Statistical and Algorithmic Foundations of Reinforcement Learning

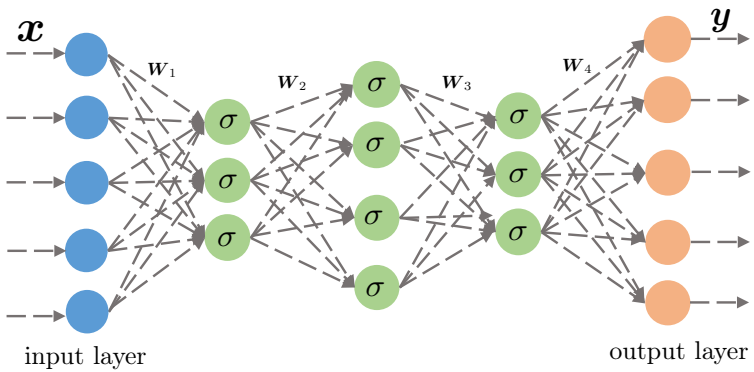
Yuejie Chi
Yale University

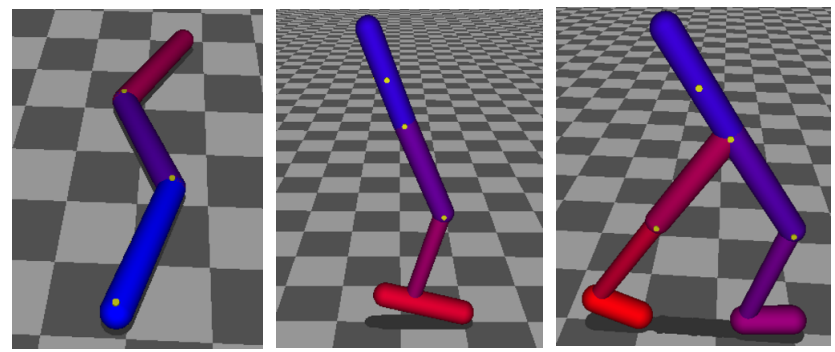
DeepLearn 2026
Tutorial Part 3

Policy optimization

$$\text{maximize}_\theta \quad \text{value}(\text{policy}(\theta))$$

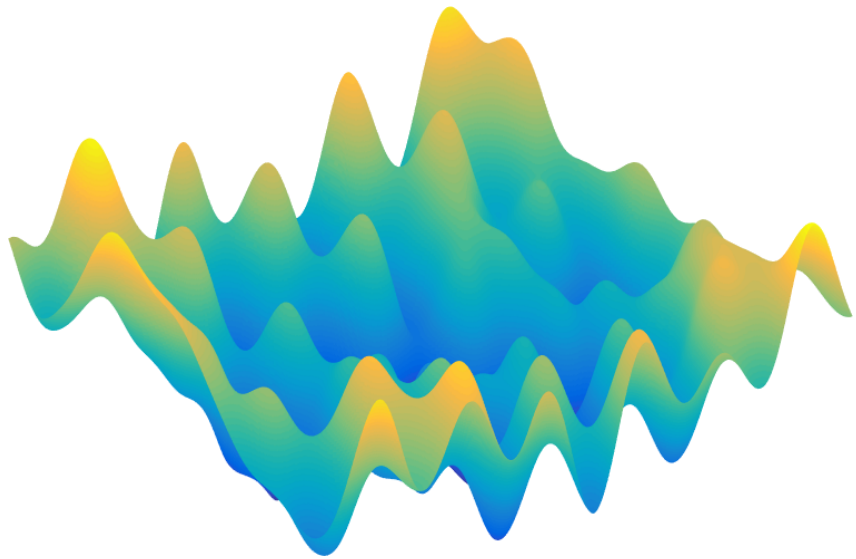
- directly optimize the policy, which is the quantity of interest;
- allow flexible differentiable parameterizations of the policy;
- work with both continuous and discrete problems.





Theoretical challenges: non-concavity

Little understanding on the global convergence of policy gradient methods until very recently, e.g. (Fazel et al., 2018; Bhandari and Russo, 2019; Agarwal et al., 2019; Mei et al. 2020), and many more.



Our goal:

- understand finite-time convergence rates of popular heuristics;
- design fast-convergent algorithms in multi-agent and federated settings.

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_\pi \quad V^\pi(\rho) := \mathbb{E}_{s\sim\rho}\left[V^\pi(s)\right]$$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_\pi \quad V^\pi(\rho) := \mathbb{E}_{s\sim\rho}\left[V^\pi(s)\right]$$

Parameterization:

$$\pi := \pi_\theta$$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_{\pi} \quad V^{\pi}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi}(s)\right]$$

Parameterization:

$$\pi := \pi_{\theta}$$

$$\text{maximize}_{\theta} \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_\pi \quad V^\pi(\rho) := \mathbb{E}_{s\sim\rho}\left[V^\pi(s)\right]$$

Parameterization:

$$\pi := \pi_\theta$$

$$\text{maximize}_\theta \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$$

Policy gradient method (Sutton et al., 2000)

For $t = 0, 1, \cdots$

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

where η *is the learning rate.*

The policy gradient theorem

Theorem (Policy gradient theorem, Sutton et al., 2000)

The policy gradient can be evaluated via

$$\begin{aligned}\nabla_\theta V^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}\Big[Q^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)\Big] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}\Big[A^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)\Big],\end{aligned}$$

where

- $d_\rho^{\pi_\theta}$ *is the discounted state visitation distribution,*
- $\psi_\theta(s,a) := \nabla\log\pi_\theta(a|s)$ *is the score function, and*
- $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ *is the advantage function.*

Provides a general scheme for policy gradient evaluation (e.g., REINFORCE).

Examples of policy parameterization

Discrete action space: softmax parameterization with function approximation

$$\pi_\theta(a|s) \propto \exp(\phi(s,a)^\top \theta)$$

- $\phi(s,a)$ is the feature vector of each state-action pair;
- the score function $\nabla \log \pi_\theta(a|s) = \phi(s,a) - \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\phi(s,\cdot)]$.

Examples of policy parameterization

Discrete action space: softmax parameterization with function approximation

$$\pi_\theta(a|s) \propto \exp(\phi(s,a)^\top \theta)$$

- $\phi(s,a)$ is the feature vector of each state-action pair;
- the score function $\nabla \log \pi_\theta(a|s) = \phi(s,a) - \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\phi(s,\cdot)]$.

Continuous action space: Gaussian policy

$$a \sim \mathcal{N}(\mu(s), \sigma^2), \quad \mu(s) = \phi(s)^\top \theta$$

- $\phi(s)$ is the feature of each state;
- σ^2 is the variance (kept constant for simplicity);
- the score function $\nabla \log \pi_\theta(a|s) = \frac{(a-\mu(s))\phi(s)}{\sigma^2}$.

Softmax policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_\pi \quad V^\pi(\rho) := \mathbb{E}_{s\sim\rho}\left[V^\pi(s)\right]$$

softmax parameterization:

$$\pi_\theta(a|s) \propto \exp(\theta(s,a))$$

$$\text{maximize}_\theta \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$$

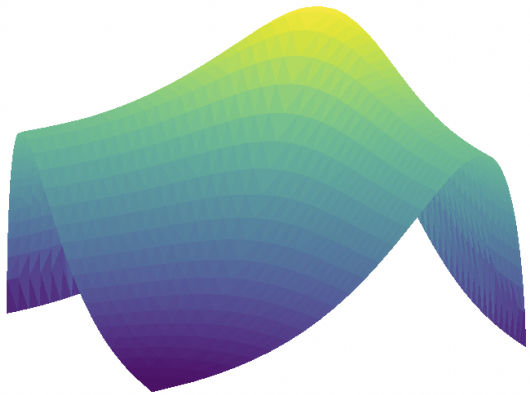
Policy gradient method (Sutton et al., 2000)

For $t = 0, 1, \cdots$

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

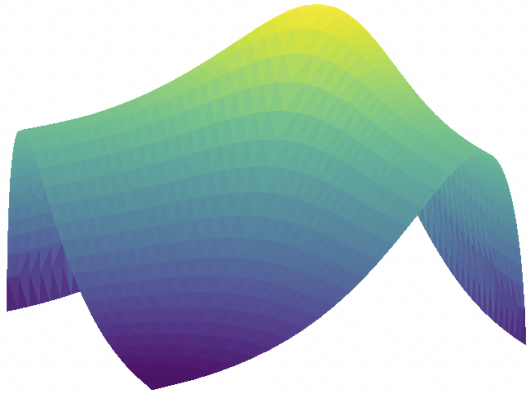
where η *is the learning rate.*

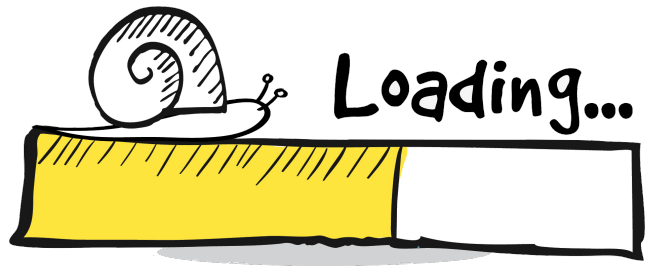
Global convergence of the PG method?



- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.

Global convergence of the PG method?

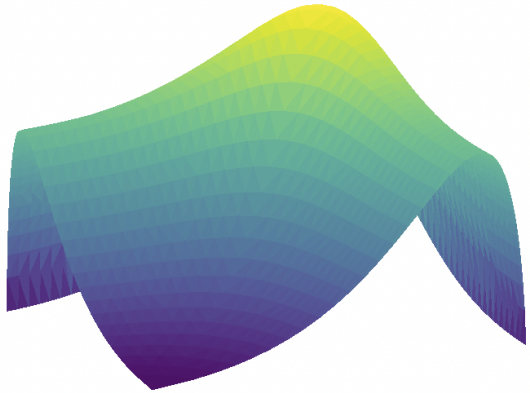


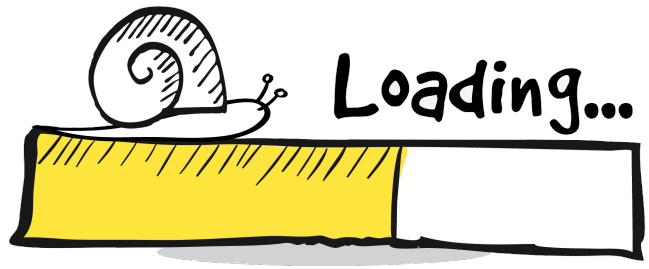


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

$$O\left(\frac{1}{\epsilon}\right) \text{ iterations}$$

Global convergence of the PG method?

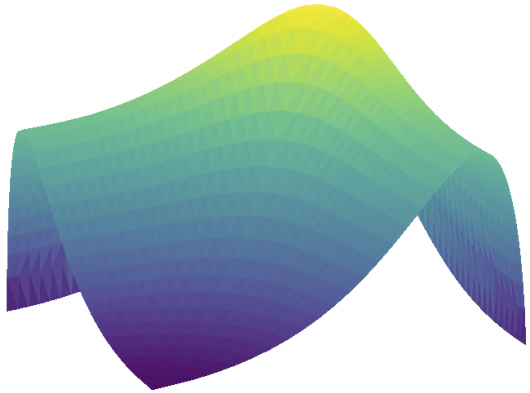


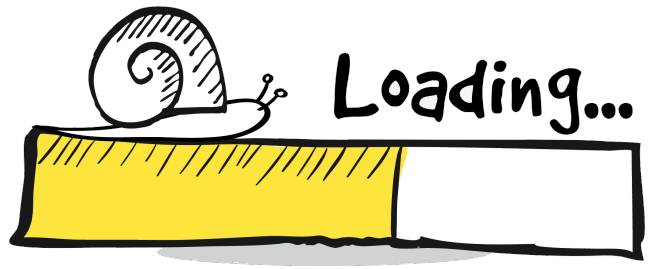


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

$$c(|\mathcal{S}|, |\mathcal{A}|, \tfrac{1}{1-\gamma}, \cdots)\, O\left(\tfrac{1}{\epsilon}\right) \text{ iterations}$$

Global convergence of the PG method?





- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

$$c(|\mathcal{S}|, |\mathcal{A}|, \tfrac{1}{1-\gamma}, \cdots)\, O\left(\tfrac{1}{\epsilon}\right) \text{ iterations}$$

Is the rate of PG good, bad or ugly?

A negative message

Theorem (Li, Wei, Chi, Chen, MP 2023)

There exists an MDP s.t. it takes softmax PG at least

$$\frac{1}{\eta}\,|\mathcal{S}|^{2^{\Theta(\frac{1}{1-\gamma})}} \quad \textit{iterations}$$

to achieve $\|V^{(t)} - V^\star\|_\infty \leq 0.15$.

A negative message

Theorem (Li, Wei, Chi, Chen, MP 2023)

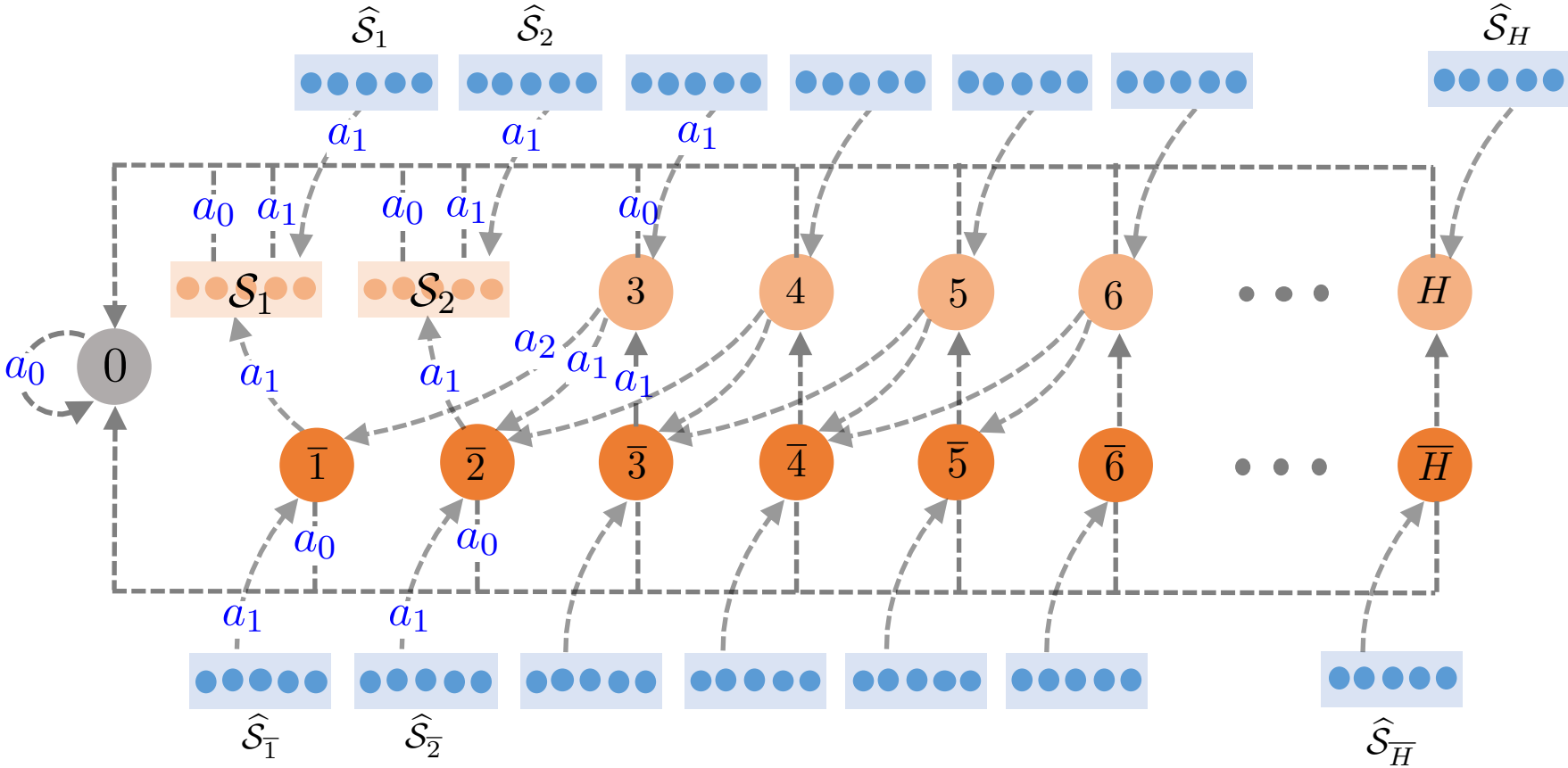
There exists an MDP s.t. it takes softmax PG at least

$$\frac{1}{\eta}\,|\mathcal{S}|^{2^{\Theta(\frac{1}{1-\gamma})}} \quad \textit{iterations}$$

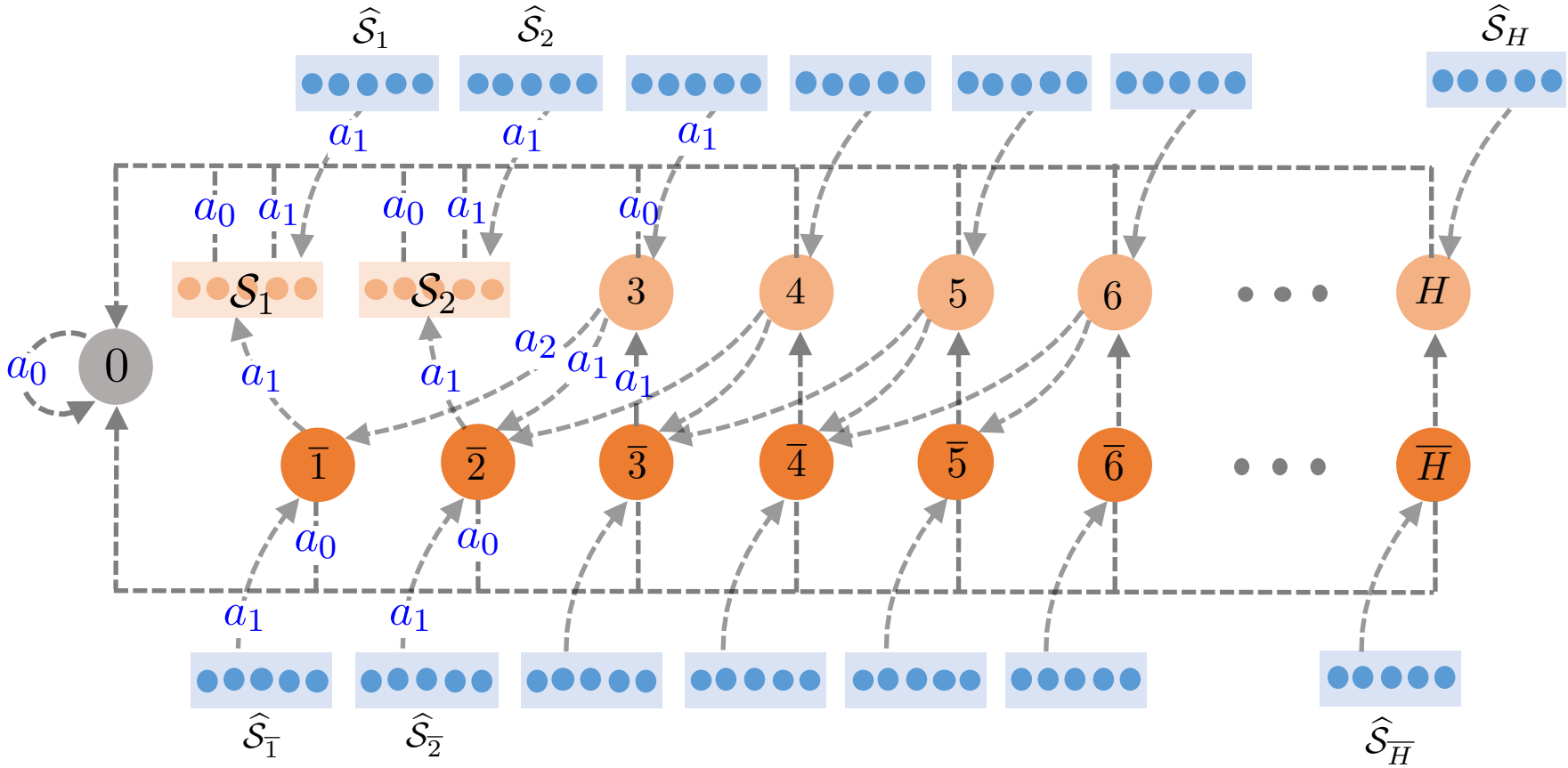
to achieve $\|V^{(t)} - V^\star\|_\infty \leq 0.15$.

- Softmax PG can take (super)-exponential time to converge (in problems w/ large state space $\&$ long effective horizon)!
- Also hold for average sub-opt gap $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\left[V^{(t)}(s) - V^\star(s)\right]$.
- In the constructed MDP, the convergence time for a sequence of states grows geometrically.

MDP construction for our lower bound

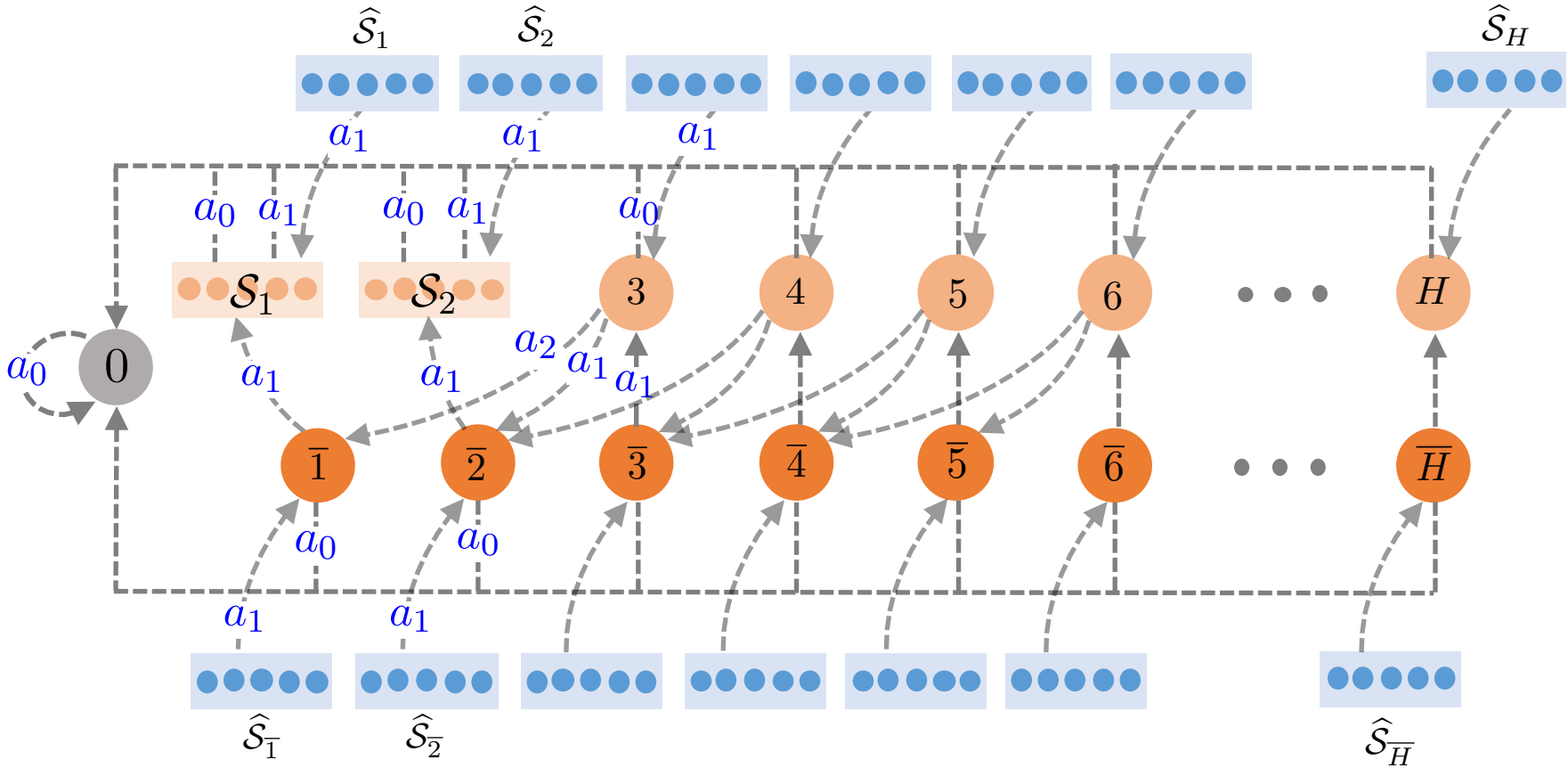


MDP construction for our lower bound



Key ingredients: for $3 \leq s \leq H \asymp \frac{1}{1-\gamma}$,

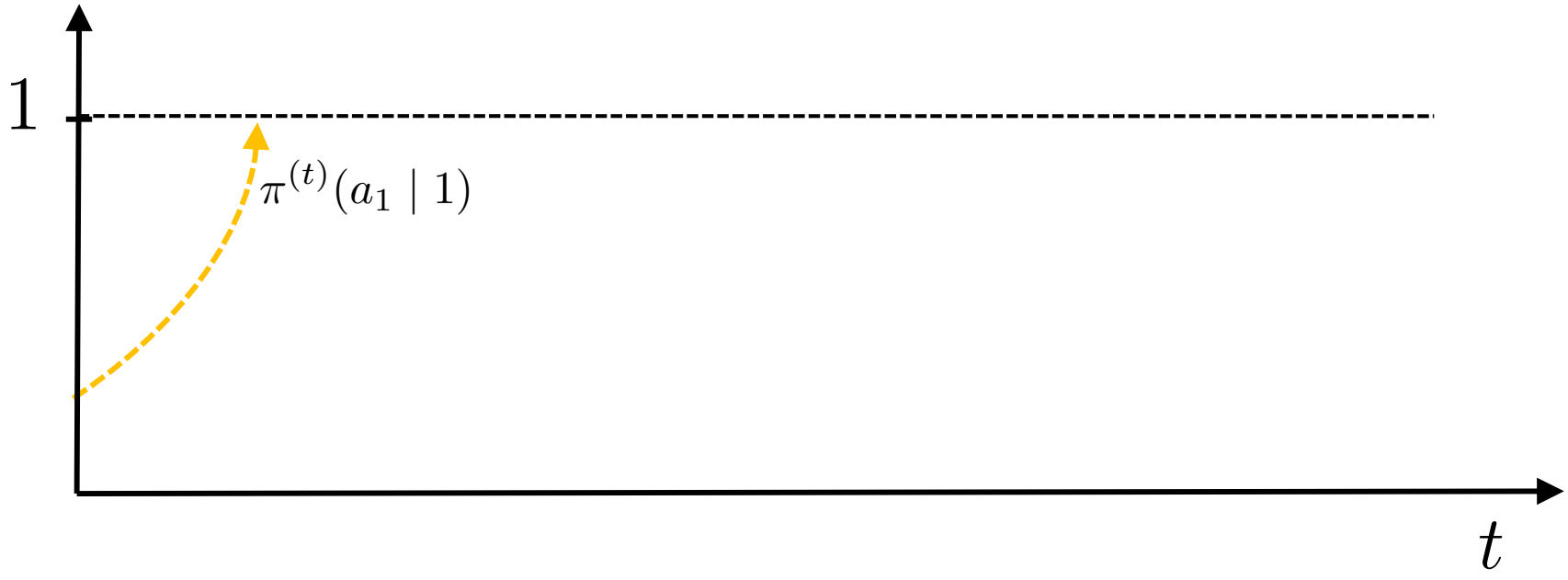
MDP construction for our lower bound



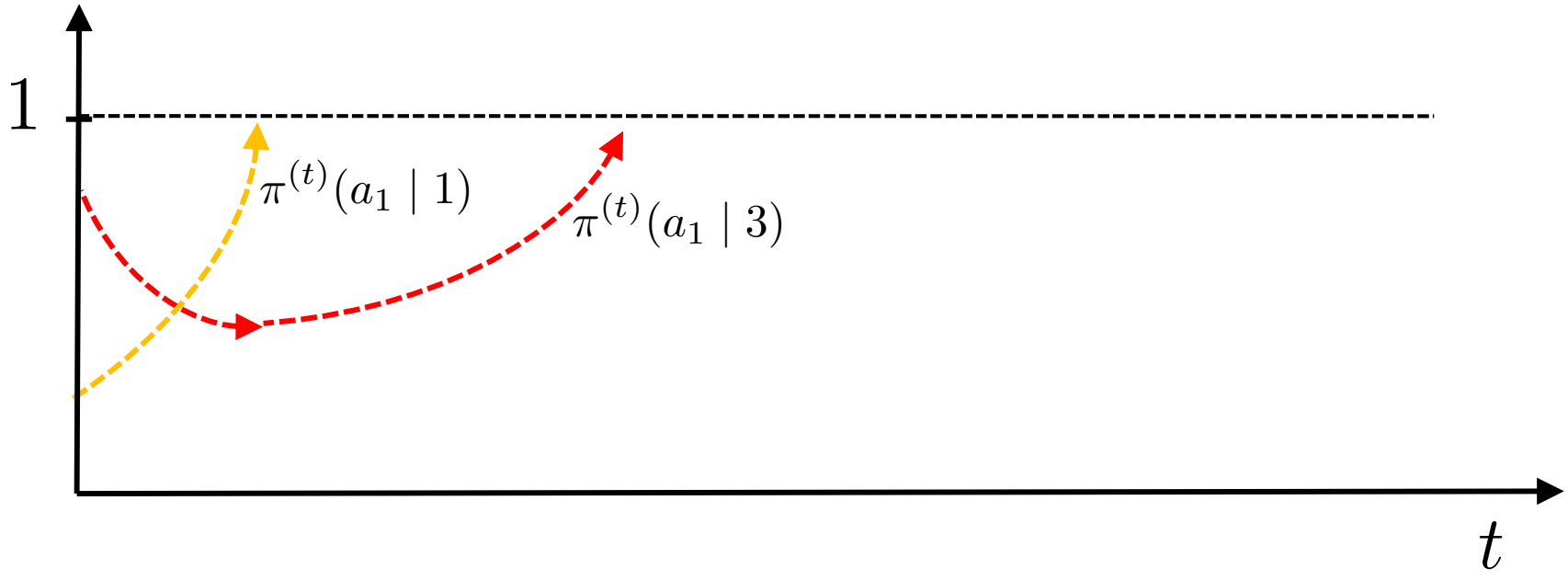
Key ingredients: for $3 \le s \le H \asymp \frac{1}{1-\gamma}$,

- $\pi^{(t)}(a_{\mathsf{opt}} \,|\, s)$ keeps decreasing until $\pi^{(t)}(a_{\mathsf{opt}} \,|\, s-2) \approx 1$

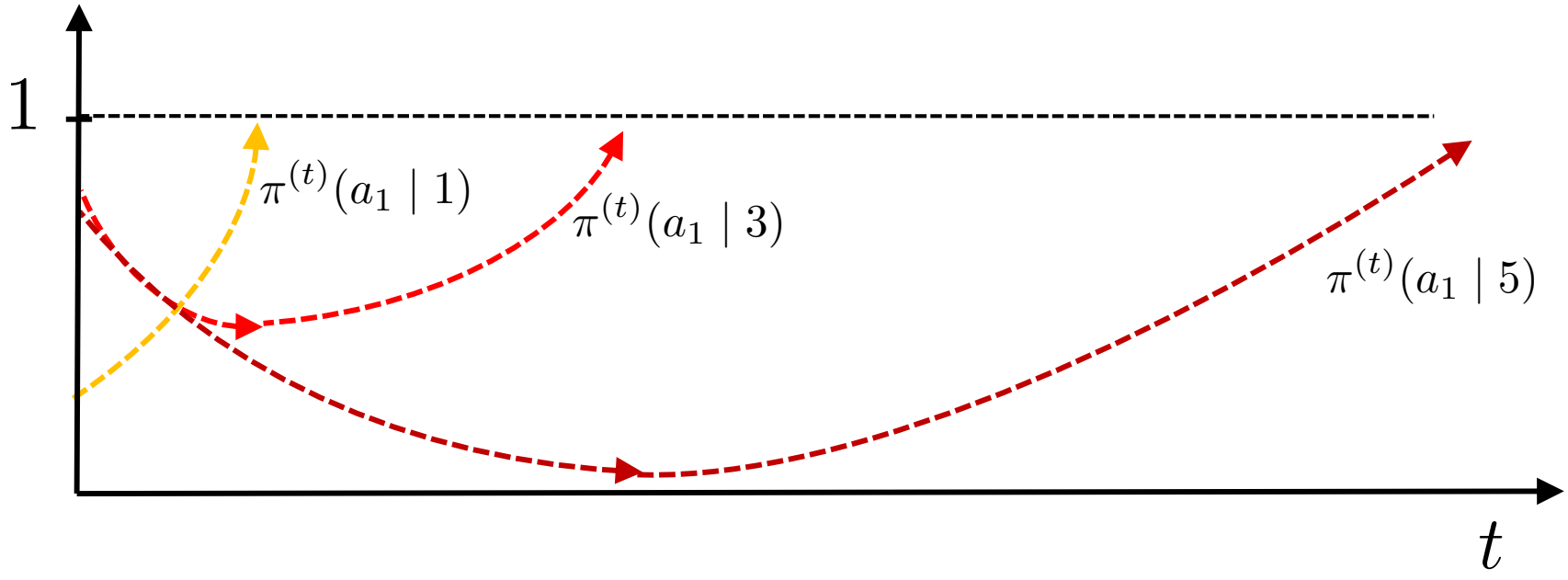
What is happening in our constructed MDP?



What is happening in our constructed MDP?

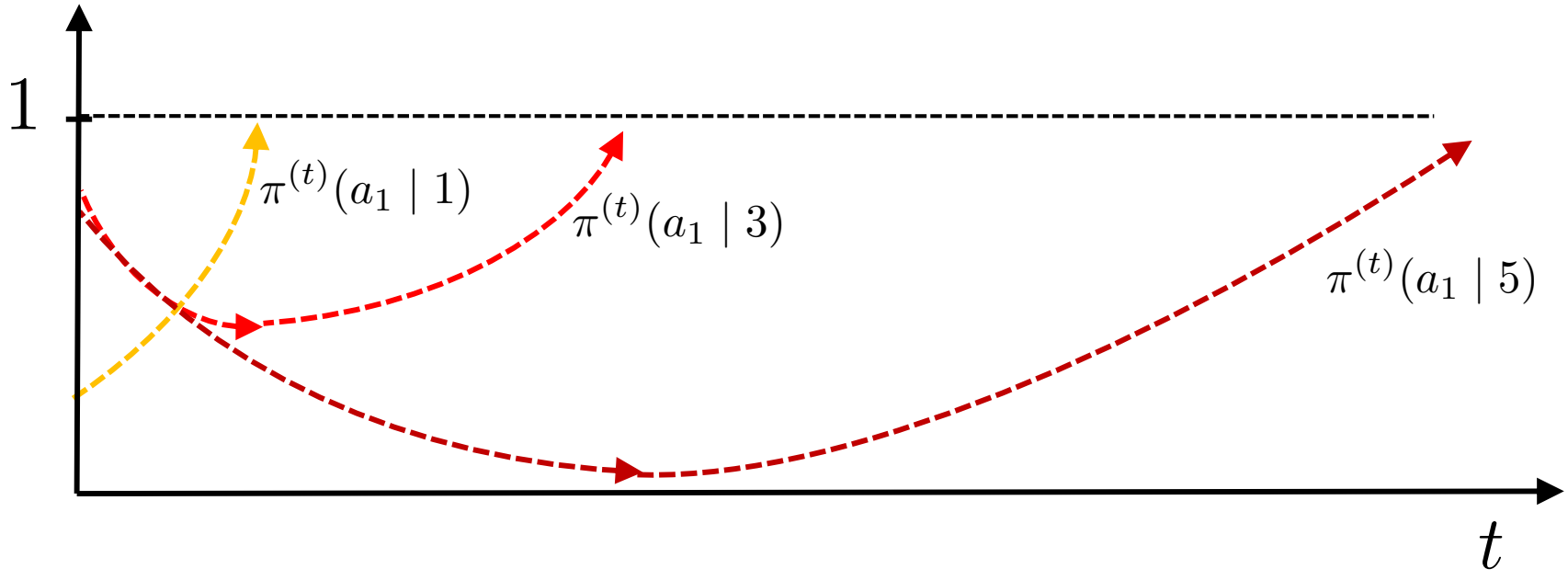


What is happening in our constructed MDP?



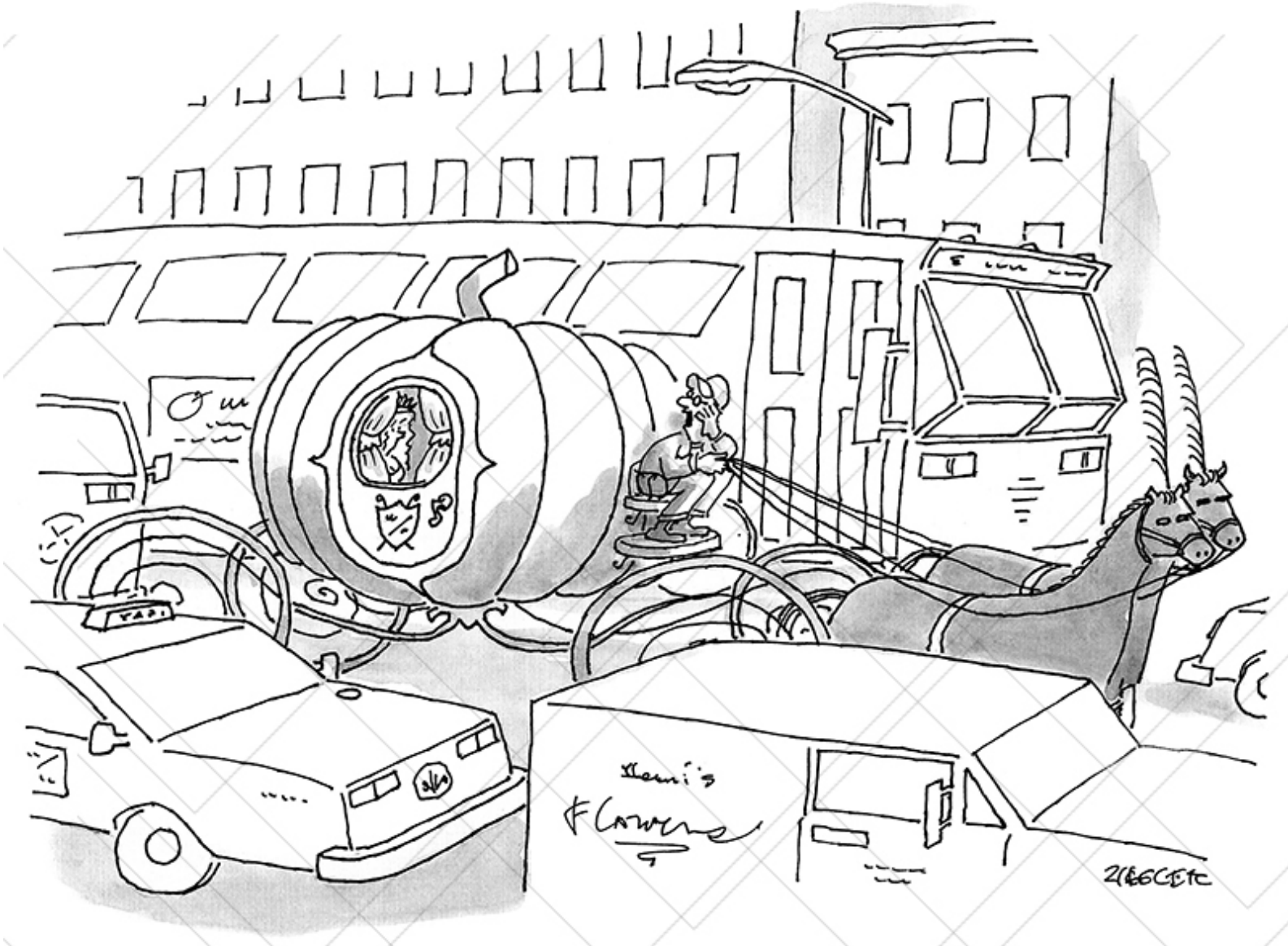
Convergence time for state s grows geometrically as s increases

What is happening in our constructed MDP?



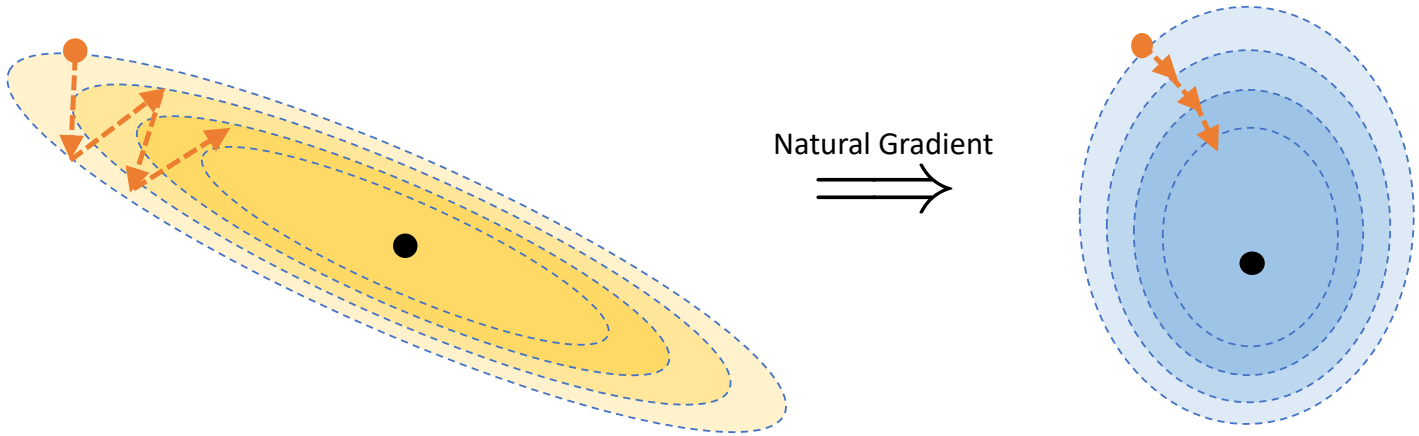
Convergence time for state s grows geometrically as s increases

$$\text{convergence-time}(s) \gtrsim \left(\text{convergence-time}(s-2)\right)^{1.5}$$



"Seriously, lady, at this hour you'd make a lot better time taking the subway."

Booster #1: natural policy gradient



Natural policy gradient (NPG) method (Kakade, 2002)

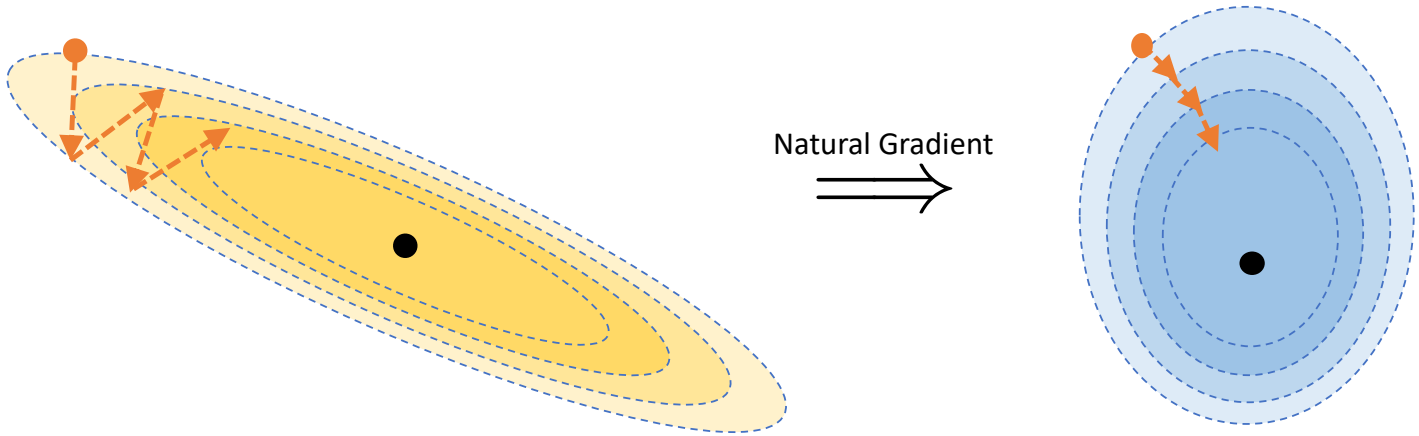
For $t = 0, 1, \cdots$

$$\theta^{(t+1)} = \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

where η *is the learning rate and* $\mathcal{F}_\rho^\theta$ *is the Fisher information matrix:*

$$\mathcal{F}_\rho^\theta := \mathbb{E}\left[\big(\nabla_\theta \log \pi_\theta(a|s)\big)\big(\nabla_\theta \log \pi_\theta(a|s)\big)^\top\right].$$

Booster #1: natural policy gradient



Natural policy gradient (NPG) method (Kakade, 2002)

For $t = 0, 1, \cdots$

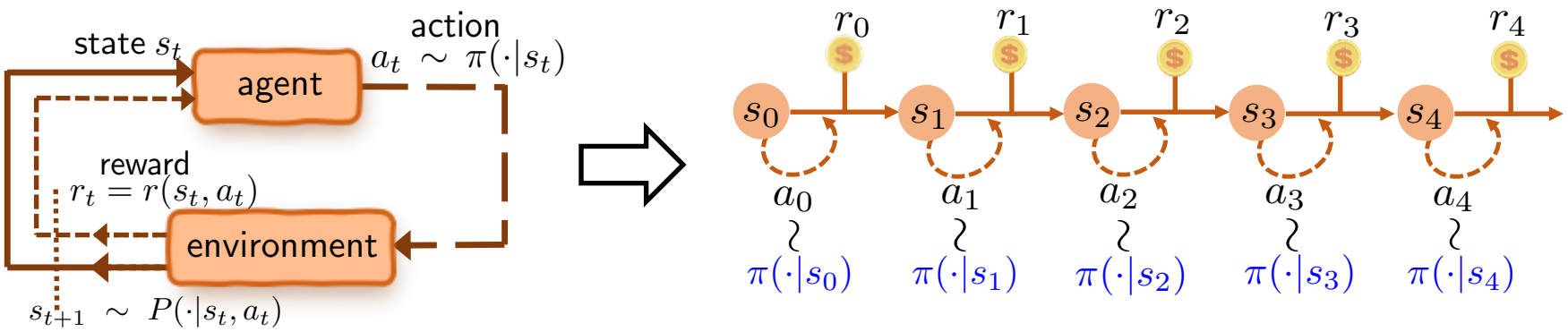
$$\theta^{(t+1)} = \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

where η *is the learning rate and* $\mathcal{F}_\rho^\theta$ *is the Fisher information matrix:*

$$\mathcal{F}_\rho^\theta := \mathbb{E}\left[\big(\nabla_\theta \log \pi_\theta(a|s)\big)\big(\nabla_\theta \log \pi_\theta(a|s)\big)^\top\right].$$

In fact, popular heuristic TRPO (Schulman et al., 2015) = NPG + line search.

Booster #2: entropy regularization

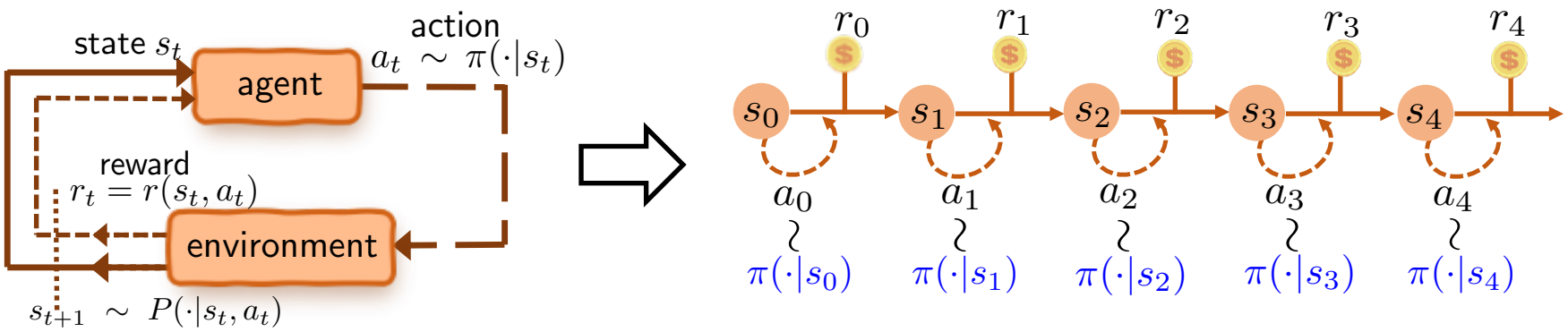


To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function (Williams and Peng, 1991):

$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \big(r_t + \tau \mathcal{H}(\pi(\cdot|s_t)) \,\big|\, s_0 = s\right]$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

Booster #2: entropy regularization



To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function (Williams and Peng, 1991):

$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \big(r_t + \tau \mathcal{H}(\pi(\cdot|s_t)) \,\big|\, s_0 = s\right]$$

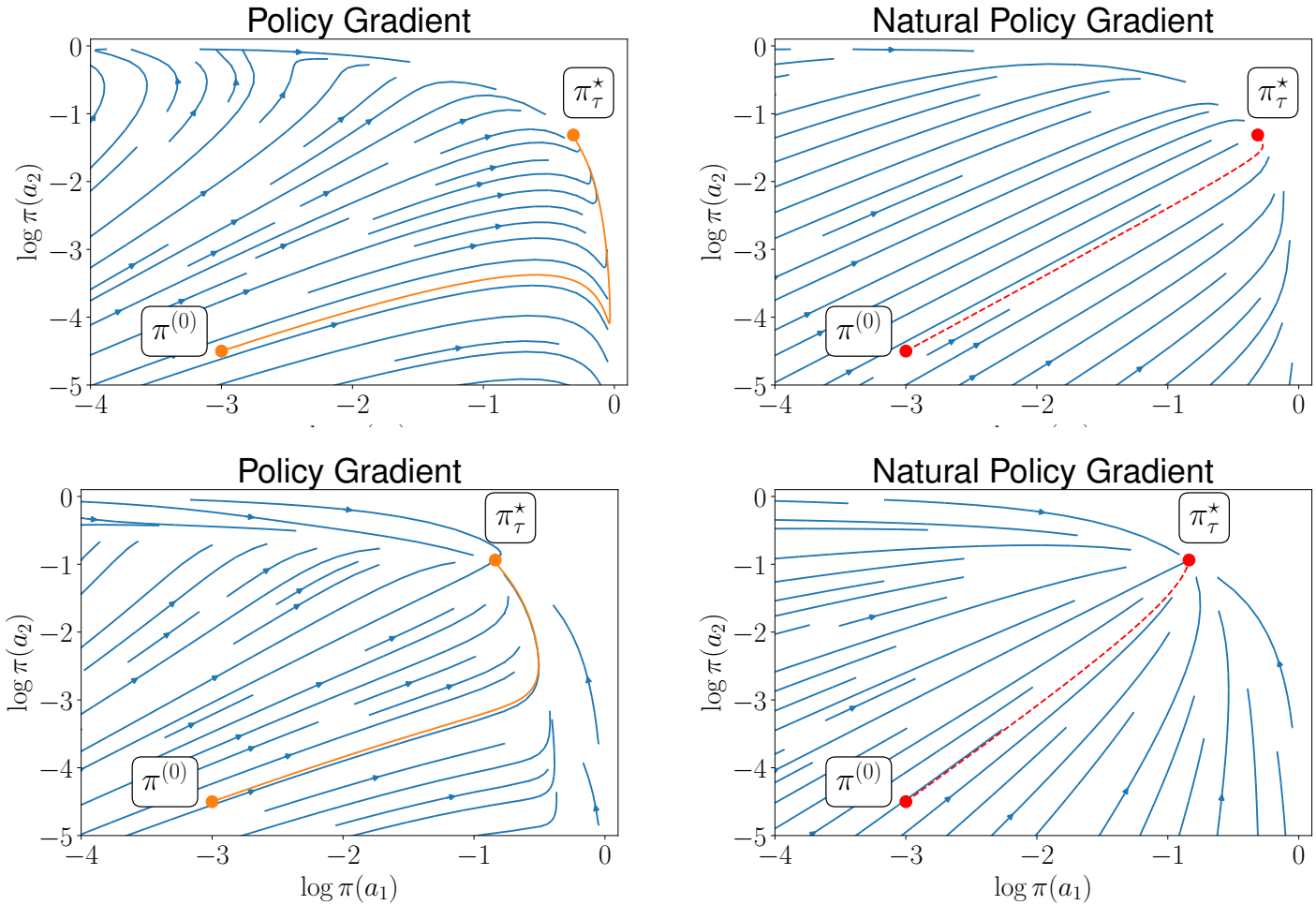
where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

$$\text{maximize}_\theta \quad V_\tau^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V_\tau^{\pi_\theta}(s)\right]$$

Entropy-regularized natural gradient helps!

Toy example: a bandit with 3 arms of rewards 1, 0.9 and 0.1.

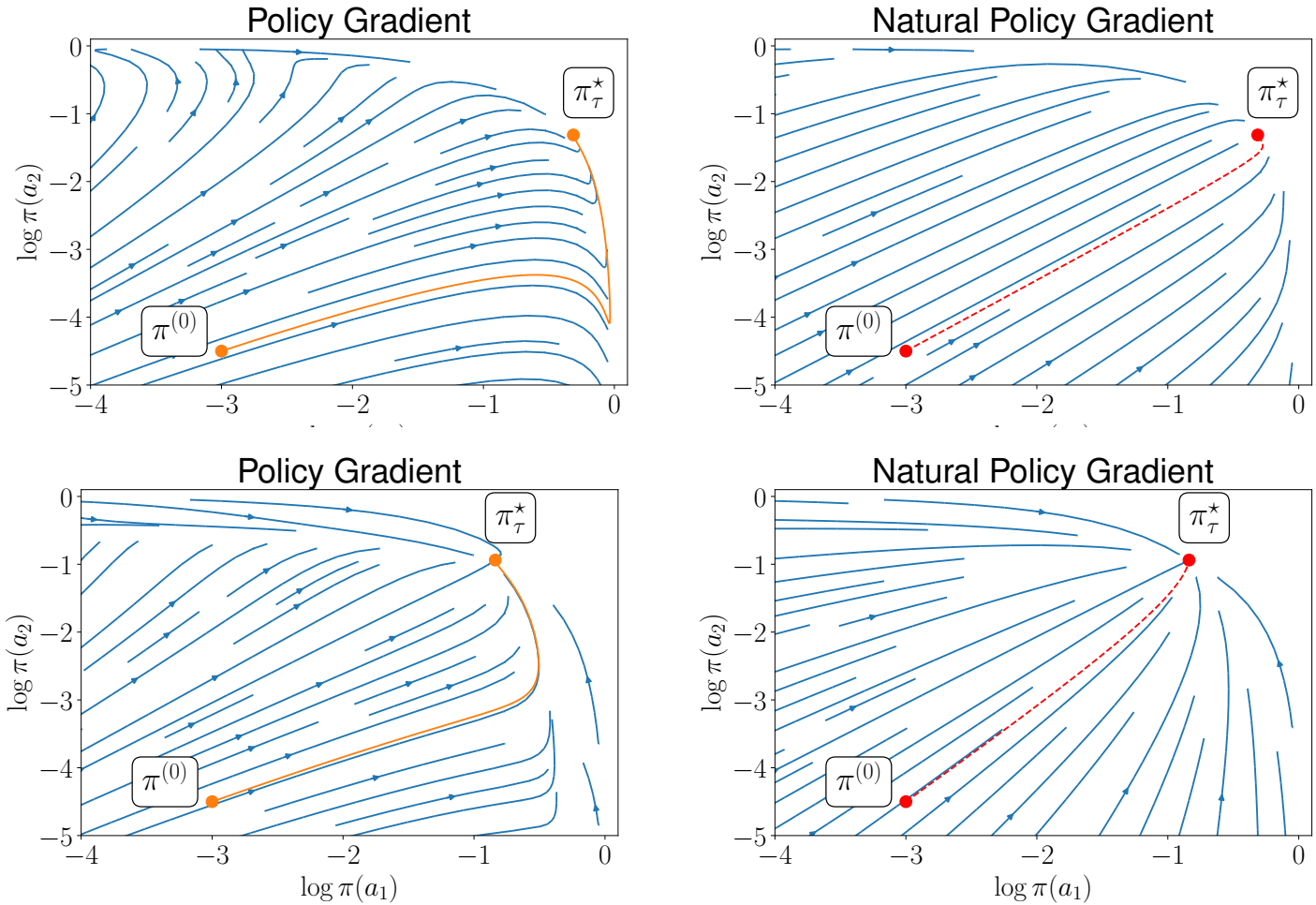
increase regularization



Entropy-regularized natural gradient helps!

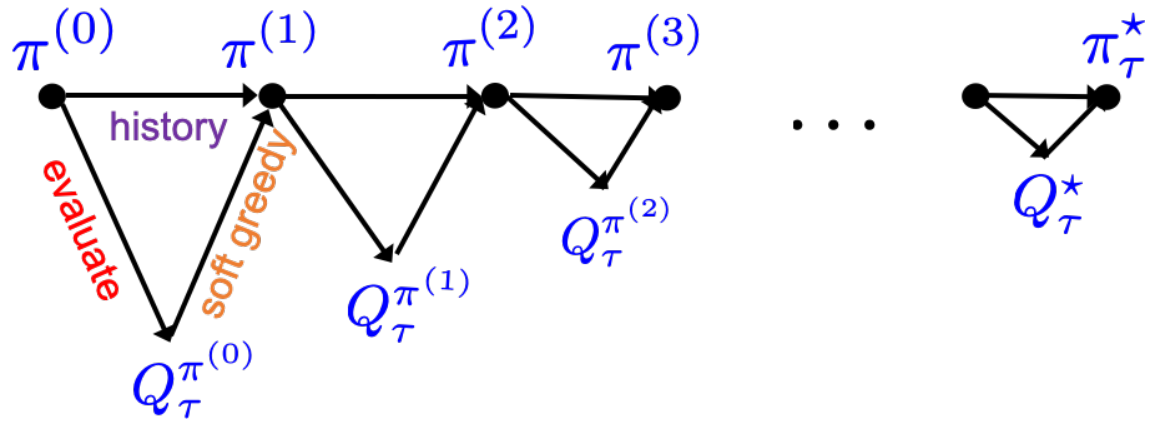
Toy example: a bandit with 3 arms of rewards 1, 0.9 and 0.1.

increase regularization



Can we justify the efficacy of entropy-regularized NPG?

Entropy-regularized NPG in the tabular setting



Entropy-regularized NPG (Tabular setting)

For $t = 0, 1, \cdots$, *the policy is updated via*

$$\pi^{(t+1)}(\cdot|s) \propto \underbrace{\pi^{(t)}(\cdot|s)}_{\text{current policy}}{}^{1-\frac{\eta\tau}{1-\gamma}} \underbrace{\exp(Q_\tau^{(t)}(s,\cdot)/\tau)}_{\text{soft greedy}}{}^{\frac{\eta\tau}{1-\gamma}}$$

where $Q_\tau^{(t)} := Q_\tau^{\pi^{(t)}}$ *is the soft Q-function of* $\pi^{(t)}$, *and* $0 < \eta \leq \frac{1-\gamma}{\tau}$.

- invariant with the choice of ρ
- Reduces to soft policy iteration (SPI) when $\eta = \frac{1-\gamma}{\tau}$.

Linear convergence with exact gradient

Exact oracle: perfect evaluation of $Q_\tau^{\pi^{(t)}}$ given $\pi^{(t)}$;

Theorem (Cen, Cheng, Chen, Wei, Chi, OR 2022)

For any learning rate $0 < \eta \leq (1-\gamma)/\tau$, the entropy-regularized NPG updates satisfy

- **Linear convergence of soft Q-functions:**

$$\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq C_1 \gamma \, (1-\eta\tau)^t$$

for all $t \geq 0$, where $Q_\tau^\star$ is the optimal soft Q-function, and

$$C_1 = \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + 2\tau \left(1 - \frac{\eta\tau}{1-\gamma}\right) \|\log \pi_\tau^\star - \log \pi^{(0)}\|_\infty.$$

Linear convergence with exact gradient

Exact oracle: perfect evaluation of $Q_\tau^{\pi^{(t)}}$ given $\pi^{(t)}$;

Theorem (Cen, Cheng, Chen, Wei, Chi, OR 2022)

For any learning rate $0 < \eta \leq (1-\gamma)/\tau$, the entropy-regularized NPG updates satisfy

- **Linear convergence of soft Q-functions:**

$$\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq C_1 \gamma \, (1-\eta\tau)^t$$

for all $t \geq 0$, where $Q_\tau^\star$ is the optimal soft Q-function, and

$$C_1 = \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + 2\tau \left(1 - \frac{\eta\tau}{1-\gamma}\right) \|\log \pi_\tau^\star - \log \pi^{(0)}\|_\infty.$$

- Robustness: continues to hold under inexact $\left\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\right\|_\infty$.

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq \epsilon$, the iteration complexity is at most

- **General learning rates ($0 < \eta < \frac{1-\gamma}{\tau}$):**

$$\frac{1}{\eta\tau} \log\left(\frac{C_1\gamma}{\epsilon}\right)$$

- **Soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$):**

$$\frac{1}{1-\gamma} \log\left(\frac{\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty \gamma}{\epsilon}\right)$$

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq \epsilon$, the iteration complexity is at most

- **General learning rates ($0 < \eta < \frac{1-\gamma}{\tau}$):**

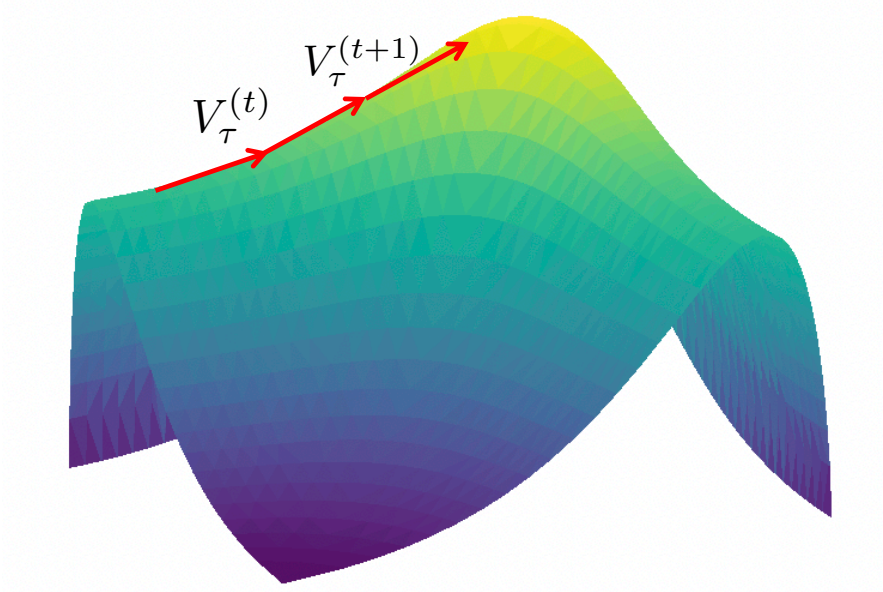
$$\frac{1}{\eta\tau} \log\left(\frac{C_1\gamma}{\epsilon}\right)$$

- **Soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$):**

$$\frac{1}{1-\gamma} \log\left(\frac{\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty \gamma}{\epsilon}\right)$$

Global linear convergence of entropy-regularized NPG
at a rate independent of $|\mathcal{S}|$, $|\mathcal{A}|$!

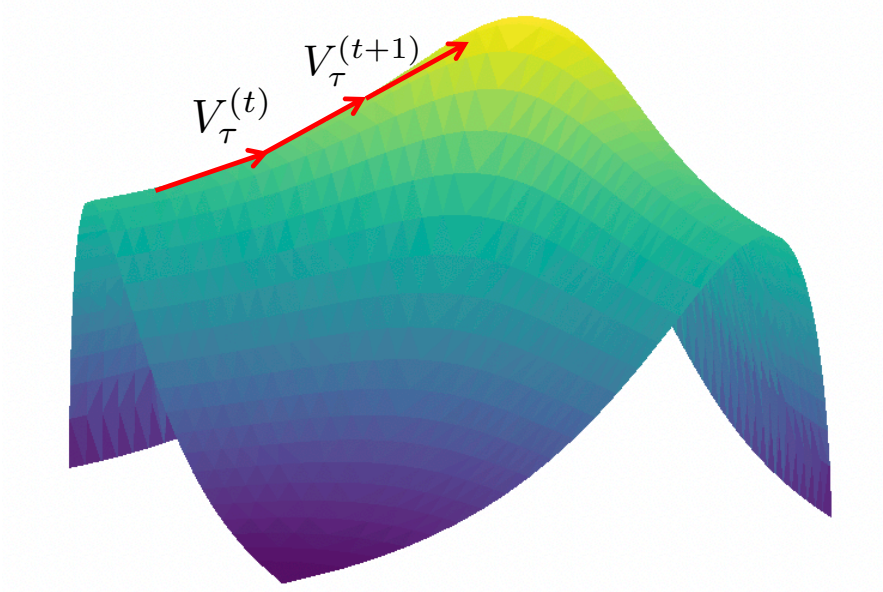
A key lemma: monotonic performance improvement



$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) = \mathbb{E}_{s\sim d_\rho^{(t+1)}}\Bigg[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma}\right)\underbrace{\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s)\,\middle\|\,\pi^{(t)}(\cdot|s)\right)}_{\text{KL divergence}}$$

$$+\frac{1}{\eta}\underbrace{\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\middle\|\,\pi^{(t+1)}(\cdot|s)\right)}_{\text{KL divergence}}\Bigg]$$

A key lemma: monotonic performance improvement



$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) = \mathbb{E}_{s\sim d_\rho^{(t+1)}}\Bigg[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma}\right)\underbrace{\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s)\,\middle\|\,\pi^{(t)}(\cdot|s)\right)}_{\text{KL divergence}}$$

$$+\frac{1}{\eta}\underbrace{\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\middle\|\,\pi^{(t+1)}(\cdot|s)\right)}_{\text{KL divergence}}\Bigg]$$

Implication: monotonic improvement of $V_\tau(s)$ and $Q_\tau(s,a)$.

A key operator: soft Bellman operator

Soft Bellman operator

$$\mathcal{T}_\tau(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}}$$

$$+ \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{\pi(\cdot|s')} \mathop{\mathbb{E}}_{a' \sim \pi(\cdot|s')} \left[\underbrace{Q(s',a')}_{\text{next state's value}} - \underbrace{\tau \log \pi(a'|s')}_{\text{entropy}} \right] \right],$$

A key operator: soft Bellman operator

Soft Bellman operator

$$\mathcal{T}_\tau(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}}$$

$$+\gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\max_{\pi(\cdot|s')} \mathop{\mathbb{E}}_{a'\sim\pi(\cdot|s')}\left[\underbrace{Q(s',a')}_{\text{next state's value}} - \underbrace{\tau\log\pi(a'|s')}_{\text{entropy}}\right]\right],$$

Soft Bellman equation: $Q^\star_\tau$ is *unique* solution to

$$\mathcal{T}_\tau(Q^\star_\tau) = Q^\star_\tau$$

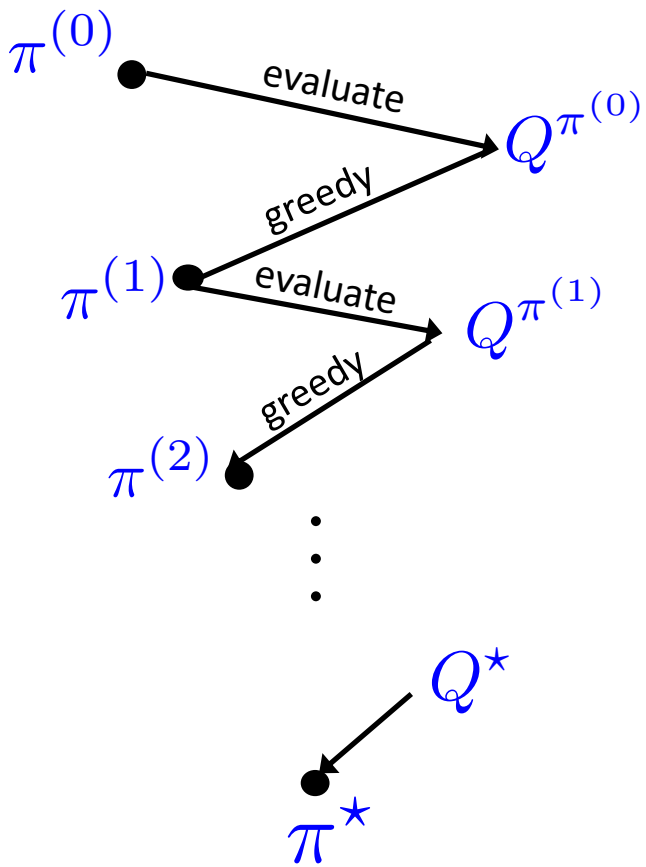
γ-contraction of soft Bellman operator:

$$\|\mathcal{T}_\tau(Q_1) - \mathcal{T}_\tau(Q_2)\|_\infty \le \gamma\|Q_1 - Q_2\|_\infty$$

Richard Bellman

Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

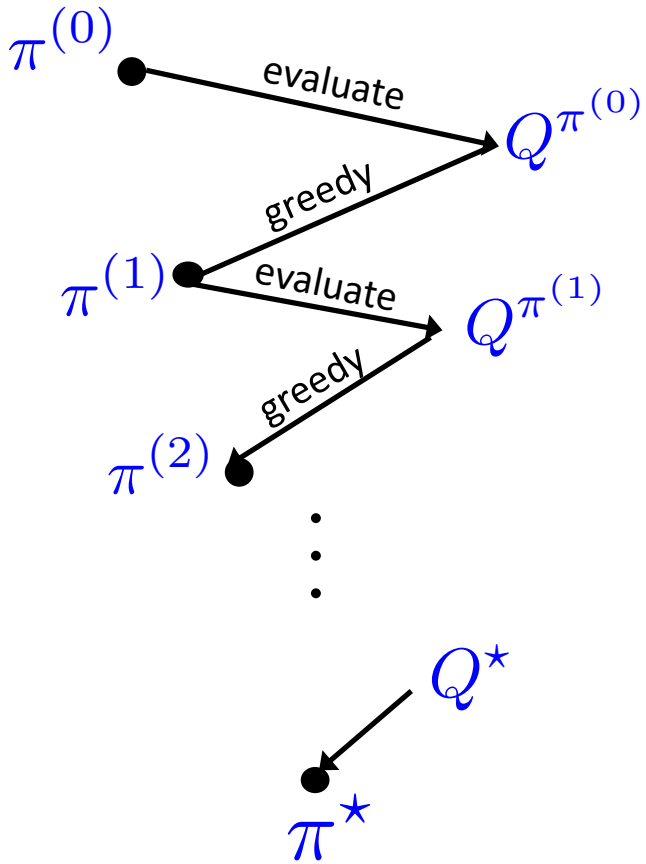
Policy iteration



Bellman operator

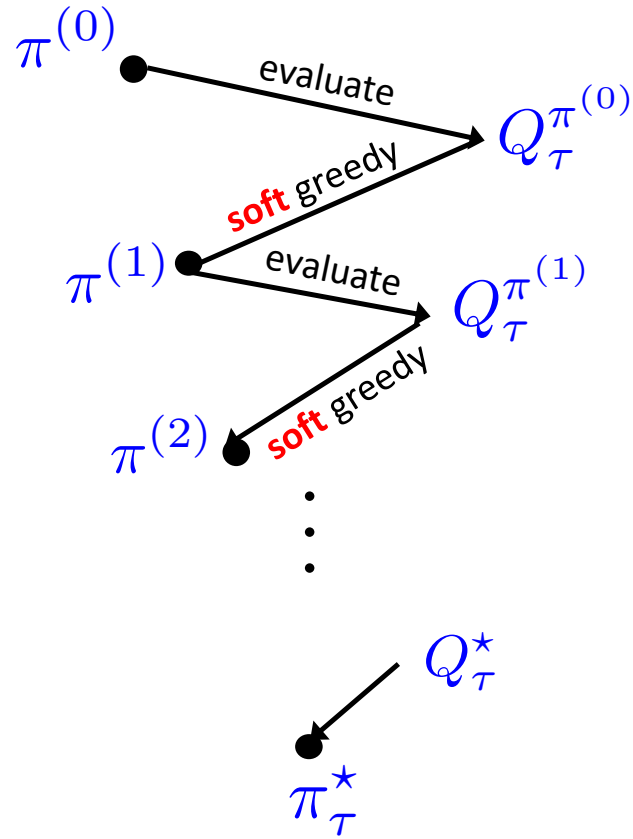
Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

Policy iteration



Bellman operator

Soft policy iteration



Soft Bellman operator

A key linear system: general learning rates

Let $x_t := \begin{bmatrix} \|Q_\tau^\star - Q_\tau^{(t)}\|_\infty \\ \|Q_\tau^\star - \tau \log \xi^{(t)}\|_\infty \end{bmatrix}$ and $y := \begin{bmatrix} \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty \\ 0 \end{bmatrix}$,

where $\xi^{(t)} \propto \pi^{(t)}$ is an auxiliary sequence, then

A key linear system: general learning rates

Let $x_t := \begin{bmatrix} \|Q_\tau^\star - Q_\tau^{(t)}\|_\infty \\ \|Q_\tau^\star - \tau \log \xi^{(t)}\|_\infty \end{bmatrix}$ and $y := \begin{bmatrix} \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty \\ 0 \end{bmatrix}$,

where $\xi^{(t)} \propto \pi^{(t)}$ is an auxiliary sequence, then

$$x_{t+1} \leq A x_t + \gamma \left(1 - \frac{\eta\tau}{1-\gamma}\right)^{t+1} y,$$

where

$$A := \begin{bmatrix} \gamma \\ 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{\eta\tau}{1-\gamma} & 1 - \frac{\eta\tau}{1-\gamma} \end{bmatrix}$$

is a rank-1 matrix with a non-zero eigenvalue $\underbrace{1-\eta\tau}_{\text{contraction rate!}}$.

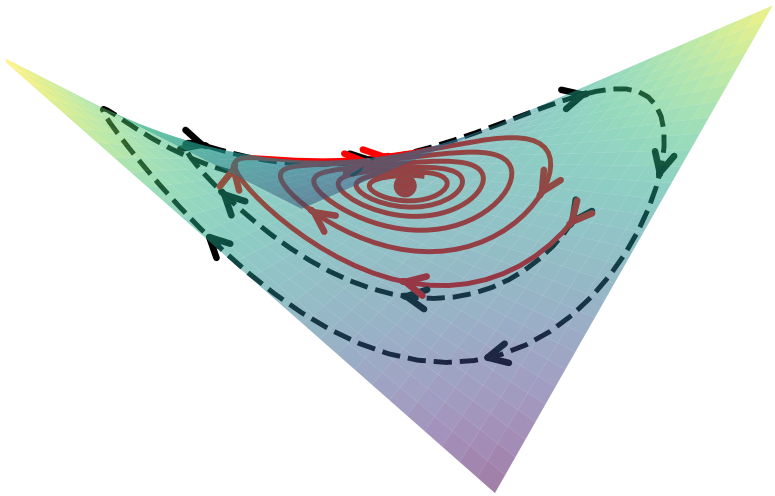
Policy Optimization for Zero-Sum Markov Games

Policy optimization: saddle-point optimization

Zero-sum two-player Markov game

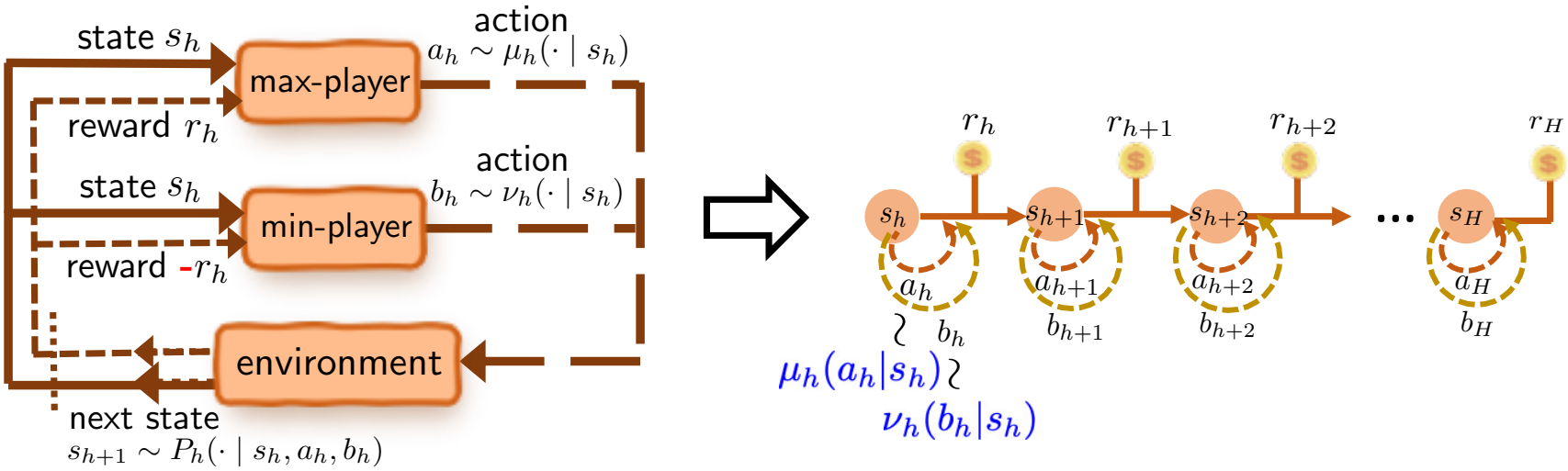
Given an initial state distribution $s \sim \rho$, find policy π such that

$$\max_{\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V^{\mu,\nu}(\rho) := \mathbb{E}_{s\sim\rho}[V^{\mu,\nu}(s)]$$



Can we design a policy optimization method that guarantees fast *last-iterate* convergence?

Entropy regularization in MARL

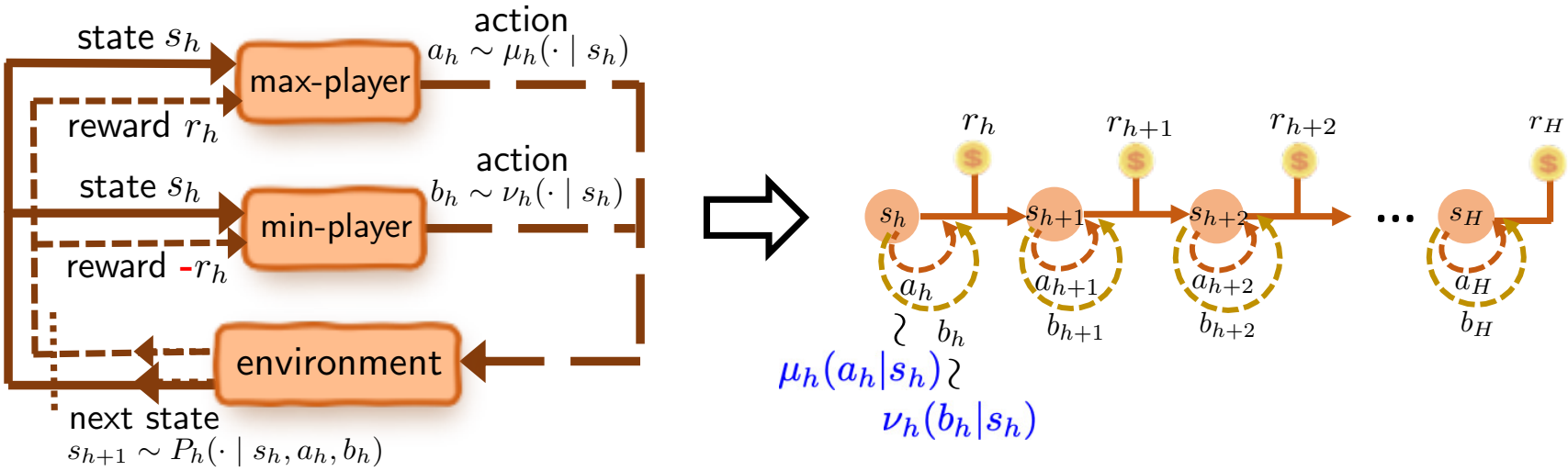


Promote the stochasticity of the policy pair using the **"soft"** value function (Williams and Peng, 1991; Cen et al., 2020):

$$V_\tau^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{h=1}^{H}\left(r_t + \tau\mathcal{H}(\mu_t(\cdot|s_t) - \tau\mathcal{H}(\nu_t(\cdot|s_t)\right)\,\middle|\, s_0 = s\right],$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

Entropy regularization in MARL



Promote the stochasticity of the policy pair using the **"soft"** value function (Williams and Peng, 1991; Cen et al., 2020):

$$V_\tau^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{h=1}^{H}\left(r_t + \tau\mathcal{H}(\mu_t(\cdot|s_t) - \tau\mathcal{H}(\nu_t(\cdot|s_t)\right)\,\middle|\, s_0 = s\right],$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

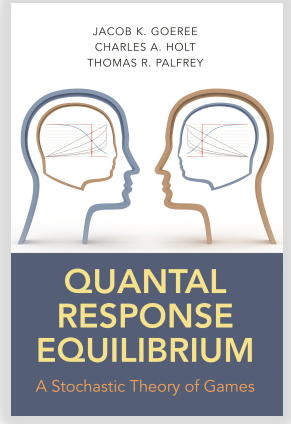
$$\max_{\mu\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu\in\Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho)$$

Quantal response equilibrium (QRE)

Quantal response equilibrium (McKelvey and Palfrey, 1995)

The quantal response equilibrium (QRE) is the policy pair $(\mu_\tau^\star, \nu_\tau^\star)$ *that is the unique solution to*

$$\max_{\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho).$$



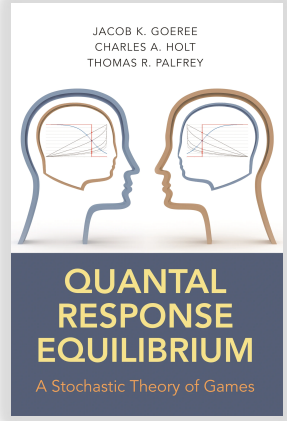
- Unlike NE, QRE assumes bounded rationality: action probability follows the logit function.

Quantal response equilibrium (QRE)

Quantal response equilibrium (McKelvey and Palfrey, 1995)

The quantal response equilibrium (QRE) is the policy pair $(\mu_\tau^\star, \nu_\tau^\star)$ *that is the unique solution to*

$$\max_{\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho).$$



- Unlike NE, QRE assumes bounded rationality: action probability follows the logit function.

Translating to an ϵ-NE: setting $\tau \asymp \widetilde{O}(\epsilon/H)$.

Soft value iteration

Soft value iteration: for $h = H, \dots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \left[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s') \nu(s') + \tau \mathcal{H}(\mu(s')) - \tau \mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \right],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

Soft value iteration

Soft value iteration: for $h = H, \dots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Bigg[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s') \nu(s') + \tau \mathcal{H}(\mu(s')) - \tau \mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Bigg],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

Entropy-regularized matrix game

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} \mu^\top A \nu + \tau \mathcal{H}(\mu) - \tau \mathcal{H}(\nu)$$

Motivation: an implicit update method

Implicit update (IU) method

For $t = 0, 1, \cdots,$

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\nu^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \mu^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

Motivation: an implicit update method

Implicit update (IU) method

For $t = 0, 1, \cdots,$

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\Big([A\nu^{(t+1)}]/\tau\Big)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\Big(-[A^\top \mu^{(t+1)}]/\tau\Big)^{\eta\tau} \end{cases}$$

Theorem (Cen, Wei, Chi, 2021)

Suppose that $0 < \eta \le 1/\tau$, *then for all* $t \ge 0$,

$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \le (1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big),$$

where $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) = \mathsf{KL}\big(\mu_\tau^\star \| \mu^{(t)}\big) + \mathsf{KL}\big(\nu_\tau^\star \| \nu^{(t)}\big).$

Motivation: an implicit update method

Implicit update (IU) method

For $t = 0, 1, \cdots,$

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\Big([A\nu^{(t+1)}]/\tau\Big)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\Big(-[A^\top \mu^{(t+1)}]/\tau\Big)^{\eta\tau} \end{cases}$$

Theorem (Cen, Wei, Chi, 2021)

Suppose that $0 < \eta \le 1/\tau$, *then for all* $t \ge 0$,

$$\mathsf{KL}(\zeta_\tau^\star \,\|\, \zeta^{(t)}) \le (1-\eta\tau)^t \mathsf{KL}(\zeta_\tau^\star \,\|\, \zeta^{(0)}),$$

where $\mathsf{KL}(\zeta_\tau^\star \,\|\, \zeta^{(t)}) = \mathsf{KL}(\mu_\tau^\star \| \mu^{(t)}) + \mathsf{KL}(\nu_\tau^\star \| \nu^{(t)})$.

Can we make this practical?

Policy extragradient methods: predict and update

Predictive update (PU) method

For $t = 0, 1, \cdots,$

2. *update:*

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \bar{\mu}^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

Policy extragradient methods: predict and update

Predictive update (PU) method

For $t = 0, 1, \cdots,$

1. *extrapolate/predict:*

$$\begin{cases} \bar{\mu}^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\nu^{(t)}]/\tau\right)^{\eta\tau} \\ \bar{\nu}^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \mu^{(t)}]/\tau\right)^{\eta\tau} \end{cases}$$

2. *update:*

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \bar{\mu}^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

Last-iterate convergence

Theorem (Cen, Wei, Chi, JMLR 2024)

Suppose that $\eta \le \min\left\{\frac{1}{2\tau+2\|A\|_\infty}, \frac{1}{4\|A\|_\infty}\right\}$, *then for all* $t \ge 0$, *the last-iterate converges to* ϵ*-QRE within* $\widetilde{O}\left(\frac{1}{\eta\tau}\log\frac{1}{\epsilon}\right)$ *iterations.*

- **Entropy-regularized matrix game:** To get an ϵ-QRE, the iteration complexity is at most

$$\widetilde{O}\left(\left(1+\frac{\|A\|_\infty}{\tau}\right)\log\frac{1}{\epsilon}\right).$$

- **Unregularized matrix game:** To get an ϵ-NE, the iteration complexity is at most

$$\widetilde{O}\left(\frac{\|A\|_\infty}{\epsilon}\right).$$

No need to assume unique Nash equilibrium!

Soft value iteration via nested-loop OMWU

Soft value iteration: for $h = H, \dots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Bigg[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s') \nu(s') + \tau \mathcal{H}(\mu(s')) - \tau \mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Bigg],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

Soft value iteration via nested-loop OMWU

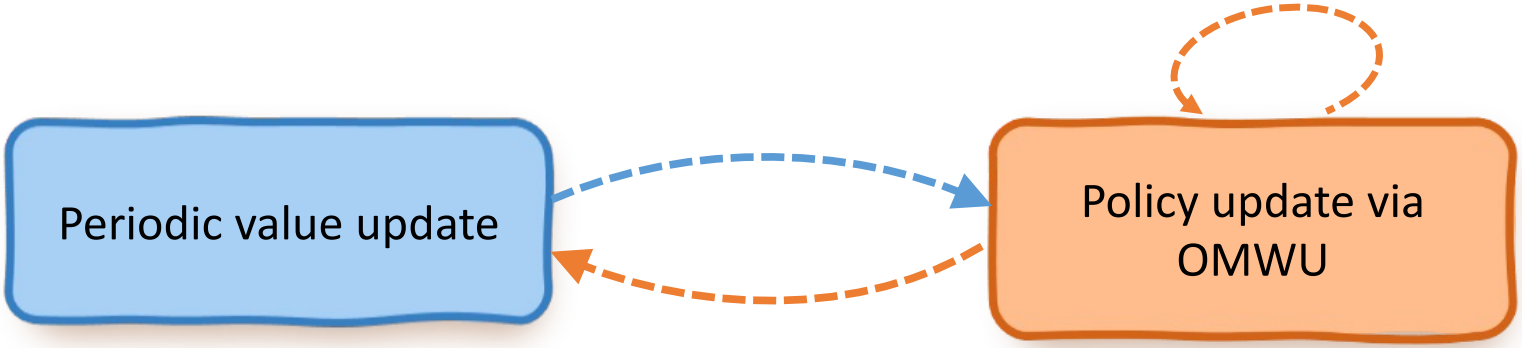
Soft value iteration: for $h = H, \ldots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Big[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s')\nu(s') + \tau\mathcal{H}(\mu(s')) - \tau\mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Big],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

Nested-loop approach:

$(\mu_h^{(t)}, \nu_h^{(t)}) \leftarrow \texttt{OMWU}(Q_h)$



$Q_h \leftarrow \texttt{SVI}(Q_{h+1})$

Soft value iteration via nested-loop OMWU

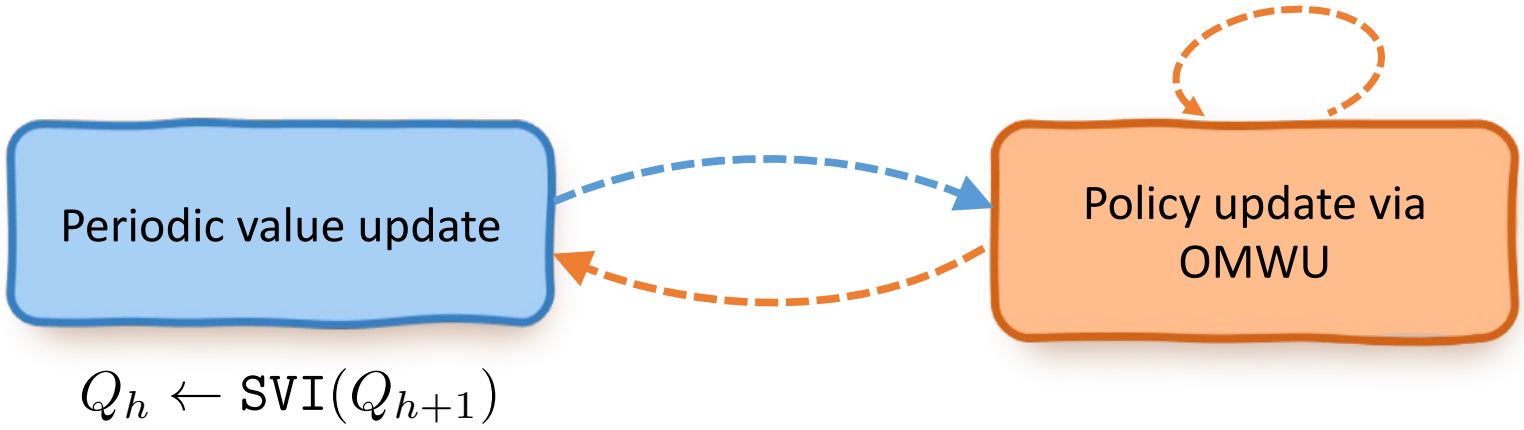
Soft value iteration: for $h = H, \ldots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Big[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s')\nu(s') + \tau\mathcal{H}(\mu(s')) - \tau\mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Big],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

Nested-loop approach:

$$(\mu_h^{(t)}, \nu_h^{(t)}) \leftarrow \texttt{OMWU}(Q_h)$$



However, not easy to use in online settings...

A two-timescale single-loop approach?

Soft value iteration: for $h = H, \ldots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Bigg[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s') \nu(s') + \tau \mathcal{H}(\mu(s')) - \tau \mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Bigg],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

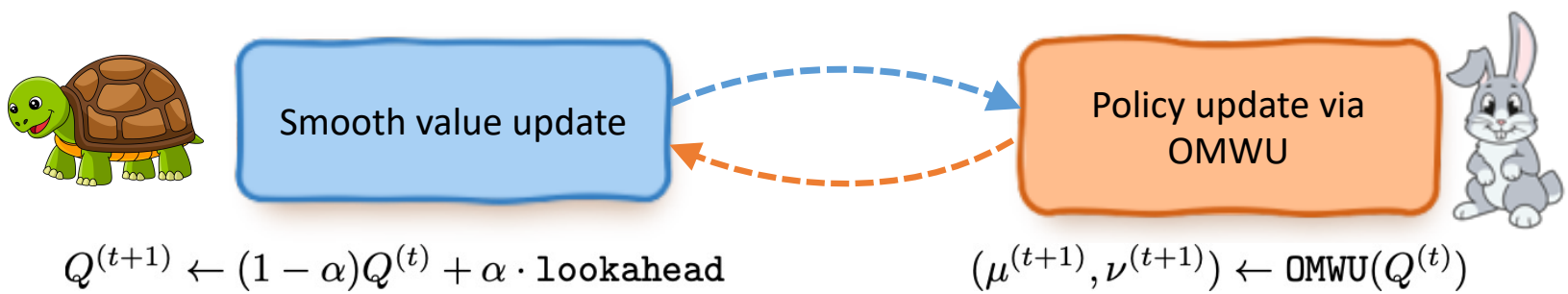
A two-timescale single-loop approach?

Soft value iteration: for $h = H, \ldots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Big[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s')\nu(s') + \tau\mathcal{H}(\mu(s')) - \tau\mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Big],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

Single-loop, two-timescale approach:



Main result: episodic setting

Theorem (Cen, Chi, Du, Xiao, ICLR 2022)

The last-iterate of the two-timescale single-loop algorithm finds an ϵ-QRE in

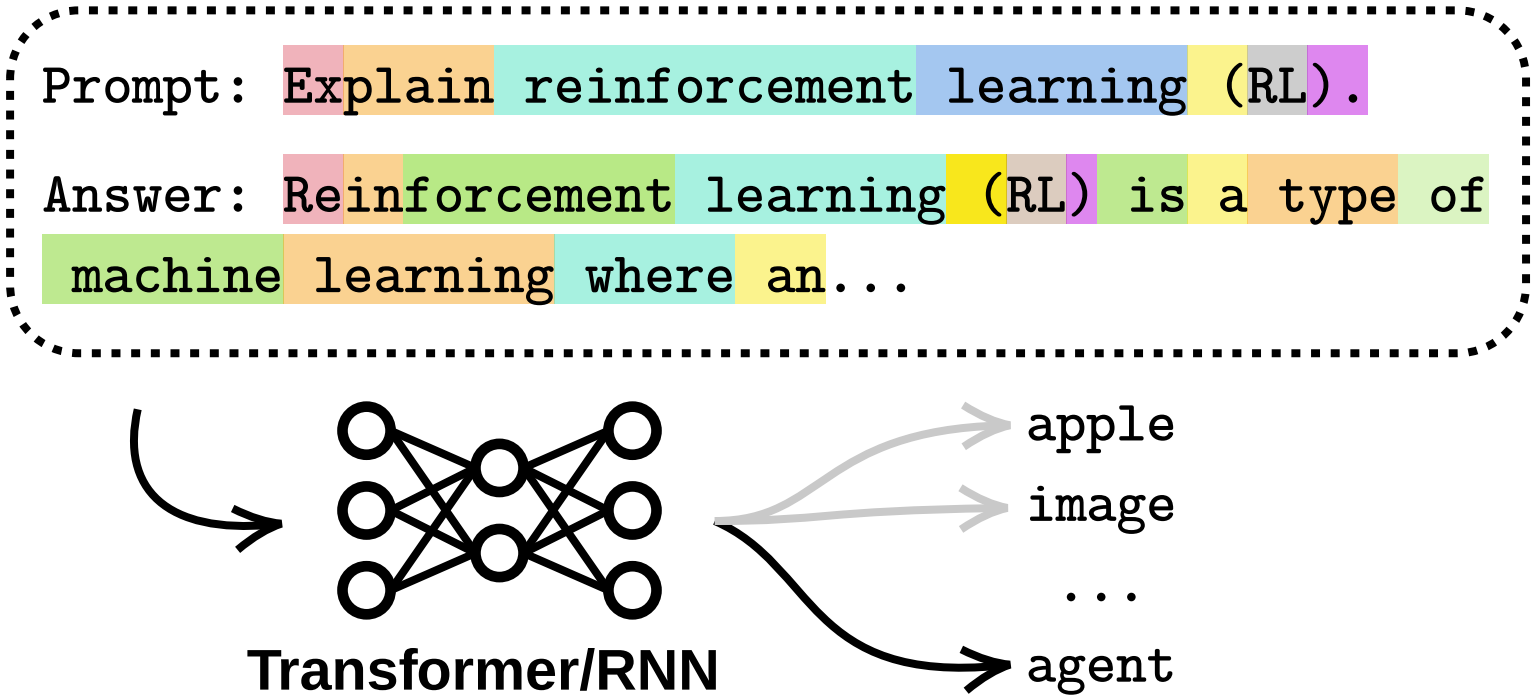
$$\widetilde{O}\left(\frac{H^2}{\tau}\log\frac{1}{\epsilon}\right)$$

iterations, corresponding to $\widetilde{O}\left(\frac{H^3}{\epsilon}\right)$ iterations for finding an ϵ-NE.

- First last-iterate convergence result for the episodic setting.
- **Almost dimension-free:** independent of the size of the state-action space.
- For the infinite-horizon γ-discounted setting, the complexity reads $\widetilde{O}\left(\frac{S}{(1-\gamma)^4\tau}\log\frac{1}{\epsilon}\right)$.

A crumb of RLHF

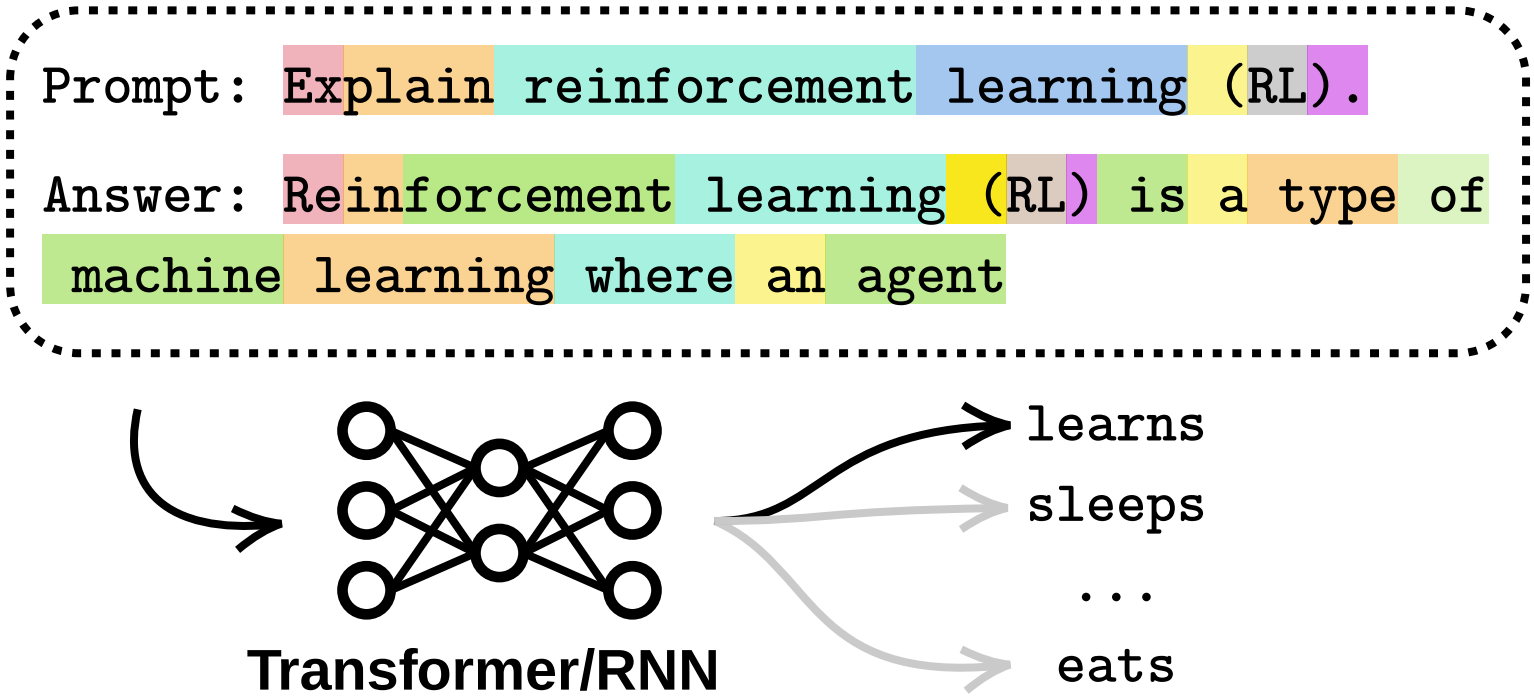
Language models as policies



Given prompt $x \in \mathcal{X}$, a language model generates an answer:

$$y \sim \underbrace{\pi(\cdot|x)}_{\text{parameterized by LLM}}$$

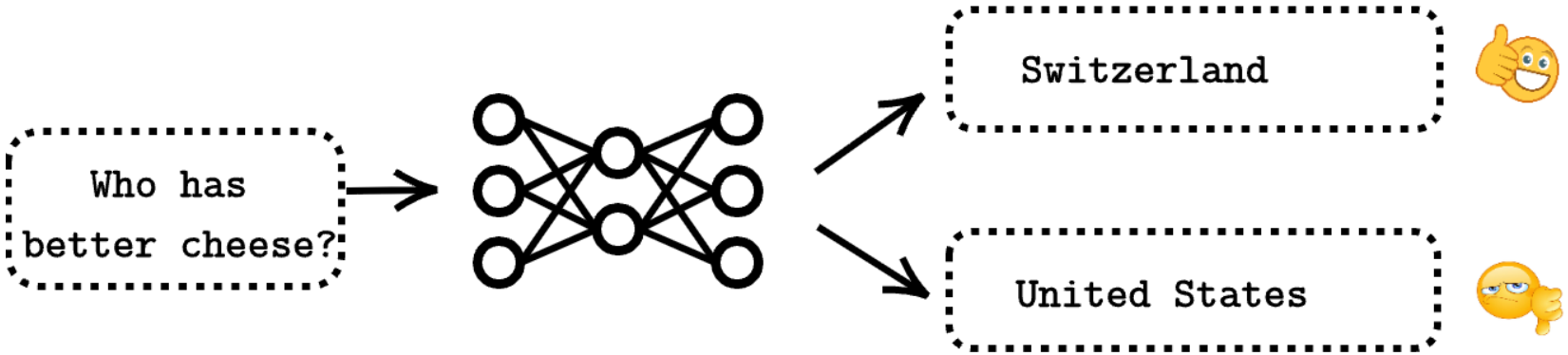
Language models as policies



Given prompt $x \in \mathcal{X}$, a language model generates an answer:

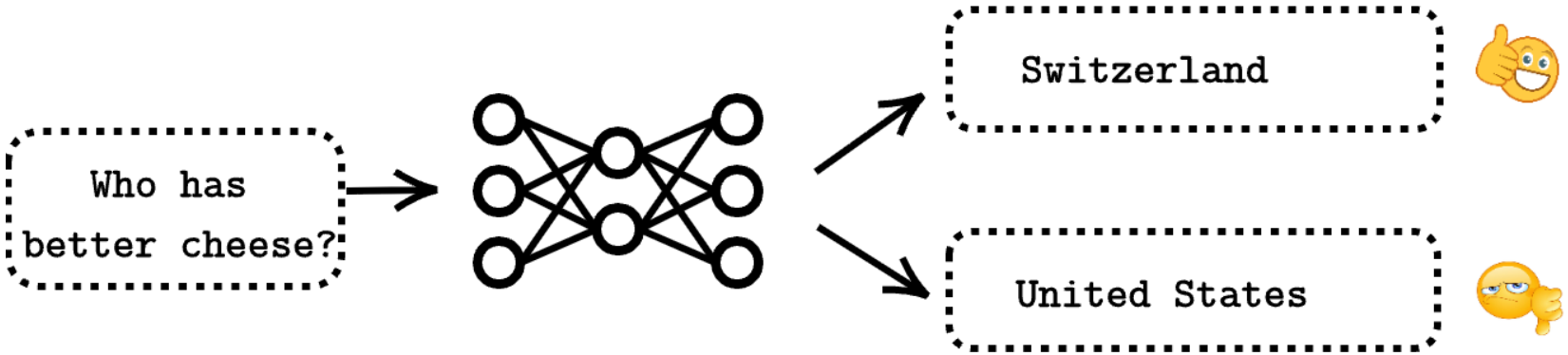
$$y \sim \underbrace{\pi(\cdot|x)}_{\text{parameterized by LLM}}$$

RL with human feedback (RLHF)



Goal: finetune the LLM to align with human preference

RL with human feedback (RLHF)



Goal: finetune the LLM to align with human preference

Prototypical pipeline:

- *Reward learning*: learn a reward model from preference data;
- *Policy optimization*: optimize the LLM to maximize the reward.

RLHF: reward learning

Bradly-Terry model

The probability of pairwise comparison $i \succ j$ is modeled by

$$\mathbb{P}(i \succ j) = \frac{\exp(r_i^\star)}{\exp(r_i^\star) + \exp(r_j^\star)} = \sigma(r_i^\star - r_j^\star),$$

where $r_i^\star \in \mathbb{R}$ is the score associated with item i.

RLHF: reward learning

Bradly-Terry model

The probability of pairwise comparison $i \succ j$ is modeled by

$$\mathbb{P}(i \succ j) = \frac{\exp(r_i^\star)}{\exp(r_i^\star) + \exp(r_j^\star)} = \sigma(r_i^\star - r_j^\star),$$

where $r_i^\star \in \mathbb{R}$ is the score associated with item i.

- **Reward model:** $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, evaluating the quality of a prompt-answer pair (x, y) that aligns with human preference;

RLHF: reward learning

Bradly-Terry model

The probability of pairwise comparison $i \succ j$ is modeled by

$$\mathbb{P}(i \succ j) = \frac{\exp(r_i^\star)}{\exp(r_i^\star) + \exp(r_j^\star)} = \sigma(r_i^\star - r_j^\star),$$

where $r_i^\star \in \mathbb{R}$ is the score associated with item i.

- **Reward model:** $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, evaluating the quality of a prompt-answer pair (x, y) that aligns with human preference;
- **Reward learning:** Given comparison data $\mathcal{D} = \{(x^i, y_+^i, y_-^i)\}_{i=1}^N$, the MLE of the reward function is given by

$$r_{\mathsf{MLE}} = \mathsf{argmin}_r \ell(r, \mathcal{D}),$$

where $$\ell(r, \mathcal{D}) = -\sum_{i=1}^N \log \sigma(r(x^i, y_+^i) - r(x^i, y_-^i)).$$

RLHF: policy optimization

Policy optimization via reward maximization

Find π that (approximately) maximizes the objective w.r.t. r

$$J(r, \pi) = \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y \sim \pi(\cdot|x)}} [r(x, y)] - \beta \mathop{\mathbb{E}}_{x \sim \rho} [\mathsf{KL}\pi(\cdot|x)\pi_{\text{ref}}(\cdot|x)]$$

- $\beta > 0$: KL regularization parameter;
- π_{ref}: a reference policy, typically the model after SFT;
- $\rho \in \Delta(\mathcal{X})$: prompt distribution.

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

1. Reward learning: $\widehat{r} \leftarrow \text{argmin}_r \ell(r, \mathcal{D})$
2. Policy learning: $\widehat{\pi} \leftarrow \text{argmax}_\pi J(\widehat{r}, \pi)$

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

Observation: the optimal π w.r.t. r admits a closed-form solution

$$\pi_r = \text{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

Observation: the optimal π w.r.t. r admits a closed-form solution

$$\pi_r = \text{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x,y)/\beta)}{Z(r,x)}.$$

- The reward function r in terms of its optimal π_r is

$$r(x,y) = \underbrace{\beta\big(\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x) + \log Z(r,x)\big)}_{=:r(\pi)}.$$

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

- The reward function r in terms of its optimal π_r is

$$r(x,y) = \underbrace{\beta\big(\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x) + \log Z(r,x)\big)}_{=:r(\pi)}.$$

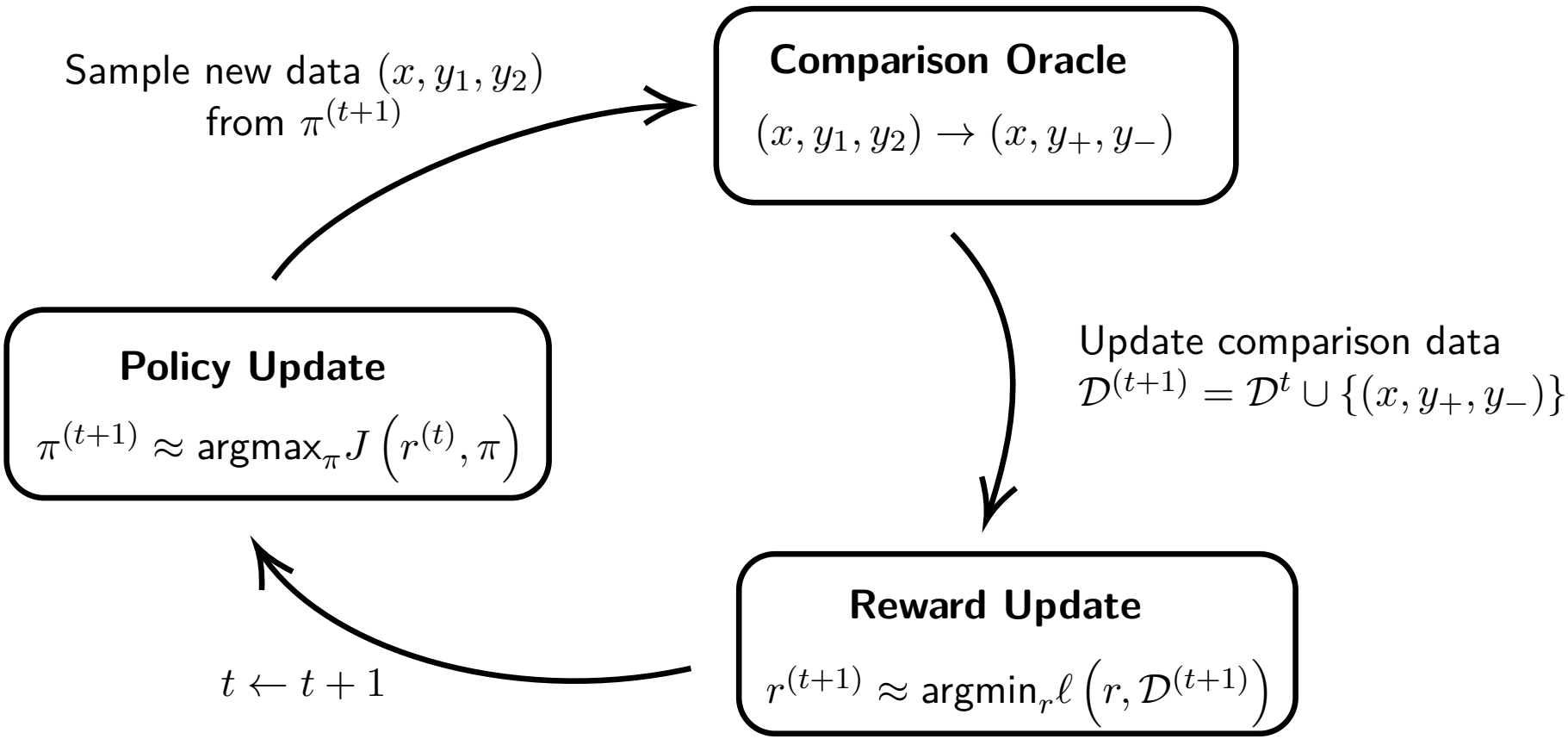
- The two-step procedure is equivalent to

$$\begin{aligned}\widehat{\pi} &\leftarrow \mathsf{argmin}_\pi \ell(r(\pi), \mathcal{D}) \\ &= -\sum_{i=1}^{N} \log \sigma\left(\beta \log \frac{\pi(y_+^i|x^i)}{\pi_{\text{ref}}(y_+^i|x^i)} - \beta \log \frac{\pi(y_-^i|x^i)}{\pi_{\text{ref}}(y_-^i|x^i)}\right).\end{aligned}$$

Single-step and **policy-only**! Very popular in practice.

Online RLHF

Leverage online data collection to improve data coverage - how do we perform **exploration** in the policy space directly?



Exploration via optimistic MLE

Optimistic MLE: Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \mathsf{argmin}_r \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

Exploration via optimistic MLE

Optimistic MLE: Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \mathsf{argmin}_r \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

- *Small caveat:* the update is not well-defined, since the BT model cannot distinguish between r and $r + c \cdot \mathbf{1}$, while

$$J^\star(r + c \cdot \mathbf{1}) = J^\star(r) + c.$$

Exploration via optimistic MLE

Optimistic MLE: Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \mathsf{argmin}_{r \in \mathcal{R}} \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

- *Small caveat:* the update is not well-defined, since the BT model cannot distinguish between r and $r + c \cdot \mathbf{1}$, while

$$J^\star(r + c \cdot \mathbf{1}) = J^\star(r) + c.$$

- We can resolve the shift ambiguity by focusing on the following equivalent class of reward functions:

$$\mathcal{R} = \left\{ r : \mathop{\mathbb{E}}_{x \sim \rho, y \sim \pi_{\mathrm{cal}}(\cdot|x)} [r(x, y)] = 0 \right\}.$$

Exploration via optimistic MLE

Optimistic MLE: Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \mathsf{argmin}_{r \in \mathcal{R}} \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

- *Small caveat:* the update is not well-defined, since the BT model cannot distinguish between r and $r + c \cdot \mathbf{1}$, while

$$J^\star(r + c \cdot \mathbf{1}) = J^\star(r) + c.$$

- We can resolve the shift ambiguity by focusing on the following equivalent class of reward functions:

$$\mathcal{R} = \left\{ r : \mathop{\mathbb{E}}_{x \sim \rho, y \sim \pi_{\mathrm{cal}}(\cdot|x)} [r(x, y)] = 0 \right\}.$$

Can we avoid solving a bilevel optimization problem?

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \mathsf{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \mathsf{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[\log Z(r, x) \right] \end{aligned}$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \text{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} [\log Z(r, x)] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log Z(r, x)] \end{aligned}$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \mathsf{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} [\log Z(r, x)] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log Z(r, x)] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \end{aligned}$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \mathsf{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[\log Z(r, x) \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \left[\log Z(r, x) \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \left[\cancel{r(x, y)} - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \end{aligned}$$

Value-incentivized preference optimization (VPO)

$$\pi^{(t+1)} \leftarrow \mathsf{argmin}_\pi\{\ell(r(\pi), \mathcal{D}^{(t)}) - \alpha J^\star(r(\pi))\}.$$

- The negative log-likelihood term reformulates into DPO loss:

$$\ell(r(\pi), \mathcal{D}^{(t)})$$
$$= -\sum_{(x,y_+,y_-)\in\mathcal{D}^{(t)}} \log\sigma\Big(\beta\Big(\log\frac{\pi(y_+|x)}{\pi_{\mathrm{ref}}(y_+|x)} - \log\frac{\pi(y_-|x)}{\pi_{\mathrm{ref}}(y_-|x)}\Big)\Big).$$

- The reward bias term can be written as:

$$J^\star(r(\pi)) = -\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} \left[\log\pi(y|x) - \log\pi_{\mathrm{ref}}(y|x)\right],$$

  which is essentially becomes a reverse-KL regularization that maximizes $\mathsf{KL}\pi_{\mathrm{cal}}(\cdot|x)\pi(\cdot|x)$.

Main results - online VPO

Theorem (Cen et al., ICLR 2025)

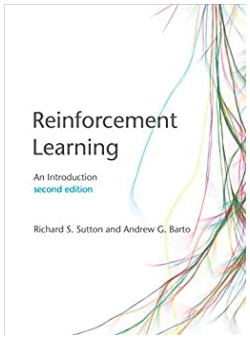
Assume that reward estimates $\|r^{(t)}\|_\infty \le B$ *and* $\|r^\star\|_\infty \le B$ *for some* $B > 0$. *With high probability we have*

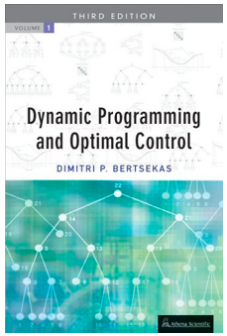
$$\sum_{t=1}^{T} \left[J^\star(r^\star) - J(r^\star, \pi^{(t)})\right] \le \widetilde{O}(\sqrt{T}).$$

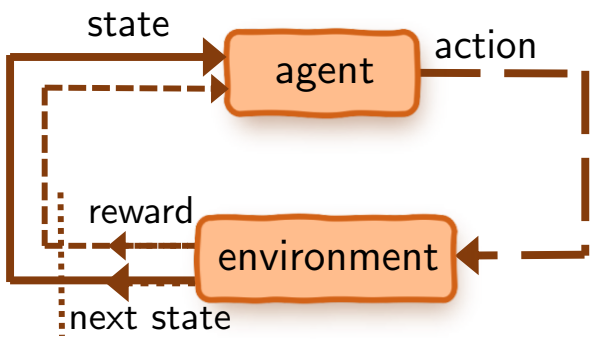
- We can obtain similar regret bounds under general function approximation of the reward model.
- Consistent with the $\widetilde{O}(\sqrt{T})$ regret for online RL with UCB bonus.
- Offline RL: flipping the sign of α leads to a pessimistic algorithm.

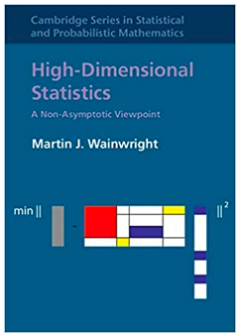
Concluding Remarks

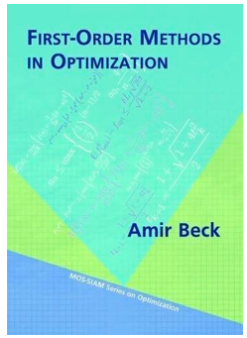
Concluding remarks











Understanding non-asymptotic performances of RL algorithms is a fruitful playground!

Promising directions:

- function approximation
- multi-agent/federated RL
- hybrid RL
- many more...

Beyond the tabular setting

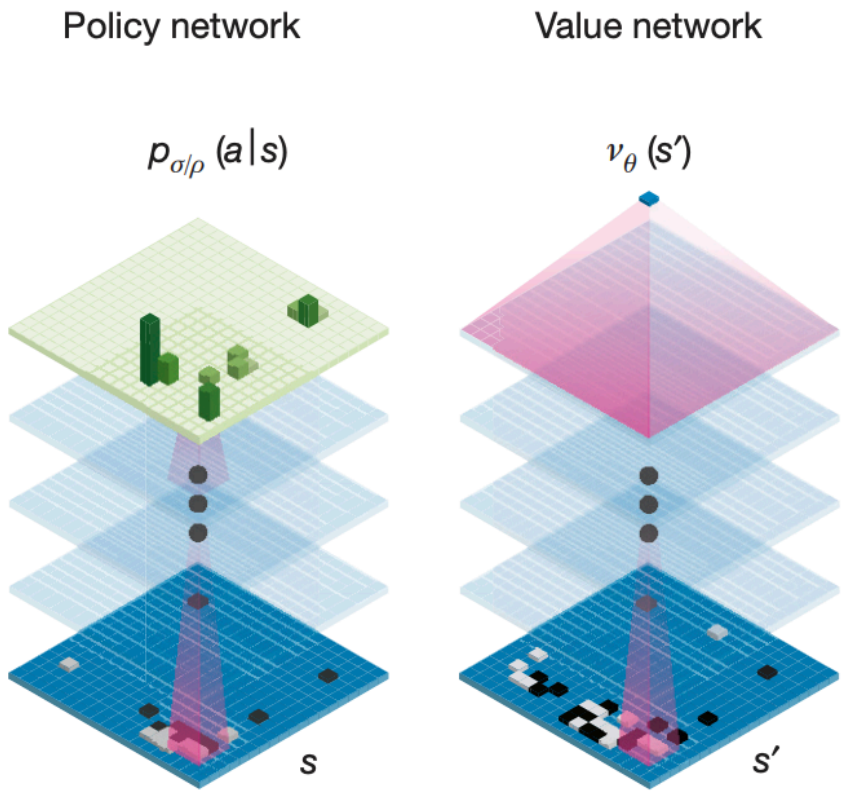
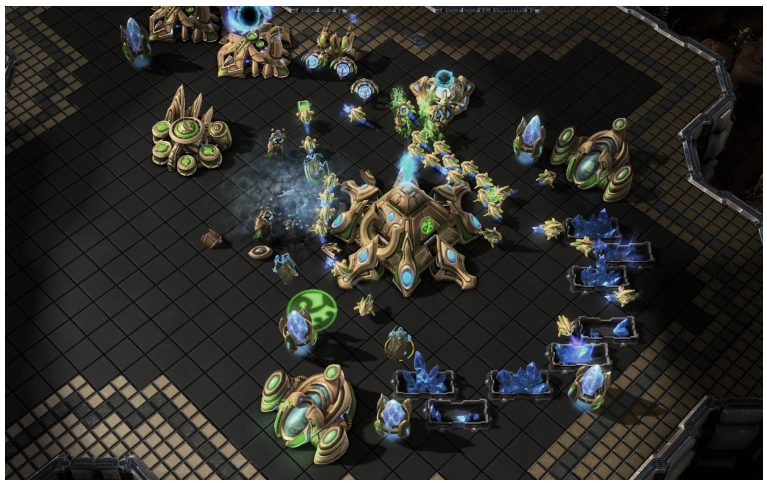


Figure credit: (Silver et al., 2016)

- function approximation for dimensionality reduction
- Provably efficient RL algorithms under minimal assumptions

(Osband and Van Roy, 2014; Dai et al., 2018; Du et al., 2019; Jin et al., 2020)

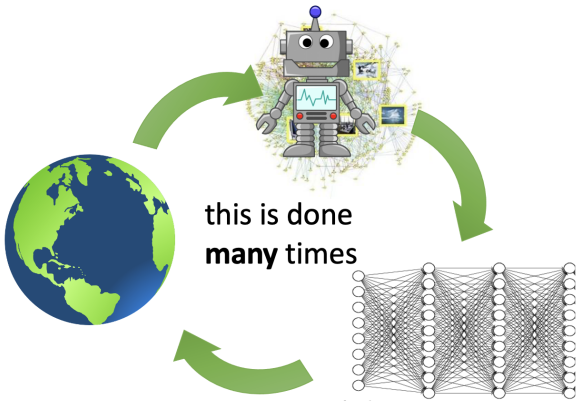
Multi-agent RL



- **Competitive setting:** finding Nash equilibria for Markov games
- **Collaborative setting:** multiple agents jointly optimize the policy to maximize the total reward

(Zhang, Yang, and Basar, 2021; Cen, Wei, and Chi, 2021)

Hybrid RL

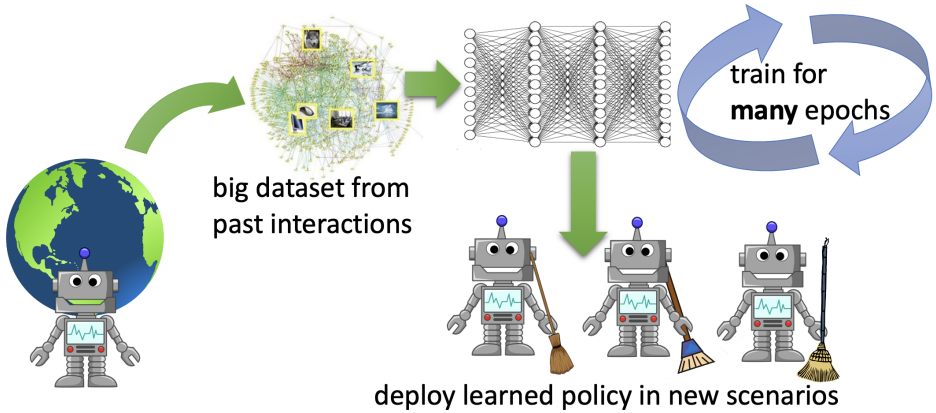


Online RL

- interact with environment
- actively collect new data

Offline/Batch RL

- no interaction
- data is given

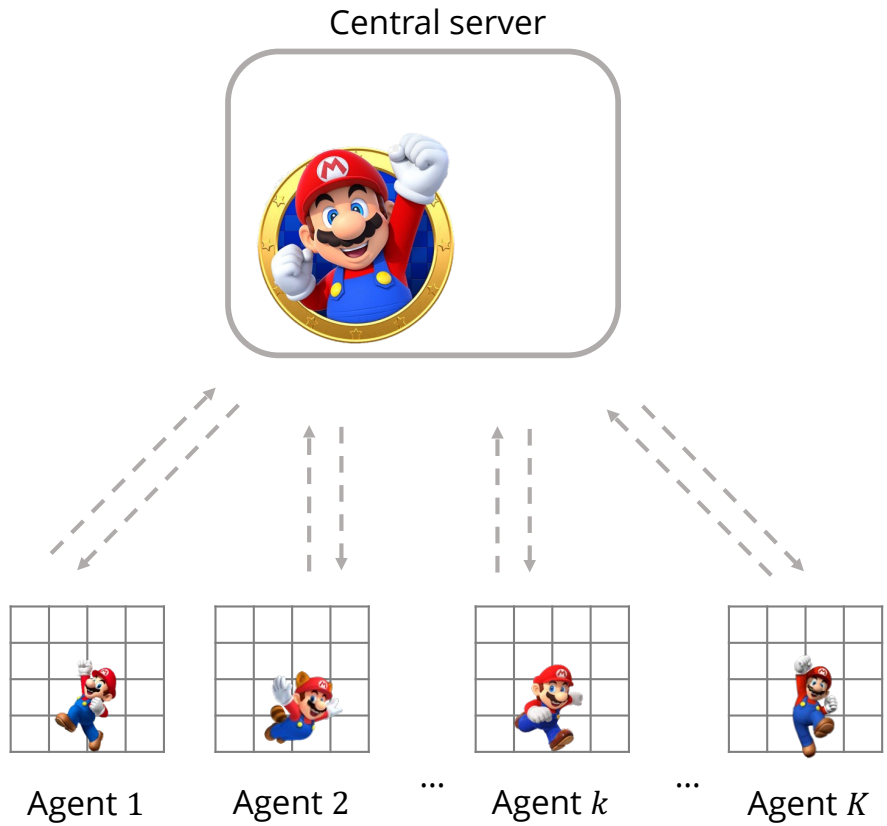


Can we achieve the best of both worlds?

(Wagenmaker and Pacchiano, 2022; Song et al., 2022; Li et al., 2023)

RL meets federated learning

Federated reinforcement learning enables multiple agents to collaboratively learn a global model without sharing datasets.



Can we achieve linear speedup via federated learning?

(Khodadadian et al., 2022; Woo et al., 2023)

Reference: policy optimization I

- "*Simple statistical gradient-following algorithms for connectionist reinforcement learning*," Williams, Machine Learning, 1992.
- "*Policy gradient methods for reinforcement learning with function approximation*," Sutton, McAllester, Singh, and Mansour, NeurIPS 1999.
- "*A natural policy gradient*," Kakade, NeurIPS 2001.
- "*Global convergence of policy gradient methods for the linear quadratic regulator*," Fazel, Ge, Kakade, and Mesbahi, ICML 2018.
- "*On the theory of policy gradient methods: Optimality, approximation, and distribution shift*," Agarwal, Kakade, Lee, and Mahajan, JMLR, 2021.
- "*On the global convergence rates of softmax policy gradient methods*," Mei, Xiao, Szepesvári, and Schuurmans, ICML 2020.
- "*Global optimality guarantees for policy gradient methods*," Bhandari and Russo, Operations Research, 2024.
- "*Provably efficient exploration in policy optimization*," Cai, Yang, Jin, and Wang, ICML 2020.

Reference: policy optimization II

- "*Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs*," Shani, Efroni, and Mannor, AAAI 2020.
- "*Softmax policy gradient methods can take exponential time to converge*," Li, Wei, Chi, Gu, and Chen, COLT 2021.
- "*Fast global convergence of natural policy gradient methods with entropy regularization*," Cen, Cheng, Chen, Wei, and Chi, Operations Research, 2021.
- "*Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence*," Zhan et. al., SIAM Journal on Optimization, 2023.
- "*Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes*," Lan, Mathematical Programming, 2023.
- "*An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods*," Liu, Zhang, Basar, and Yin, NeurIPS 2020.

Reference: policy optimization III

- "*Variational policy gradient method for reinforcement learning with general utilities*," Zhang, Koppel, Bedi, Szepesvári, and Wang, NeurIPS 2020.
- "*Fast policy extragradient methods for competitive games with entropy regularization*," Cen, Wei, and Chi, JMLR, 2024.

Reference: RLHF I

- "*Training language models to follow instructions with human feedback*," Ouyang et al., NeurIPS 2022.
- "*Direct preference optimization: Your language model is secretly a reward model*," Rafailov et. al., NeurIPS 2023.
- "*Principled reinforcement learning with human feedback from pairwise or K-wise comparisons*," Zhu, Jordan, and Jiao, ICML 2023.
- "*A general theoretical paradigm to understand learning from human preferences*," Azar et. al., AISTATS 2024.
- "*Nash learning from human feedback*," Munos et. al., ICML 2024.
- "*BoNBoN alignment for large language models and the sweetness of best-of-n sampling*," Gui, et. al., NeurIPS 2024.
- "*Value-incentivized preference optimization: A unified approach to online and offline RLHF*," Cen et. al., ICLR 2025.

Thanks!

yuejiechi.github.io