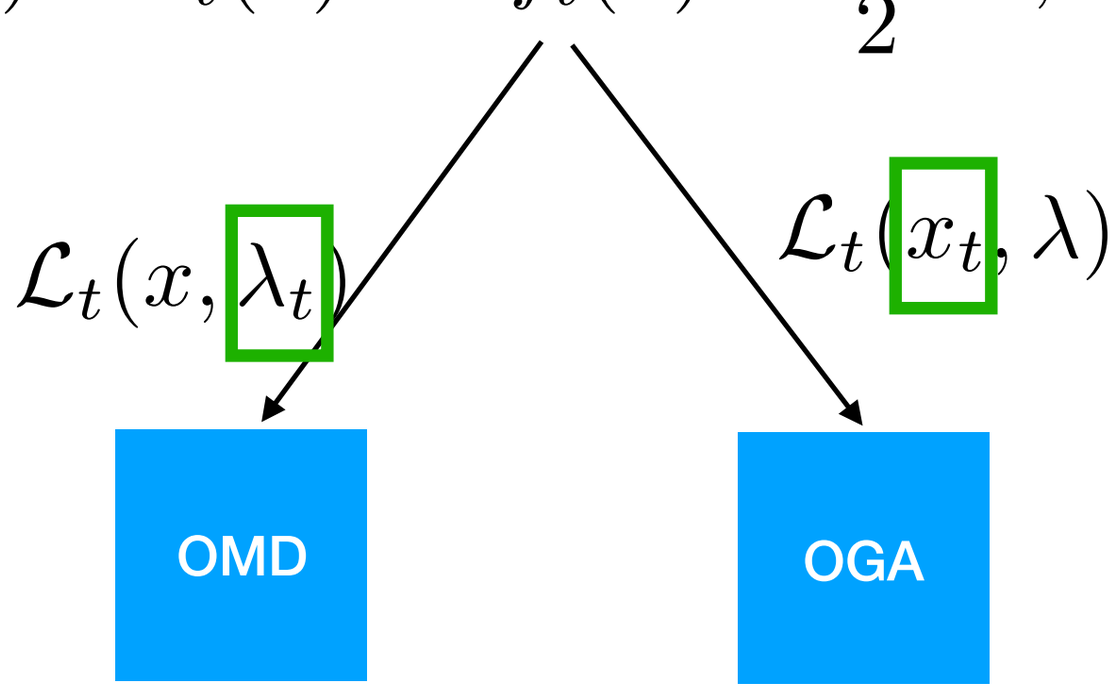
Algorithm

Strongly convex regularizer $R(x)$

Reduction to Online Mirror Descent + Online Gradient Ascent

At iteration t, x_t, λ_t

$$\mathcal{L}_t(x, \lambda) = \ell_t(x) + \lambda f_t(x) - \frac{\delta\mu}{2}\lambda^2, \quad \delta \in \mathbb{R}^+$$



$$\nabla R(\tilde{x}_{t+1}) = \nabla R(x_t) - \mu \nabla \mathcal{L}_t(x, \lambda_t)|_{x=x_t}$$

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} D_R(x, \tilde{x}_{t+1})$$

Algorithm

Strongly convex regularizer $R(x)$

Reduction to Online Mirror Descent + Online Gradient Ascent

At iteration t, x_t, λ_t

$$\mathcal{L}_t(x, \lambda) = \ell_t(x) + \lambda f_t(x) - \frac{\delta\mu}{2}\lambda^2, \quad \delta \in \mathbb{R}^+$$

$\mathcal{L}_t(x, \lambda_t)$ → OMD

$\mathcal{L}_t(x_t, \lambda)$ → OGA

$$\nabla R(\tilde{x}_{t+1}) = \nabla R(x_t) - \mu \nabla \mathcal{L}_t(x, \lambda_t)|_{x=x_t}$$

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} D_R(x, \tilde{x}_{t+1})$$

$$\lambda_{t+1} = \max\{0, \lambda_t + \mu \nabla \mathcal{L}_t(x_t, \lambda)|_{\lambda=\lambda_t}\}$$

Analysis

Analysis

Apply classic OMD analysis on $\{\mathcal{L}_t(x, \lambda_t)\}$

$$\sum_{t=1}^{T}(\mathcal{L}_t(x_t, \lambda_t) - \mathcal{L}_t(x, \lambda_t)) \leq \frac{D_R(x, x_1)}{\mu} + \frac{\mu}{2\alpha}\sum_{t=1}^{T}\|\nabla_x \mathcal{L}_t(x_t, \lambda_t)\|^2$$

Analysis

Apply classic OMD analysis on $\{\mathcal{L}_t(x, \lambda_t)\}$

$$\sum_{t=1}^{T}(\mathcal{L}_t(x_t, \lambda_t) - \mathcal{L}_t(x, \lambda_t)) \leq \frac{D_R(x, x_1)}{\mu} + \frac{\mu}{2\alpha}\sum_{t=1}^{T}\|\nabla_x \mathcal{L}_t(x_t, \lambda_t)\|^2$$

Apply classic OGD analysis on $\{\mathcal{L}_t(x_t, \lambda)\}$

$$\sum_{t=1}^{T}\mathcal{L}_t(x_t, \lambda) - \sum_{t=1}^{T}\mathcal{L}_t(x_t, \lambda_t) \leq \frac{\lambda^2}{\mu} + \frac{\mu}{2}\sum_{t=1}^{T}(\frac{\partial \mathcal{L}_t(x_t, \lambda_t)}{\partial \lambda_t})^2$$

Analysis

Apply classic OMD analysis on $\{\mathcal{L}_t(x, \lambda_t)\}$

$$\sum_{t=1}^{T}(\mathcal{L}_t(x_t, \lambda_t) - \mathcal{L}_t(x, \lambda_t)) \leq \frac{D_R(x, x_1)}{\mu} + \frac{\mu}{2\alpha}\sum_{t=1}^{T}\|\nabla_x \mathcal{L}_t(x_t, \lambda_t)\|^2$$

Apply classic OGD analysis on $\{\mathcal{L}_t(x_t, \lambda)\}$

$$\sum_{t=1}^{T}\mathcal{L}_t(x_t, \lambda) - \sum_{t=1}^{T}\mathcal{L}_t(x_t, \lambda_t) \leq \frac{\lambda^2}{\mu} + \frac{\mu}{2}\sum_{t=1}^{T}(\frac{\partial \mathcal{L}_t(x_t, \lambda_t)}{\partial \lambda_t})^2$$

Combine them:

Analysis

Apply classic OMD analysis on $\{\mathcal{L}_t(x, \lambda_t)\}$

$$\sum_{t=1}^{T}(\mathcal{L}_t(x_t, \lambda_t) - \mathcal{L}_t(x, \lambda_t)) \leq \frac{D_R(x, x_1)}{\mu} + \frac{\mu}{2\alpha}\sum_{t=1}^{T}\|\nabla_x \mathcal{L}_t(x_t, \lambda_t)\|^2$$

Apply classic OGD analysis on $\{\mathcal{L}_t(x_t, \lambda)\}$

$$\sum_{t=1}^{T}\mathcal{L}_t(x_t, \lambda) - \sum_{t=1}^{T}\mathcal{L}_t(x_t, \lambda_t) \leq \frac{\lambda^2}{\mu} + \frac{\mu}{2}\sum_{t=1}^{T}(\frac{\partial \mathcal{L}_t(x_t, \lambda_t)}{\partial \lambda_t})^2$$

Combine them:

$$\frac{1}{T}\Big[\sum_{t=1}^{T}\ell_t(x_t) - \min_{x^* \in \mathcal{O}}\sum_{t=1}^{T}\ell_t(x^*)\Big] \leq O(1/\sqrt{T})$$

$$\frac{1}{T}\Big[\sum_{t=1}^{T}f_t(x_t)\Big] \leq O(T^{-1/4})$$

Special Case

Special Case

Set regularizer: $R(x) = \sum_i x[i] \ln x[i]$

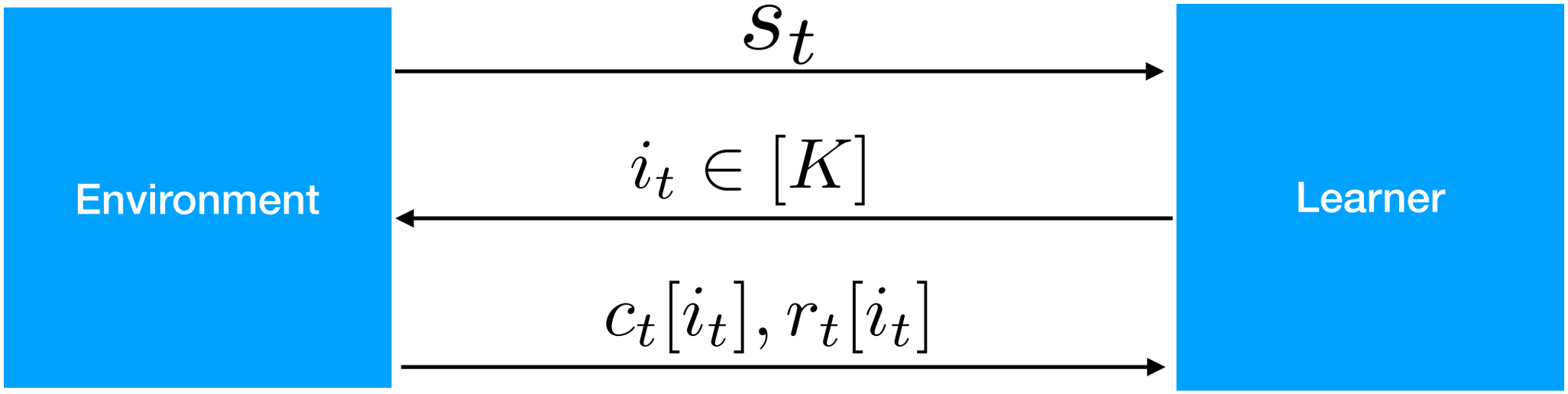
Special Case

Set regularizer: $R(x) = \sum_i x[i] \ln x[i]$

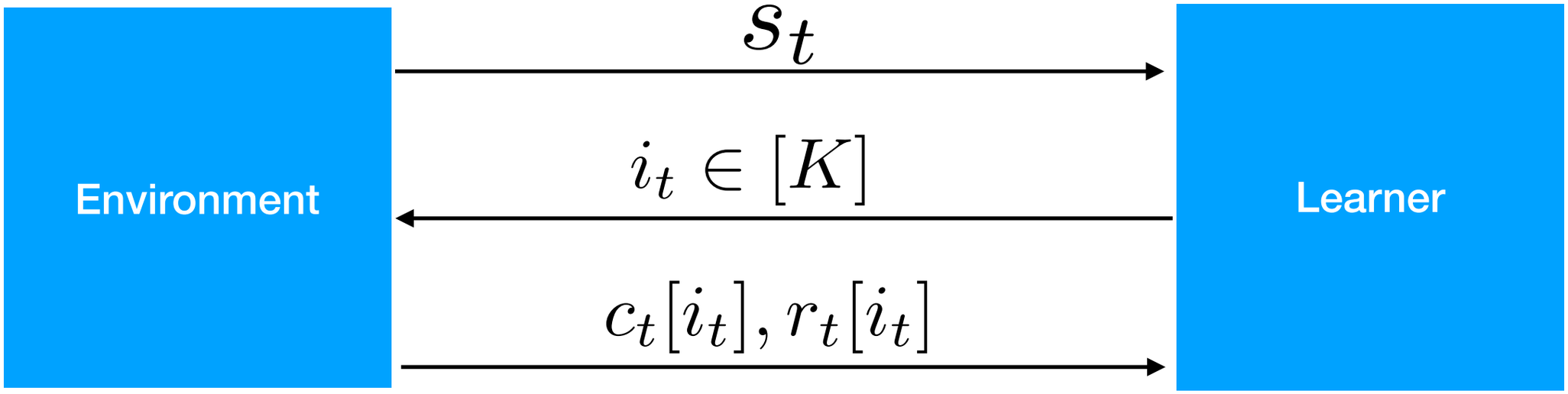
$$x_{t+1}[i] = \frac{x_t[i] \exp(-\mu \nabla_x \mathcal{L}_t(x_t, \lambda_t)[i])}{\sum_{j=1}^{d} x_t[j] \exp(-\mu \nabla_x \mathcal{L}_t(x_t, \lambda_t)[j])}$$

$$\lambda_{t+1} = \max\{0, \lambda_t + \mu \nabla \mathcal{L}_t(x_t, \lambda)|_{\lambda=\lambda_t}\}$$

Contextual Bandit Setting

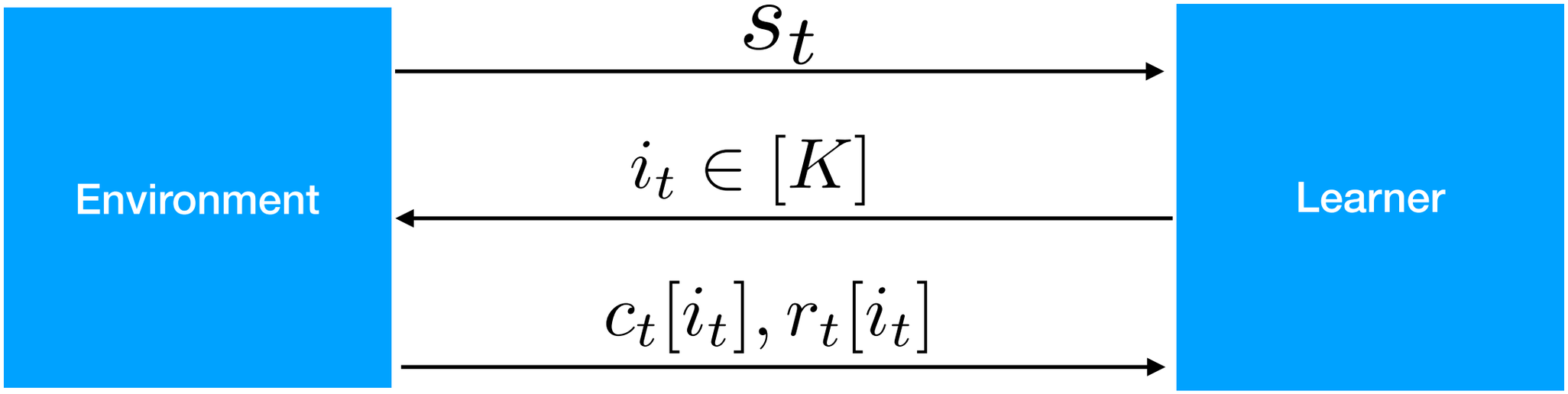


Contextual Bandit Setting



Expert Setting:

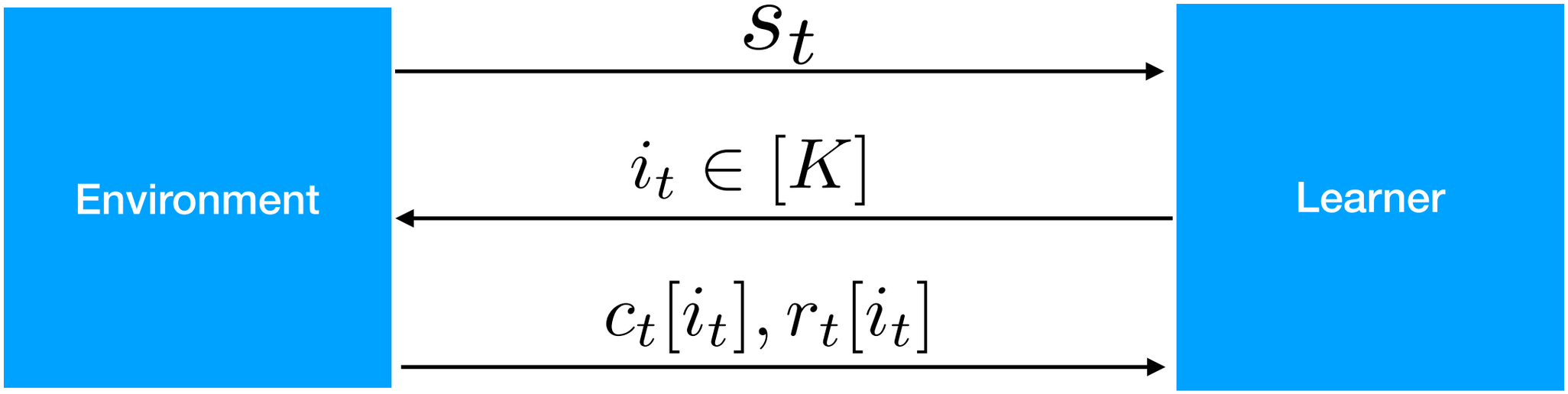
Contextual Bandit Setting



Expert Setting:

Finite Expert Set: $\Pi : \pi(s) : \mathcal{S} \rightarrow \Delta(K)$

Contextual Bandit Setting

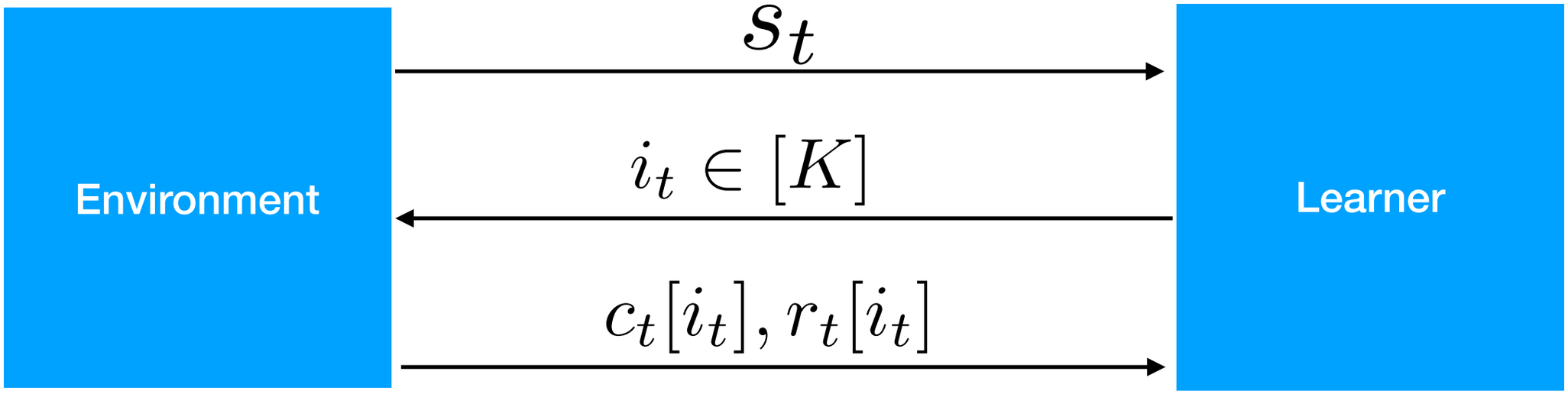


Expert Setting:

Finite Expert Set: $\Pi : \pi(s) : \mathcal{S} \rightarrow \Delta(K)$

Decision Set: $\Delta(\Pi)$ (all distributions over expert set)

Contextual Bandit Setting



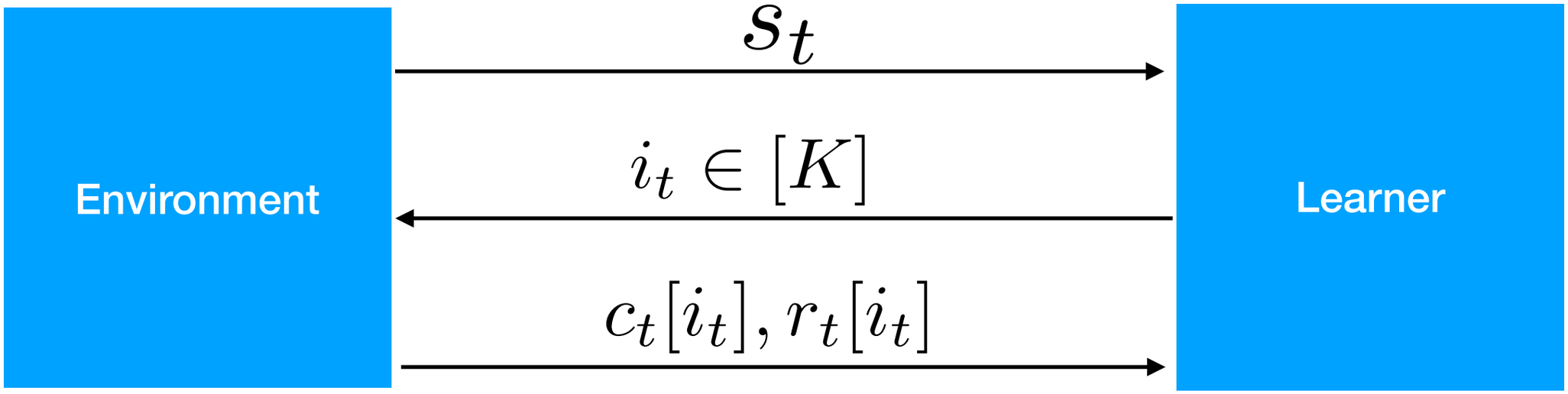
Expert Setting:

Finite Expert Set: $\Pi : \pi(s) : \mathcal{S} \rightarrow \Delta(K)$

Decision Set: $\Delta(\Pi)$ (all distributions over expert set)

Competing Against: $P = \{w \in \Delta(\Pi) : \mathbb{E}_{i \sim w, j \sim \pi_i(s_t)} r_t[j] \leq \beta, \forall t\}$

Contextual Bandit Setting



Expert Setting:

Finite Expert Set: $\Pi : \pi(s) : \mathcal{S} \rightarrow \Delta(K)$

Decision Set: $\Delta(\Pi)$ (all distributions over expert set)

Competing Against: $P = \{w \in \Delta(\Pi) : \mathbb{E}_{i \sim w, j \sim \pi_i(s_t)} r_t[j] \leq \beta, \forall t\}$

Algorithm

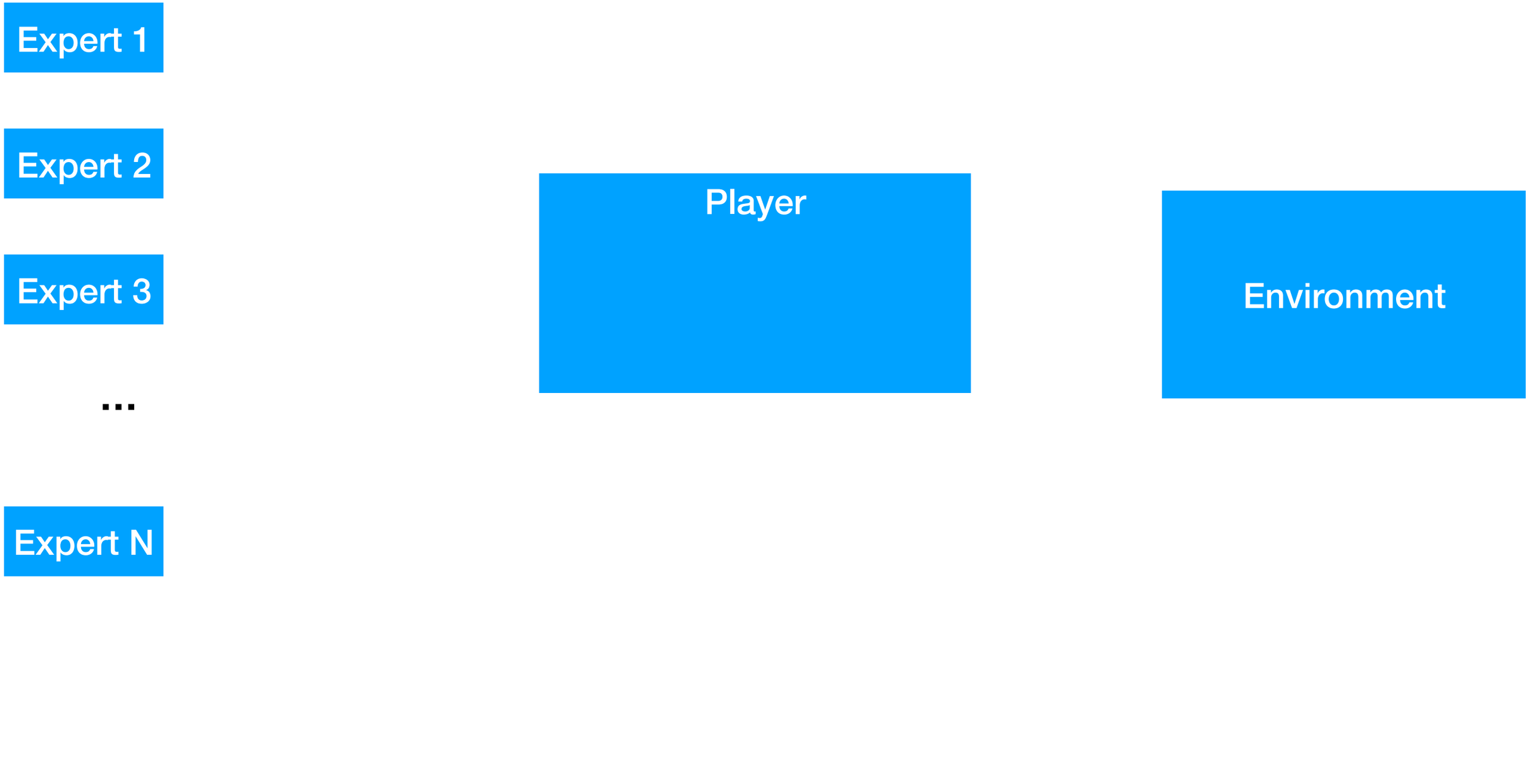
Reduction to Full Information Setting

(i.e., EXP4 [Auer et al., 02,], EXP4.P [Beygelzimer et al., 11])

Algorithm

Reduction to Full Information Setting

(i.e., EXP4 [Auer et al., 02,], EXP4.P [Beygelzimer et al., 11])



Algorithm

Reduction to Full Information Setting

(i.e., EXP4 [Auer et al., 02,], EXP4.P [Beygelzimer et al., 11])

s_t

Expert 1

Expert 2

Expert 3

...

Expert N

Player

Environment

Algorithm

Reduction to Full Information Setting
(i.e., EXP4 [Auer et al., 02,], EXP4.P [Beygelzimer et al., 11])

s_t

Expert 1 $p_1 = \pi_1(s_t)$

Expert 2

Expert 3

...

Expert N

Player

Environment

Algorithm

Reduction to Full Information Setting
(i.e., EXP4 [Auer et al., 02,], EXP4.P [Beygelzimer et al., 11])

s_t

Expert 1 $p_1 = \pi_1(s_t)$

Expert 2

Expert 3

...

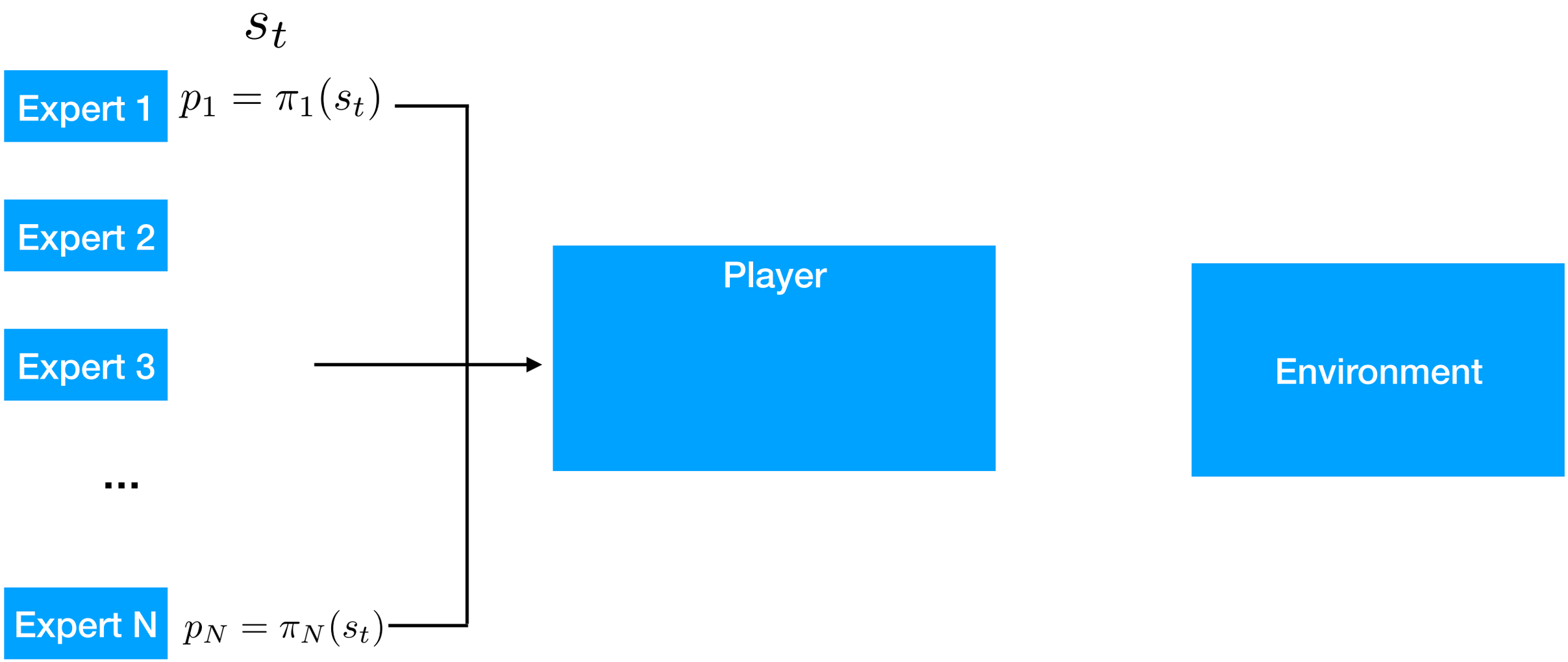
Expert N $p_N = \pi_N(s_t)$

Player

Environment

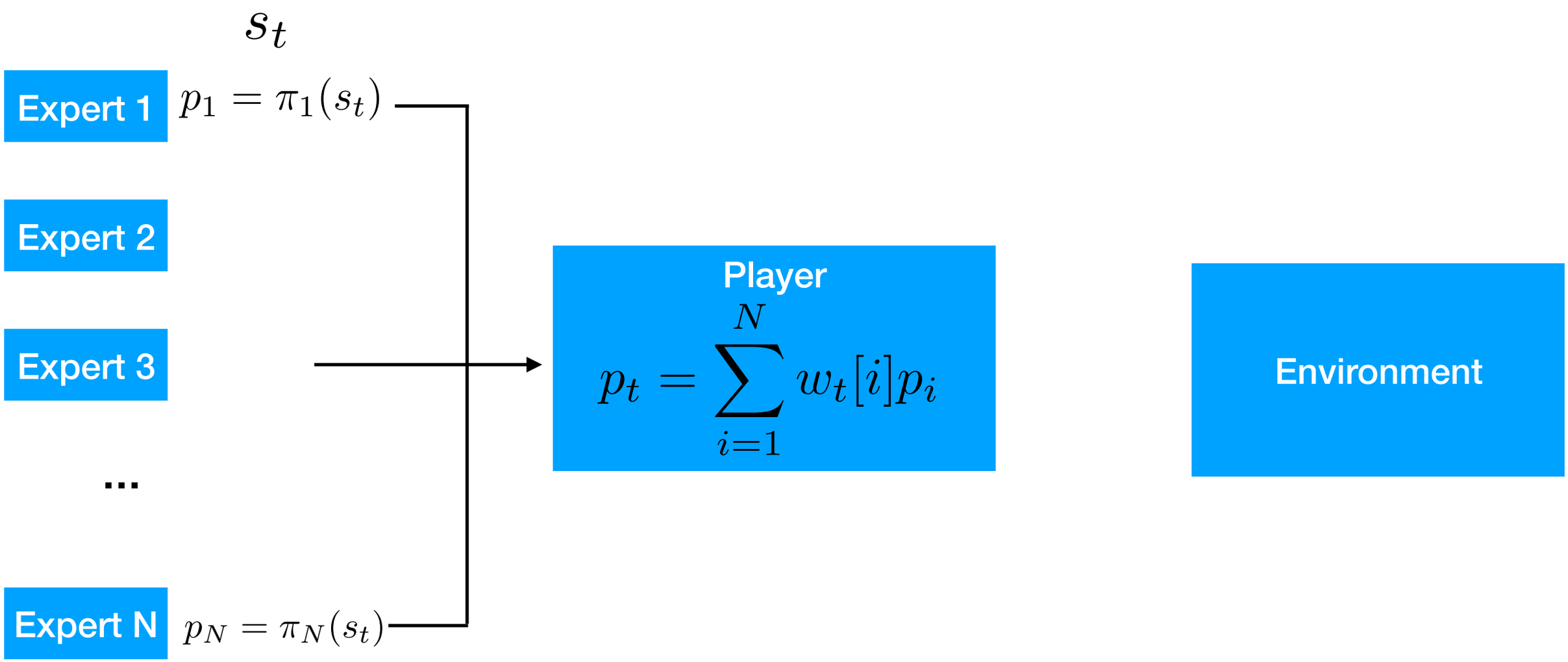
Algorithm

Reduction to Full Information Setting
(i.e., EXP4 [Auer et al., 02,], EXP4.P [Beygelzimer et al., 11])



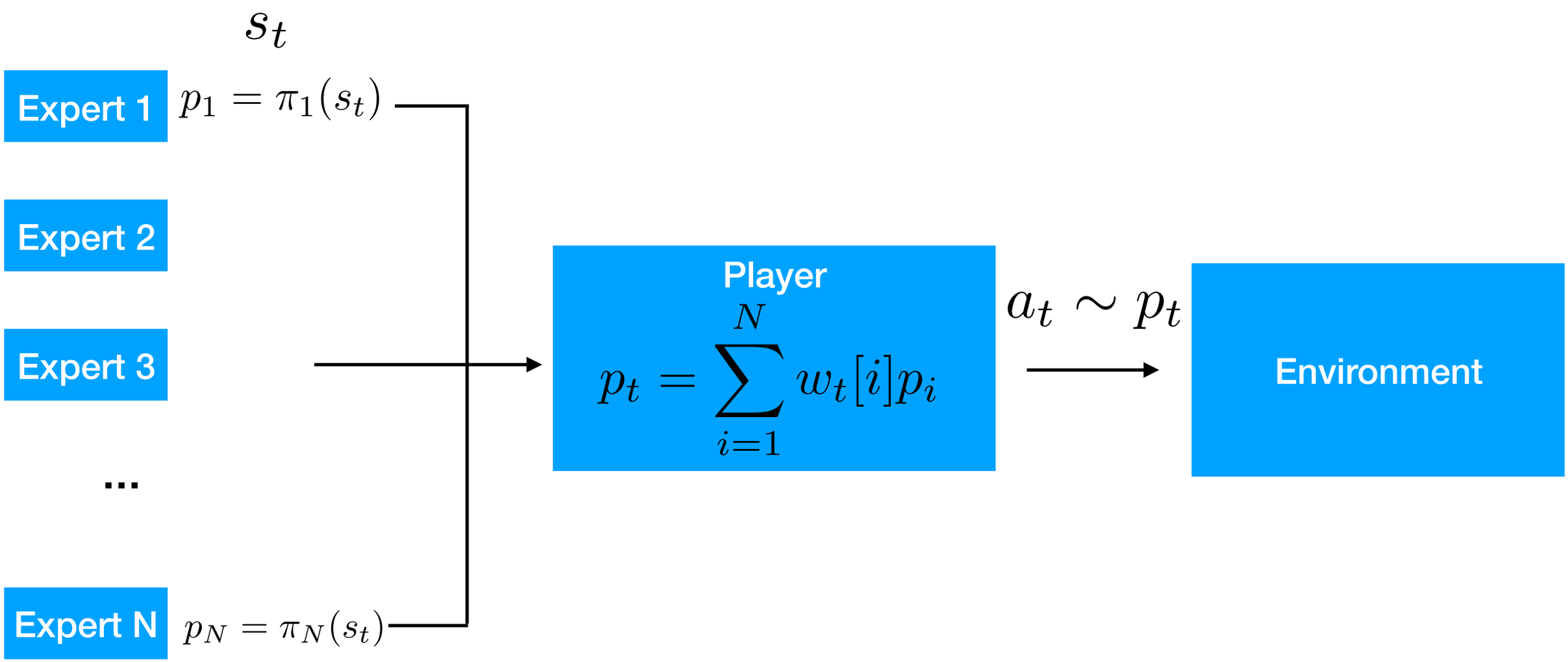
Algorithm

Reduction to Full Information Setting
(i.e., EXP4 [Auer et al., 02,], EXP4.P [Beygelzimer et al., 11])



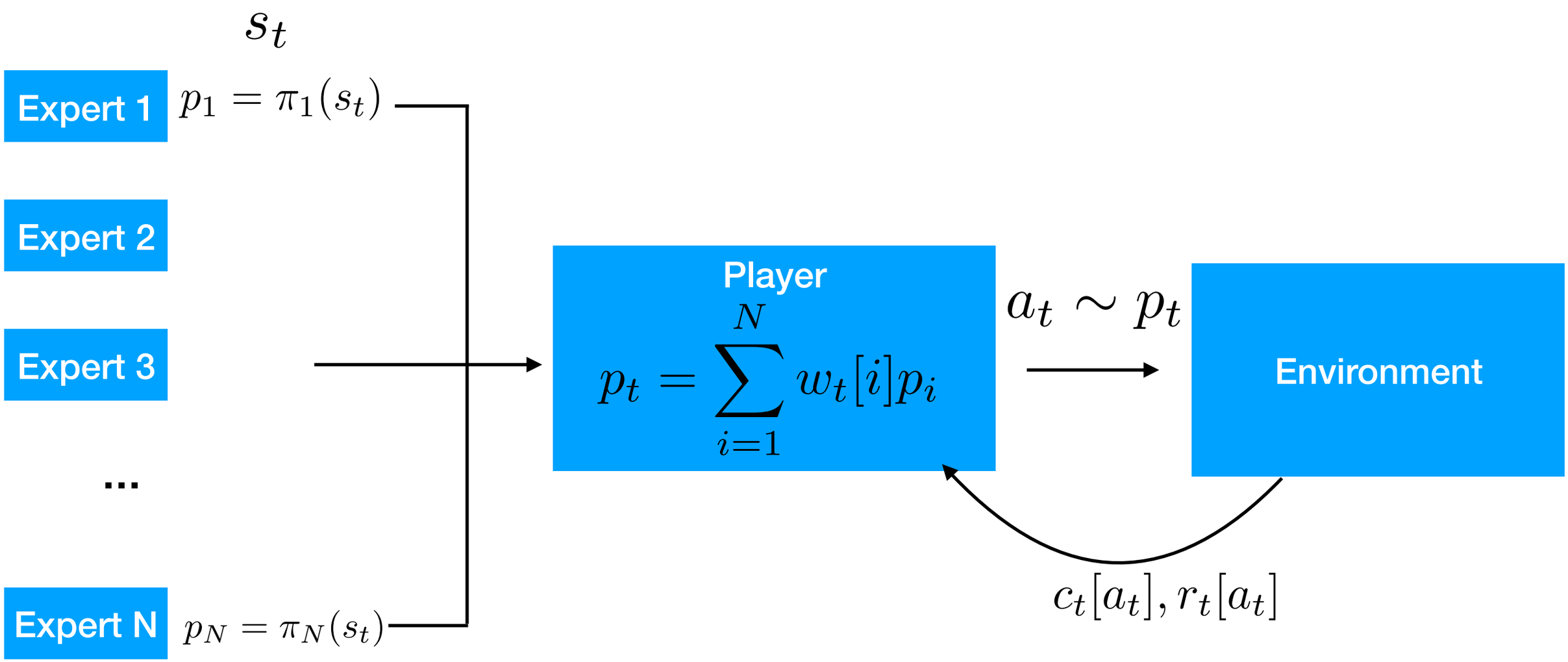
Algorithm

Reduction to Full Information Setting
(i.e., EXP4 [Auer et al., 02,], EXP4.P [Beygelzimer et al., 11])



Algorithm

Reduction to Full Information Setting
(i.e., EXP4 [Auer et al., 02,], EXP4.P [Beygelzimer et al., 11])



Algorithm

Player

$$p_t = \sum_{i=1}^{N} w_t[i] p_i \rightarrow a_t \in [K]$$

$$c_t[a_t], r_t[a_t]$$

Black-Box Learner of OCP
with Constraints
(R(w) set to negative entropy)

Algorithm

Player

$$p_t = \sum_{i=1}^{N} w_t[i] p_i \rightarrow a_t \in [K]$$

$$c_t[a_t], r_t[a_t]$$

$$\hat{c}_t = [0, 0, ..., c_t[a_t]/p_t[a_t], 0...], \quad \hat{r}_t = [0, 0, ..., r_t[a_t]/p_t[a_t], 0, ...]$$

Black-Box Learner of OCP with Constraints (R(w) set to negative entropy)

Algorithm

Player

$$p_t = \sum_{i=1}^{N} w_t[i] p_i \rightarrow a_t \in [K]$$

$$c_t[a_t], r_t[a_t]$$

$$\hat{c}_t = [0, 0, ..., c_t[a_t]/p_t[a_t], 0...], \quad \hat{r}_t = [0, 0, ..., r_t[a_t]/p_t[a_t], 0, ...]$$

$$\hat{y}_t[j] = p_j \cdot \hat{c}_t, \hat{z}_t[j] = p_j \cdot \hat{r}_t \quad \text{for expert j}$$

Black-Box Learner of OCP with Constraints (R(w) set to negative entropy)

Algorithm

Player

$$p_t = \sum_{i=1}^{N} w_t[i] p_i \rightarrow a_t \in [K]$$

$$c_t[a_t], r_t[a_t]$$

$$\hat{c}_t = [0, 0, ..., c_t[a_t]/p_t[a_t], 0...], \quad \hat{r}_t = [0, 0, ..., r_t[a_t]/p_t[a_t], 0, ...]$$

$$\hat{y}_t[j] = p_j \cdot \hat{c}_t, \hat{z}_t[j] = p_j \cdot \hat{r}_t \quad \text{for expert j}$$

$$\hat{\ell}_t(w) = w \cdot \hat{y}_t, \hat{f}_t(w) = w \cdot \hat{z}_t$$

Black-Box Learner of OCP with Constraints (R(w) set to negative entropy)

Algorithm

Player

$$p_t = \sum_{i=1}^{N} w_t[i] p_i \rightarrow a_t \in [K]$$

$$c_t[a_t], r_t[a_t]$$

$$\hat{c}_t = [0, 0, ..., c_t[a_t]/p_t[a_t], 0...], \quad \hat{r}_t = [0, 0, ..., r_t[a_t]/p_t[a_t], 0, ...]$$

$$\hat{y}_t[j] = p_j \cdot \hat{c}_t, \hat{z}_t[j] = p_j \cdot \hat{r}_t \quad \text{for expert j}$$

$$\hat{\ell}_t(w) = w \cdot \hat{y}_t, \hat{f}_t(w) = w \cdot \hat{z}_t$$

$$\downarrow \beta$$

Black-Box Learner of OCP with Constraints (R(w) set to negative entropy)

Algorithm

Player

$$p_t = \sum_{i=1}^{N} w_t[i] p_i \rightarrow a_t \in [K]$$

$$c_t[a_t], r_t[a_t]$$

$$\hat{c}_t = [0, 0, ..., c_t[a_t]/p_t[a_t], 0...], \quad \hat{r}_t = [0, 0, ..., r_t[a_t]/p_t[a_t], 0, ...]$$

$$\hat{y}_t[j] = p_j \cdot \hat{c}_t, \hat{z}_t[j] = p_j \cdot \hat{r}_t \quad \text{for expert j}$$

$$\hat{\ell}_t(w) = w \cdot \hat{y}_t, \hat{f}_t(w) = w \cdot \hat{z}_t$$

$\downarrow \beta$

Black-Box Learner of OCP with Constraints (R(w) set to negative entropy)

$\downarrow$

Update to w_{t+1}

Algorithm

Player

$$p_t = \sum_{i=1}^{N} w_t[i] p_i \rightarrow a_t \in [K]$$

$$c_t[a_t], r_t[a_t]$$

$$\hat{c}_t = [0, 0, ..., c_t[a_t]/p_t[a_t], 0...], \quad \hat{r}_t = [0, 0, ..., r_t[a_t]/p_t[a_t], 0, ...]$$

$$\hat{y}_t[j] = p_j \cdot \hat{c}_t, \hat{z}_t[j] = p_j \cdot \hat{r}_t \quad \text{for expert j}$$

$$\hat{\ell}_t(w) = w \cdot \hat{y}_t, \hat{f}_t(w) = w \cdot \hat{z}_t$$

$$\downarrow \beta$$

Black-Box Learner of OCP
with Constraints
(R(w) set to negative entropy)

$$\downarrow$$

Update to w_{t+1}

***EXP4.R* (*EXP4* with *R*isk constraints)**

Analysis

Under the assumption that $P \neq \emptyset$, for any sequence of cost and risk vectors, EXP4.R has the following guarantees:

$$\mathbb{E}[\sum_{t=1}^{T} c_t[a_t] - \sum_{t=1}^{T} \mathbb{E}_{i\sim w^*, j\sim\pi_i(s_t)} c_t[j]] \ \leq O(\sqrt{TK\ln(|\Pi|)})$$

$$\mathbb{E}[\sum_{t=1}^{T} r_t[a_t] - \beta] \ \leq O(T^{3/4}(K\ln(|\Pi|))^{1/4})$$

Where $w^* \in \{w \in \Delta(\Pi) : \mathbb{E}_{i\sim w, j\sim\pi_i(s_t)} r_t[j] \leq \beta, \forall t\}$

Algorithm

We also want high probability statement.
Use the trick in EXP3.P (and EXP4.P)

Algorithm

We also want high probability statement.
Use the trick in EXP3.P (and EXP4.P)

$$\hat{\mathcal{L}}_t(w, \lambda) = w \cdot (\hat{y}_t + \lambda \hat{z}_t) - \delta \mu \lambda^2 / 2$$

Algorithm

We also want high probability statement.
Use the trick in EXP3.P (and EXP4.P)

$$\hat{\mathcal{L}}_t(w, \lambda) = w \cdot (\hat{y}_t + \lambda \hat{z}_t) - \delta \mu \lambda^2 / 2$$

$$\mathbb{E}_t[\hat{y}_t + \lambda \hat{z}_t] = y_t + \lambda z_t$$

Algorithm

We also want high probability statement.
Use the trick in EXP3.P (and EXP4.P)

$$\hat{\mathcal{L}}_t(w,\lambda) = w \cdot (\hat{y}_t + \lambda \hat{z}_t) - \delta\mu\lambda^2/2$$

$$\mathbb{E}_t[\hat{y}_t + \lambda \hat{z}_t] = y_t + \lambda z_t$$

$$\hat{y}_t + \lambda \hat{z}_t - \kappa \sum_{k=1}^{K} \frac{\pi_j(s_t)[k]}{p_t[k]}$$

Algorithm

We also want high probability statement.
Use the trick in EXP3.P (and EXP4.P)

$$\hat{\mathcal{L}}_t(w,\lambda) = w \cdot (\hat{y}_t + \lambda \hat{z}_t) - \delta\mu\lambda^2/2$$

$$\mathbb{E}_t[\hat{y}_t + \lambda \hat{z}_t] = y_t + \lambda z_t$$

$$\tilde{\mathcal{L}}_t(w,\lambda) = w \cdot \left(\hat{y}_t + \lambda \hat{z}_t - \kappa \sum_{k=1}^{K} \frac{\pi_j(s_t)[k]}{p_t[k]} \right) - \delta\mu\lambda^2/2$$

Algorithm

We also want high probability statement.
Use the trick in EXP3.P (and EXP4.P)

$$\hat{\mathcal{L}}_t(w,\lambda) = w \cdot (\hat{y}_t + \lambda \hat{z}_t) - \delta\mu\lambda^2/2$$

$$\mathbb{E}_t[\hat{y}_t + \lambda \hat{z}_t] = y_t + \lambda z_t$$

$$\tilde{\mathcal{L}}_t(w,\lambda) = w \cdot \left(\hat{y}_t + \lambda \hat{z}_t - \kappa \sum_{k=1}^{K} \frac{\pi_j(s_t)[k]}{p_t[k]} \right) - \delta\mu\lambda^2/2$$

EXP4.P.R (EXP4.P with Risk Constraints)

Analysis

Theorem: Under the assumption that $P \neq \emptyset$*, for any sequence of cost, risk vectors and any* $\epsilon \in (0, 0.5)$*, we have with high probability 1- v:*

$$\sum_{t=1}^{T} c_t[a_t] - \sum_{t=1}^{T} \mathbb{E}_{i\sim w^*, j\sim \pi_i(s_t)} c_t[j] \leq O(\sqrt{T^{\epsilon+1/2} K \ln(\Pi/v)})$$

$$\sum_{t=1}^{T} (r_t[a_t] - \beta) \leq O(T^{1-\epsilon/2} \sqrt{K \ln(\Pi)})$$

Analysis

Theorem: Under the assumption that $P \neq \emptyset$*, for any sequence of cost, risk vectors and any* $\epsilon \in (0, 0.5)$*, we have with high probability 1- v:*

$$\sum_{t=1}^{T} c_t[a_t] - \sum_{t=1}^{T} \mathbb{E}_{i\sim w^*, j\sim\pi_i(s_t)} c_t[j] \leq O(\sqrt{T^{\epsilon+1/2} K \ln(\Pi/v)})$$

$$\sum_{t=1}^{T} (r_t[a_t] - \beta) \leq O(T^{1-\epsilon/2}\sqrt{K \ln(\Pi)})$$

When $\epsilon \longrightarrow 0$: Average Regret-> $O(1/\sqrt{T})$ Avg constraint violation-> $O(1)$

Analysis

Theorem: Under the assumption that $P \neq \emptyset$*, for any sequence of cost, risk vectors and any* $\epsilon \in (0, 0.5)$*, we have with high probability 1- v:*

$$\sum_{t=1}^{T} c_t[a_t] - \sum_{t=1}^{T} \mathbb{E}_{i\sim w^*, j\sim\pi_i(s_t)} c_t[j] \leq O(\sqrt{T^{\epsilon+1/2} K \ln(\Pi/v)})$$

$$\sum_{t=1}^{T} (r_t[a_t] - \beta) \leq O(T^{1-\epsilon/2}\sqrt{K \ln(\Pi)})$$

When $\epsilon \to 0$: Average Regret-> $O(1/\sqrt{T})$ Avg constraint violation-> $O(1)$

When $\epsilon \to 0.5$: Average Regret-> $O(1)$ Avg constrain violation-> $O(T^{-1/4})$

Analysis

Theorem: Under the assumption that $P \neq \emptyset$*, for any sequence of cost, risk vectors and any* $\epsilon \in (0, 0.5)$*, we have with high probability 1- v:*

$$\sum_{t=1}^{T} c_t[a_t] - \sum_{t=1}^{T} \mathbb{E}_{i\sim w^*, j\sim\pi_i(s_t)} c_t[j] \leq O(\sqrt{T^{\epsilon+1/2} K \ln(\Pi/v)})$$

$$\sum_{t=1}^{T} (r_t[a_t] - \beta) \leq O(T^{1-\epsilon/2}\sqrt{K \ln(\Pi)})$$

When $\epsilon \to 0$: Average Regret-> $O(1/\sqrt{T})$ Avg constraint violation-> $O(1)$

When $\epsilon \to 0.5$: Average Regret-> $O(1)$ Avg constrain violation-> $O(T^{-1/4})$

Challenge: $\mathbb{E}_t[\hat{y}_t + \lambda_t \hat{z}_t] = y_t + \lambda_t z_t$

Analysis

Theorem: Under the assumption that $P \neq \emptyset$*, for any sequence of cost, risk vectors and any* $\epsilon \in (0, 0.5)$*, we have with high probability 1- v:*

$$\sum_{t=1}^{T} c_t[a_t] - \sum_{t=1}^{T} \mathbb{E}_{i\sim w^*, j\sim\pi_i(s_t)} c_t[j] \leq O(\sqrt{T^{\epsilon+1/2} K \ln(\Pi/v)})$$

$$\sum_{t=1}^{T} (r_t[a_t] - \beta) \leq O(T^{1-\epsilon/2}\sqrt{K \ln(\Pi)})$$

When $\epsilon \to 0$: Average Regret-> $O(1/\sqrt{T})$ Avg constraint violation-> $O(1)$

When $\epsilon \to 0.5$: Average Regret-> $O(1)$ Avg constrain violation-> $O(T^{-1/4})$

Challenge: $\mathbb{E}_t[\hat{y}_t + \lambda_t \hat{z}_t] = y_t + \lambda_t z_t$

$$Var(\lambda_t \hat{z}_t) = \frac{\lambda_t^2}{p}$$

Analysis

Theorem: Under the assumption that $P \neq \emptyset$*, for any sequence of cost, risk vectors and any* $\epsilon \in (0, 0.5)$*, we have with high probability 1- v:*

$$\sum_{t=1}^{T} c_t[a_t] - \sum_{t=1}^{T} \mathbb{E}_{i\sim w^*, j\sim\pi_i(s_t)} c_t[j] \leq O(\sqrt{T^{\epsilon+1/2} K \ln(\Pi/v)})$$

$$\sum_{t=1}^{T} (r_t[a_t] - \beta) \leq O(T^{1-\epsilon/2}\sqrt{K \ln(\Pi)})$$

When $\epsilon \to 0$: Average Regret-> $O(1/\sqrt{T})$ Avg constraint violation-> $O(1)$

When $\epsilon \to 0.5$: Average Regret-> $O(1)$ Avg constrain violation-> $O(T^{-1/4})$

Challenge: $\mathbb{E}_t[\hat{y}_t + \lambda_t \hat{z}_t] = y_t + \lambda_t z_t$

$$Var(\lambda_t \hat{z}_t) = \frac{\lambda_t^2}{p} \qquad \lambda_t = O(\frac{\beta}{\delta\mu})$$

Analysis

Theorem: Under the assumption that $P \neq \emptyset$*, for any sequence of cost, risk vectors and any* $\epsilon \in (0, 0.5)$*, we have with high probability 1- v:*

$$\sum_{t=1}^{T} c_t[a_t] - \sum_{t=1}^{T} \mathbb{E}_{i\sim w^*, j\sim \pi_i(s_t)} c_t[j] \leq O(\sqrt{T^{\epsilon+1/2} K \ln(\Pi/v)})$$

$$\sum_{t=1}^{T} (r_t[a_t] - \beta) \leq O(T^{1-\epsilon/2}\sqrt{K \ln(\Pi)})$$

When $\epsilon \to 0$: Average Regret-> $O(1/\sqrt{T})$ Avg constraint violation-> $O(1)$

When $\epsilon \to 0.5$: Average Regret-> $O(1)$ Avg constrain violation-> $O(T^{-1/4})$

Challenge: $\mathbb{E}_t[\hat{y}_t + \lambda_t \hat{z}_t] = y_t + \lambda_t z_t$

$$Var(\lambda_t \hat{z}_t) = \frac{\lambda_t^2}{p} \qquad \lambda_t = O(\frac{\beta}{\delta\mu})\sqrt{T}$$

Analysis

Theorem: Under the assumption that $P \neq \emptyset$*, for any sequence of cost, risk vectors and any* $\epsilon \in (0, 0.5)$*, we have with high probability 1- v:*

$$\sum_{t=1}^{T} c_t[a_t] - \sum_{t=1}^{T} \mathbb{E}_{i\sim w^*, j\sim\pi_i(s_t)} c_t[j] \leq O(\sqrt{T^{\epsilon+1/2} K \ln(\Pi/v)})$$

$$\sum_{t=1}^{T} (r_t[a_t] - \beta) \leq O(T^{1-\epsilon/2}\sqrt{K \ln(\Pi)})$$

When $\epsilon \to 0$: Average Regret-> $O(1/\sqrt{T})$ Avg constraint violation-> $O(1)$

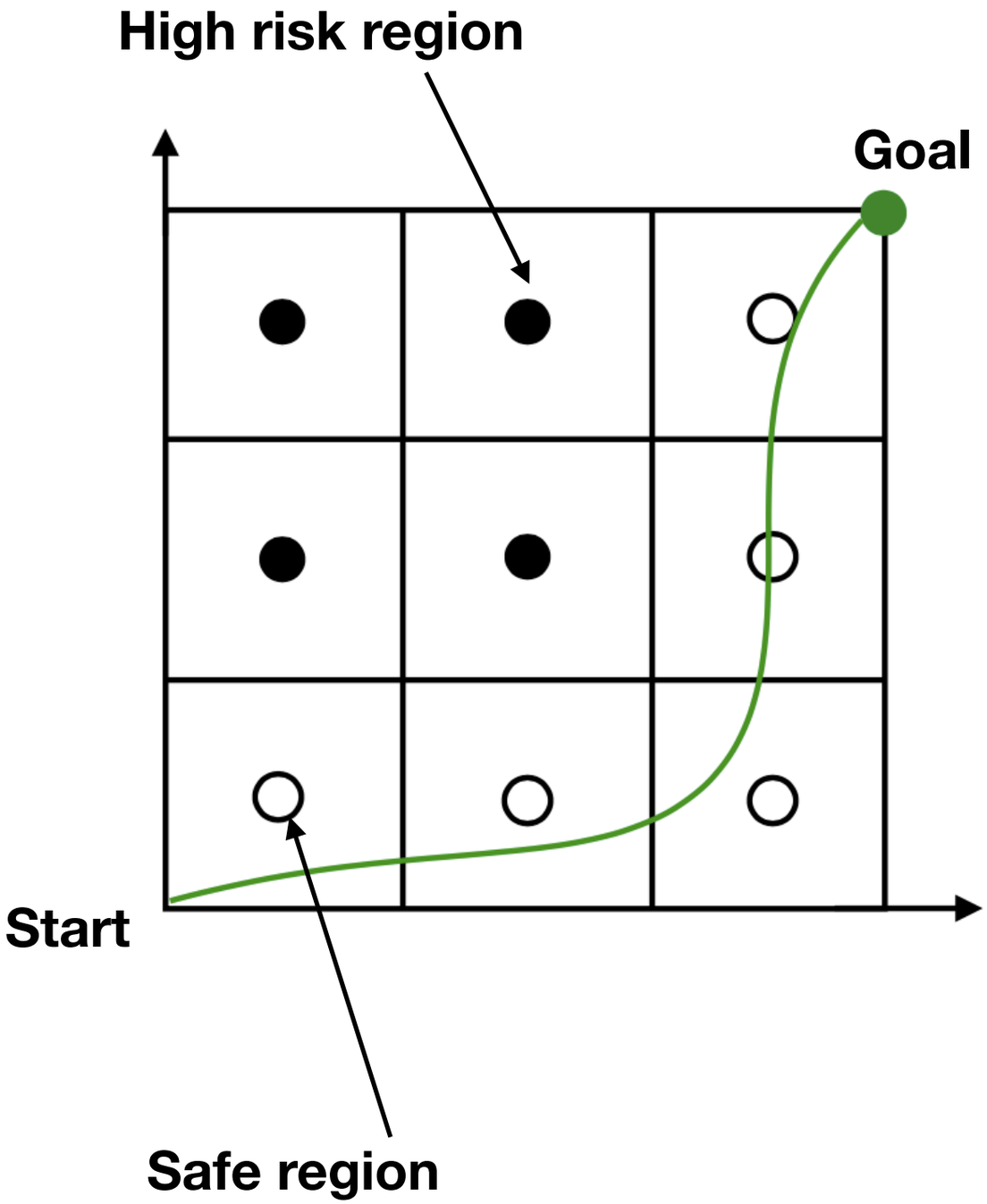
When $\epsilon \to 0.5$: Average Regret-> $O(1)$ Avg constrain violation-> $O(T^{-1/4})$

Challenge: $\mathbb{E}_t[\hat{y}_t + \lambda_t \hat{z}_t] = y_t + \lambda_t z_t$

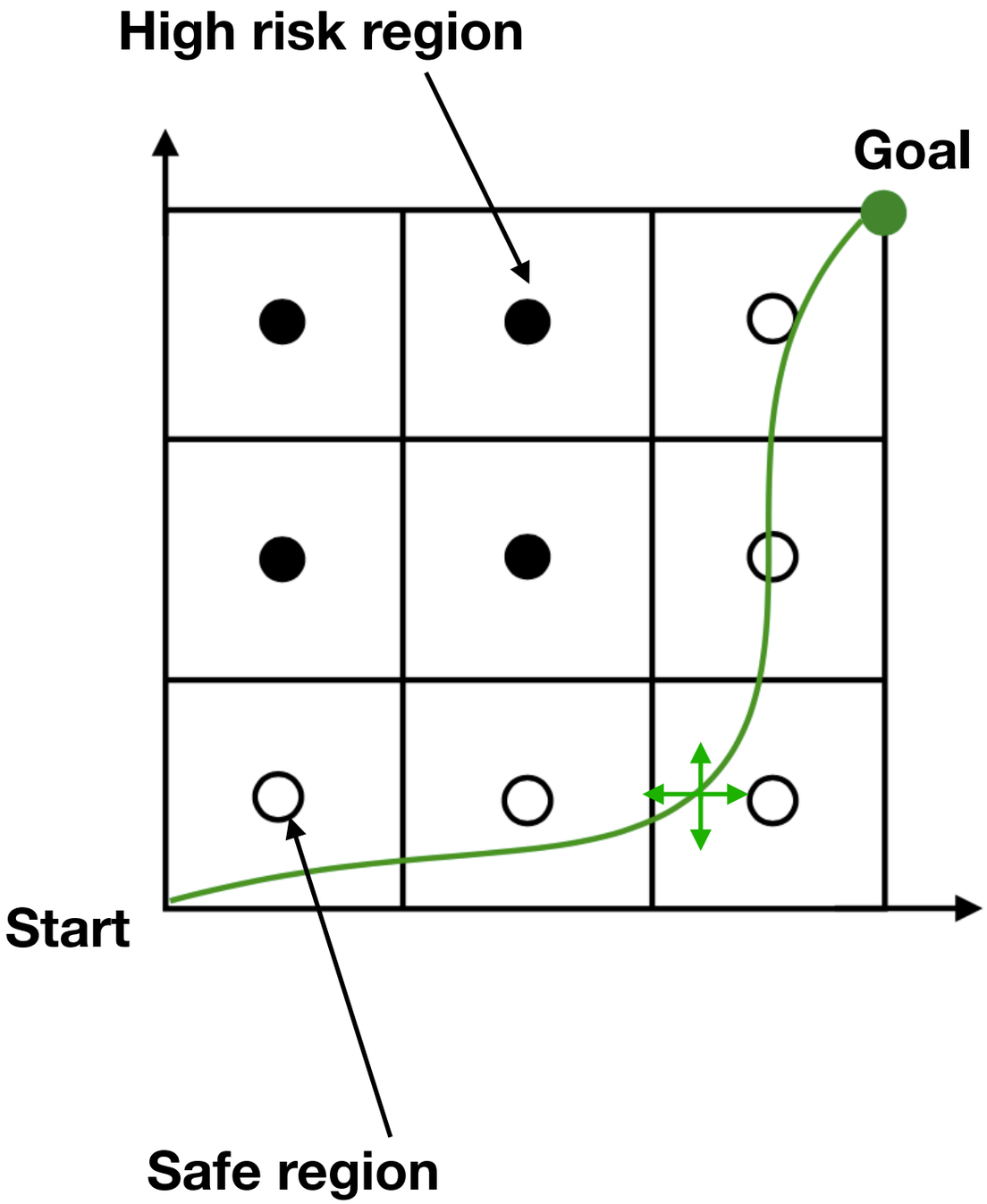
$$Var(\lambda_t \hat{z}_t) = \frac{\lambda_t^2}{p} \quad \boxed{\lambda_t = O(\frac{\beta}{\delta\mu})} \quad \to Var(\lambda_t \hat{z}_t) = \frac{T}{p}$$

$\sqrt{T}$

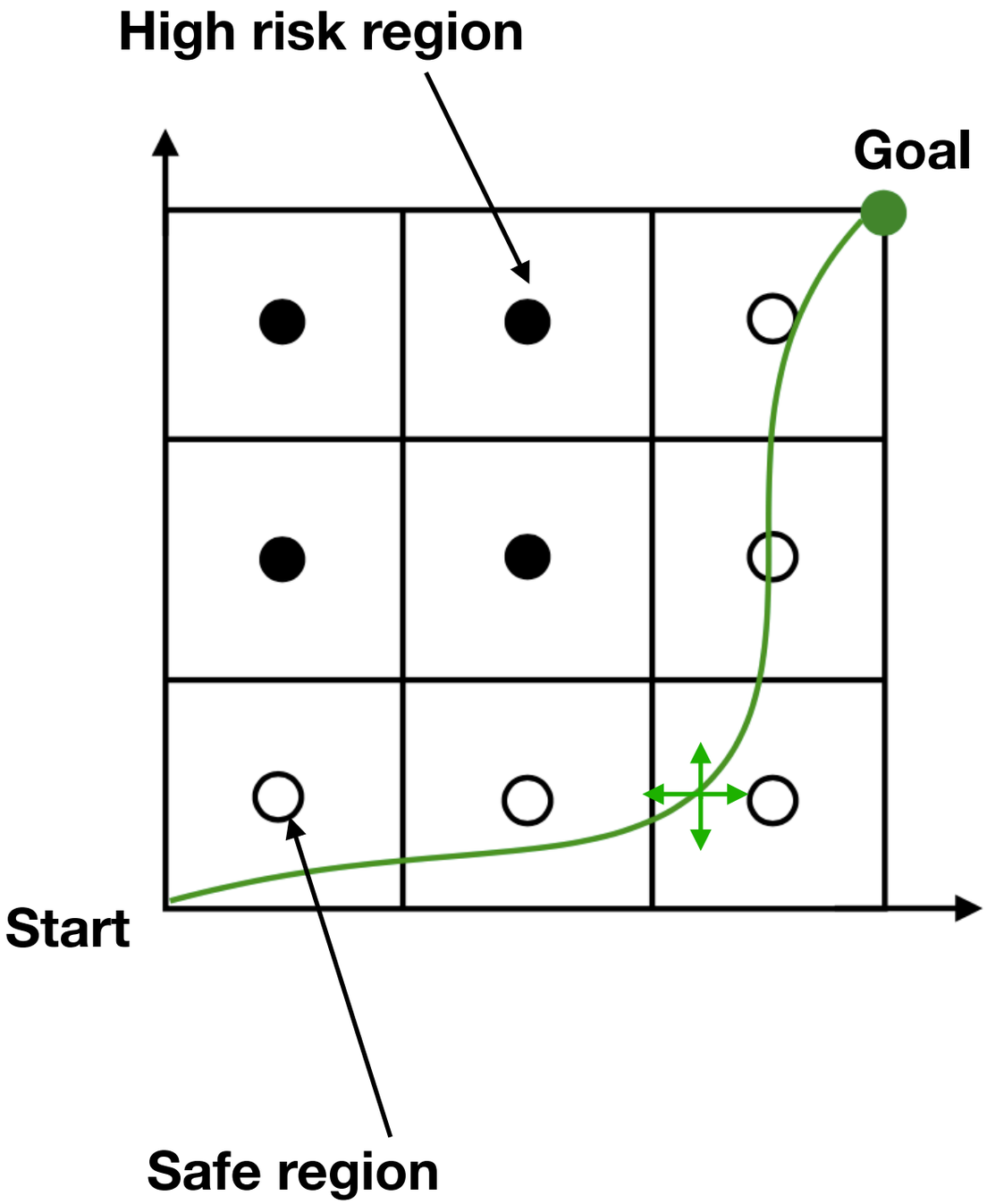
Simulation



Simulation



Simulation

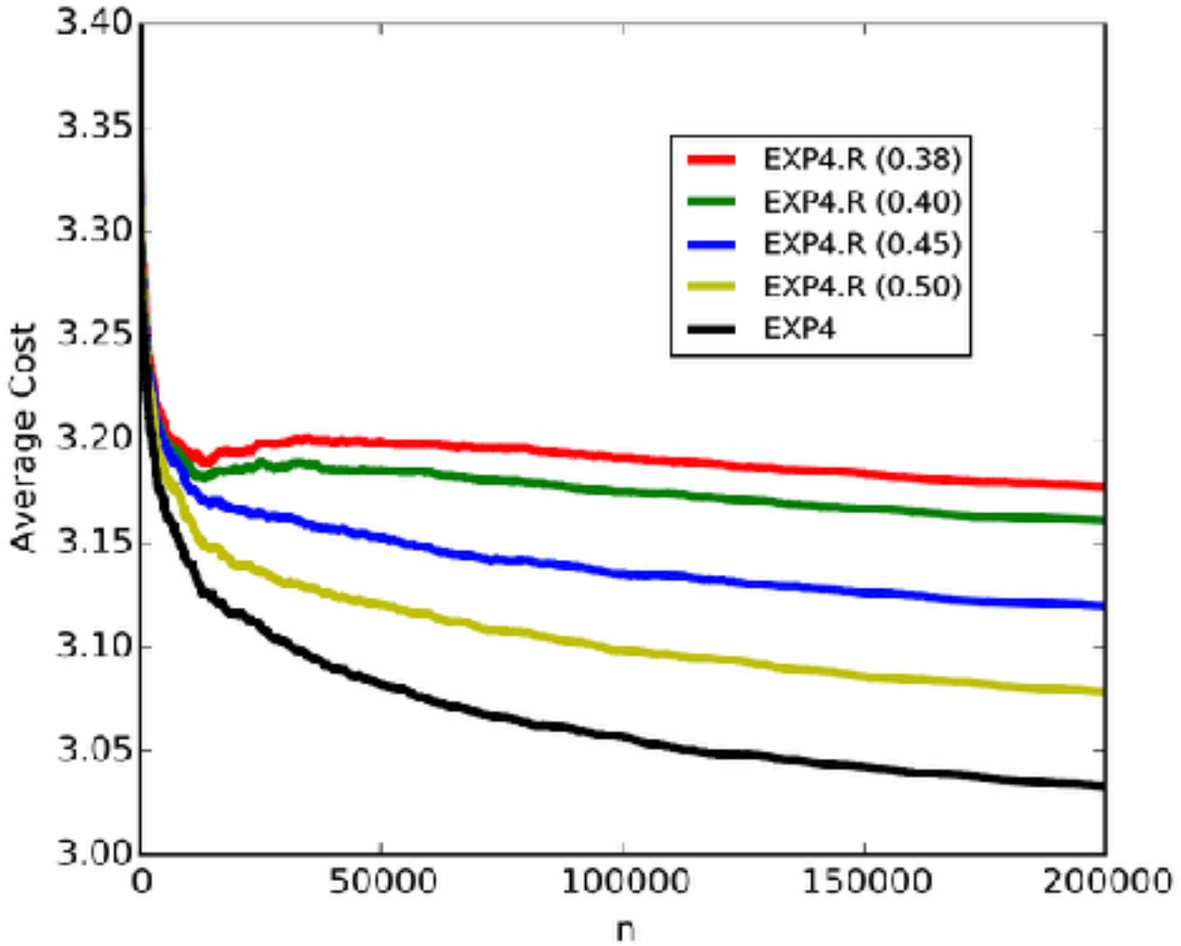


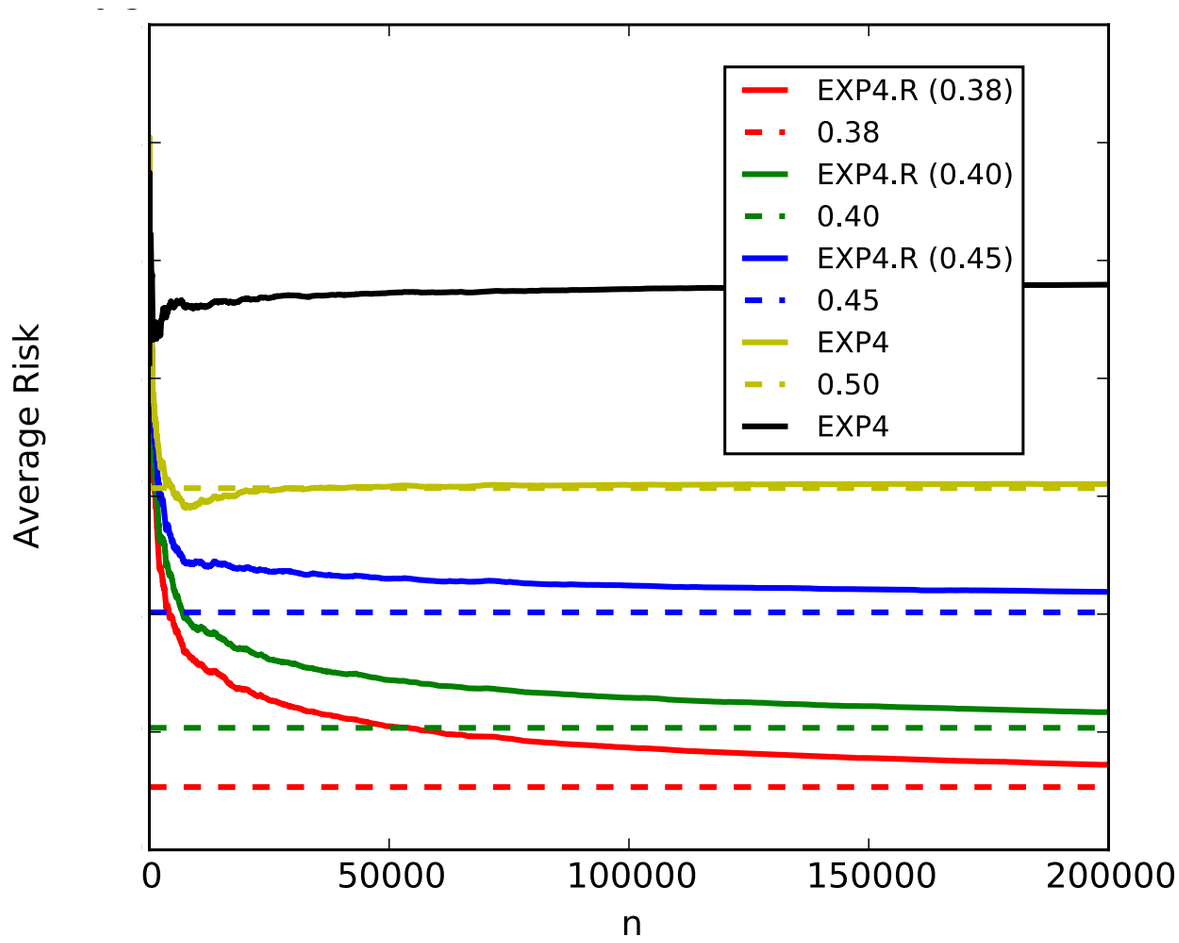
Context is the RBF feature with respect to the nine way points.

We have 4^9 experts. Namely each expert suggests one action at each waypoint

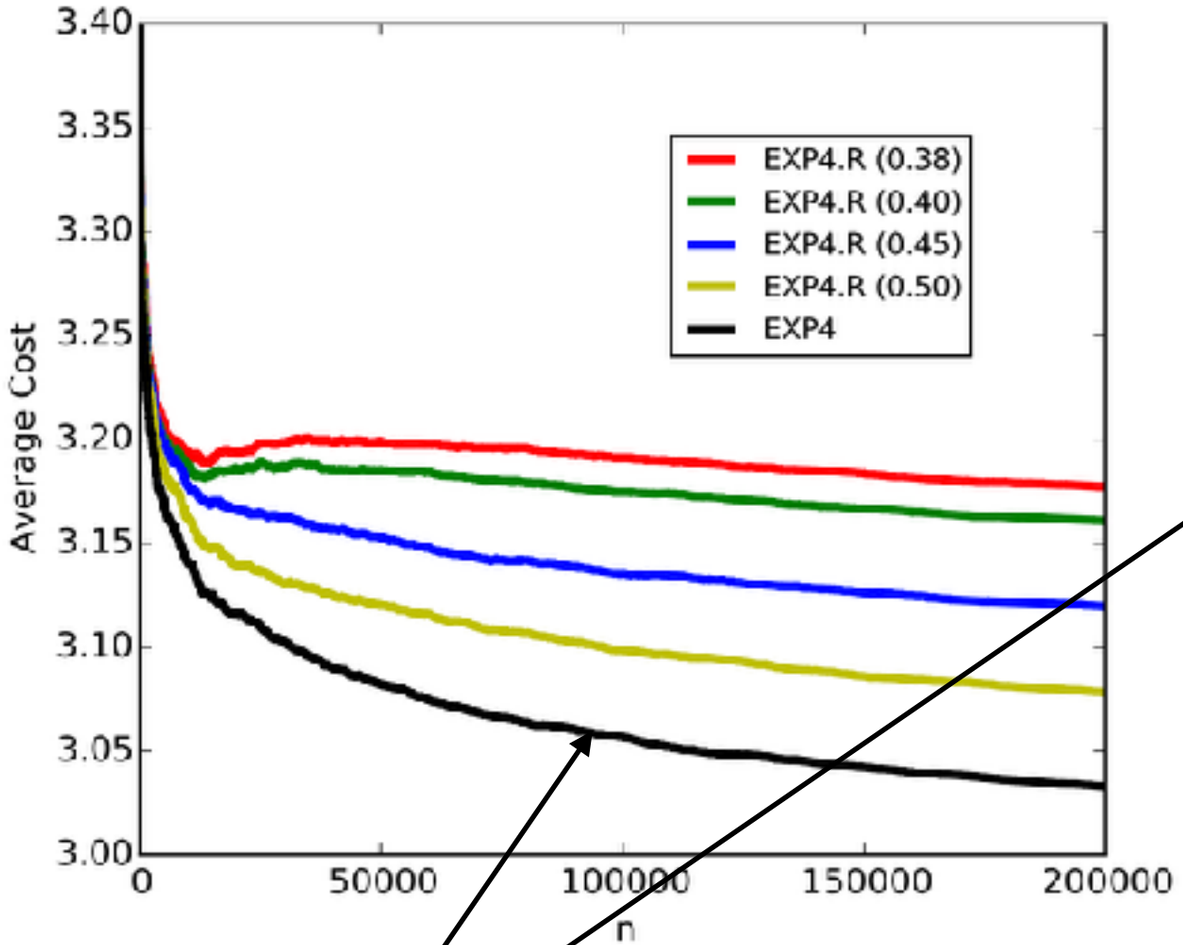
We ran the EXP4.R with different risk thresholds

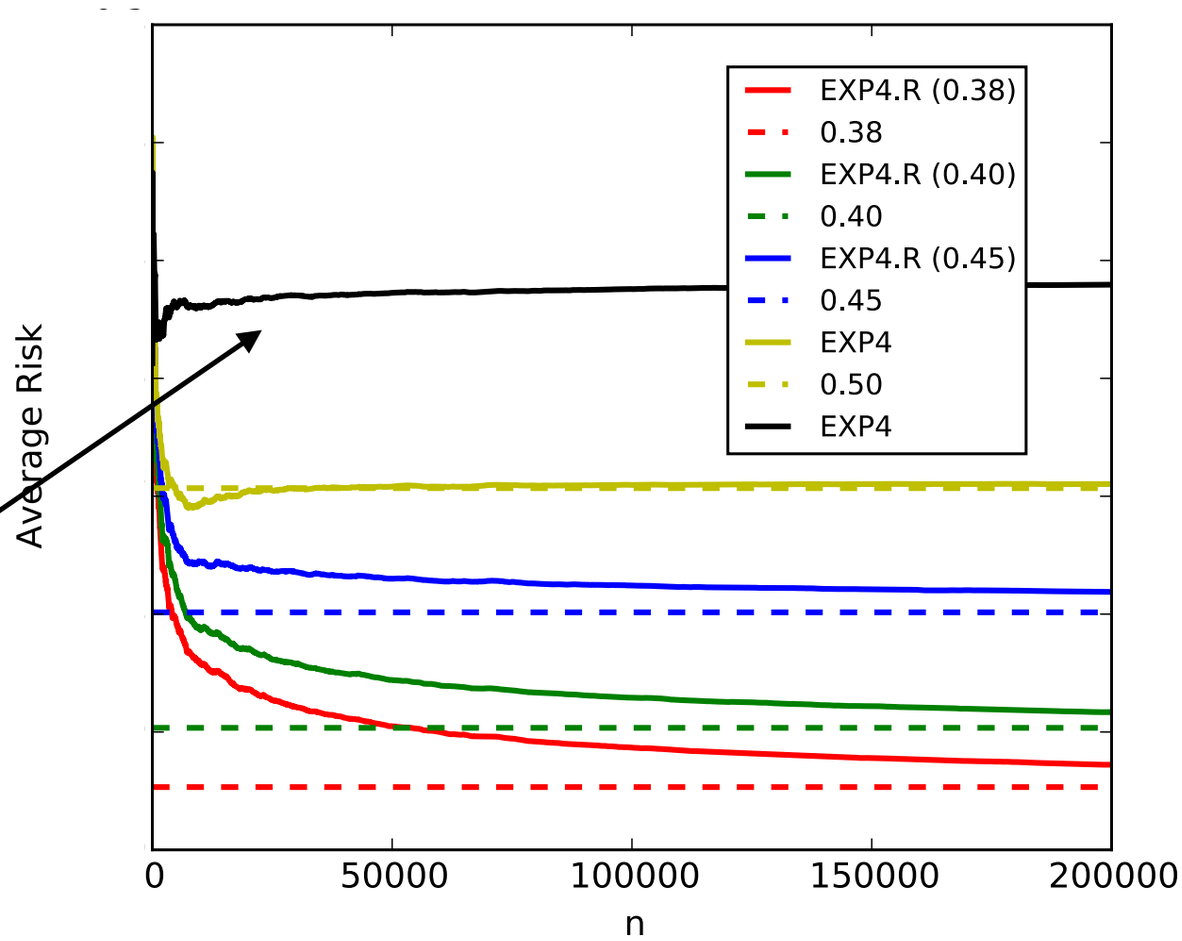
Simulation


3.40
3.35
3.30
3.25
3.20
3.15
3.10
3.05
3.00
Average Cost
0
50000
100000
150000
200000
n
EXP4.R (0.38)
EXP4.R (0.40)
EXP4.R (0.45)
EXP4.R (0.50)
EXP4


Average Risk
0
50000
100000
150000
200000
n
EXP4.R (0.38)
0.38
EXP4.R (0.40)
0.40
EXP4.R (0.45)
0.45
EXP4
0.50
EXP4

Simulation

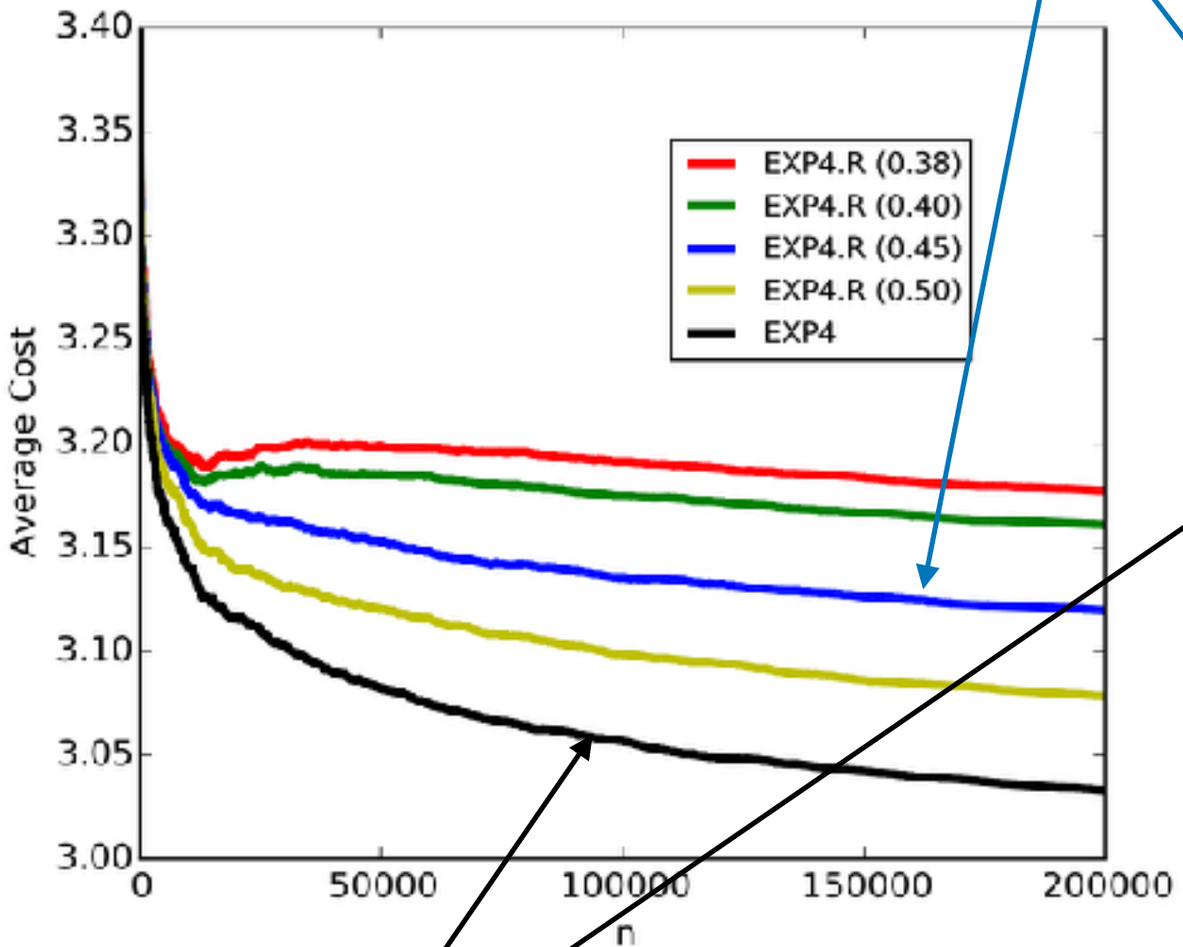


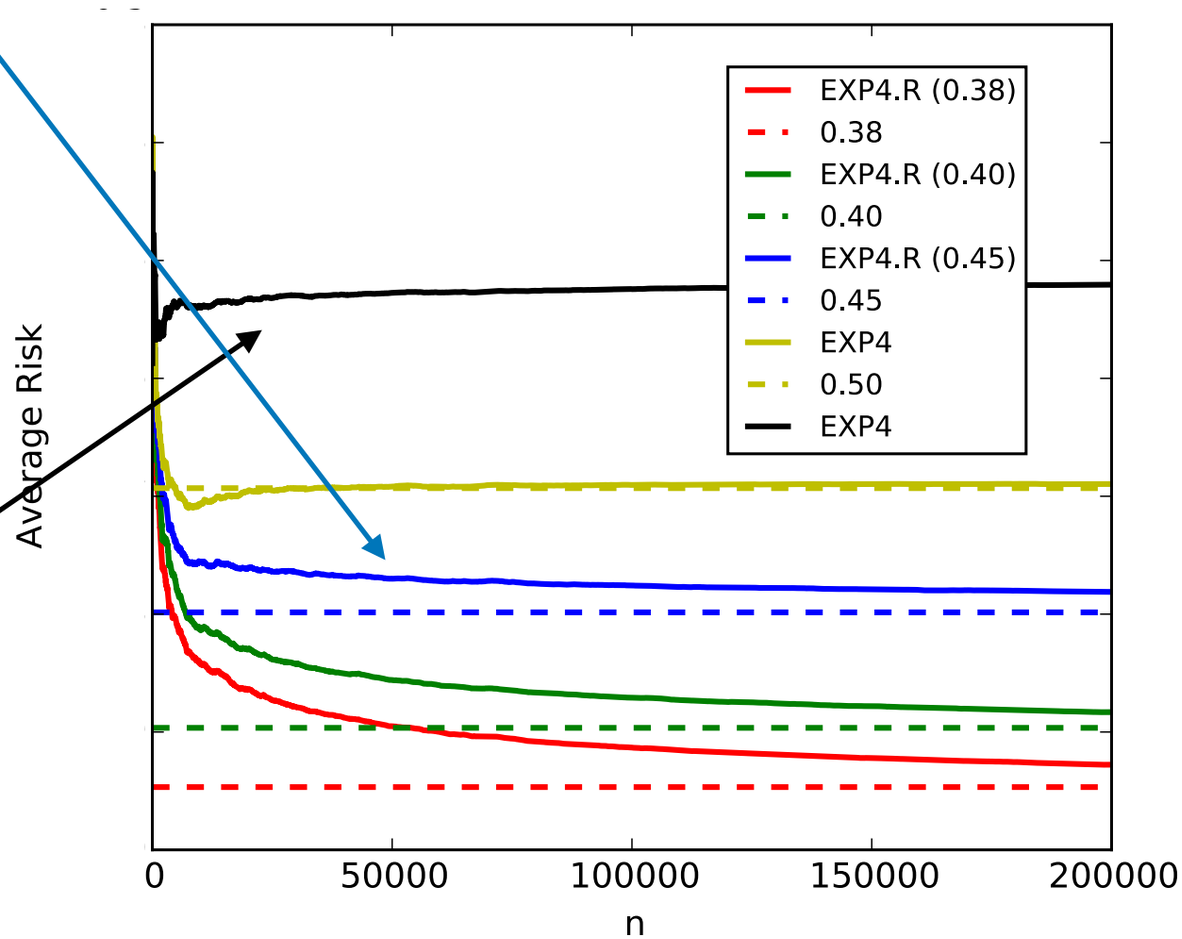


No risk constraint

Simulation

Threshold 0.45





No risk constraint

Conclusion and Future Work

1. We consider sequential decision making problem with additional adversarial constraints.

2. In our applications these constraints are used to model safety related issues in decision making process.

3. Is there any algorithm that can achieve $\sqrt{T}$ total regret and $\sqrt{T}$ total constraint violation simultaneously?

4. Is there better heuristic we can leverage to achieve tighter regret and constrain violation in high probability?

Thanks

Wen Sun
[wensun@cs.cmu.edu]

The Robotics Institute
Carnegie Mellon University

Microsoft Research