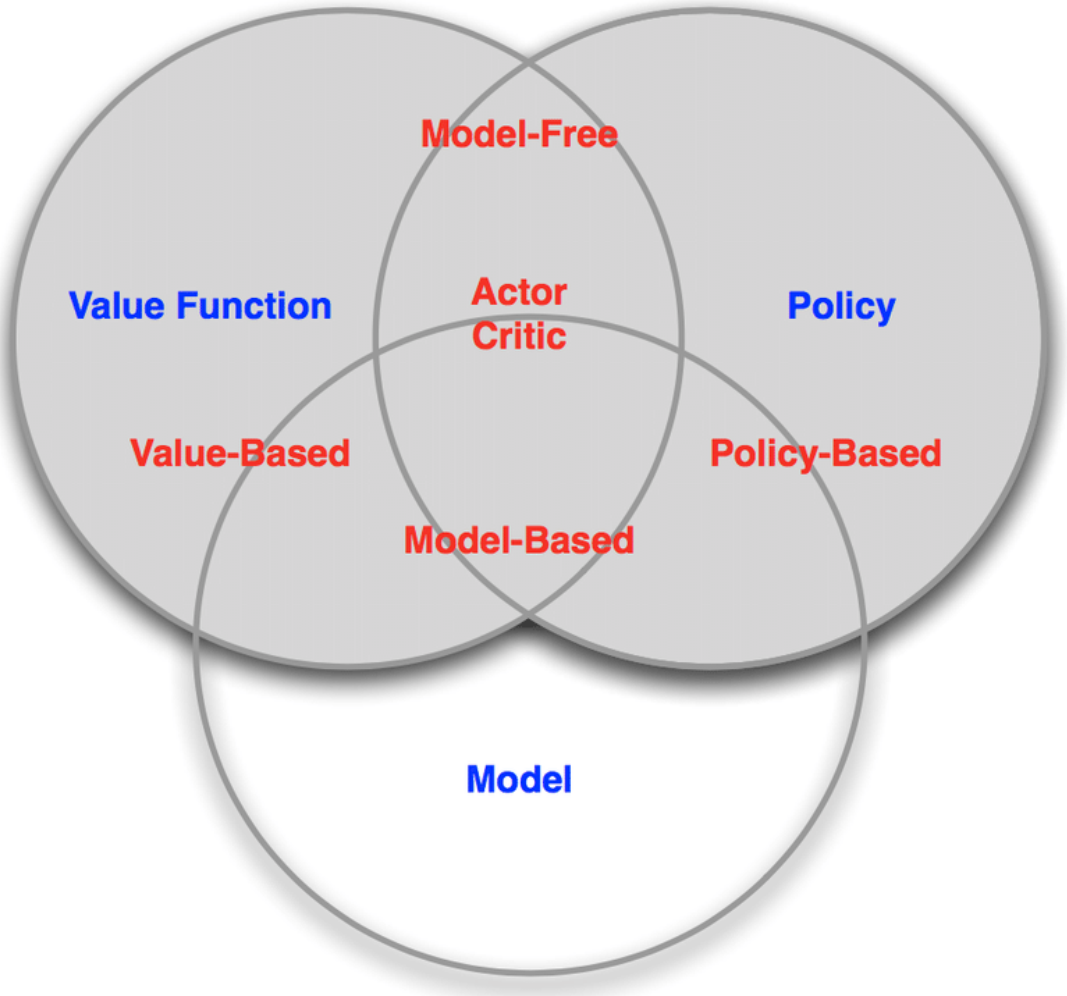
# Non-asymptotic Analysis for Reinforcement Learning (Part 3)

Yuejie Chi

Carnegie Mellon University

Sigmetrics Tutorial
June 2023

# A triad of RL approaches

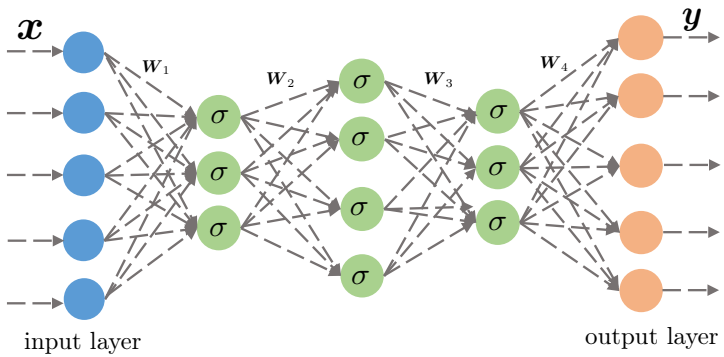


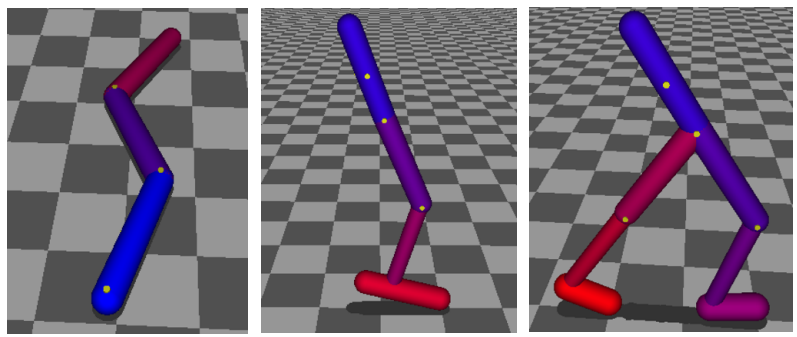

— *Figure credit: D. Silver*

# Policy optimization in practice

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

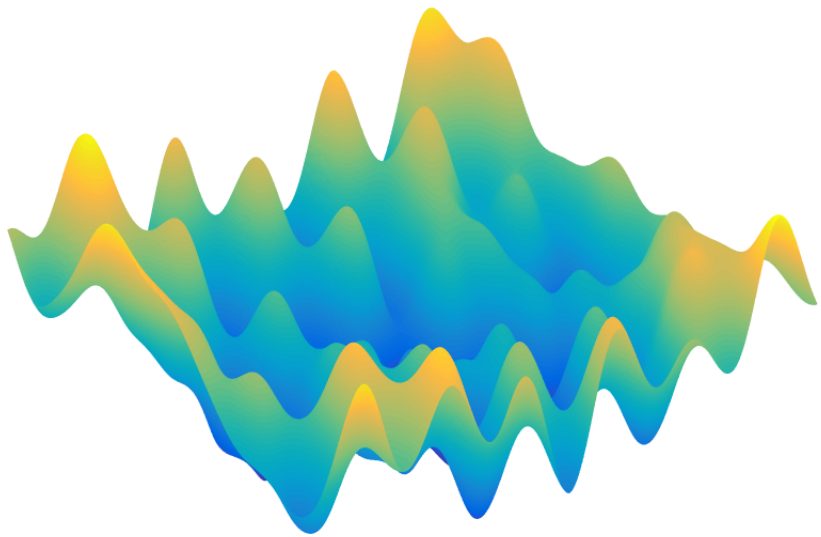
- directly optimize the policy, which is the quantity of interest;
- allow flexible differentiable parameterizations of the policy;
- work with both continuous and discrete problems.



# 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:**

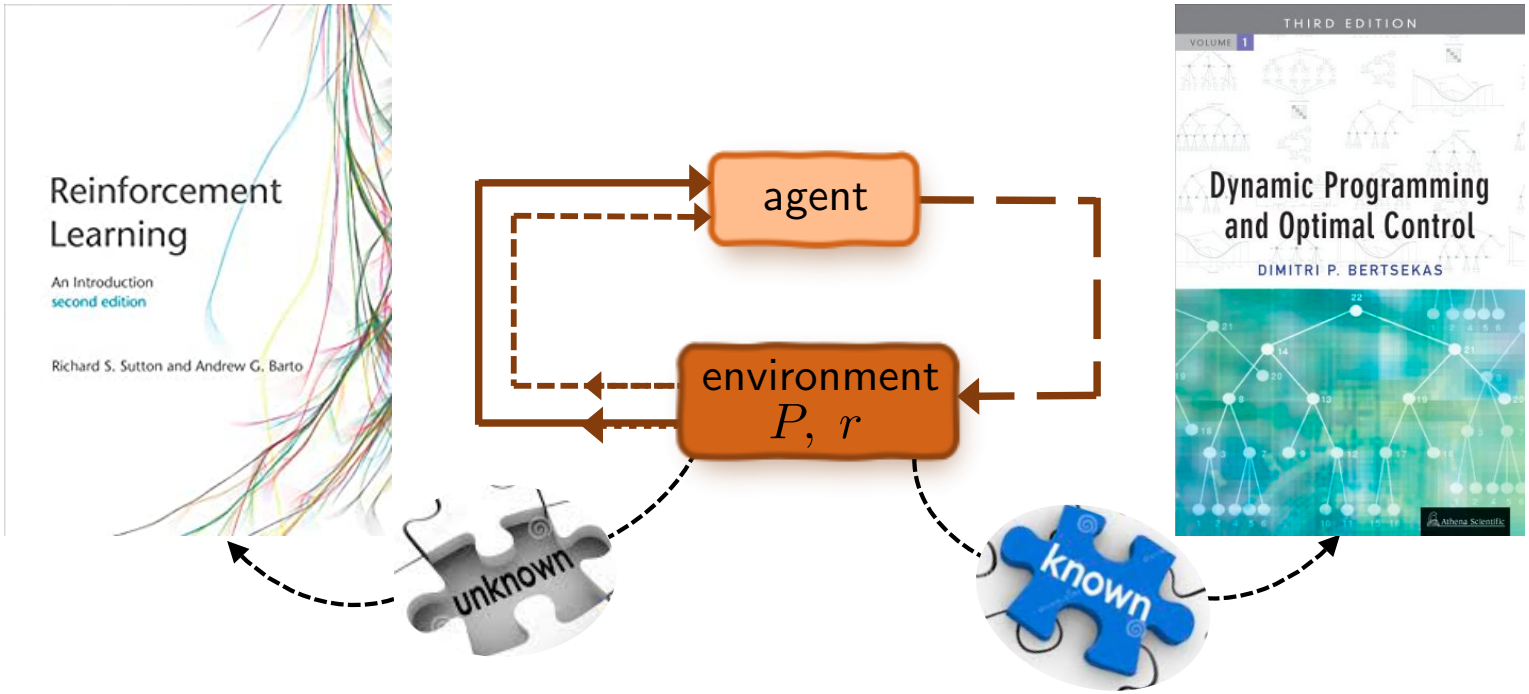
- understand finite-time convergence rates of popular heuristics;
- design fast-convergent algorithms that scale for finding policies with desirable properties.

# Outline

- Backgrounds and basics
  - policy gradient method
- Convergence guarantees of single-agent policy optimization
  - (natural) policy gradient methods
  - finite-time rate of global convergence
  - entropy regularization and beyond
- Multi-agent policy optimization: two-player zero-sum games
  - Matrix game
  - Markov game
- Concluding remarks and further pointers

*Backgrounds: policy optimization in tabular Markov decision processes*

# Searching for the optimal policy




**Goal:** find the optimal policy $\pi^\star$ that maximize $V^\pi(s)$

- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$

# Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy $\pi$ 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* $\eta$ *is the learning rate.*

# Softmax policy gradient methods

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

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

$\Downarrow$

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)**

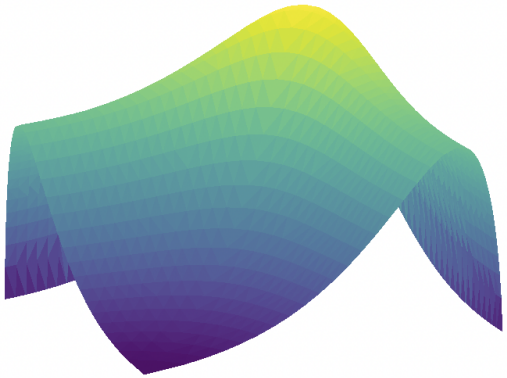
*For* $t = 0, 1, \cdots$

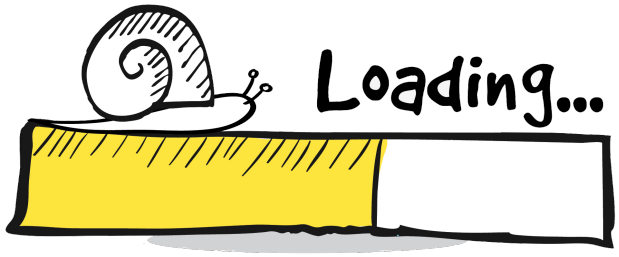
$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

*where* $\eta$ *is the learning rate.*

*Finite-time global convergence guarantees*

# 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

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

Is the rate of PG good, bad or ugly?

# A negative message

**Theorem (Li, Wei, Chi, Chen, 2021)**

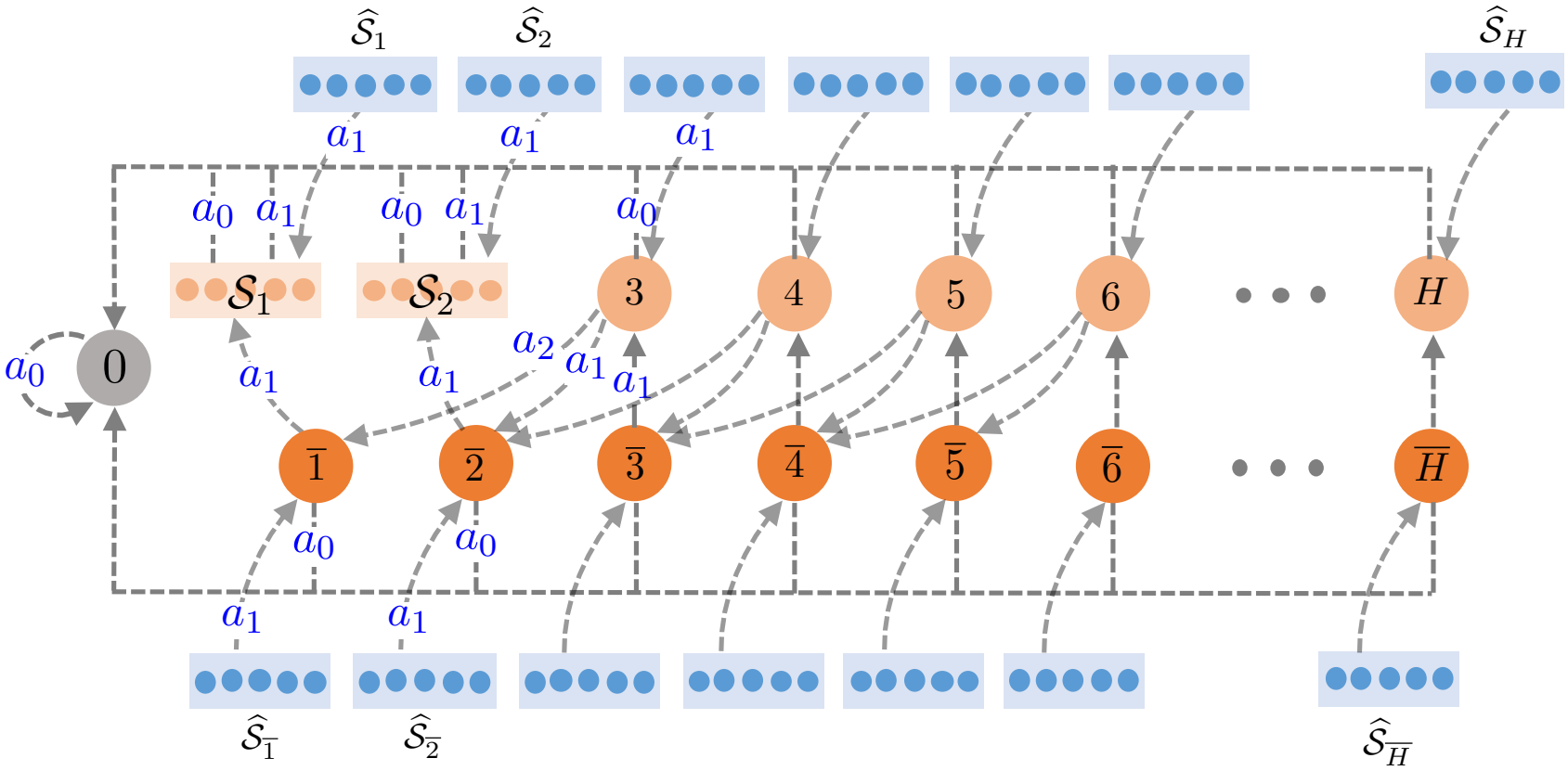
*There exists an MDP s.t. it takes softmax PG at least*

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

*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 $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\left[V^{(t)}(s) - V^\star(s)\right]$.

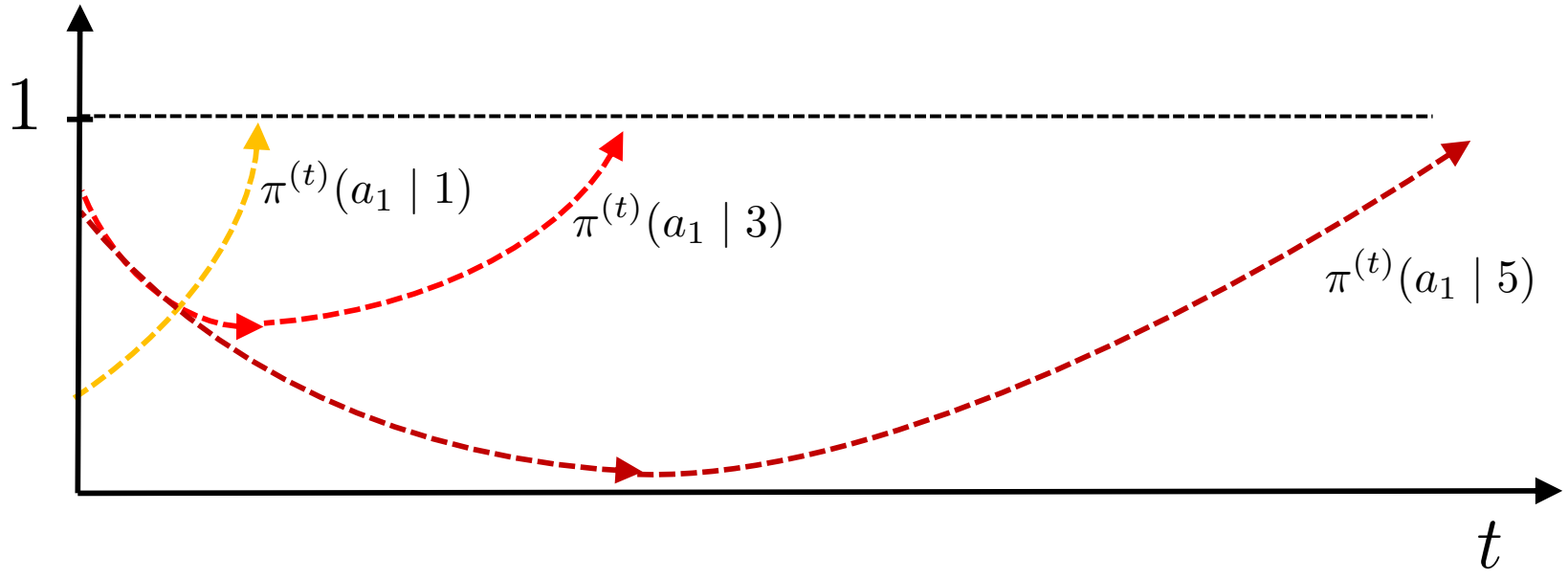
# MDP construction for our lower bound




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

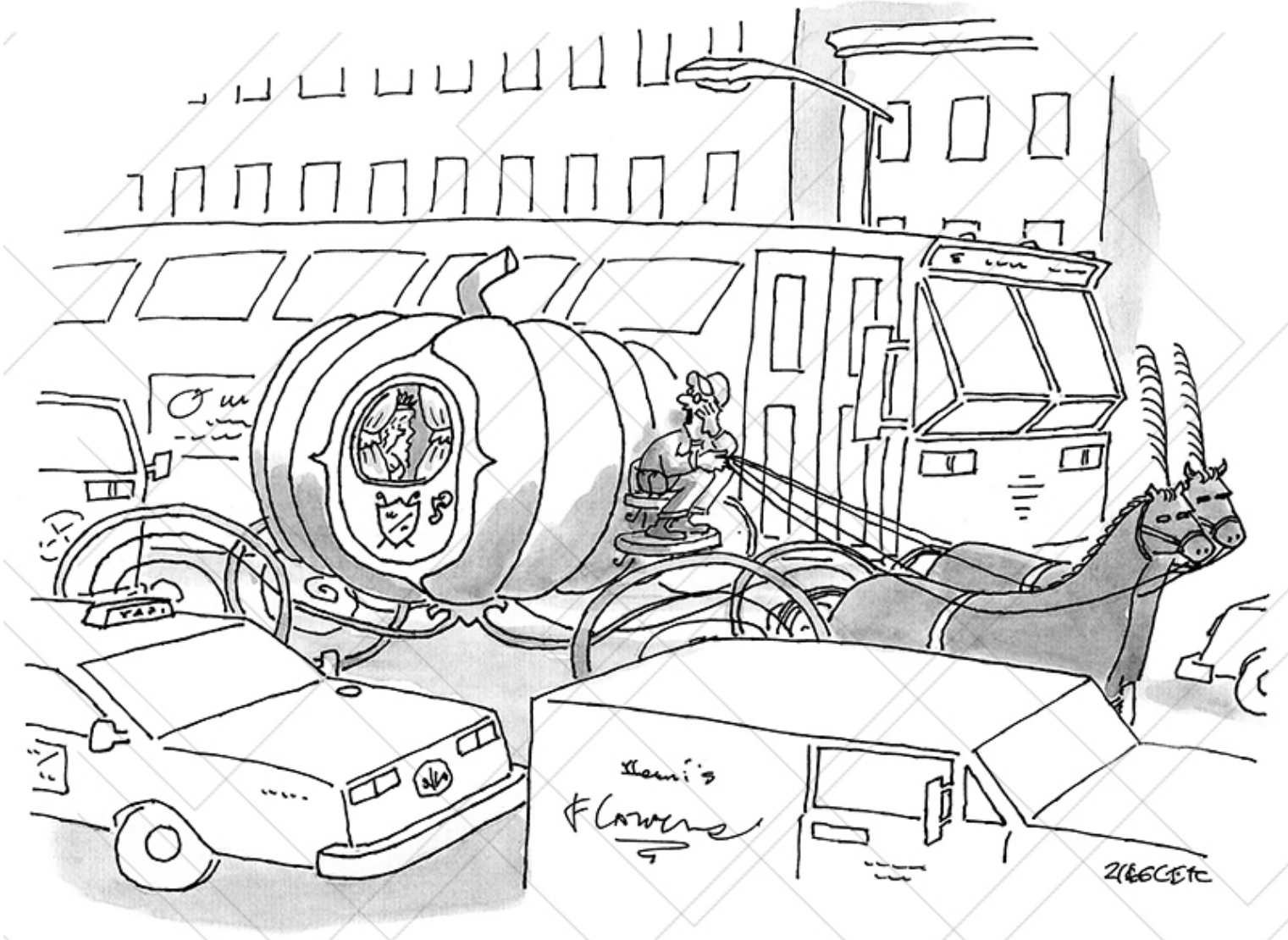
- $\pi^{(t)}(a_{\mathsf{opt}} \,|\, s)$ keeps decreasing until $\pi^{(t)}(a_{\mathsf{opt}} \,|\, s-2) \approx 1$

# What is happening in our constructed MDP?



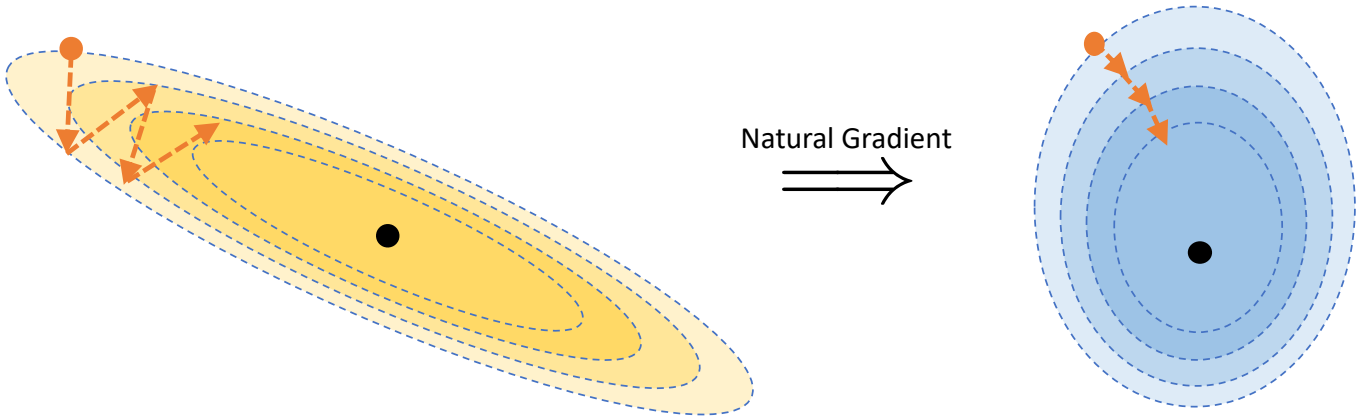

Convergence time for state $s$ grows geometrically as $s$ increases

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



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

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

# Connection with TRPO/PPO

TRPO/PPO (Schulman et al., 2015; 2017) are popular heuristics in training RL algorithms, with **KL regularization**

$$\mathrm{KL}(\pi_\theta^{(t)} \| \pi_\theta) \approx \frac{1}{2}(\theta - \theta^{(t)})^\top \mathcal{F}_\rho^\theta (\theta - \theta^{(t)})$$

via constrained or proximal terms:

$$\begin{aligned}\theta^{(t+1)} &= \operatorname*{argmax}_\theta V^{\pi_\theta^{(t)}}(\rho) + (\theta - \theta^{(t)})^\top \nabla_\theta V^{\pi_\theta^{(t)}}(\rho) - \eta \mathrm{KL}(\pi_\theta^{(t)} \| \pi_\theta) \\ &\approx \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho),\end{aligned}$$

leading to exactly NPG!

NPG $\approx$ TRPO/PPO!

# NPG in the tabular setting

**Natural policy gradient (NPG) method (Tabular setting)**

*For $t = 0, 1, \cdots$, NPG updates the policy via*

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

*where $Q^{(t)} := Q^{\pi^{(t)}}$ is the Q-function of $\pi^{(t)}$, and $\eta > 0$.*

- invariant with the choice of $\rho$
- Reduces to policy iteration (PI) when $\eta = \infty$.

# Global convergence of NPG

**Theorem (Agarwal et al., 2019)**

*Set $\pi^{(0)}$ as a uniform policy. For all $t \geq 0$, we have*

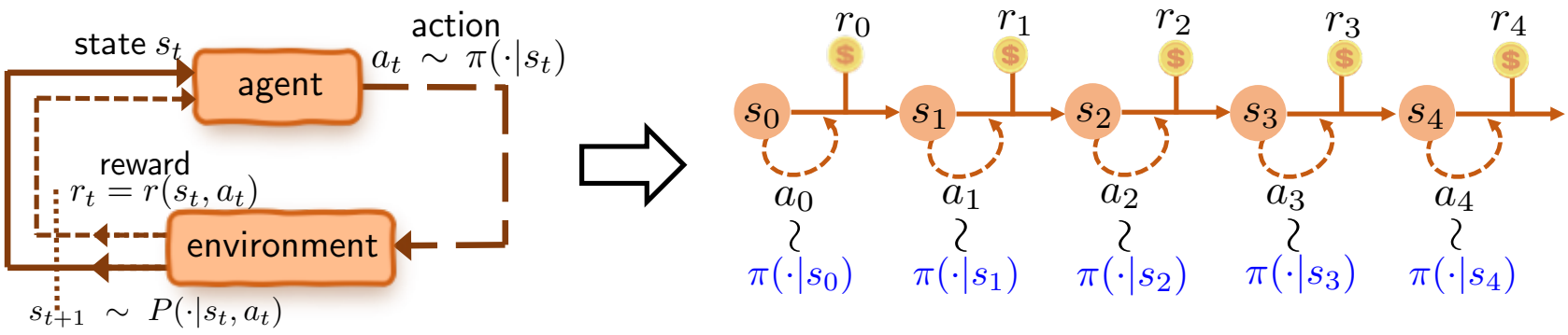
$$V^{(t)}(\rho) \geq V^{\star}(\rho) - \left(\frac{\log|\mathcal{A}|}{\eta} + \frac{1}{(1-\gamma)^2}\right)\frac{1}{t}.$$

**Implication:** set $\eta \geq (1-\gamma)^2 \log|\mathcal{A}|$, we find an $\epsilon$-optimal policy within at most

$$\frac{2}{(1-\gamma)^2\epsilon} \text{ iterations.}$$

Global convergence at a sublinear rate independent of $|\mathcal{S}|$, $|\mathcal{A}|$!

# Booster #2: entropy regularization




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

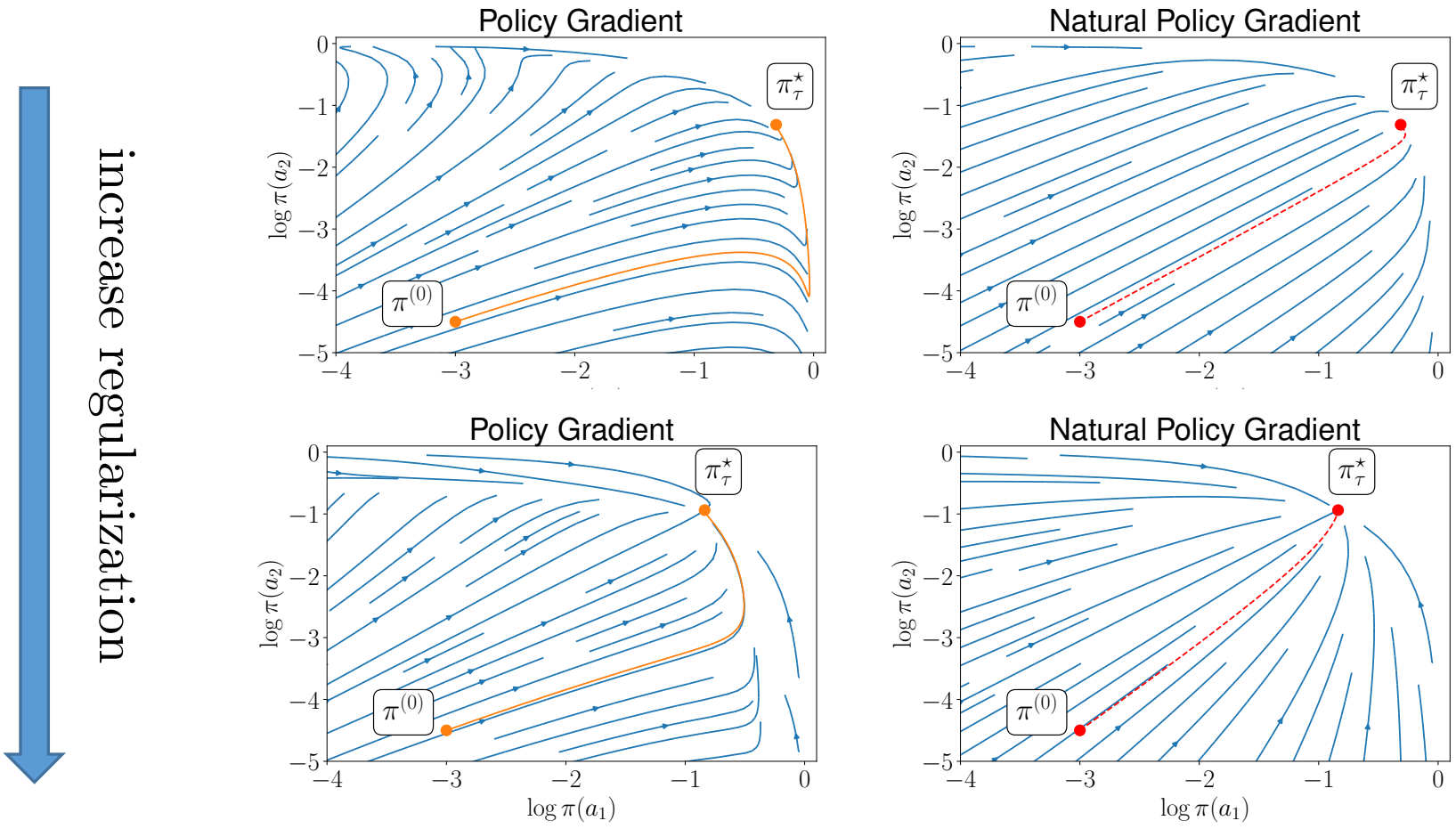
$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \big(r_t + \tau \mathcal{H}(\pi(\cdot|s_t)\big) \,\middle|\, s_0 = s\right]$$

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

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

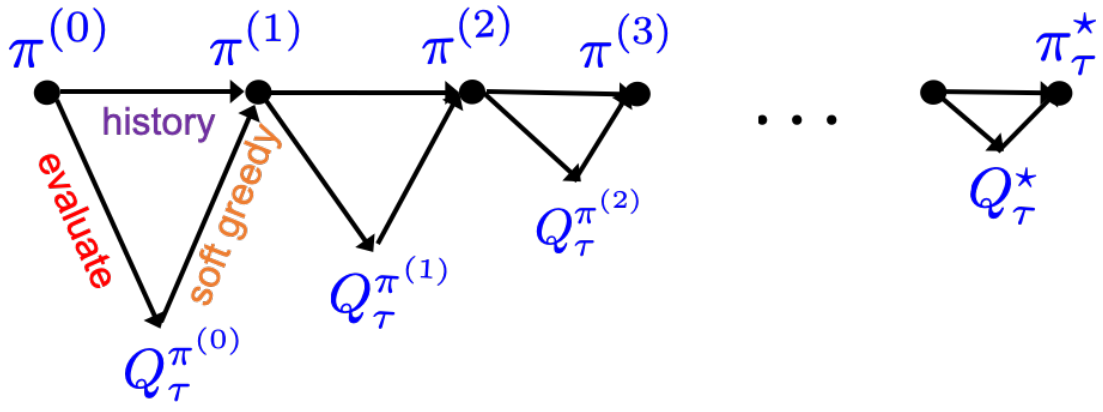
# Entropy-regularized natural gradient helps!

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




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 $\rho$
- 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 $\pi^{(t)}$;

*—Read the paper for the inexact case*

**Theorem (Cen, Cheng, Chen, Wei, Chi, 2020)**

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

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

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

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

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

# Implications

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

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

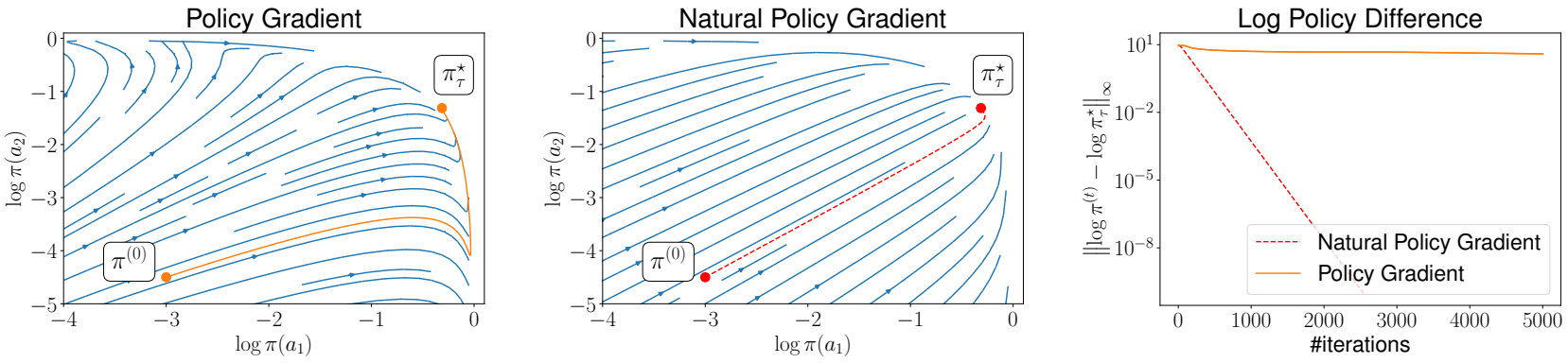
$$\frac{1}{\eta\tau} \log\left(\frac{C_1\gamma}{\epsilon}\right)$$

- **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 $|\mathcal{S}|$, $|\mathcal{A}|$!

# Comparisons with entropy-regularized PG




**(Mei et al., 2020)** showed entropy-regularized PG achieves

$$V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \le \left(V_\tau^\star(\rho) - V_\tau^{(0)}(\rho)\right)$$

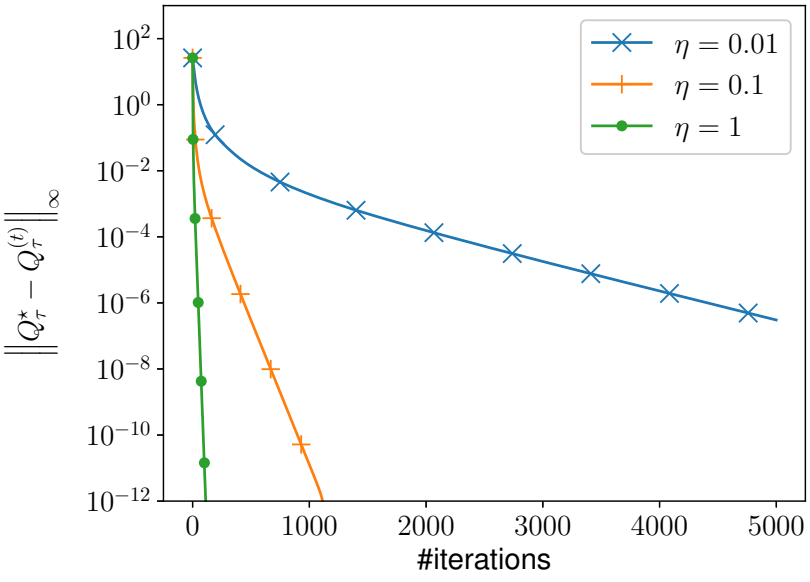
$$\cdot \exp\left( -\frac{(1-\gamma)^4 t}{(8/\tau + 4 + 8\log|\mathcal{A}|)|\mathcal{S}|} \left\| \frac{d_\rho^{\pi_\tau^\star}}{\rho} \right\|_\infty^{-1} \min_s \rho(s) \underbrace{\left( \inf_{0\le k\le t-1} \min_{s,a} \pi^{(k)}(a|s) \right)^2}_{\text{can be exponential in } |\mathcal{S}| \text{ and } \frac{1}{1-\gamma}} \right)$$

Much faster convergence of entropy-regularized NPG at a **dimension-free** rate!

# Comparison with unregularized NPG

**Regularized NPG**

$\tau = 0.001$

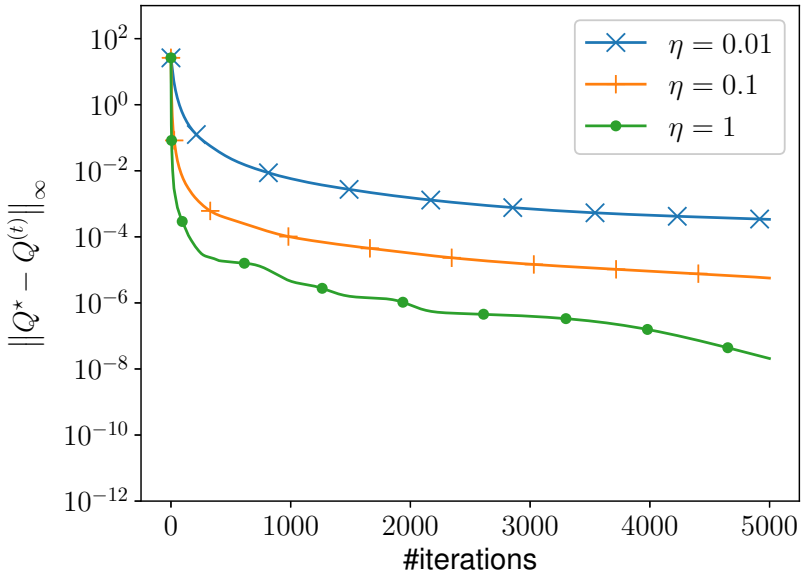



**Linear rate:** $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

**Ours**

**Vanilla NPG**

$\tau = 0$




**Sublinear rate:** $\frac{1}{\min\{\eta,(1-\gamma)^2\}\epsilon}$

**(Agarwal et al. 2019)**

Entropy regularization enables fast convergence!

# 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:** $Q_\tau^\star$ is *unique* solution to

$$\mathcal{T}_\tau(Q_\tau^\star) = Q_\tau^\star$$

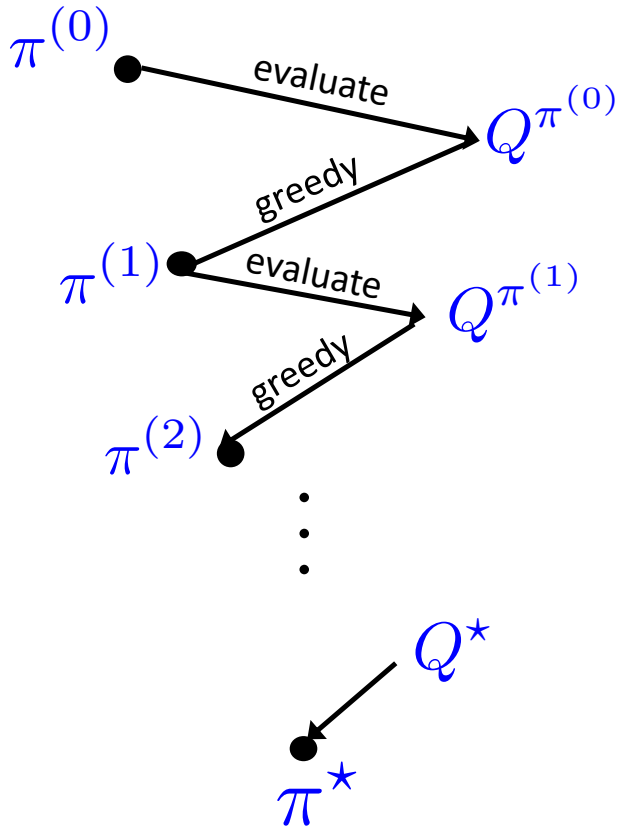
**$\gamma$-contraction of soft Bellman operator:**

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

*Richard Bellman*

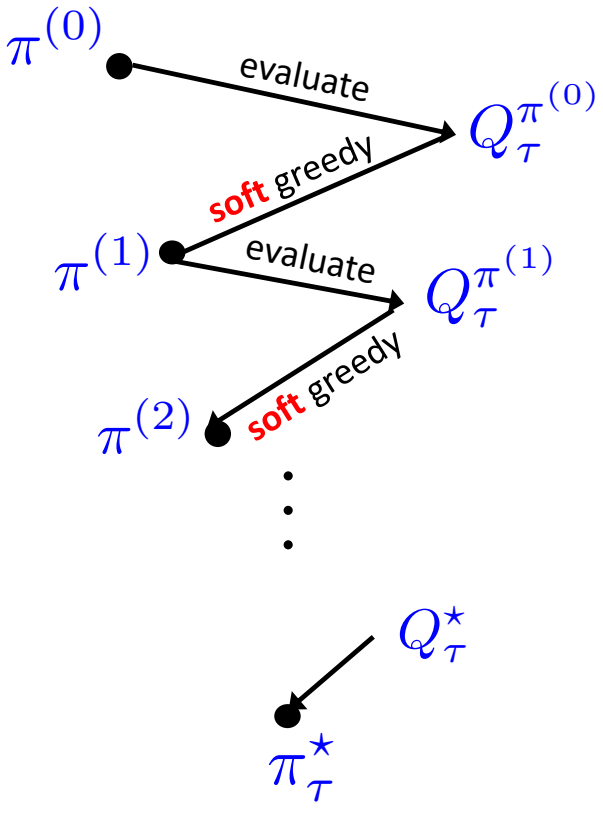
# Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

**Policy iteration**




Bellman operator

**Soft policy iteration**




Soft Bellman operator

# A key linear system: general learning rates

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

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 $\underbrace{1-\eta\tau}_{\text{contraction rate!}}$.

# Beyond entropy regularization

Leverage regularization to promote structural properties of the learned policy.

**cost-sensitive RL**

weighted 1-norm

**sparse exploration**

Tsallis entropy

**constrained and safe RL**

log-barrier

*For further details, see: (Lan, PMD 2021) and (Zhan et al, GPMD 2021)*

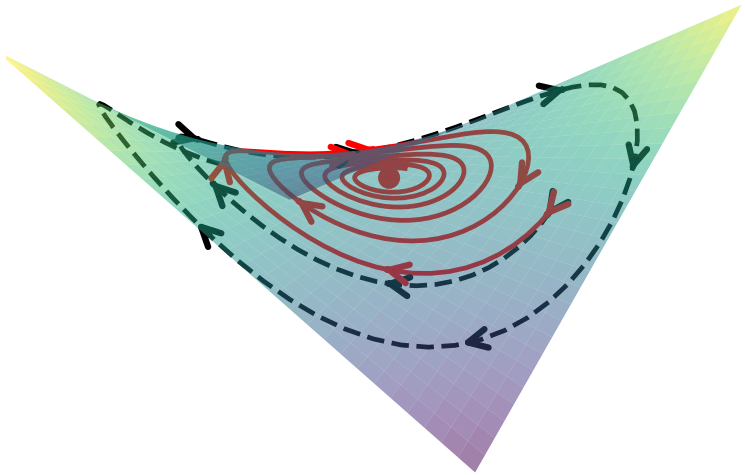
*Policy optimization for games*

# Policy optimization: saddle-point optimization

**Zero-sum two-player Markov game**

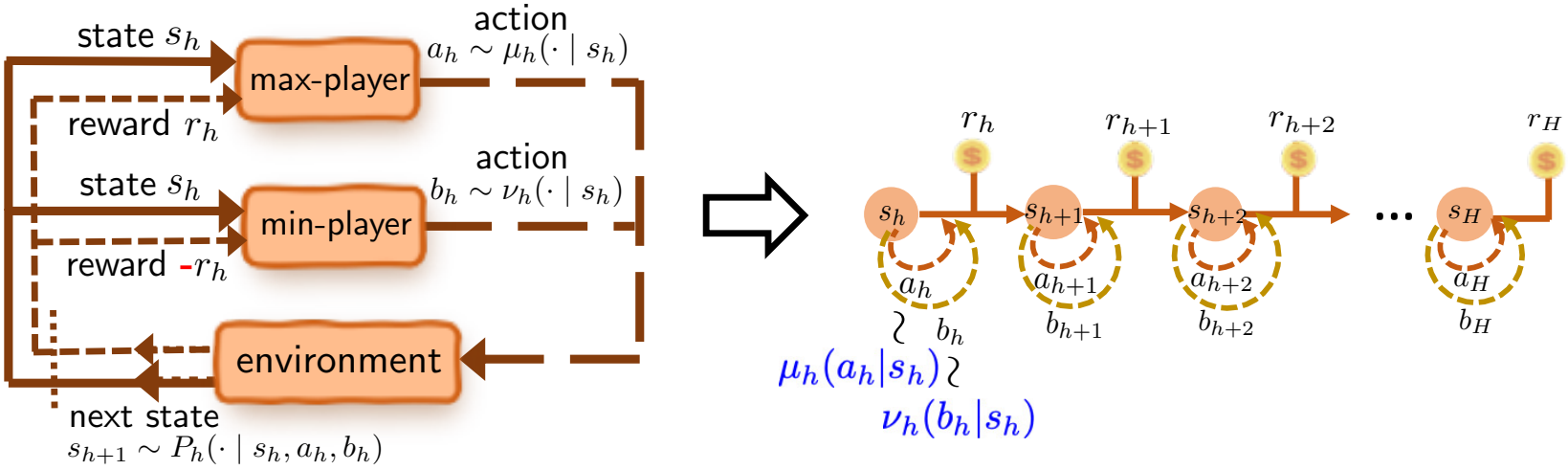
*Given an initial state distribution $s \sim \rho$, find policy $\pi$ such that*

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



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):

$$V_\tau^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{h=1}^{H}\left(r_h + \tau\mathcal{H}(\mu_h(\cdot|s_h)) - \tau\mathcal{H}(\nu_h(\cdot|s_h))\right)\Big|\, s_0 = s\right],$$

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

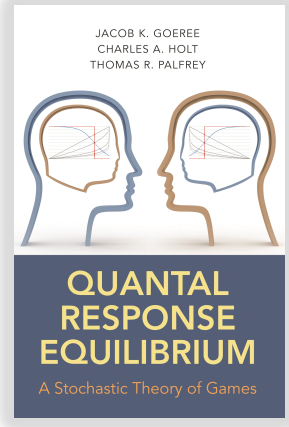
$$\max_{\mu\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu\in\Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho)$$

# 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.

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

# Soft value iteration

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

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

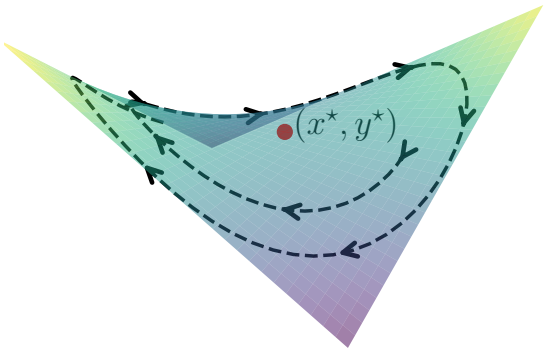
where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

**Entropy-regularized matrix game**

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

# Failure of NPG/MWU methods

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




- Multiplicative Weights Update (**MWU**):

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

- $\eta > 0$: step size;
- The trajectory may cycle/diverge!

# Motivation: an implicit update method

**Implicit update (IU) method**

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

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

**Theorem (Cen, Wei, Chi, 2021)**

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

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

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

Can we make this practical?

# From implicit updates to policy extragradient methods

**Optimistic multiplicative weights update (OMWU) method**

**(Related to OMD, Rakhlin and Sridharan, 2013):** for $t = 0, 1, \cdots$,

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

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

**Theorem (Cen, Wei, Chi, 2021)**

*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* $\epsilon$*-QRE within* $\widetilde{O}\left(\frac{1}{\eta\tau}\log\frac{1}{\epsilon}\right)$ *iterations.*

*Linear, last-iterate convergence to the QRE!*

# Soft value iteration via nested-loop OMWU

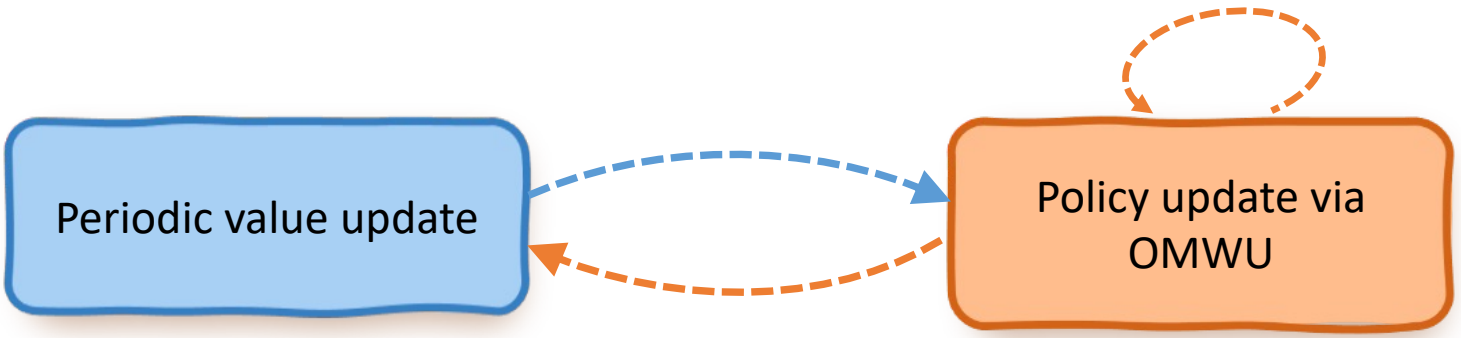
**Soft value iteration:** for $h = H, \dots, 1$

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

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...*

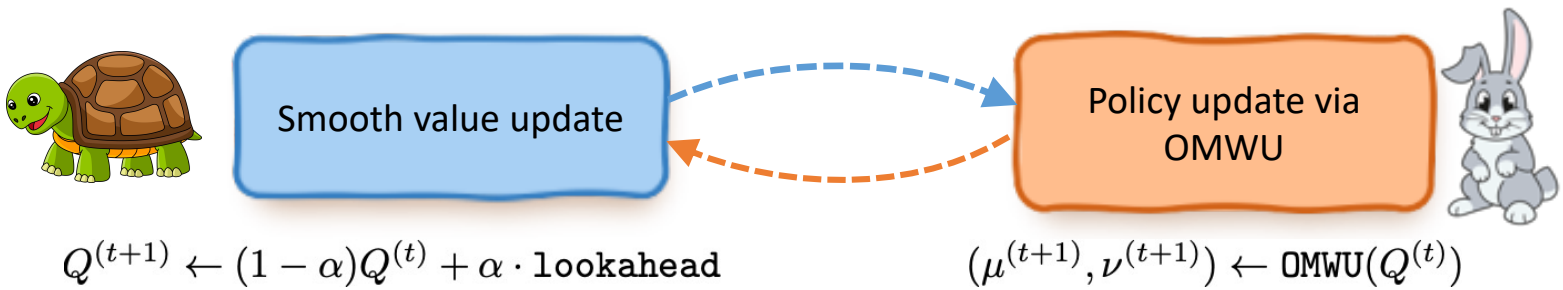
# A two-timescale single-loop approach?

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

**Single-loop, two-timescale approach:**



# Main result: episodic setting

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

*The last-iterate of the two-timescale single-loop algorithm finds an $\epsilon$-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 $\epsilon$-NE.*

- First last-iterate convergence result for the episodic setting.
- **Almost dimension-free:** independent of the size of the state-action space.

# Main result: discounted setting

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

*For the infinite-horizon $\gamma$-discounted setting, the last-iterate of the single-loop algorithm finds an $\epsilon$-QRE in*

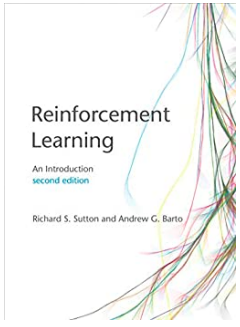
$$\widetilde{O}\left(\frac{S}{(1-\gamma)^4\tau}\log\frac{1}{\epsilon}\right)$$

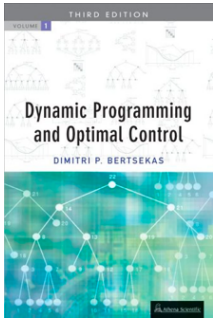
*iterations, and in $\widetilde{O}\left(\frac{S}{(1-\gamma)^5\epsilon}\right)$ iterations for finding an $\epsilon$-NE.*

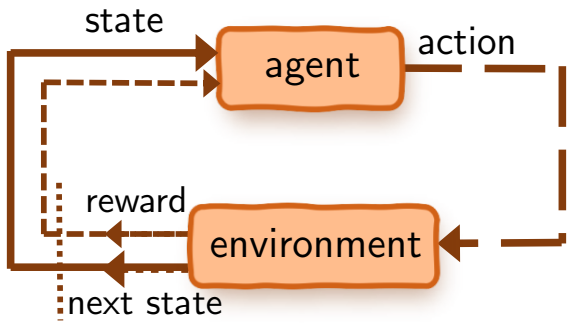
- This significantly improves upon the prior art $\widetilde{O}\left(\frac{S^5(A+B)^{1/2}}{(1-\gamma)^{16}c^4\epsilon^2}\right)$ of (Wei et al., 2021) and $\widetilde{O}\left(\frac{S^2\|1/\rho\|^5}{(1-\gamma)^{14}c^4\epsilon^3}\right)$ of (Zeng et al., 2022) in *all* parameter dependencies.

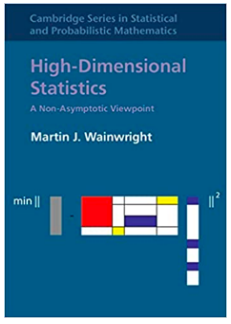
*Concluding Remarks*

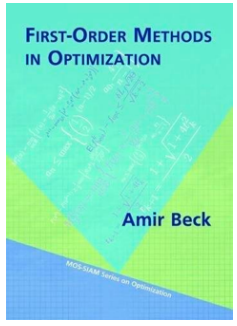
# Concluding remarks




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

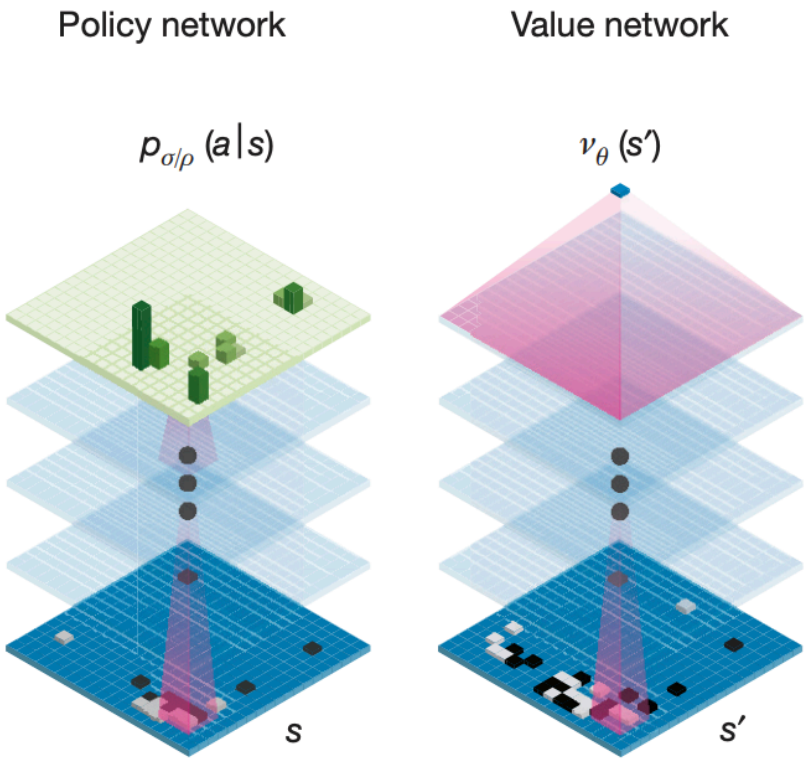
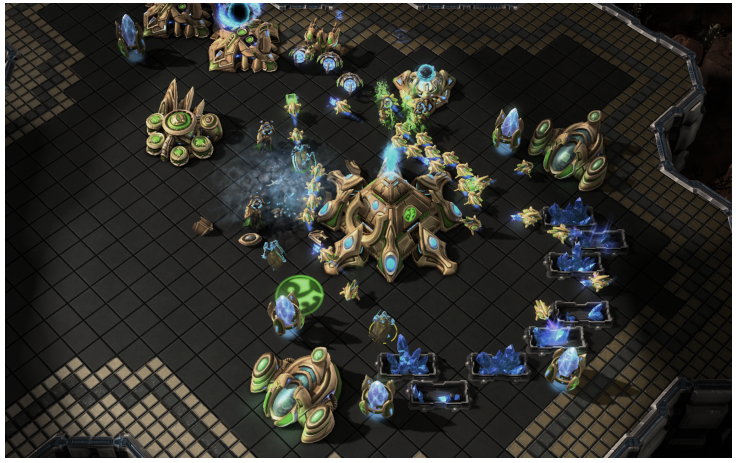



Figure credit: (Silver et al., 2016)

- 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)

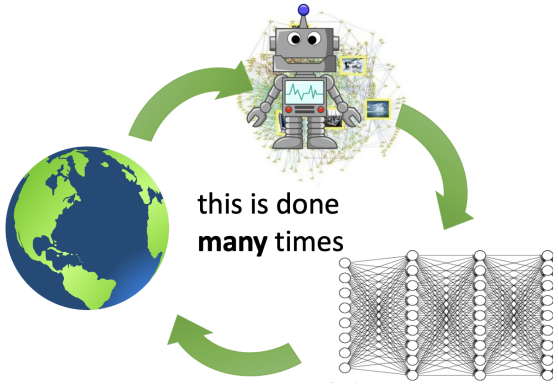
# Multi-agent RL



- **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

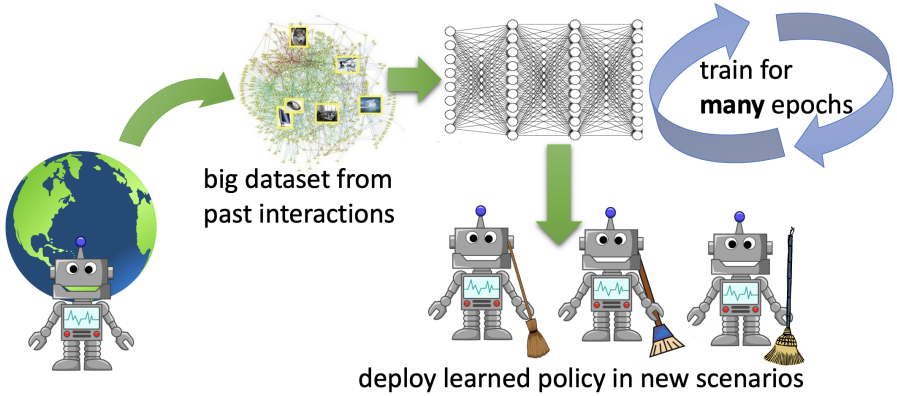



*Online RL*

- interact with environment
- actively collect new data

*Offline/Batch RL*

- no interaction
- data is given

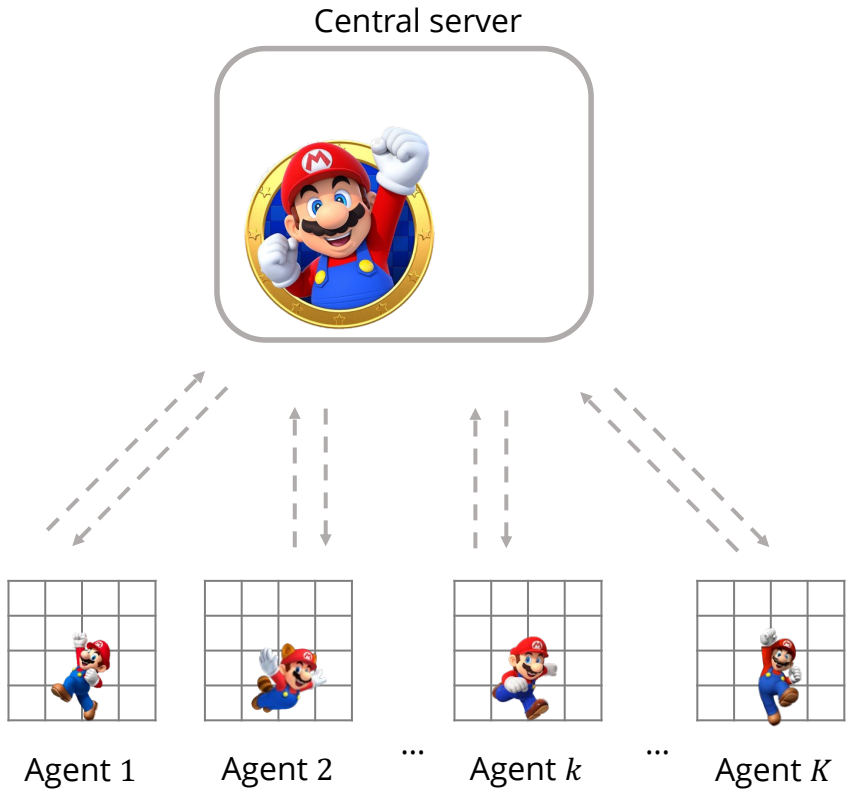



**Can we achieve the best of both worlds?**

(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)

# Bibliography I

**Disclaimer:** this straw-man list is by no means exhaustive (in fact, it is quite the opposite given the fast pace of the field), and biased towards materials most related to this tutorial; readers are invited to further delve into the references therein to gain a more complete picture.

## Books and monographs:


- Sutton and Barto. *Reinforcement learning: An introduction, 2nd edition*. MIT press, 2018.
- Agarwal, Jiang, Kakade, and Sun. *Reinforcement learning: Theory and algorithms*, monograph, 2021+.
- Bertsekas. *Reinforcement learning and optimal control*. Athena Scientific, 2019.
- Szepesvári. *Algorithms for reinforcement learning*. Synthesis lectures on artificial intelligence and machine learning, 2010.
- Bertsekas and Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

# Bibliography II


**Policy optimization:**

- Williams. "*Simple statistical gradient-following algorithms for connectionist reinforcement learning.*" Machine Learning, 1992.
- Sutton, McAllester, Singh, and Mansour. "*Policy gradient methods for reinforcement learning with function approximation.*" NeurIPS 1999.
- Kakade. "*A natural policy gradient.*" NeurIPS 2001.
- Fazel, Ge, Kakade, and Mesbahi. "*Global convergence of policy gradient methods for the linear quadratic regulator.*" ICML 2018.
- Agarwal, Kakade, Lee, and Mahajan. "*On the theory of policy gradient methods: Optimality, approximation, and distribution shift.*" Journal of Machine Learning Research, 2021.
- Mei, Xiao, Szepesvári, and Schuurmans. "*On the global convergence rates of softmax policy gradient methods.*" ICML 2020.
- Bhandari and Russo. "*Global optimality guarantees for policy gradient methods.*" arXiv preprint arXiv:1906.01786, 2019.

# Bibliography III


- Cai, Yang, Jin, and Wang. "*Provably efficient exploration in policy optimization.*" ICML 2020.
- Shani, Efroni, and Mannor."*Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs.*" AAAI 2020.
- Li, Gen, Wei, Chi, and Chen. "*Softmax policy gradient methods can take exponential time to converge.*" arXiv preprint arXiv:2102.11270, 2021.
- Cen, Cheng, Chen, Wei, and Chi. "*Fast global convergence of natural policy gradient methods with entropy regularization.*" Operations Research, 2021+.
- Zhan, Cen, Huang, Chen, Lee, and Chi."*Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence.*" arXiv preprint arXiv:2105.11066, 2021.
- Lan. "*Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes.*" arXiv preprint arXiv:2102.00135, 2021.
- Liu, Zhang, Basar, and Yin. "*An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods.*" NeurIPS 2020.

# Bibliography IV


- Zhang, Koppel, Bedi, Szepesvári, and Wang. "*Variational policy gradient method for reinforcement learning with general utilities.*" NeurIPS 2020.
- Cen, Wei, and Chi. "*Fast policy extragradient methods for competitive games with entropy regularization.*" arXiv preprint arXiv:2105.15186, 2021.
- Cen, Chi, Du, and Xiao, "*Faster last-iterate convergence of policy optimization in zero-sum Markov games.*" arXiv preprint arXiv:2210.01050, 2022.


**Additional ad-hoc pointers:**


- Neu, Jonsson, and Gómez. "*A unified view of entropy-regularized Markov Decision Processes.*" arXiv preprint arXiv:1705.07798, 2017.
- Dai, Shaw, Li, Xiao, He, Liu, Chen, and Song."*SBEED: Convergent reinforcement learning with nonlinear function approximation.*" ICML 2018.
- Geist, Scherrer, and Pietquin. "*A theory of regularized Markov Decision Processes.*" ICML 2019.

# Bibliography V


- Du, Kakade, Wang, and Yang. "*Is a good representation sufficient for sample efficient reinforcement learning?*" ICLR 2019.
- Jin, Yang, Wang, and Jordan. "*Provably efficient reinforcement learning with linear function approximation.*" COLT 2020.
- Zhang, Yang, and Basar. "*Multi-agent reinforcement learning: A selective overview of theories and algorithms.*" Handbook of Reinforcement Learning and Control, 2021.
- Woo, Joshi, and Chi. "*The Blessing of Heterogeneity in Federated Q-learning: Linear Speedup and Beyond.*" ICML 2023.

# Thanks!





https://users.ece.cmu.edu/~yuejiec/