# First-Order Regret in Reinforcement Learning with Linear Function Approximation: A Robust Estimation Approach

**Andrew Wagenmaker**[1], Yifang Chen[1], Max Simchowitz[2], Simon S. Du[1], Kevin Jamieson[1]

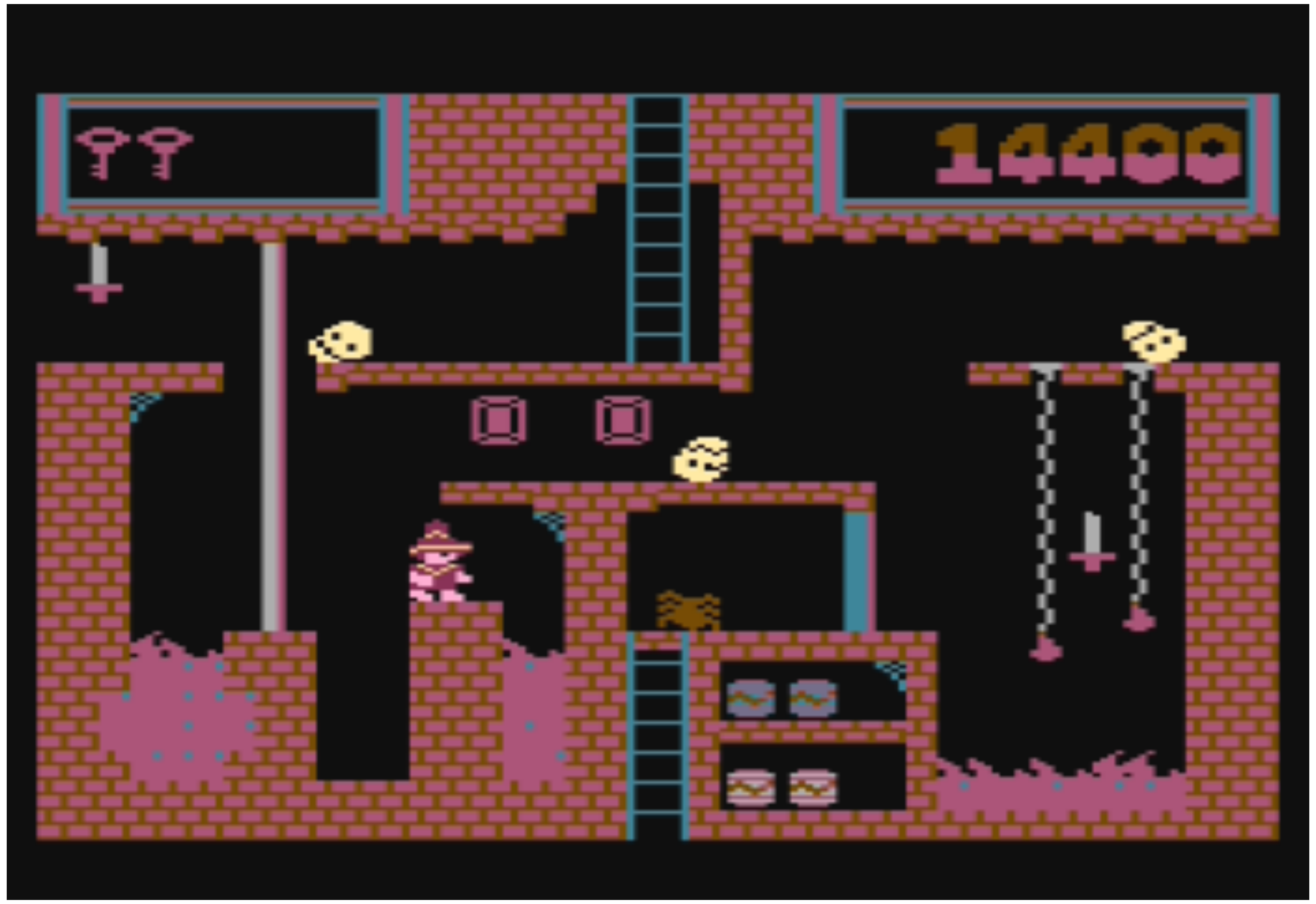
1. University of Washington, 2. MIT

# Motivation

In practical RL settings, rewards may be **sparse** and **hard-to-reach**

# Motivation

In practical RL settings, rewards may be **sparse** and **hard-to-reach**

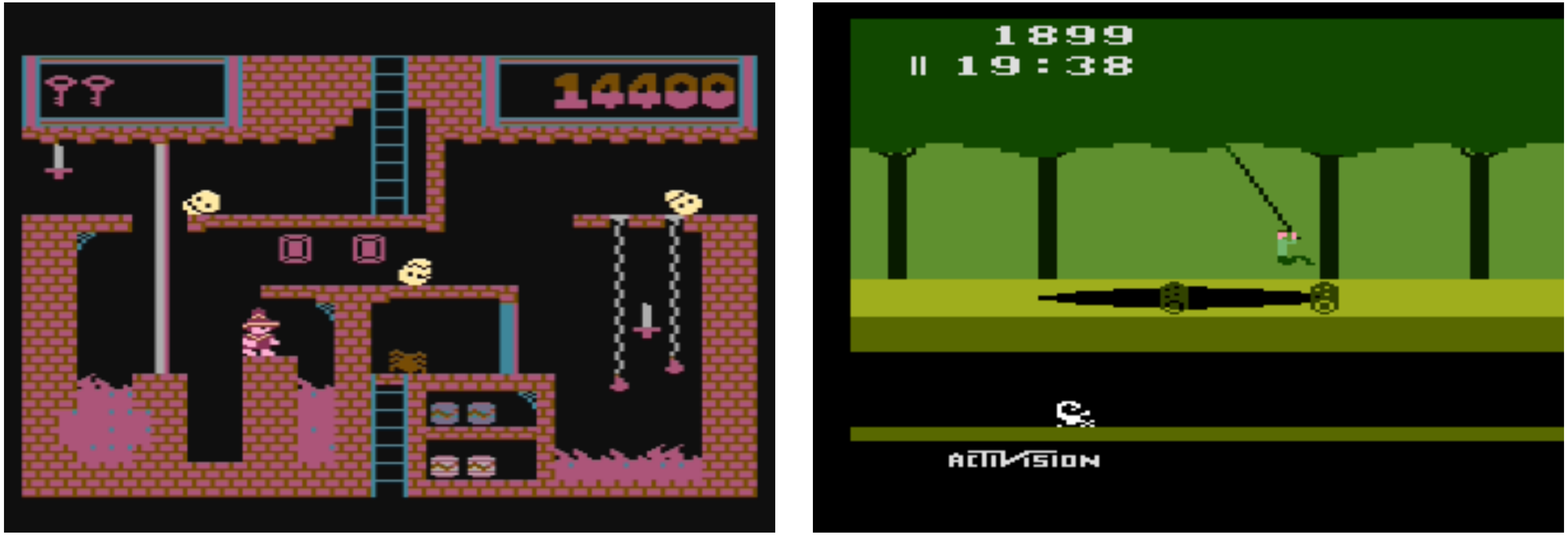




**Atari's Montezuma's Revenge & Pitfall**

# Motivation

In practical RL settings, rewards may be **sparse** and **hard-to-reach**

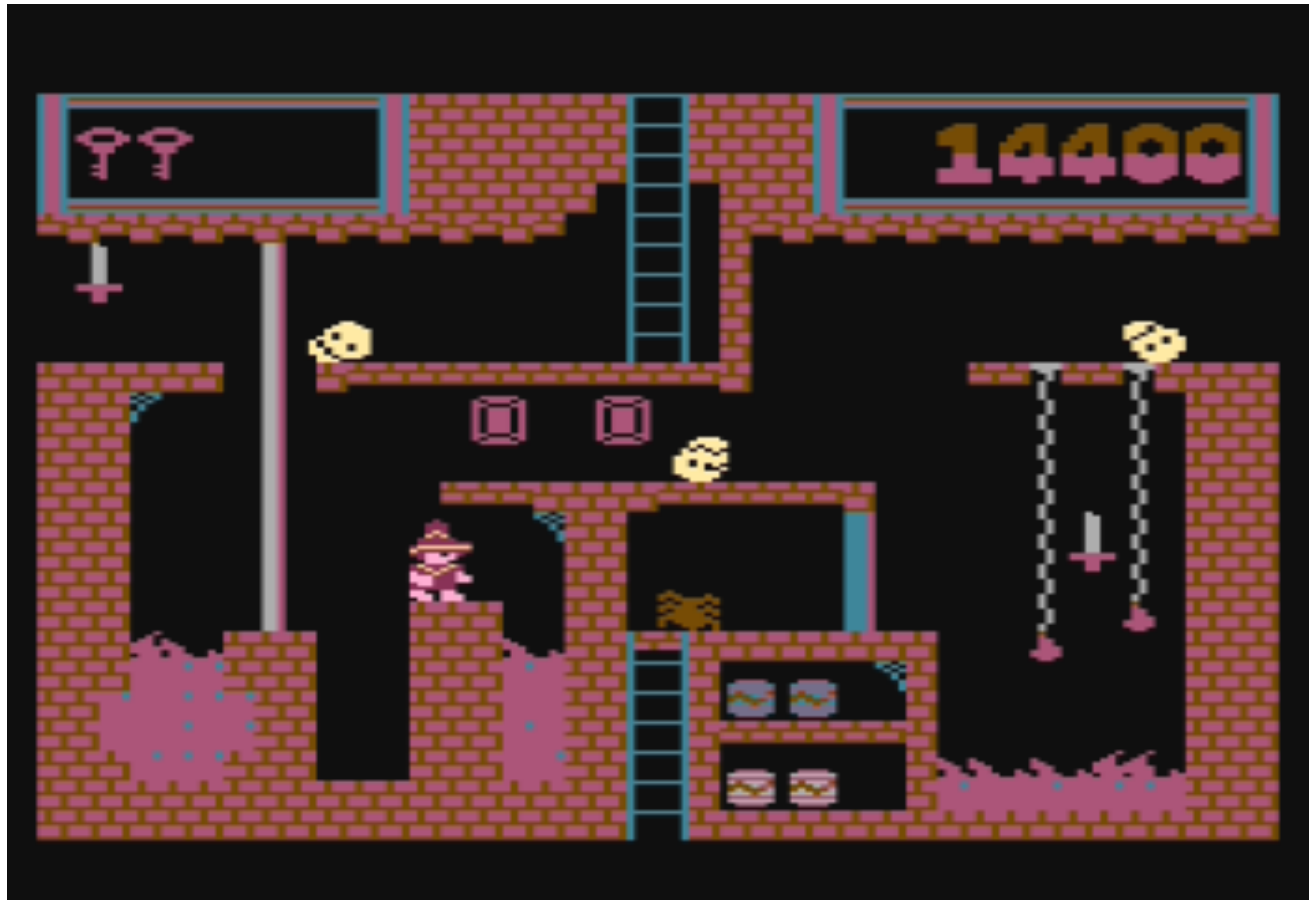



**Atari's Montezuma's Revenge & Pitfall**

In such settings, could have $V_1^\star \ll 1$, for $V_1^\star$ the maximum expected reward

# Motivation

The maximum attainable reward gives a **scale** to the problem

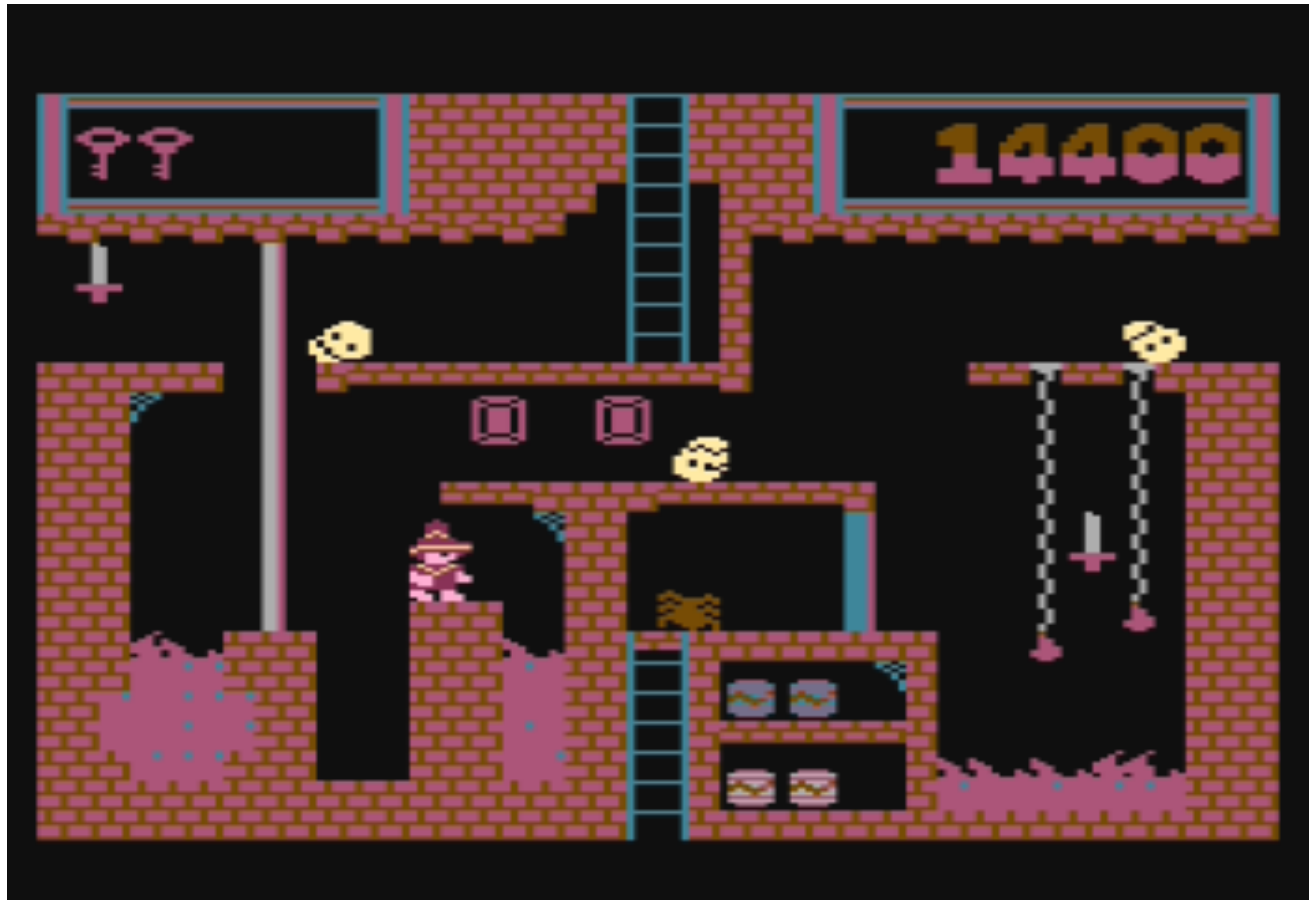


# Motivation

The maximum attainable reward gives a **scale** to the problem

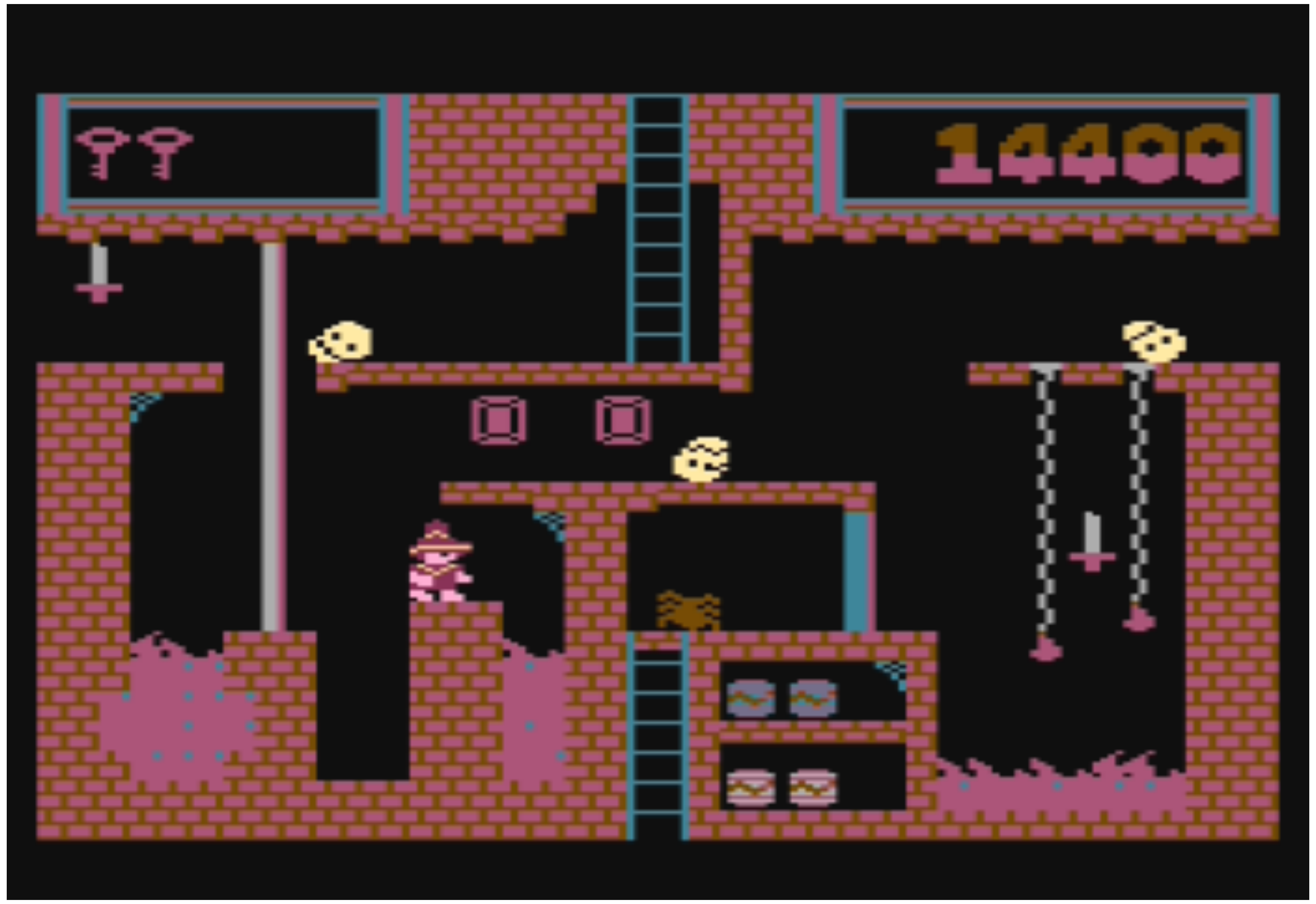
In general in RL, we are interested in finding some policy $\hat{\pi}$ with performance close to that of the optimal policy:

$$V_1^\star - V_1^{\hat{\pi}} \leq \epsilon$$



# Motivation

The maximum attainable reward gives a **scale** to the problem

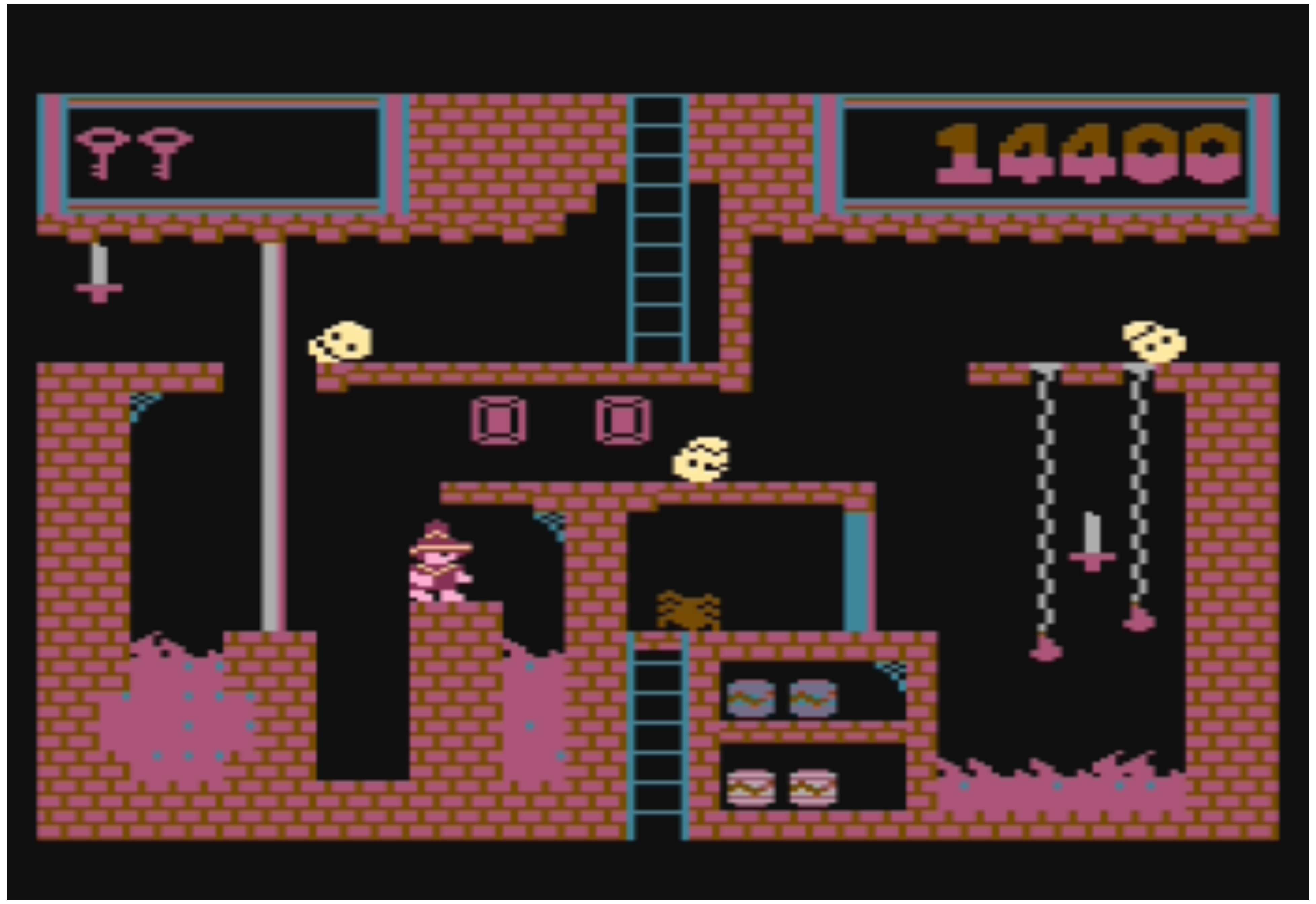



In general in RL, we are interested in finding some policy $\hat{\pi}$ with performance close to that of the optimal policy:

$$V_1^\star - V_1^{\hat{\pi}} \leq \epsilon$$

Say we want $V_1^{\hat{\pi}} \geq 0.9 \cdot V_1^\star$, then need $\epsilon \sim 0.1 \cdot V_1^\star$

# Motivation




The maximum attainable reward gives a **scale** to the problem

In general in RL, we are interested in finding some policy $\hat{\pi}$ with performance close to that of the optimal policy:

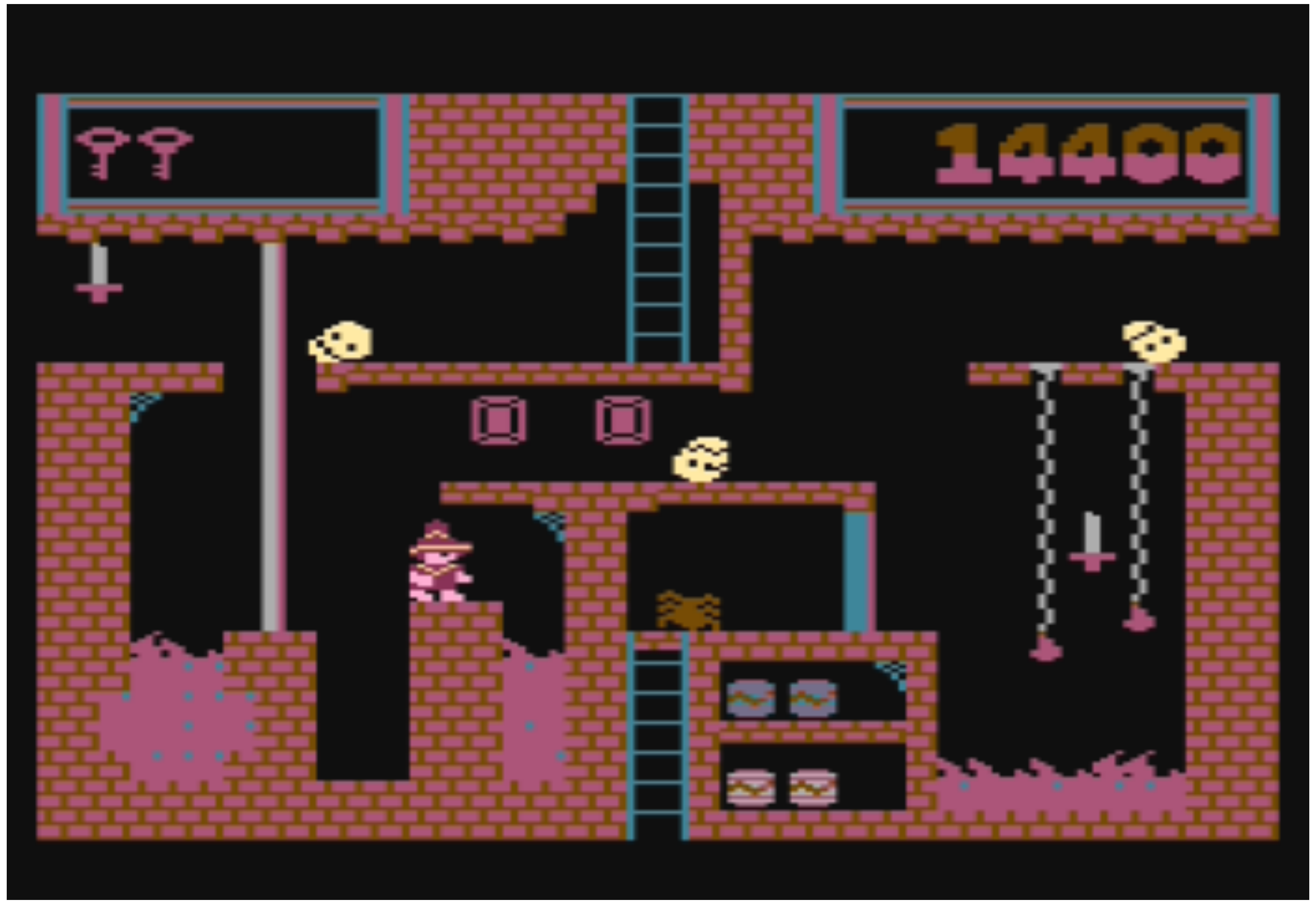
$$V_1^\star - V_1^{\hat{\pi}} \leq \epsilon$$

Say we want $V_1^{\hat{\pi}} \geq 0.9 \cdot V_1^\star$, then need $\epsilon \sim 0.1 \cdot V_1^\star$

Standard guarantees scale as $O(1/\epsilon^2) = O(1/(V_1^\star)^2)$

# Can we do better?

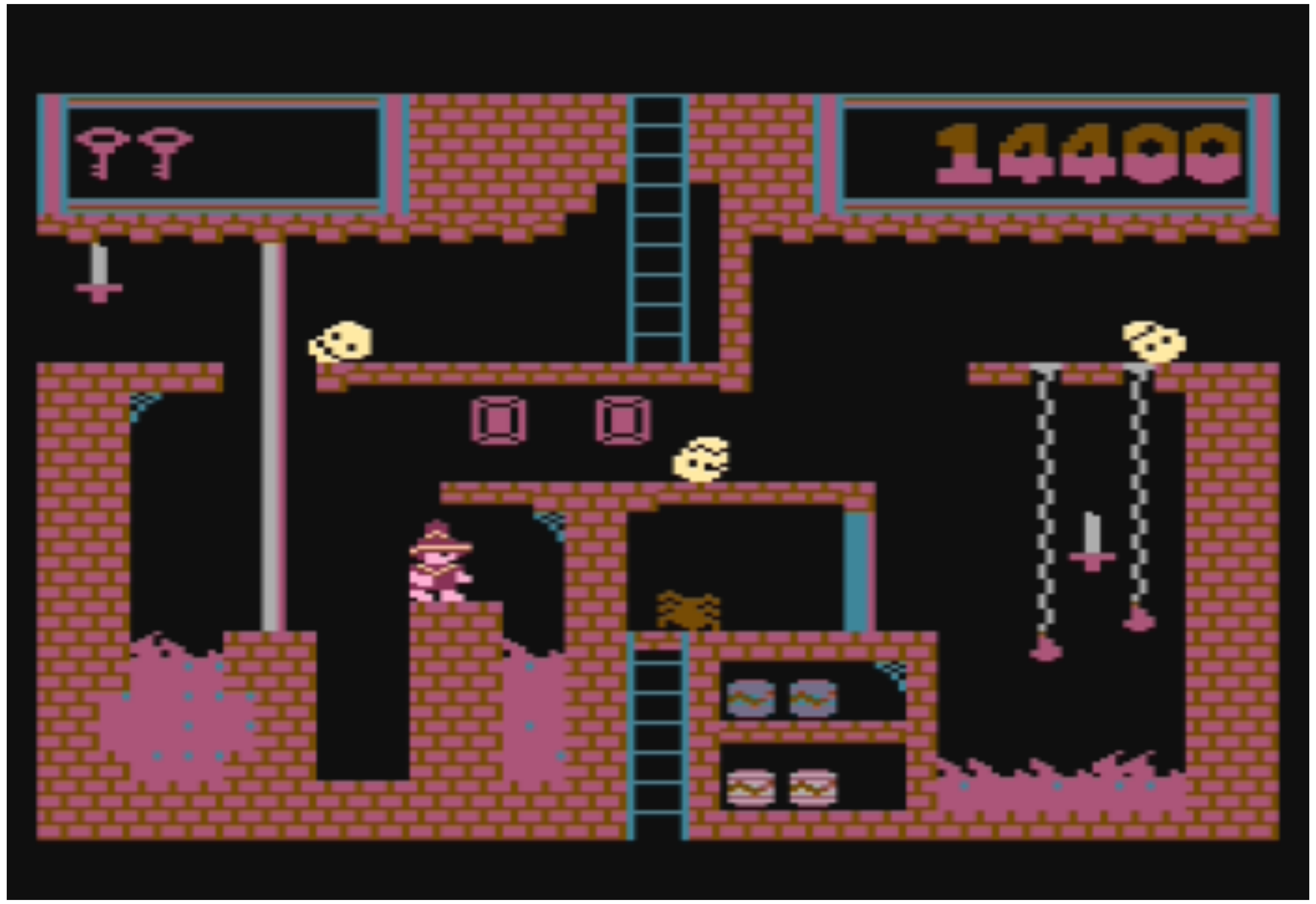
It is known that in online and statistical learning, **first-order** guarantees are possible—guarantees that scale with the **value of the optimal policy**



# Can we do better?

It is known that in online and statistical learning, **first-order** guarantees are possible —guarantees that scale with the **value of the optimal policy**

Taking inspiration from this literature, we might hope that we can obtain a scaling of:

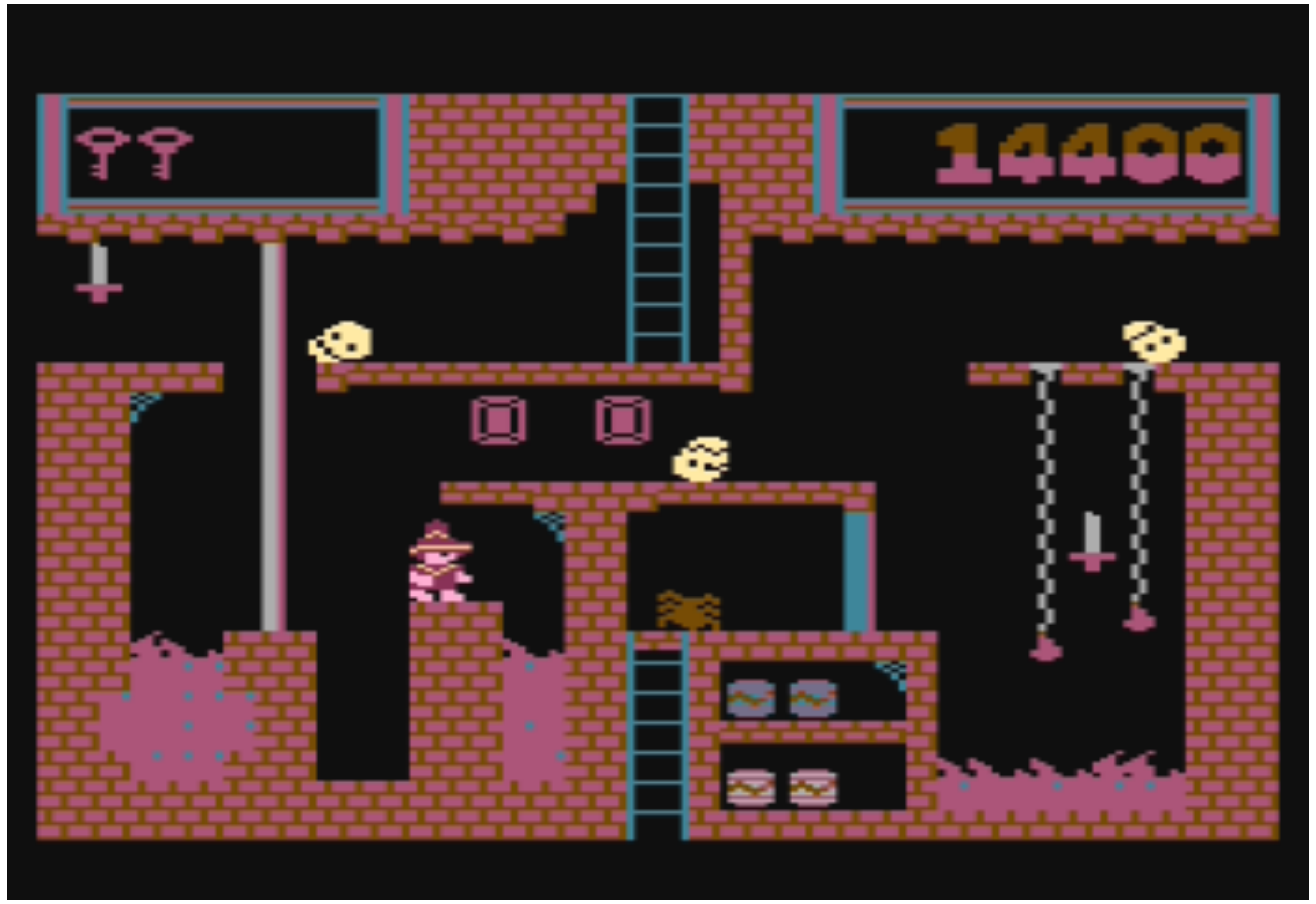


# Can we do better?

It is known that in online and statistical learning, **first-order** guarantees are possible—guarantees that scale with the **value of the optimal policy**

Taking inspiration from this literature, we might hope that we can obtain a scaling of:

$$O(V_1^\star/\epsilon^2)$$



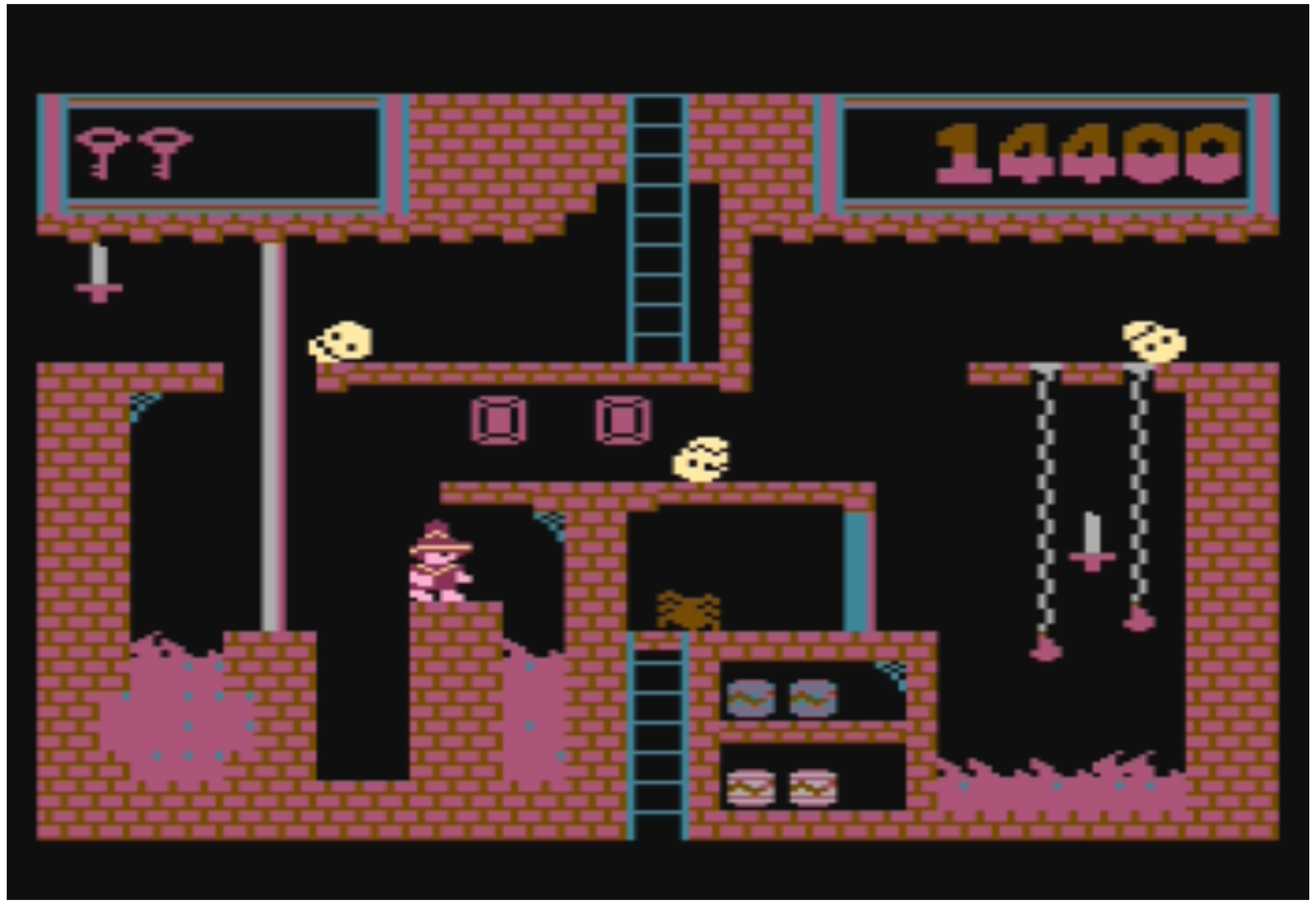
# Can we do better?

It is known that in online and statistical learning, **first-order** guarantees are possible—guarantees that scale with the **value of the optimal policy**

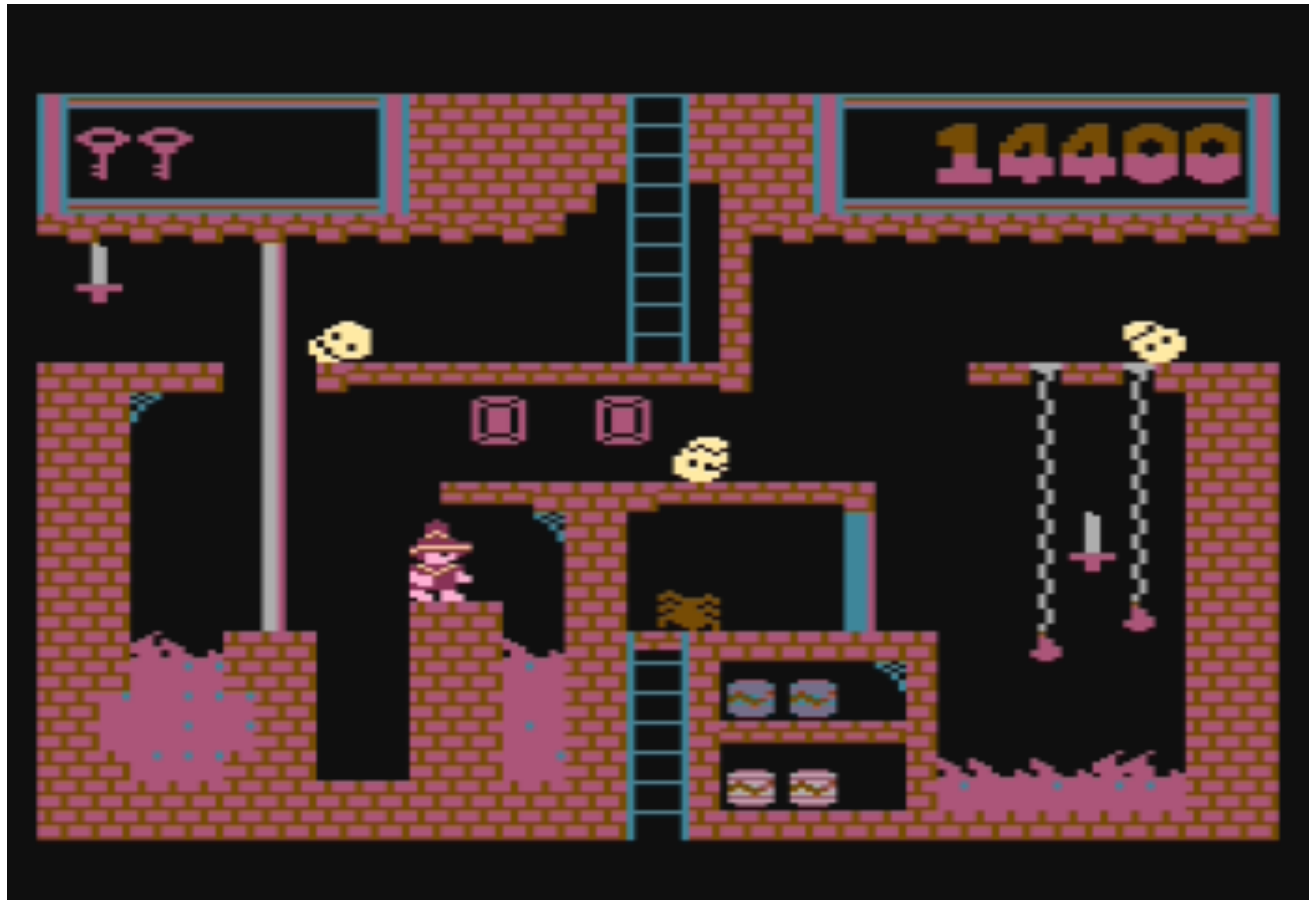
Taking inspiration from this literature, we might hope that we can obtain a scaling of:

$$O(V_1^\star/\epsilon^2)$$

→ Improves $O(1/(V_1^\star)^2)$ scaling to $O(1/V_1^\star)$



# Can we do better?




It is known that in online and statistical learning, **first-order** guarantees are possible —guarantees that scale with the **value of the optimal policy**

Taking inspiration from this literature, we might hope that we can obtain a scaling of:

$$O(V_1^\star/\epsilon^2)$$

→ Improves $O(1/(V_1^\star)^2)$ scaling to $O(1/V_1^\star)$

More importantly, to obtain such guarantees, algorithms must explore more efficiently, yielding better practical performance

# Our Contributions

We are particularly interested in studying this problem in the setting of MDPs with **large state spaces** using **function approximation**

# Our Contributions

We are particularly interested in studying this problem in the setting of MDPs with **large state spaces** using **function approximation**

- We obtain a ***first-order-style*** regret bound in linear MDPs of $O(\sqrt{V_1^\star K})$, which translates to a PAC guarantee of $O(V_1^\star/\epsilon^2)$

# Our Contributions

We are particularly interested in studying this problem in the setting of MDPs with **large state spaces** using **function approximation**

- We obtain a ***first-order-style*** regret bound in linear MDPs of $O(\sqrt{V_1^\star K})$, which translates to a PAC guarantee of $O(V_1^\star/\epsilon^2)$
- Our algorithm critically relies on a novel extension of the **robust Catoni estimator** to correlated, heteroscedastic data

# Our Contributions

We are particularly interested in studying this problem in the setting of MDPs with **large state spaces** using **function approximation**

- We obtain a ***first-order-style*** regret bound in linear MDPs of $O(\sqrt{V_1^\star K})$, which translates to a PAC guarantee of $O(V_1^\star/\epsilon^2)$
- Our algorithm critically relies on a novel extension of the **robust Catoni estimator** to correlated, heteroscedastic data

To our knowledge, ours is the first result to show first-order regret in RL with large state spaces

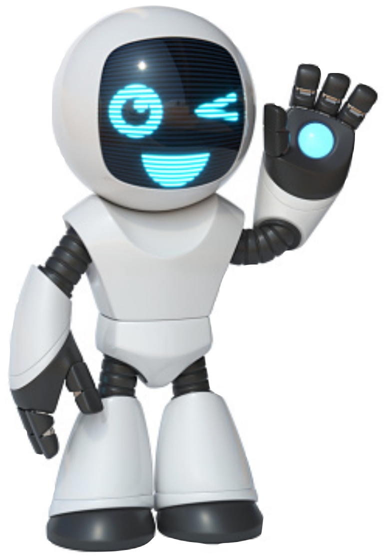
# Preliminaries

# Markov Decision Processes (MDPs)

**Episodic RL:**

# Markov Decision Processes (MDPs)

**Episodic RL:**

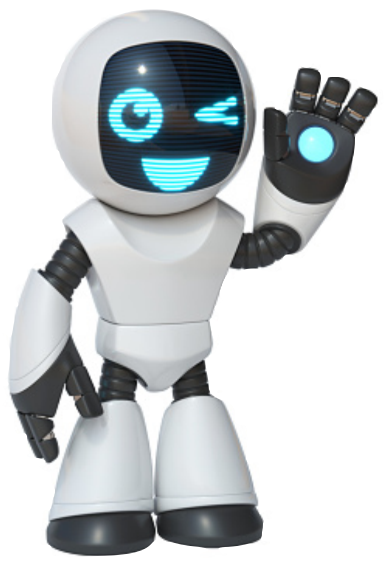


Agent

# Markov Decision Processes (MDPs)

**Episodic RL:**

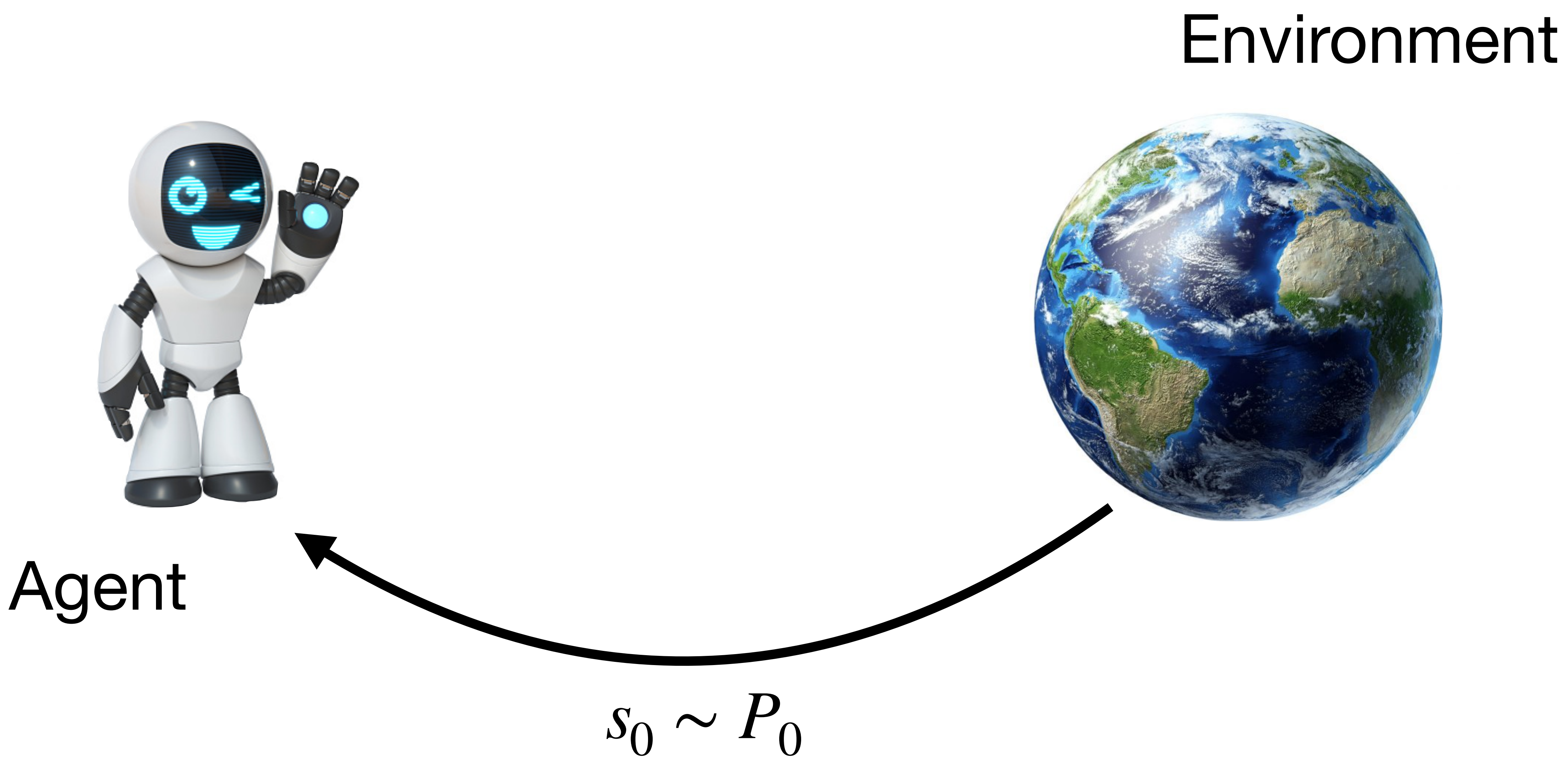
Environment



Agent

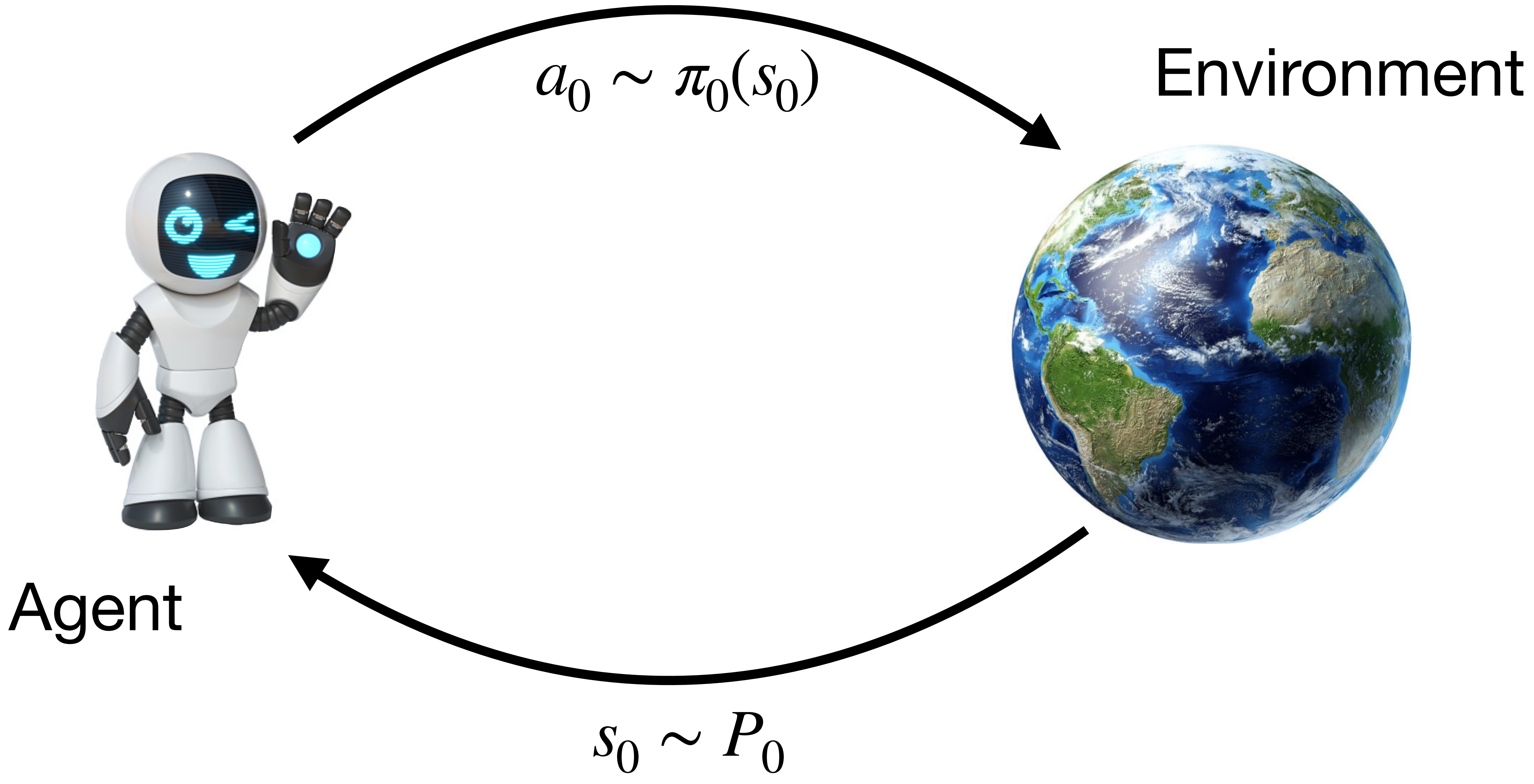
# Markov Decision Processes (MDPs)

**Episodic RL:**



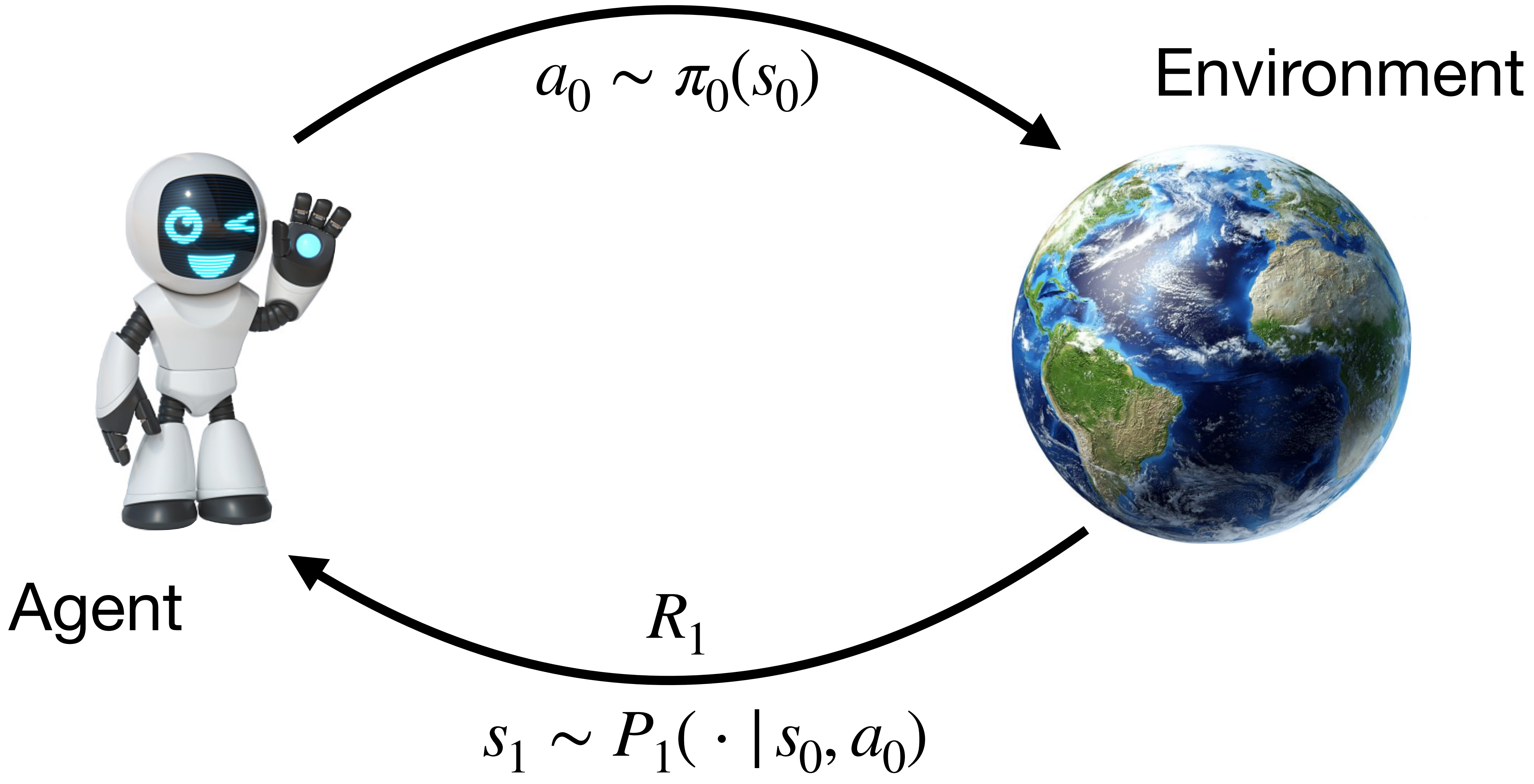
# Markov Decision Processes (MDPs)

**Episodic RL:**



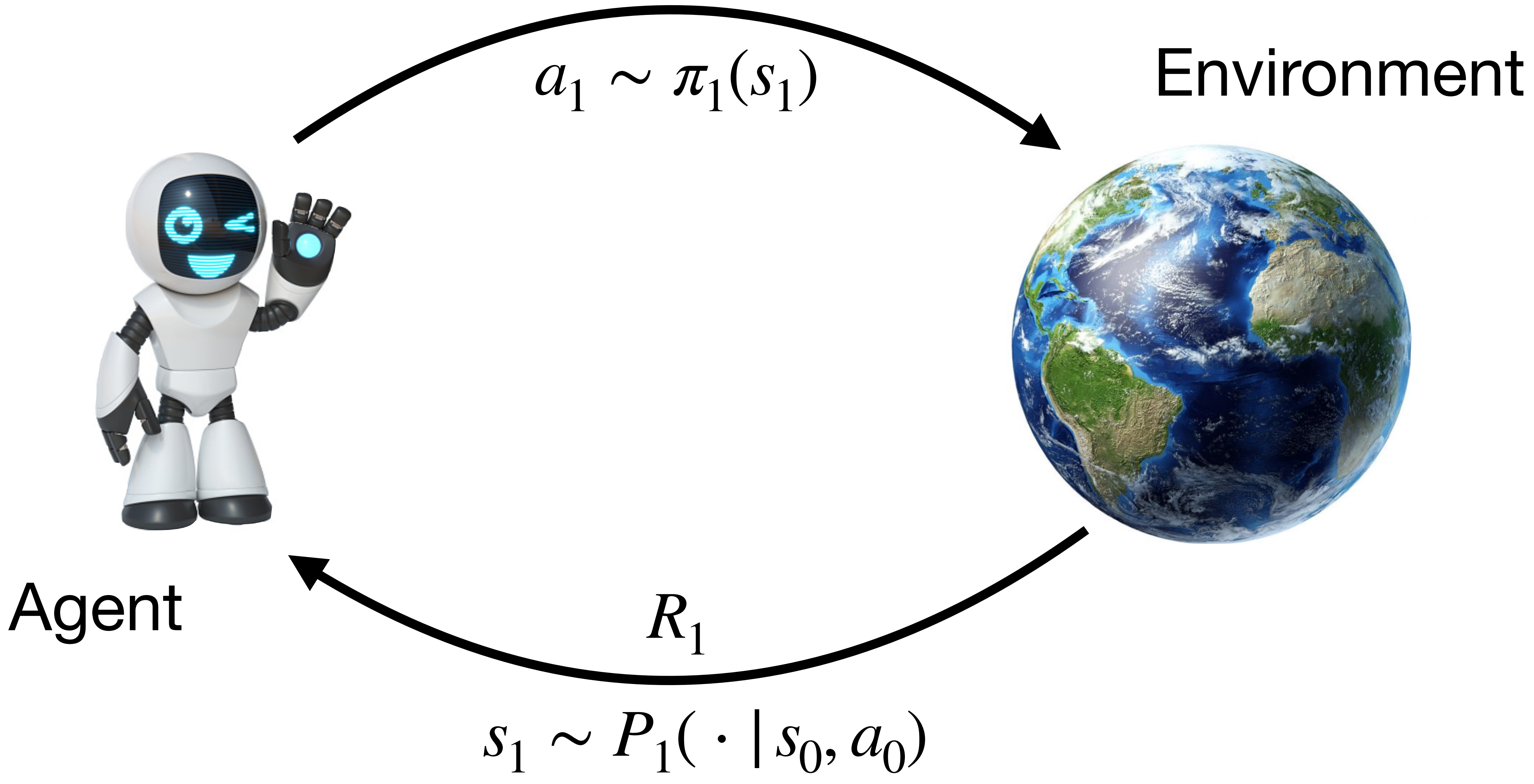
# Markov Decision Processes (MDPs)

**Episodic RL:**



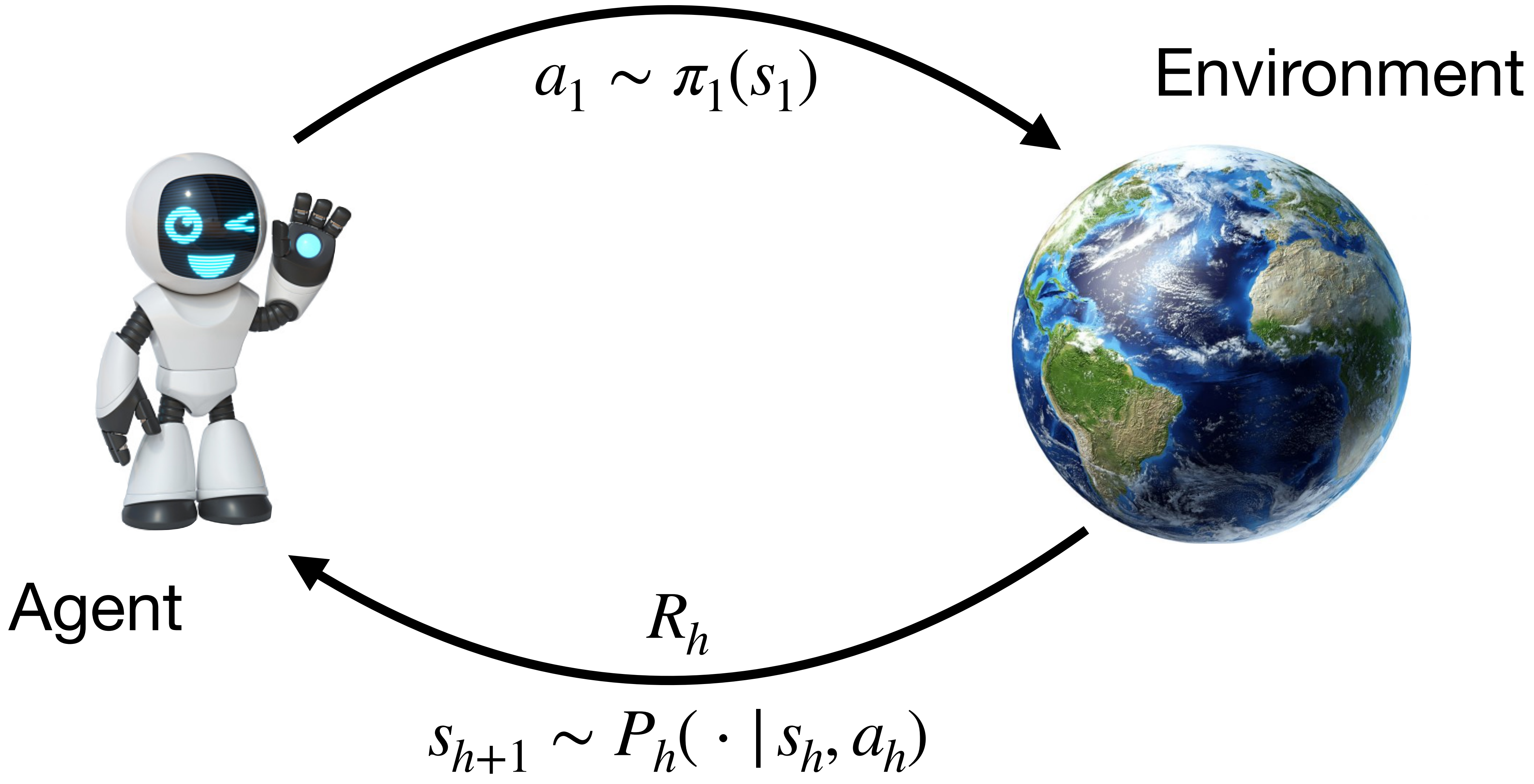
# Markov Decision Processes (MDPs)

**Episodic RL:**



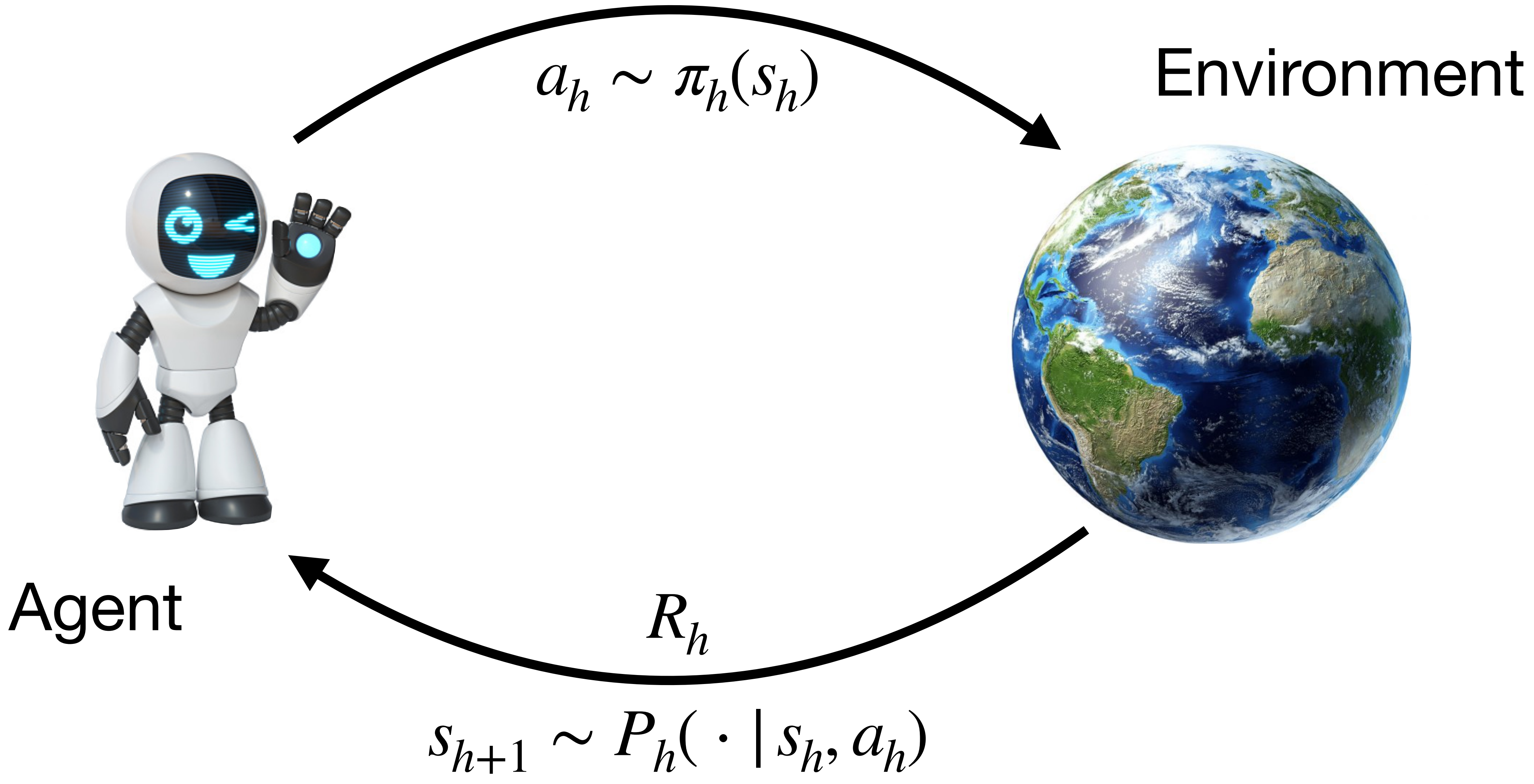
# Markov Decision Processes (MDPs)

**Episodic RL:**



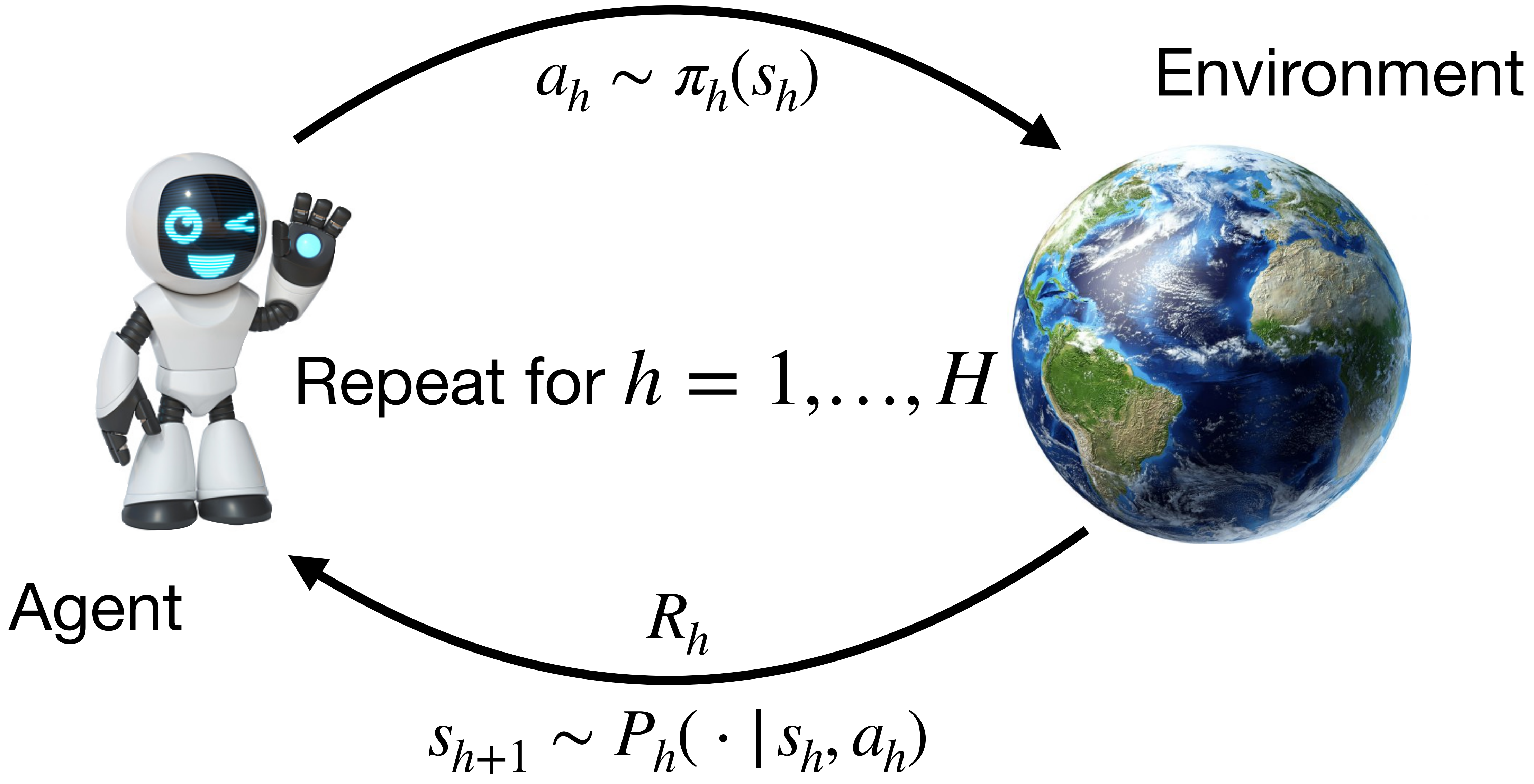
# Markov Decision Processes (MDPs)

**Episodic RL:**



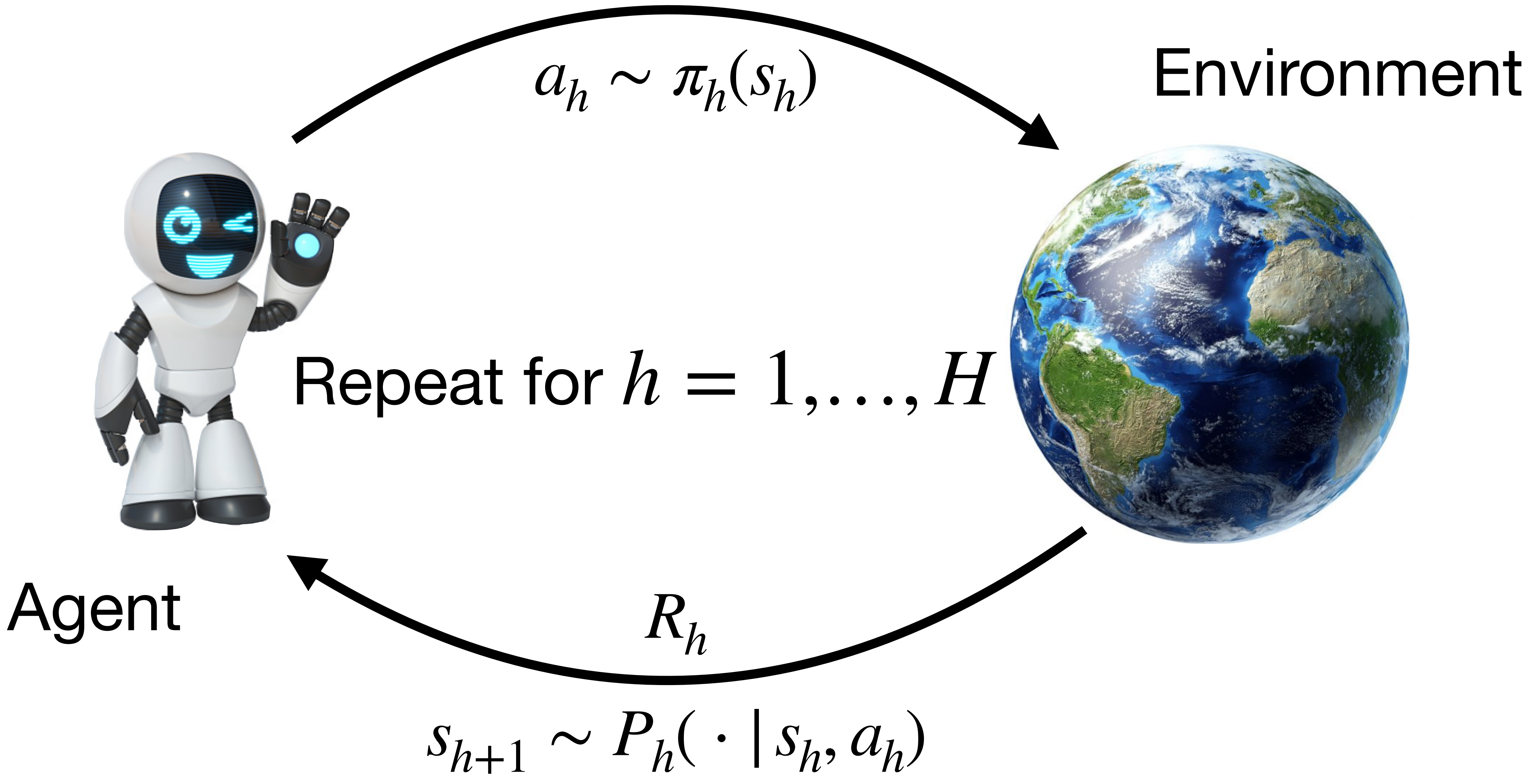
# Markov Decision Processes (MDPs)

**Episodic RL:**



# Markov Decision Processes (MDPs)

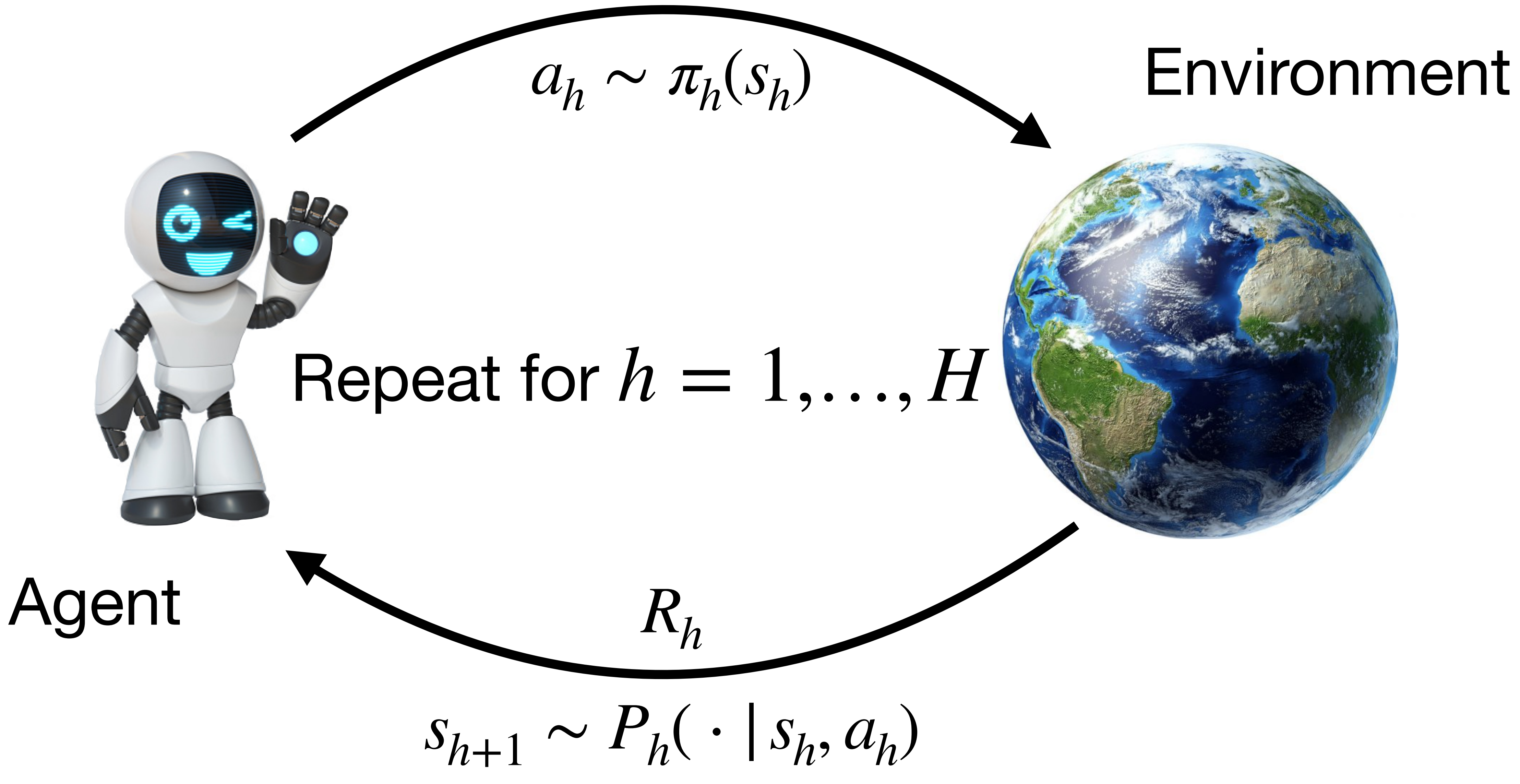
**Episodic RL:**




We assume $\{P_h\}_{h=1}^{H}$ is **unknown**, and $\{R_h\}_{h=1}^{H}$ **known**

# Markov Decision Processes (MDPs)

**Episodic RL:**

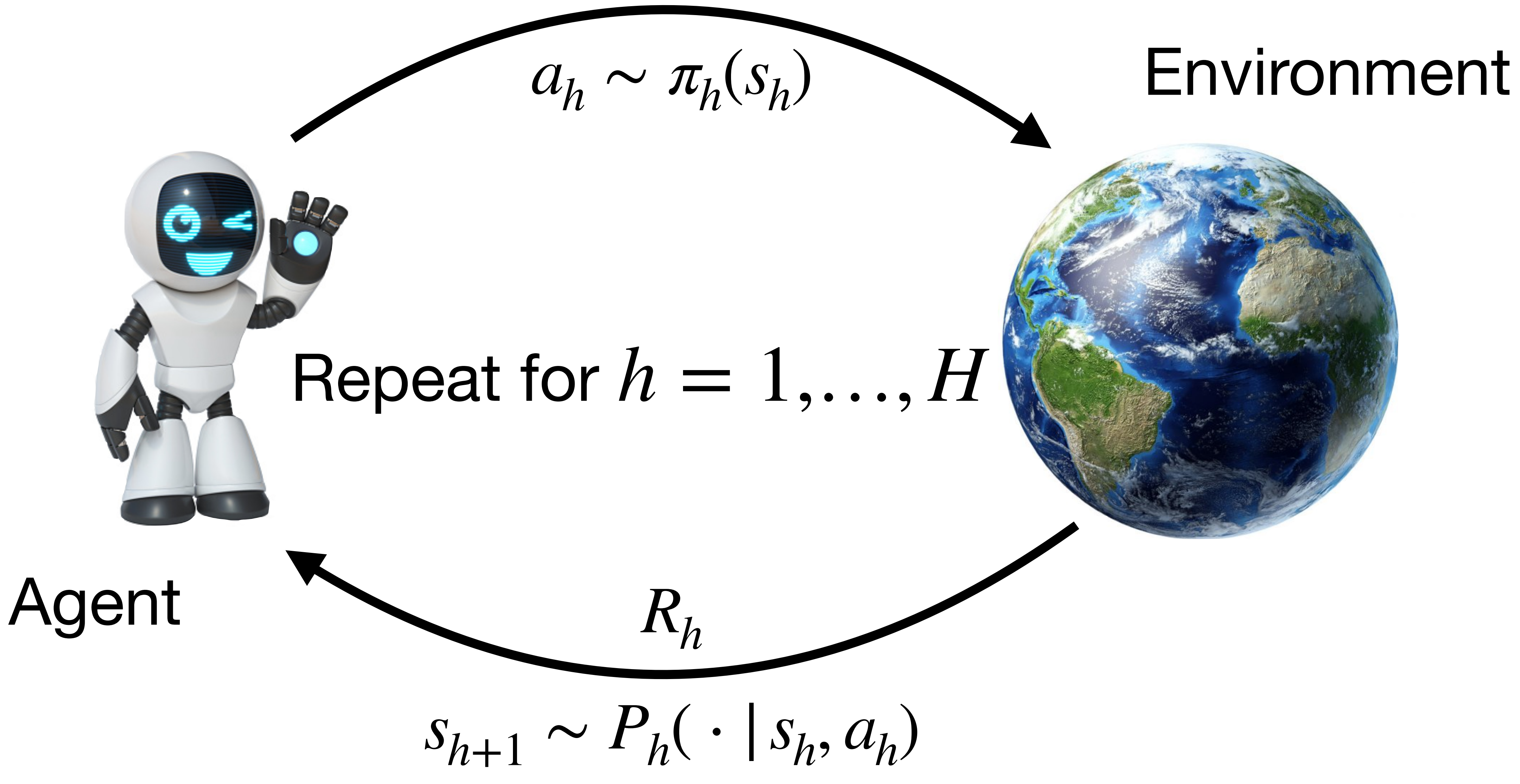



We consider **Linear MDPs**:

We assume $\{P_h\}_{h=1}^{H}$ is **unknown**, and $\{R_h\}_{h=1}^{H}$ **known**

# Markov Decision Processes (MDPs)

**Episodic RL:**



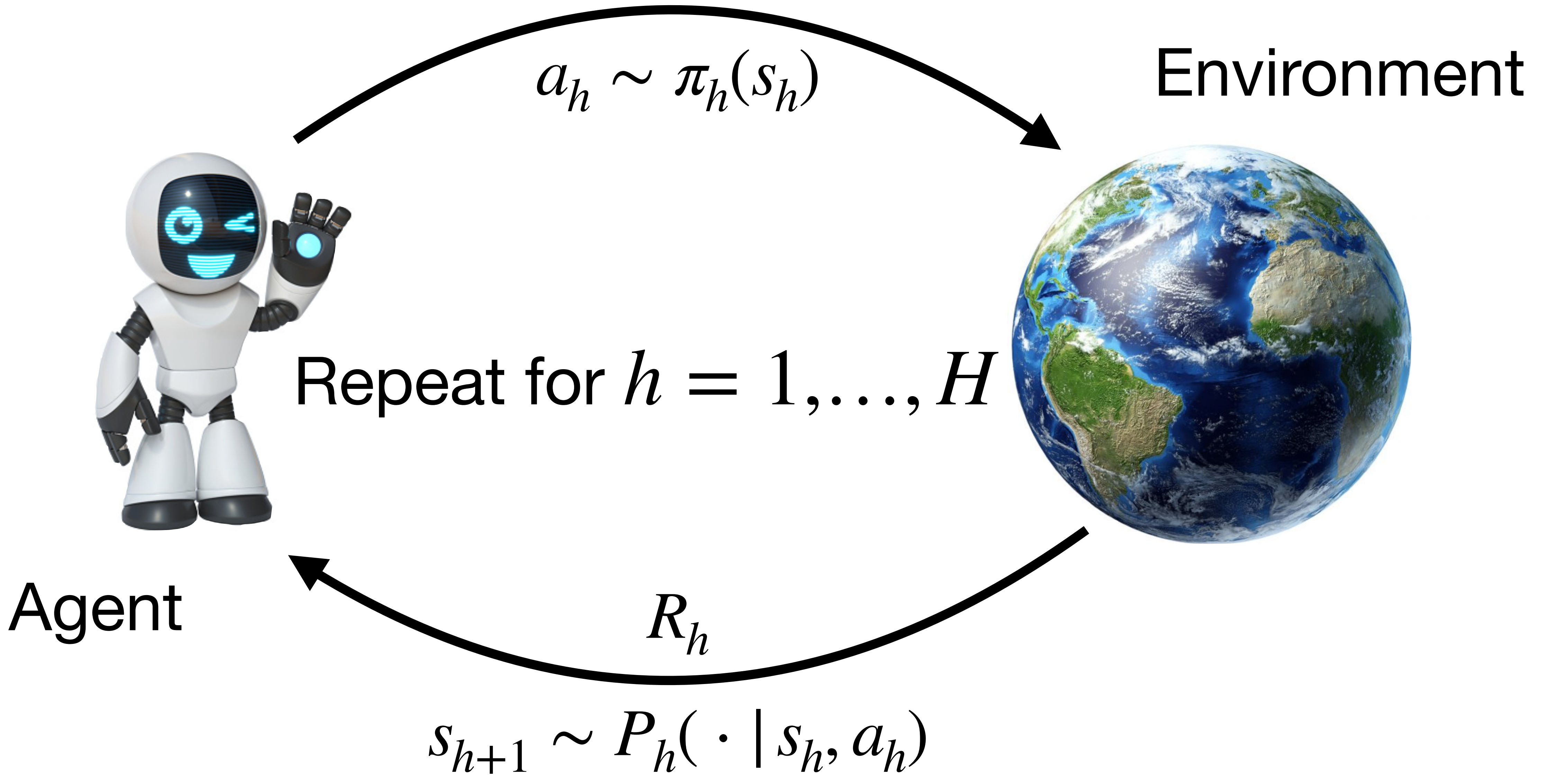

We consider **Linear MDPs**:

- Known feature vectors $\phi(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$

We assume $\{P_h\}_{h=1}^H$ is **unknown**, and $\{R_h\}_{h=1}^H$ **known**

# Markov Decision Processes (MDPs)

**Episodic RL:**




We consider **Linear MDPs**:

- Known feature vectors $\phi(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$
- $H$ unknown signed measures $\mu_h \in \mathbb{R}^d$ over $\mathcal{S}$ such that: $P_h(\cdot \mid s, a) = \langle \phi(s, a), \mu_h(\cdot) \rangle$

We assume $\{P_h\}_{h=1}^{H}$ is **unknown**, and $\{R_h\}_{h=1}^{H}$ **known**

# Preliminaries: Regret

Let:

# Preliminaries: Regret

Let:

- $V_1^\pi := \mathbb{E}_\pi[\sum_{h=1}^H R_h(s_h, a_h)]$, the expected reward of policy $\pi$

# Preliminaries: Regret

Let:

- $V_1^{\pi} := \mathbb{E}_{\pi}[\sum_{h=1}^{H} R_h(s_h, a_h)]$, the expected reward of policy $\pi$
- $V_1^{\star} := \sup_{\pi} V_1^{\pi}$

# Preliminaries: Regret

Let:

- $V_1^\pi := \mathbb{E}_\pi[\sum_{h=1}^H R_h(s_h, a_h)]$, the expected reward of policy $\pi$
- $V_1^\star := \sup_\pi V_1^\pi$

Consider playing some algorithm for $K$ episodes where at episode $k$ we play policy $\pi_k$. Then the **regret** is defined as:

$$\mathcal{R}_K := \sum_{k=1}^K (V_1^\star - V_1^{\pi_k})$$

# Preliminaries: Regret

Let:

- $V_1^\pi := \mathbb{E}_\pi[\sum_{h=1}^H R_h(s_h, a_h)]$, the expected reward of policy $\pi$
- $V_1^\star := \sup_\pi V_1^\pi$

**Goal:** Obtain regret scaling with $V_1^\star$

Consider playin … sode $k$ we play policy $\pi_k$. Then the **regret** is defined as:

$$\mathscr{R}_K := \sum_{k=1}^K (V_1^\star - V_1^{\pi_k})$$

# Main Result

# Main Result

**Theorem.** There exists an algorithm, FORCE, which, with probability at least $1 - \delta$, has regret bounded as

$$\mathscr{R}_K \lesssim \sqrt{d^3 H^3 V_1^\star K \cdot \log^3(HK/\delta)} + d^{7/2} H^3 \log^{7/2}(HK/\delta)$$

# Main Result

**Theorem.** There exists an algorithm, FORCE, which, with probability at least $1-\delta$, has regret bounded as

$$\mathscr{R}_K \lesssim \sqrt{d^3 H^3 V_1^\star K \cdot \log^3(HK/\delta)} + d^{7/2} H^3 \log^{7/2}(HK/\delta)$$

This corresponds to a PAC guarantee of: $O\left(\frac{d^3 H^3 \cdot V_1^\star}{\epsilon^2}\right)$

# Main Result

**Theorem.** There exists an algorithm, FORCE, which, with probability at least $1-\delta$, has regret bounded as

$$\mathcal{R}_K \lesssim \sqrt{d^3 H^3 V_1^\star K \cdot \log^3(HK/\delta)} + d^{7/2} H^3 \log^{7/2}(HK/\delta)$$

This corresponds to a PAC guarantee of: $O\left(\frac{d^3H^3 \cdot V_1^\star}{\epsilon^2}\right)$

**Existing Work**: $O(\sqrt{d^3H^4K})$ computationally efficient (Jin et al., 2020), $O(\sqrt{d^2H^4K})$ computationally inefficient (Zanette et al., 2020)

# Computationally Efficient Force

FORCE is computationally inefficient, but we obtain a computationally efficient version with the following guarantee

# Computationally Efficient Force

FORCE is computationally inefficient, but we obtain a computationally efficient version with the following guarantee

**Corollary.** There exists a computationally efficient version of FORCE, which, with probability at least $1-\delta$, has regret bounded as

$$\mathscr{R}_K \lesssim \sqrt{d^4 H^3 V_1^\star K \cdot \log^3(HK/\delta)} + d^4 H^3 \log^{7/2}(HK/\delta)$$

# Can Existing Approaches Achieve First-Order Regret?

# Optimistic LSVI (Jin et al., 2020)

In linear MDPs, for any $\pi$, there exists $w_h^\pi$ such that

$$Q_h^\pi(s, a) = \langle \phi(s, a), w_h^\pi \rangle$$

# Optimistic LSVI (Jin et al., 2020)

In linear MDPs, for any $\pi$, there exists $w_h^\pi$ such that

$$Q_h^\pi(s,a) = \langle \phi(s,a), w_h^\pi \rangle$$

Existing approaches have used this fact to fit a $w_h^k$ using least-squares regression:

$$w_h^k \leftarrow \text{argmin}_w \sum_{\tau=1}^{k-1} (r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2 + \|w\|_2^2$$

# Optimistic LSVI (Jin et al., 2020)

In linear MDPs, for any $\pi$, there exists $w_h^\pi$ such that

$$Q_h^\pi(s,a) = \langle \phi(s,a), w_h^\pi \rangle$$

Existing approaches have used this fact to fit a $w_h^k$ using least-squares regression:

$$w_h^k \leftarrow \text{argmin}_w \textstyle\sum_{\tau=1}^{k-1} (r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2 + \|w\|_2^2$$

and then form an optimistic estimate of the $Q$-value function:

$$Q_h^k(s,a) = \langle \phi(s,a), w_h^k \rangle + \beta \|\phi(s,a)\|_{\Lambda_{h,k-1}^{-1}}$$

for $\Lambda_{h,k-1}$ the covariance up to episode $k-1$

# Self-Normalized Bounds

Apply the inequality:

# Self-Normalized Bounds

Apply the inequality:

Assume $\mathbb{E}[\eta_\tau | \mathcal{F}_{\tau-1}] = 0$, $|\eta_\tau| \leq \gamma$, and $\phi_\tau \in \mathbb{R}^d$ is $\mathcal{F}_{\tau-1}$-measurable.
Then with high probability:

$$\| \sum_{\tau=1}^{k} \phi_\tau \eta_\tau \|_{\Lambda_k^{-1}} \lesssim \gamma \sqrt{d + \log 1/\delta}$$

where $\Lambda_k = \sum_{\tau=1}^{k} \phi_\tau \phi_\tau^\top + \lambda I$ are the covariates

# Self-Normalized Bounds

Apply the inequality:

$$\textbf{Statistical Deviation} \leq O(\textbf{ Absolute Magnitude of Noise })$$

# Self-Normalized Bounds

Apply the inequality:

$$\textbf{Statistical Deviation} \leq O(\textbf{ Absolute Magnitude of Noise })$$

This is fundamentally a *Hoeffding-style* bound—it scales with the magnitude of the noise

# LSVI with Bernstein Bonuses

Let $\sigma_{h,\tau}^2$ be an upper bound on the next-state variance, and now let:

$$w_h^k \leftarrow \text{argmin}_w \sum_{\tau=1}^{k-1} (r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2 / \sigma_{h,\tau}^2 + \|w\|_2^2$$

# LSVI with Bernstein Bonuses

Let $\sigma_{h,\tau}^2$ be an upper bound on the next-state variance, and now let:

$$w_h^k \leftarrow \text{argmin}_w \sum_{\tau=1}^{k-1} (r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2 / \sigma_{h,\tau}^2 + \|w\|_2^2$$

Assume $\mathbb{E}[\eta_\tau \,|\, \mathcal{F}_{\tau-1}] = 0$, $\mathbb{V}[\eta_\tau \,|\, \mathcal{F}_{\tau-1}] \leq \sigma^2$, $|\eta_\tau| \leq \gamma$ and $\phi_\tau \in \mathbb{R}^d$ is $\mathcal{F}_{\tau-1}$-measurable. Then with high probability:

$$\|\sum_{\tau=1}^k \phi_\tau \eta_\tau\|_{\Lambda_k^{-1}} \lesssim \sigma\sqrt{d + \log 1/\delta} + \gamma \log 1/\delta$$

where $\Lambda_k = \sum_{\tau=1}^k \phi_\tau \phi_\tau^\top + \lambda I$ are the covariates

# LSVI with Bernstein Bonuses

Let $\sigma^2_{h,\tau}$ be an upper bound on the next-state variance, and now let:

$$w_h^k \leftarrow \text{argmin}_w \sum_{\tau=1}^{k-1} (r_{h,\tau} + V^k_{h+1}(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2 / \sigma^2_{h,\tau} + \|w\|_2^2$$

**Statistical Deviation** $\leq O($ **Standard Deviation of Noise**

**+ Absolute Magnitude of Noise** $)$

# LSVI with Bernstein Bonuses

Let $\sigma_{h,\tau}^2$ be an upper bound on the next-state variance, and now let:

$$w_h^k \leftarrow \text{argmin}_w \sum_{\tau=1}^{k-1} (r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2 / \sigma_{h,\tau}^2 + \|w\|_2^2$$

**Statistical Deviation** $\leq O($ **Standard Deviation of Noise**

**+ Absolute Magnitude of Noise** $)$

In RL, magnitude of "noise" could be large, and regret always $\Omega(\sqrt{K})$

# Improving on Existing Approaches

**Takeaway:** Existing bounds scale in with the *magnitude* of the noise, which is prohibitively large

# Improving on Existing Approaches

**Takeaway:** Existing bounds scale in with the *magnitude* of the noise, which is prohibitively large

Can we do something better?

# Improving on Existing Approaches

**Takeaway:** Existing bounds scale in with the *magnitude* of the noise, which is prohibitively large

Can we do something better?

***Catoni Estimation***

# Catoni Estimation

# Catoni Mean Estimation

**Proposition** (Catoni, 2012). Let $X_1, \ldots, X_T$ be mean $\mu$ iid random variables with variance $\sigma^2$. Then the Catoni estimator will produce an estimate $\hat{\mu}_{\text{cat}}$ such that

$$|\hat{\mu}_{\text{cat}} - \mu| \lesssim \sqrt{\frac{\sigma^2 \log 1/\delta}{T}}$$

# Catoni Mean Estimation

**Proposition** (Catoni, 2012). Let $X_1, \ldots, X_T$ be mean $\mu$ iid random variables with variance $\sigma^2$. Then the Catoni estimator will produce an estimate $\hat{\mu}_{\text{cat}}$ such that

$$|\hat{\mu}_{\text{cat}} - \mu| \lesssim \sqrt{\frac{\sigma^2 \log 1/\delta}{T}}$$

In contrast, Bernstein assumes $|X_i| \leq \gamma$ and has guarantee

$$|\hat{\mu} - \mu| \lesssim \sqrt{\frac{\sigma^2 \log 1/\delta}{T}} + \frac{\gamma \cdot \log 1/\delta}{T}$$

# Catoni Estimation in RL

Note that

$$w_h^k \leftarrow \text{argmin}_w \sum_{\tau=1}^{k-1} (r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2/\sigma_{h,\tau}^2 + \lambda\|w\|_2^2$$

simply equals

$$w_h^k = \sum_{\tau=1}^{k-1} \Lambda_{h,k-1}^{-1} \phi_{h,\tau}(r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}))/\sigma_{h,\tau}^2$$

# Catoni Estimation in RL

Note that

$$w_h^k \leftarrow \text{argmin}_w \sum_{\tau=1}^{k-1} (r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2/\sigma_{h,\tau}^2 + \lambda \|w\|_2^2$$

simply equals

$$w_h^k = \sum_{\tau=1}^{k-1} \Lambda_{h,k-1}^{-1} \phi_{h,\tau}(r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}))/\sigma_{h,\tau}^2$$

So we could replace the least-squares estimate with a Catoni estimate.
Several issues:

# Catoni Estimation in RL

Note that

$$w_h^k \leftarrow \text{argmin}_w \sum_{\tau=1}^{k-1} (r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2/\sigma_{h,\tau}^2 + \lambda \|w\|_2^2$$

simply equals

$$w_h^k = \sum_{\tau=1}^{k-1} \Lambda_{h,k-1}^{-1} \phi_{h,\tau}(r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}))/\sigma_{h,\tau}^2$$

So we could replace the least-squares estimate with a Catoni estimate.
Several issues:

- Our data is vector-valued not scalar-valued

# Catoni Estimation in RL

Note that

$$w_h^k \leftarrow \text{argmin}_w \sum_{\tau=1}^{k-1} (r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2/\sigma_{h,\tau}^2 + \lambda \|w\|_2^2$$

simply equals

$$w_h^k = \sum_{\tau=1}^{k-1} \Lambda_{h,k-1}^{-1} \phi_{h,\tau}(r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}))/\sigma_{h,\tau}^2$$

So we could replace the least-squares estimate with a Catoni estimate.
Several issues:

- Our data is vector-valued not scalar-valued
- Our data is correlated—Catoni assumes independent data

# Catoni Estimation in RL

Note that

$$w_h^k \leftarrow \text{argmin}_w \sum_{\tau=1}^{k-1} (r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}) - w^\top \phi_{h,\tau})^2/\sigma_{h,\tau}^2 + \lambda \|w\|_2^2$$

simply equals

$$w_h^k = \sum_{\tau=1}^{k-1} \Lambda_{h,k-1}^{-1} \phi_{h,\tau}(r_{h,\tau} + V_{h+1}^k(s_{h+1,\tau}))/\sigma_{h,\tau}^2$$

So we could replace the least-squares estimate with a Catoni estimate.
Several issues:

- Our data is vector-valued not scalar-valued
- Our data is correlated—Catoni assumes independent data
- In particular, $V_{h+1}^k$ and $\Lambda_{h,k-1}$ are random and correlated with all the data

# Solution: Uniform Catoni Estimation

We prove a novel perturbation bound on the Catoni estimator that allows us to derive a **uniform convergence-style** bound on Catoni estimation

# Solution: Uniform Catoni Estimation

We prove a novel perturbation bound on the Catoni estimator that allows us to derive a **uniform convergence-style** bound on Catoni estimation

Applying this perturbation bound, we:

# Solution: Uniform Catoni Estimation

We prove a novel perturbation bound on the Catoni estimator that allows us to derive a **uniform convergence-style** bound on Catoni estimation

Applying this perturbation bound, we:

- Cover the space of directions in $\mathbb{R}^d$ to handle the vector nature of the data

# Solution: Uniform Catoni Estimation

We prove a novel perturbation bound on the Catoni estimator that allows us to derive a **uniform convergence-style** bound on Catoni estimation

Applying this perturbation bound, we:

- Cover the space of directions in $\mathbb{R}^d$ to handle the vector nature of the data
- Cover the space of functions $V_{h+1}^k$ and covariance matrices $\Lambda_{h,k-1}$ to eliminate correlations

# Solution: Uniform Catoni Estimation

We prove a novel perturbation bound on the Catoni estimator that allows us to derive a **uniform convergence-style** bound on Catoni estimation

Applying this perturbation bound, we:

- Cover the space of directions in $\mathbb{R}^d$ to handle the vector nature of the data
- Cover the space of functions $V_{h+1}^k$ and covariance matrices $\Lambda_{h,k-1}$ to eliminate correlations

Combining these innovations with a martingale version of the Catoni estimator due to Wei et al. (2020) yields the needed result

# Uniform Catoni Estimation

Consider setting where $\phi_t$ are some $\mathscr{F}_{t-1}$ vectors and

$$y_t = \langle \phi_t, \theta \rangle + \eta_t$$

for some $\theta$, $\mathbb{E}[y_t^2 \mid \mathscr{F}_{t-1}] \leq \sigma_t^2$, and $|\eta_t| \leq \gamma$

# Uniform Catoni Estimation

Consider setting where $\phi_t$ are some $\mathscr{F}_{t-1}$ vectors and

$$y_t = \langle \phi_t, \theta \rangle + \eta_t$$

for some $\theta$, $\mathbb{E}[y_t^2 \mid \mathscr{F}_{t-1}] \leq \sigma_t^2$, and $|\eta_t| \leq \gamma$

**Proposition.** Consider running the Catoni estimator on the data $X_t = Tv^\top \Lambda_T^{-1} \phi_t y_t / \sigma_t^2$. Then for all $v \in \mathcal{S}^{d-1}$ simultaneously, with probability at least $1 - \delta$,

$$|\,\mathsf{cat}[v] - v^\top \theta\,| \lesssim \|v\|_{\Lambda_T^{-1}} \cdot \sqrt{d + \log \gamma / \delta}$$

for $\Lambda_T = \sum_{\tau=1}^T \sigma_\tau^{-2} \phi_\tau \phi_\tau^\top + \lambda I$.

# Algorithm: FORCE

**Key Idea**: replace weighted least-squares estimator with Catoni estimator when forming estimate of optimistic $Q$-value function

# Algorithm: FORCE

**Key Idea**: replace weighted least-squares estimator with Catoni estimator when forming estimate of optimistic $Q$-value function

This removes the dependence on the magnitude term

# Algorithm: FORCE

**Key Idea**: replace weighted least-squares estimator with Catoni estimator when forming estimate of optimistic $Q$-value function

This removes the dependence on the magnitude term

Some calculation shows that regret then scales as

$$\mathrm{poly}(d, H) \cdot \sqrt{\sum_{\tau=1}^{K} \sum_{h=1}^{H} \sigma_{h,\tau}^2} + \mathrm{poly}(d, H)$$

For $\sigma_{h,\tau}^2$ upper bounds on the expected next-state squared value function

# Algorithm: FORCE

**Key Idea**: replace weighted least-squares estimator with Catoni estimator when forming estimate of optimistic $Q$-value function

This removes the dependence on the magnitude term

Some calculation shows that regret then scales as

$$\mathrm{poly}(d, H) \cdot \sqrt{\sum_{\tau=1}^{K} \sum_{h=1}^{H} \sigma_{h,\tau}^2} + \mathrm{poly}(d, H)$$

For $\sigma_{h,\tau}^2$ upper bounds on the expected next-state squared value function

This can be bounded as: $\mathrm{poly}(d, H) \cdot \sqrt{V_1^\star K} + \mathrm{poly}(d, H)$

# Thanks!