Provably Adaptive Average Reward Reinforcement Learning for Metric Spaces

Avik Kar
ECE

Rahul Singh
ECE

3rd April 2025

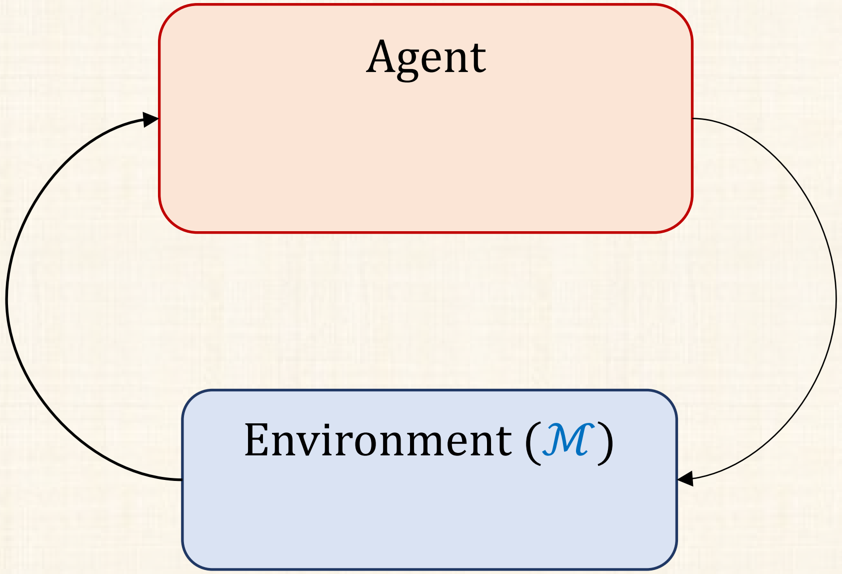
IISc, Bengaluru

This talk contains

- Average Reward Reinforcement Learning
- Lipschitz Continuity of Average Reward
- Algorithm: Zooming in Policy Space
- Algorithm: Zooming in State-Action Space

Reinforcement Learning Setup

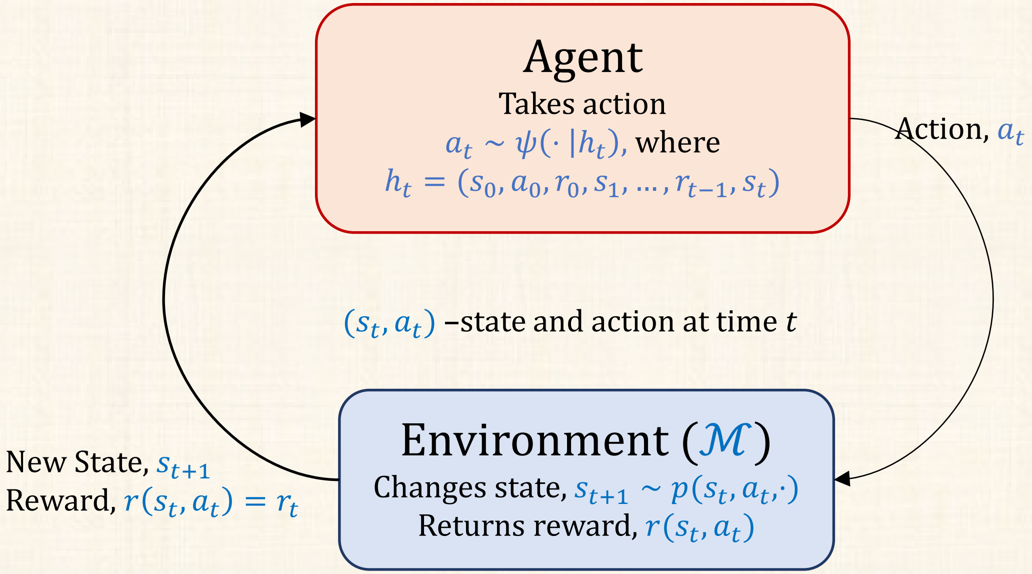
Reinforcement Learning Setup



<u>Model of the Environment:</u> $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$

- $\mathcal{S}$ – state space, $\mathcal{A}$ – action space
- p – transition kernel;
$$p(s, a, B) = \mathbb{P}(s_{t+1} \in B \mid s_t = s, a_t = a)$$
- $r: \mathcal{S} \times \mathcal{A} \to [0,1]$ – reward function

Reinforcement Learning Setup



<u>Model of the Environment:</u> $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$

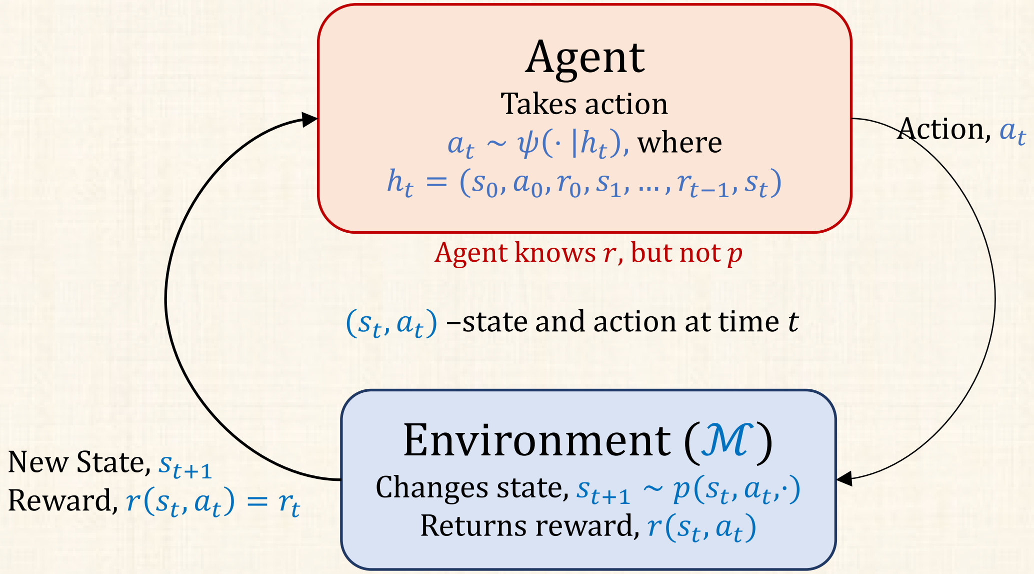
- $\mathcal{S}$ – state space, $\mathcal{A}$ – action space
- p – transition kernel;

$$p(s, a, B) = \mathbb{P}(s_{t+1} \in B \mid s_t = s, a_t = a)$$

- $r: \mathcal{S} \times \mathcal{A} \to [0,1]$ – reward function

⋆ ψ – Learning algorithm

Reinforcement Learning Setup



<u>Model of the Environment:</u> $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$

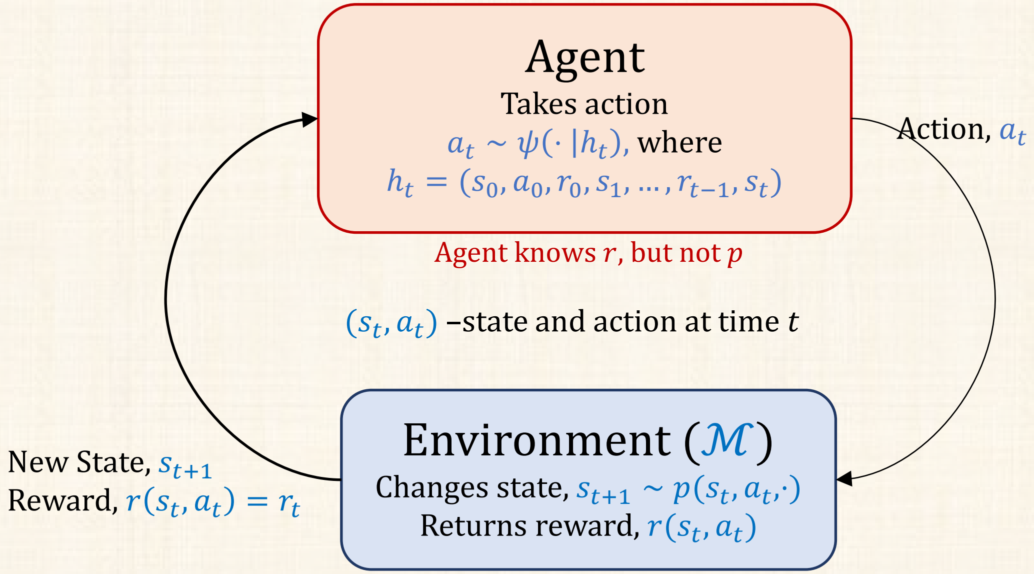
- $\mathcal{S}$ – state space, $\mathcal{A}$ – action space
- p – transition kernel;

$$p(s, a, B) = \mathbb{P}(s_{t+1} \in B \mid s_t = s, a_t = a)$$

- $r: \mathcal{S} \times \mathcal{A} \to [0,1]$ – reward function

⋆ ψ – Learning algorithm

Reinforcement Learning Setup



<u>Model of the Environment:</u> $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$

- $\mathcal{S}$ – state space, $\mathcal{A}$ – action space
- p – transition kernel;

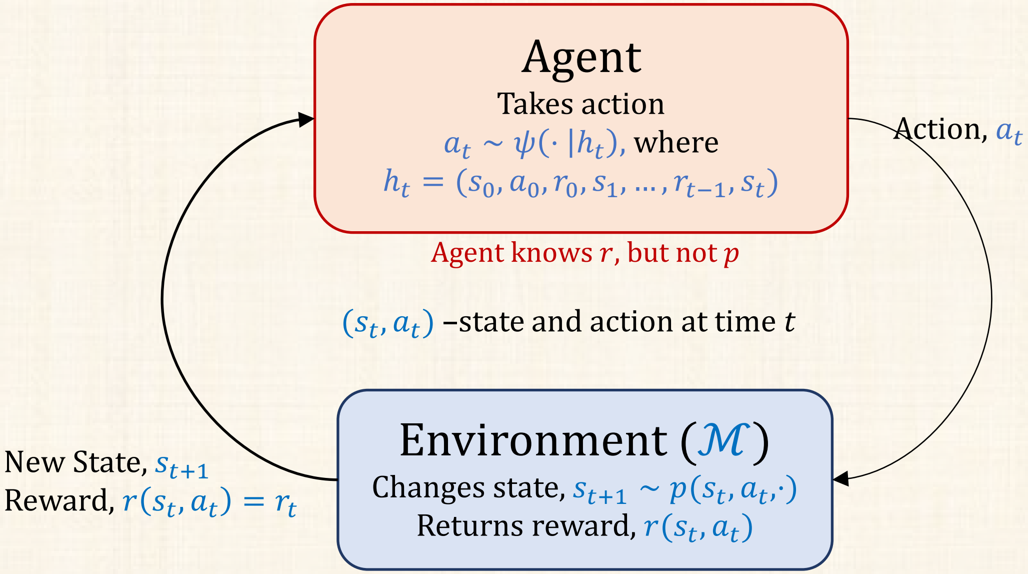
$$p(s, a, B) = \mathbb{P}(s_{t+1} \in B \mid s_t = s, a_t = a)$$

- $r: \mathcal{S} \times \mathcal{A} \to [0,1]$ – reward function

⋆ ψ – Learning algorithm

- Stationary deterministic policy: a function $\phi: \mathcal{S} \to \mathcal{A}$

Reinforcement Learning Setup



<u>Model of the Environment:</u> $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$

- $\mathcal{S}$ – state space, $\mathcal{A}$ – action space
- p – transition kernel;

$$p(s, a, B) = \mathbb{P}(s_{t+1} \in B \mid s_t = s, a_t = a)$$

- $r: \mathcal{S} \times \mathcal{A} \to [0,1]$ – reward function

- Stationary deterministic policy: a function $\phi: \mathcal{S} \to \mathcal{A}$

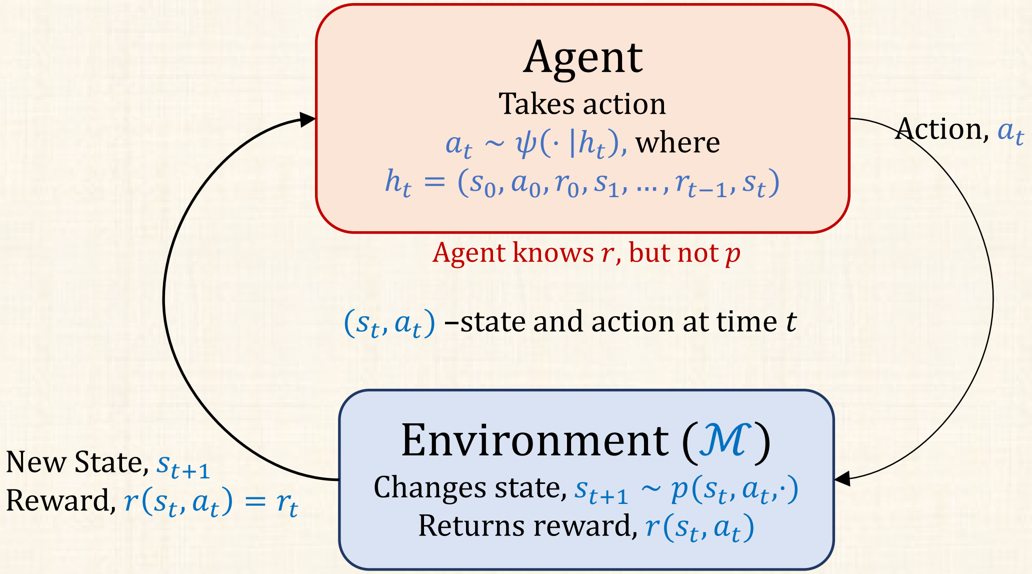
<u>Average reward criterion:</u>

- Let

$$J_{\mathcal{M}}(\phi) \coloneqq \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}_{\phi}\left[\sum_{t=1}^{T} r_t\right]$$

⋆ ψ – Learning algorithm
⋆ $\mathbb{E}_{\phi}$ – Expectation taken considering policy ϕ is played

Reinforcement Learning Setup



Model of the Environment: $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$

- $\mathcal{S}$ – state space, $\mathcal{A}$ – action space
- p – transition kernel;

$$p(s, a, B) = \mathbb{P}(s_{t+1} \in B \mid s_t = s, a_t = a)$$

- $r: \mathcal{S} \times \mathcal{A} \to [0,1]$ – reward function

- Stationary deterministic policy: a function $\phi: \mathcal{S} \to \mathcal{A}$

Average reward criterion:

- Let

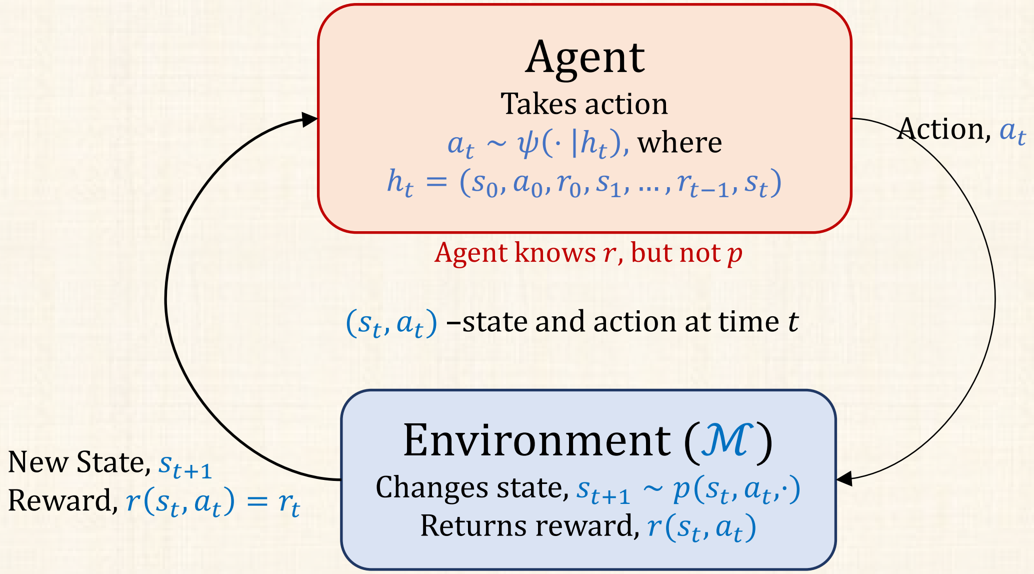
$$J_{\mathcal{M}}(\phi) \coloneqq \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}_\phi \left[\sum_{t=1}^{T} r_t \right]$$

- Optimal average reward

$$J^\star_{\mathcal{M}} = \max_\phi J_{\mathcal{M}}(\phi)$$

⋆ ψ – Learning algorithm
⋆ $\mathbb{E}_\phi$ – Expectation taken considering policy ϕ is played

Reinforcement Learning Setup



<u>Model of the Environment:</u> $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$

- $\mathcal{S}$ – state space, $\mathcal{A}$ – action space
- p – transition kernel;

$$p(s, a, B) = \mathbb{P}(s_{t+1} \in B \mid s_t = s, a_t = a)$$

- $r: \mathcal{S} \times \mathcal{A} \to [0,1]$ – reward function

- Stationary deterministic policy: a function $\phi: \mathcal{S} \to \mathcal{A}$

<u>Average reward criterion:</u>

- Let

$$J_{\mathcal{M}}(\phi) \coloneqq \liminf_{T \to \infty} \frac{1}{T} \mathbb{E}_{\phi}\left[\sum_{t=1}^{T} r_t\right]$$

- Optimal average reward

$$J^{\star}_{\mathcal{M}} = \max_{\phi} J_{\mathcal{M}}(\phi)$$

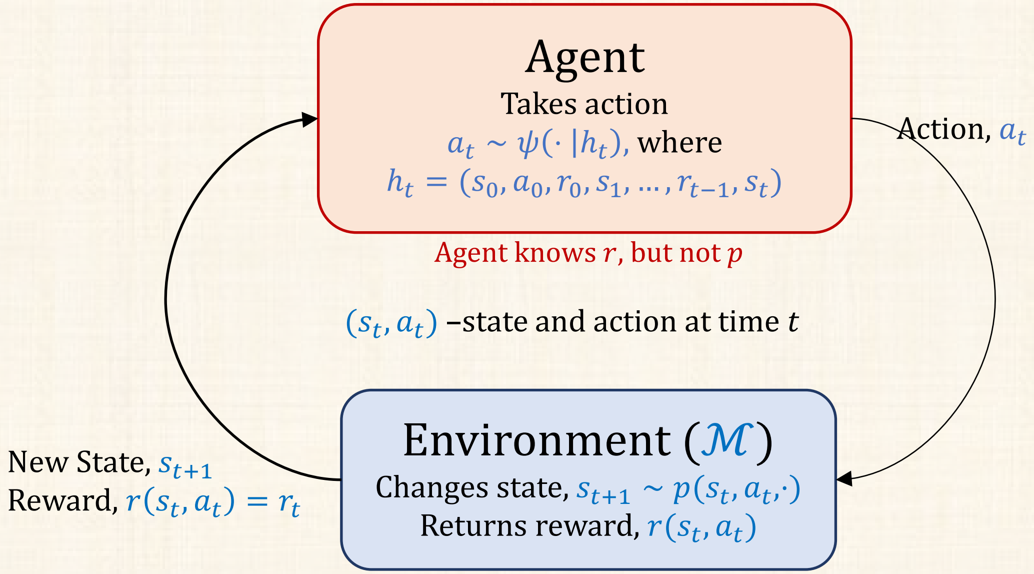
<u>Regret:</u>

- Cumulative regret of algorithm, ψ is defined as

$$R(T; \psi) \coloneqq T J^{\star}_{\mathcal{M}} - \sum_{t=1}^{T} r_t$$

⋆ ψ – Learning algorithm
⋆ $\mathbb{E}_{\phi}$ – Expectation taken considering policy ϕ is played

Reinforcement Learning Setup



<u>Model of the Environment:</u> $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$

- $\mathcal{S}$ – state space, $\mathcal{A}$ – action space
- p – transition kernel;

$$p(s, a, B) = \mathbb{P}(s_{t+1} \in B \mid s_t = s, a_t = a)$$

- $r: \mathcal{S} \times \mathcal{A} \to [0,1]$ – reward function

- Stationary deterministic policy: a function $\phi: \mathcal{S} \to \mathcal{A}$

<u>Average reward criterion:</u>

- Let

$$J_{\mathcal{M}}(\phi) \coloneqq \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}_\phi\left[\sum_{t=1}^{T} r_t\right]$$

- Optimal average reward

$$J^\star_{\mathcal{M}} = \max_{\phi} J_{\mathcal{M}}(\phi)$$

<u>Regret:</u>

- Cumulative regret of algorithm, ψ is defined as

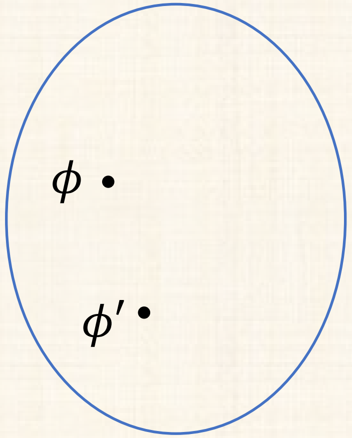
$$R(T; \psi) \coloneqq T J^\star_{\mathcal{M}} - \sum_{t=1}^{T} r_t$$

<u>Goal</u>: To propose ψ with low regret upper bound

⋆ ψ – Learning algorithm
⋆ $\mathbb{E}_\phi$ – Expectation taken considering policy ϕ is played

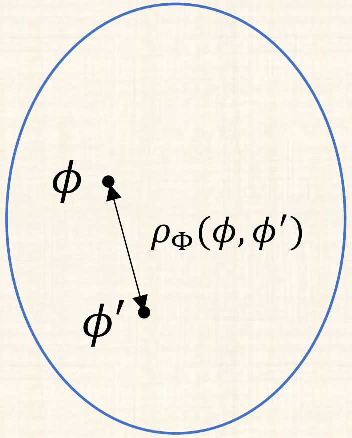
Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$

Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$



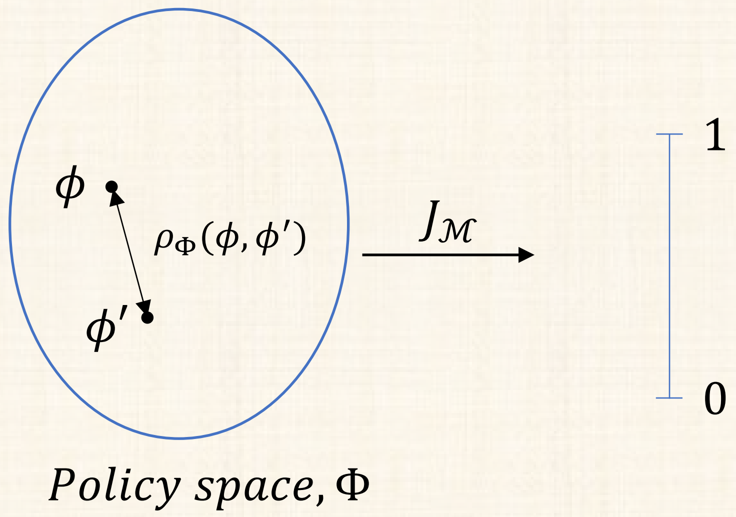
$Policy\ space, \Phi$

Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$

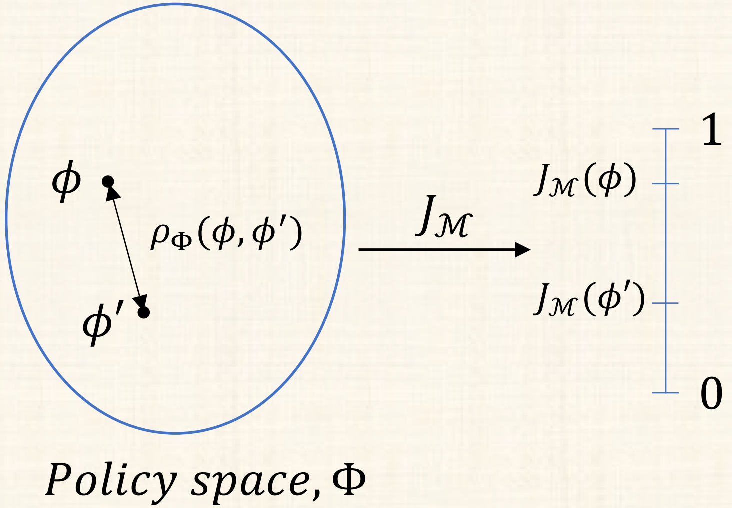


$Policy\ space, \Phi$

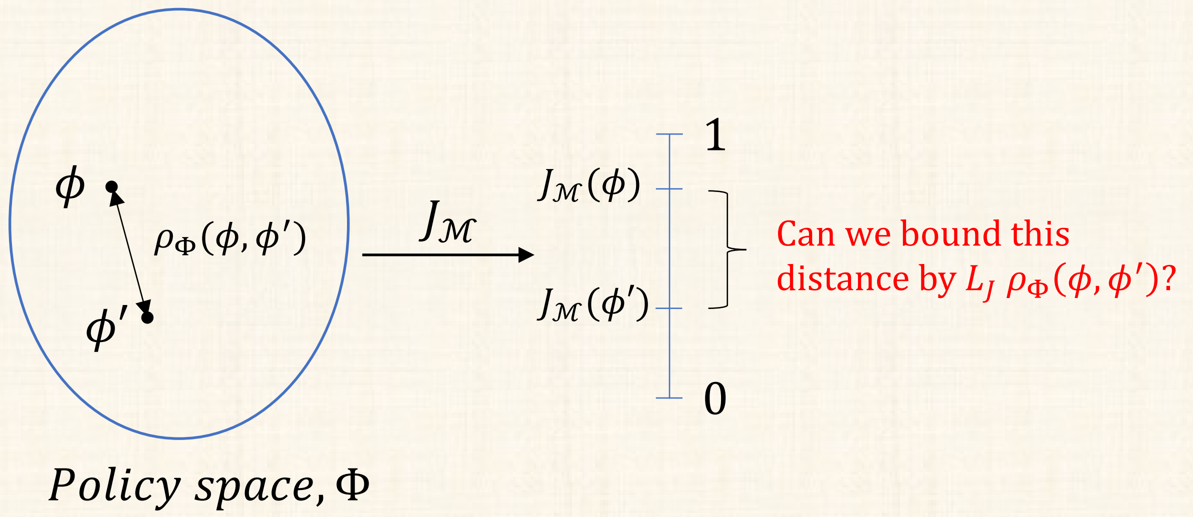
Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$



Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$

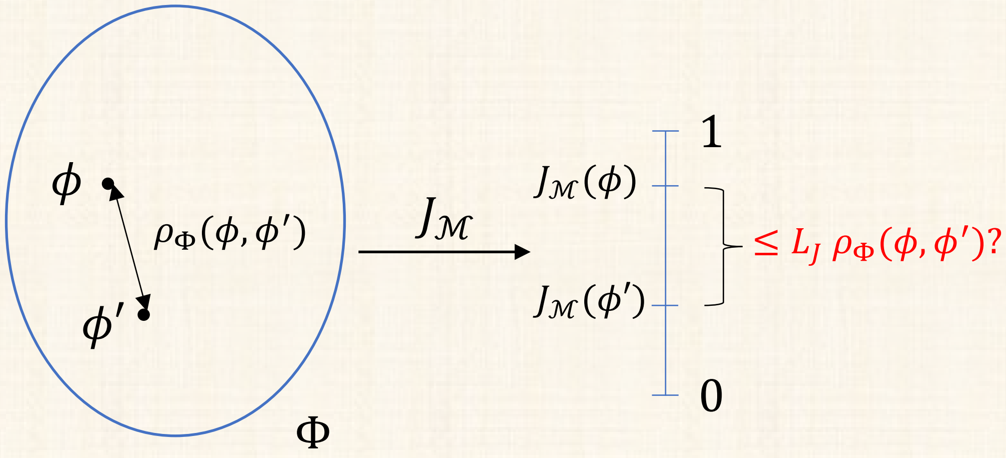


Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$



Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$

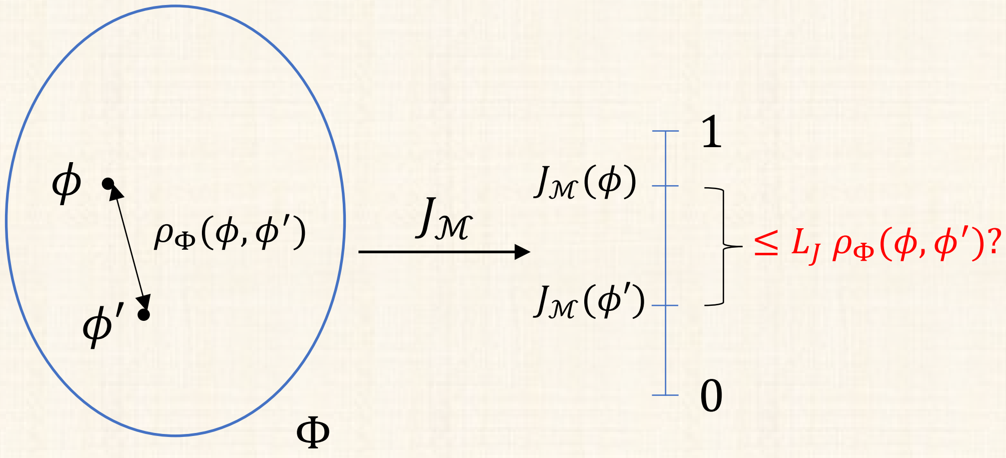
Assumptions



Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$

Assumptions

1. Lipschitz continuity: For every $(s,a),(s',a')$
$$|r(s,a)-r(s',a')| \le L_r \rho\big((s,a),(s',a')\big),$$
$$\|p(s,a,\cdot)-p(s',a',\cdot)\|_{TV} \le L_p \rho\big((s,a),(s',a')\big).$$
2. Uniform ergodicity: $\exists \alpha \in [0,1)$ and $C < \infty$ such that for every $(s,a),(s',a')$
$$\left\|\mu_{\phi,p,s_0}^{(t)} - \mu_{\phi,p}^{(\infty)}\right\|_{TV} \le C \cdot \alpha^t, \forall t \in \mathbb{N}.$$
3. Upper bound on stationary measure: $\exists \bar{\kappa} > 0$ and a probability measure ν such that for every ϕ, $\mu_{\phi,p}^{(\infty)} \le \bar{\kappa} \cdot \nu$.



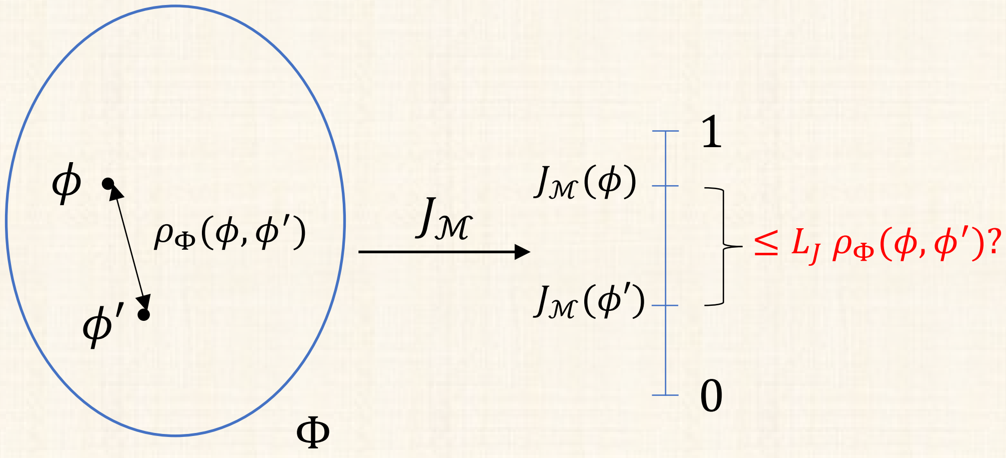
⋆ $\mu_{\phi,p,s'}^{(t)}$ - distribution of s_t when ϕ is played and initial state is s'

⋆ $\mu_{\phi,p}^{(\infty)}$ - stationary distribution of $\{s_t\}$ under application of ϕ

Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$

Assumptions

1. Lipschitz continuity: For every $(s,a),(s',a')$
$$|r(s,a) - r(s',a')| \le L_r \rho\big((s,a),(s',a')\big),$$
$$\|p(s,a,\cdot) - p(s',a',\cdot)\|_{TV} \le L_p \rho\big((s,a),(s',a')\big).$$
2. Uniform ergodicity: $\exists \alpha \in [0,1)$ and $C < \infty$ such that for every $(s,a),(s',a')$
$$\left\| \mu^{(t)}_{\phi,p,s_0} - \mu^{(\infty)}_{\phi,p} \right\|_{TV} \le C \cdot \alpha^t, \forall t \in \mathbb{N}.$$
3. Upper bound on stationary measure: $\exists \bar{\kappa} > 0$ and a probability measure ν such that for every ϕ, $\mu^{(\infty)}_{\phi,p} \le \bar{\kappa} \cdot \nu$.



Metric , ρ_Φ

⋆ $\mu^{(t)}_{\phi,p,s'}$ - distribution of s_t when ϕ is played and initial state is s'

⋆ $\mu^{(\infty)}_{\phi,p}$ - stationary distribution of $\{s_t\}$ under application of ϕ

Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$

Assumptions

1. Lipschitz continuity: For every $(s,a),(s',a')$

$$|r(s,a)-r(s',a')| \le L_r\rho\big((s,a),(s',a')\big),$$
$$\|p(s,a,\cdot)-p(s',a',\cdot)\|_{TV} \le L_p\rho\big((s,a),(s',a')\big).$$

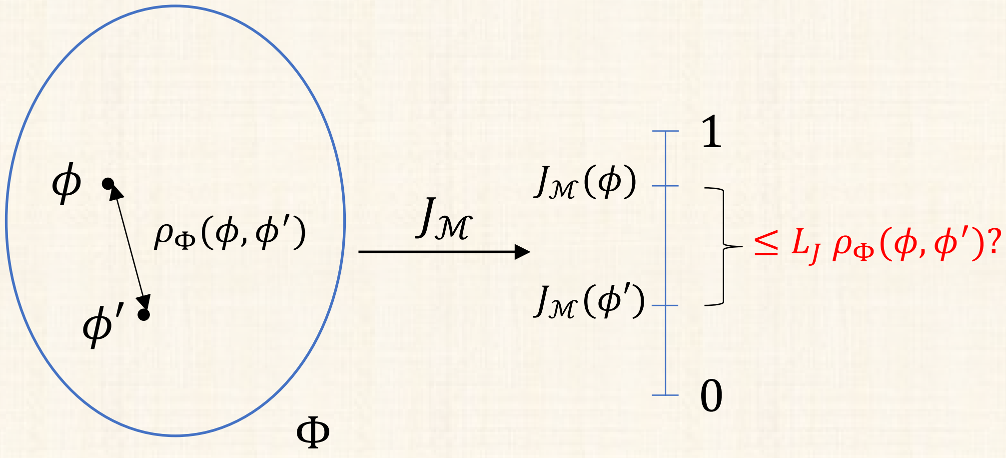
2. Uniform ergodicity: $\exists\alpha \in [0,1)$ and $C<\infty$ such that for every $(s,a),(s',a')$

$$\left\|\mu^{(t)}_{\phi,p,s_0} - \mu^{(\infty)}_{\phi,p}\right\|_{TV} \le C\cdot\alpha^t, \forall t\in\mathbb{N}.$$

3. Upper bound on stationary measure: $\exists\bar{\kappa}>0$ and a probability measure ν such that for every ϕ, $\mu^{(\infty)}_{\phi,p} \le \bar{\kappa}\cdot\nu$.

Metric , ρ_Φ

- $\rho_\Phi(\phi,\phi') \coloneqq \int \rho_{\mathcal{A}}\big(\phi(s),\phi'(s)\big)d\nu(s)$, where ν satisfies 3 and $\rho_{\mathcal{A}}$ is a metric on $\mathcal{A}$.



⋆ $\mu^{(t)}_{\phi,p,s'}$ - distribution of s_t when ϕ is played and initial state is s'

⋆ $\mu^{(\infty)}_{\phi,p}$ - stationary distribution of $\{s_t\}$ under application of ϕ

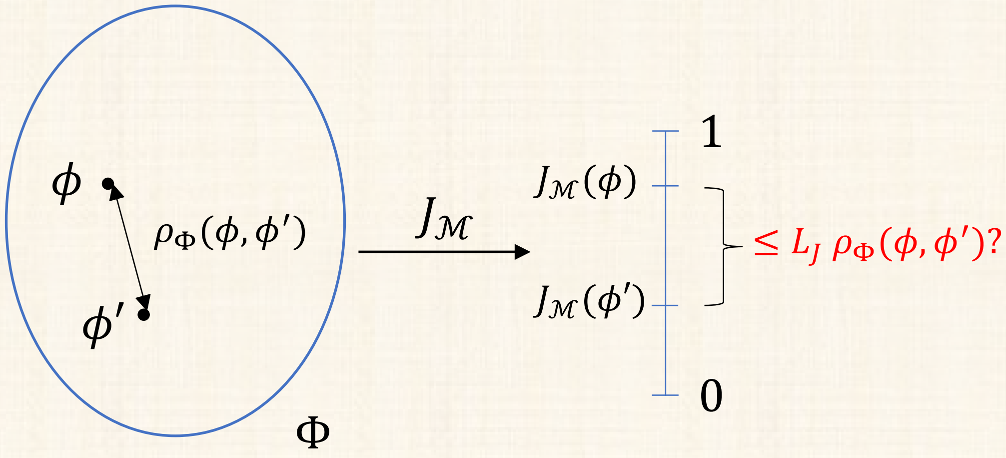
Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$

Assumptions

1. Lipschitz continuity: For every $(s,a),(s',a')$
$$|r(s,a)-r(s',a')| \le L_r \rho\big((s,a),(s',a')\big),$$
$$\|p(s,a,\cdot)-p(s',a',\cdot)\|_{TV} \le L_p \rho\big((s,a),(s',a')\big).$$
2. Uniform ergodicity: $\exists \alpha \in [0,1)$ and $C < \infty$ such that for every $(s,a),(s',a')$
$$\left\|\mu^{(t)}_{\phi,p,s_0} - \mu^{(\infty)}_{\phi,p}\right\|_{TV} \le C \cdot \alpha^t, \forall t \in \mathbb{N}.$$
3. Upper bound on stationary measure: $\exists \bar{\kappa} > 0$ and a probability measure ν such that for every ϕ, $\mu^{(\infty)}_{\phi,p} \le \bar{\kappa} \cdot \nu$.

Metric , ρ_Φ

- $\rho_\Phi(\phi,\phi') \coloneqq \int \rho_{\mathcal{A}}\big(\phi(s),\phi'(s)\big) d\nu(s)$, where ν satisfies 3 and $\rho_{\mathcal{A}}$ is a metric on $\mathcal{A}$.



Theroem: Under Assumption 1, 2 and 3, for any ϕ, ϕ'
$$\left\|\mu^{(\infty)}_{\phi,p} - \mu^{(\infty)}_{\phi',p}\right\|_{TV} \le \left(\lceil \log_{\alpha^{-1}}(C) \rceil + 1\right) \frac{\bar{\kappa} L_p}{1-\alpha} \rho_\Phi(\phi,\phi')$$

⋆ $\mu^{(t)}_{\phi,p,s'}$ - distribution of s_t when ϕ is played and initial state is s'

⋆ $\mu^{(\infty)}_{\phi,p}$ - stationary distribution of $\{s_t\}$ under application of ϕ

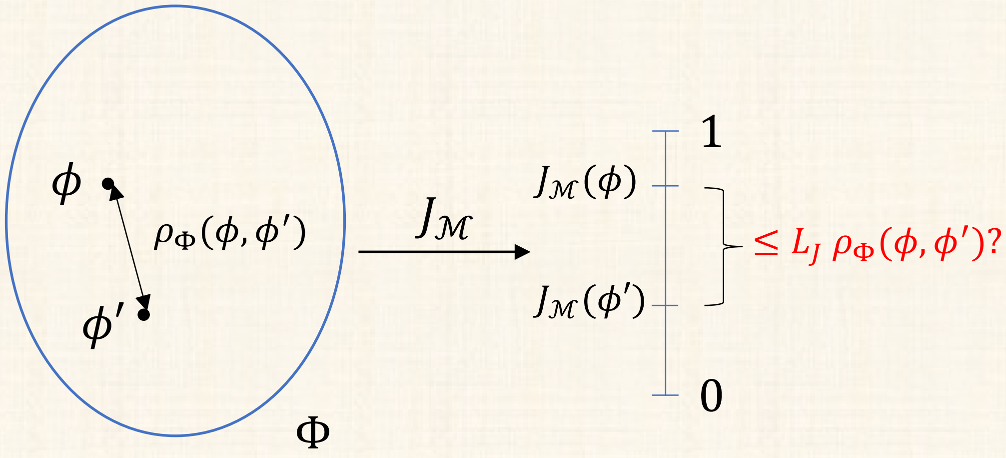
Lipschitz Continuity of $J_{\mathcal{M}}(\cdot)$

Assumptions

1. <u>Lipschitz continuity:</u> For every $(s,a),(s',a')$
$$|r(s,a) - r(s',a')| \le L_r \rho\big((s,a),(s',a')\big),$$
$$\|p(s,a,\cdot) - p(s',a',\cdot)\|_{TV} \le L_p \rho\big((s,a),(s',a')\big).$$
2. <u>Uniform ergodicity:</u> $\exists \alpha \in [0,1)$ and $C < \infty$ such that for every $(s,a),(s',a')$
$$\left\|\mu_{\phi,p,s_0}^{(t)} - \mu_{\phi,p}^{(\infty)}\right\|_{TV} \le C \cdot \alpha^t, \forall t \in \mathbb{N}.$$
3. <u>Upper bound on stationary measure:</u> $\exists \bar{\kappa} > 0$ and a probability measure ν such that for every ϕ, $\mu_{\phi,p}^{(\infty)} \le \bar{\kappa} \cdot \nu$.

Metric , ρ_Φ

- $\rho_\Phi(\phi,\phi') \coloneqq \int \rho_{\mathcal{A}}\big(\phi(s),\phi'(s)\big) d\nu(s)$, where ν satisfies 3 and $\rho_{\mathcal{A}}$ is a metric on $\mathcal{A}$.



<u>**Theroem:**</u> Under Assumption 1, 2 and 3, for any ϕ,ϕ'
$$\left\|\mu_{\phi,p}^{(\infty)} - \mu_{\phi',p}^{(\infty)}\right\|_{TV} \le \left(\lceil \log_{\alpha^{-1}}(C) \rceil + 1\right) \frac{\bar{\kappa} L_p}{1-\alpha} \rho_\Phi(\phi,\phi')$$

<u>**Corollary:**</u> Under Assumption 1, 2 and 3, for any ϕ,ϕ'
$$|J_{\mathcal{M}}(\phi) - J_{\mathcal{M}}(\phi')| \le L_J \rho_\Phi(\phi,\phi')$$
where $L_J \coloneqq \bar{\kappa}\left(L_r + \frac{(\lceil \log_{\alpha^{-1}}(C) \rceil + 1) L_p}{(1-\alpha)}\right)$.

⋆ $\mu_{\phi,p,s'}^{(t)}$ - distribution of s_t when ϕ is played and initial state is s'

⋆ $\mu_{\phi,p}^{(\infty)}$ - stationary distribution of $\{s_t\}$ under application of ϕ

Zooming in Policy Space

Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$

Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$

Finite policy set, Φ

Suppose that we have

$$|J_{\mathcal{M}}(\phi) - \bar{r}_t(\phi)| \le c_t(\phi) = \tilde{O}\left(\frac{1}{N_t(\phi)^\beta}\right)$$

where

$$N_t(\phi) = \sum_{s<t} \mathbb{I}(\phi_s = \phi)\text{, and}$$

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N_t}(\phi)} \sum_s r_s \mathbb{I}(\phi_s = \phi)$$

Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$

Finite policy set, Φ

Suppose that we have

$$|J_{\mathcal{M}}(\phi) - \bar{r}_t(\phi)| \le c_t(\phi) = \tilde{O}\left(\frac{1}{N_t(\phi)^\beta}\right)$$

where

$$N_t(\phi) = \sum_{s<t} \mathbb{I}(\phi_s = \phi)\text{, and}$$

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N}_t(\phi)} \sum_s r_s \mathbb{I}(\phi_s = \phi)$$

Choose $\phi_t \in \operatorname{argmax}_{\phi\in\Phi} Index_t(\phi)$, where

$$Index_t(\phi) \coloneqq \bar{r}_t(\phi) + \frac{const}{N_t(\phi)^\beta}$$

Then,

$$R_\Phi(T;\psi) \le \tilde{O}\left(|\Phi|^\beta T^{1-\beta}\right)$$

Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$

Finite policy set, Φ

Suppose that we have

$$|J_{\mathcal{M}}(\phi) - \bar{r}_t(\phi)| \le c_t(\phi) = \tilde{O}\left(\frac{1}{N_t(\phi)^\beta}\right)$$

where

$$N_t(\phi) = \sum_{s<t} \mathbb{I}(\phi_s = \phi)\text{, and}$$

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N_t}(\phi)} \sum_s r_s \mathbb{I}(\phi_s = \phi)$$

Choose $\phi_t \in \operatorname*{argmax}_{\phi\in\Phi} Index_t(\phi)$, where

$Index_t(\phi) \coloneqq \bar{r}_t(\phi) + \frac{const}{N_t(\phi)^\beta}$

Then,

$$R_\Phi(T;\psi) \le \tilde{O}\left(|\Phi|^\beta T^{1-\beta}\right)$$

Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$

<u>**Finite policy set, Φ**</u>

Suppose that we have

$$|J_{\mathcal{M}}(\phi) - \bar{r}_t(\phi)| \leq c_t(\phi) = \tilde{O}\left(\frac{1}{N_t(\phi)^\beta}\right)$$

where

$$N_t(\phi) = \sum_{s<t} \mathbb{I}(\phi_s = \phi)\text{, and}$$

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N_t}(\phi)} \sum_{s} r_s \mathbb{I}(\phi_s = \phi)$$

Choose $\phi_t \in \operatorname{argmax}_{\phi\in\Phi} Index_t(\phi)$, where

$$Index_t(\phi) \coloneqq \bar{r}_t(\phi) + \frac{const}{N_t(\phi)^\beta}$$

Then,

$$R_\Phi(T;\psi) \leq \tilde{O}\left(|\Phi|^\beta T^{1-\beta}\right)$$

<u>**Compact policy set, Φ**</u>

Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$

Finite policy set, Φ

Suppose that we have

$$|J_{\mathcal{M}}(\phi) - \bar{r}_t(\phi)| \le c_t(\phi) = \tilde{O}\left(\frac{1}{N_t(\phi)^\beta}\right)$$

where

$$N_t(\phi) = \sum_{s<t} \mathbb{I}(\phi_s = \phi)\text{, and}$$

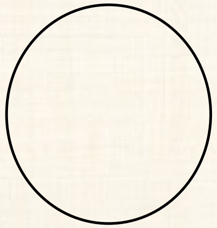
$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N_t}(\phi)} \sum_s r_s \mathbb{I}(\phi_s = \phi)$$

Choose $\phi_t \in \operatorname{argmax}_{\phi\in\Phi} Index_t(\phi)$, where

$$Index_t(\phi) \coloneqq \bar{r}_t(\phi) + \frac{const}{N_t(\phi)^\beta}$$

Then,

$$R_\Phi(T;\psi) \le \tilde{O}\left(|\Phi|^\beta T^{1-\beta}\right)$$



Policy space, Φ

Compact policy set, Φ

Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$

Finite policy set, Φ

Suppose that we have

$$|J_{\mathcal{M}}(\phi) - \bar{r}_t(\phi)| \le c_t(\phi) = \tilde{O}\left(\frac{1}{N_t(\phi)^\beta}\right)$$

where

$$N_t(\phi) = \sum_{s<t} \mathbb{I}(\phi_s = \phi)\text{, and}$$

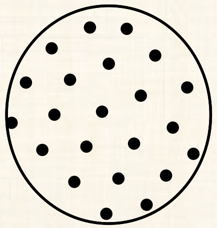
$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N_t}(\phi)} \sum_s r_s \mathbb{I}(\phi_s = \phi)$$

Choose $\phi_t \in \operatorname{argmax}_{\phi\in\Phi} Index_t(\phi)$, where

$$Index_t(\phi) \coloneqq \bar{r}_t(\phi) + \frac{const}{N_t(\phi)^\beta}$$

Then,

$$R_\Phi(T;\psi) \le \tilde{O}\left(|\Phi|^\beta T^{1-\beta}\right)$$



Policy space, Φ

Compact policy set, Φ

Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$

Finite policy set, Φ

Suppose that we have

$$|J_{\mathcal{M}}(\phi) - \bar{r}_t(\phi)| \le c_t(\phi) = \tilde{O}\left(\frac{1}{N_t(\phi)^\beta}\right)$$

where

$$N_t(\phi) = \sum_{s<t} \mathbb{I}(\phi_s = \phi)\text{, and}$$

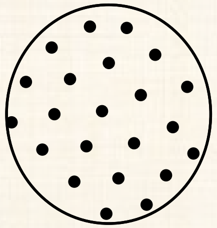
$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N_t}(\phi)} \sum_s r_s \mathbb{I}(\phi_s = \phi)$$

Choose $\phi_t \in \operatorname{argmax}_{\phi\in\Phi} Index_t(\phi)$, where

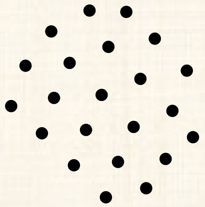
$$Index_t(\phi) \coloneqq \bar{r}_t(\phi) + \frac{const}{N_t(\phi)^\beta}$$

Then,

$$R_\Phi(T;\psi) \le \tilde{O}\left(|\Phi|^\beta T^{1-\beta}\right)$$



Policy space, Φ



ϵ-net of Φ, Φ_ϵ

Compact policy set, Φ

Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$

Finite policy set, Φ

Suppose that we have

$$|J_{\mathcal{M}}(\phi) - \bar{r}_t(\phi)| \le c_t(\phi) = \tilde{O}\left(\frac{1}{N_t(\phi)^\beta}\right)$$

where

$$N_t(\phi) = \sum_{s<t} \mathbb{I}(\phi_s = \phi)\text{, and}$$

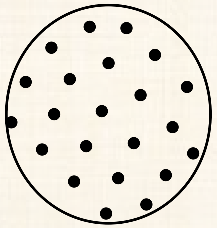
$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N_t}(\phi)} \sum_{s} r_s \mathbb{I}(\phi_s = \phi)$$

Choose $\phi_t \in \operatorname*{argmax}_{\phi\in\Phi} Index_t(\phi)$, where

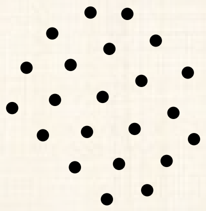
$Index_t(\phi) \coloneqq \bar{r}_t(\phi) + \frac{const}{N_t(\phi)^\beta}$

Then,

$$R_\Phi(T;\psi) \le \tilde{O}\left(|\Phi|^\beta T^{1-\beta}\right)$$



Policy space, Φ



ϵ-net of Φ, Φ_ϵ

Compact policy set, Φ

If

1. $J_{\mathcal{M}}: \Phi \to [0,1]$ is L_J-Lipschitz,
2. The algorithm for finite policy set is run with Φ_ϵ
3. $\epsilon = T^{-\frac{\beta}{d^\Phi \beta+1}}$, where d^Φ is the dimension Φ,

Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$

Finite policy set, Φ

Suppose that we have

$$|J_{\mathcal{M}}(\phi) - \bar{r}_t(\phi)| \le c_t(\phi) = \tilde{O}\left(\frac{1}{N_t(\phi)^\beta}\right)$$

where

$$N_t(\phi) = \sum_{s<t} \mathbb{I}(\phi_s = \phi)\text{, and}$$

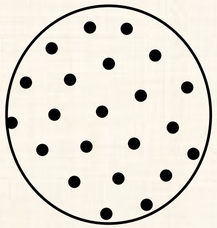
$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N}_t(\phi)} \sum_{s} r_s \mathbb{I}(\phi_s = \phi)$$

Choose $\phi_t \in \operatorname{argmax}_{\phi\in\Phi} Index_t(\phi)$, where

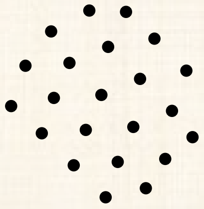
$$Index_t(\phi) \coloneqq \bar{r}_t(\phi) + \frac{const}{N_t(\phi)^\beta}$$

Then,

$$R_\Phi(T;\psi) \le \tilde{O}\left(|\Phi|^\beta T^{1-\beta}\right)$$



Policy space, Φ



ϵ-net of Φ, Φ_ϵ

Compact policy set, Φ

If

1. $J_{\mathcal{M}}: \Phi \to [0,1]$ is L_J-Lipschitz,
2. The algorithm for finite policy set is run with Φ_ϵ
3. $\epsilon = T^{-\frac{\beta}{d^\Phi \beta+1}}$, where d^Φ is the dimension Φ,

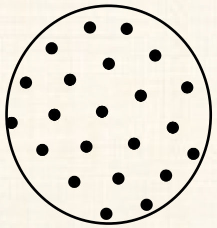
Then,

$$R_\Phi(T;\psi) \le \tilde{O}\left(T^{\frac{(d^\Phi-1)\beta+1}{d^\Phi \beta+1}}\right)$$

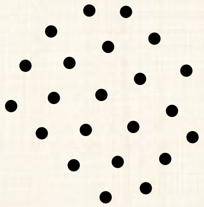
Zooming in Policy Space

Define Regret w.r.t. Φ as,

$$R_\Phi(T;\psi) \coloneqq T \max_{\phi\in\Phi} J_{\mathcal{M}}(\phi) - \sum_{t=1}^{T} r_t$$



Policy space, Φ



ϵ-net of Φ, Φ_ϵ

Finite policy set, Φ

Suppose that we have

$$|J_{\mathcal{M}}(\phi) - \bar{r}_t(\phi)| \le c_t(\phi) = \tilde{O}\left(\frac{1}{N_t(\phi)^\beta}\right)$$

where

$$N_t(\phi) = \sum_{s<t} \mathbb{I}(\phi_s = \phi)\text{, and}$$

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N}_t(\phi)} \sum_{s} r_s \mathbb{I}(\phi_s = \phi)$$

Choose $\phi_t \in \operatorname{argmax}_{\phi\in\Phi} Index_t(\phi)$, where

$$Index_t(\phi) \coloneqq \bar{r}_t(\phi) + \frac{const}{N_t(\phi)^\beta}$$

Then,

$$R_\Phi(T;\psi) \le \tilde{O}\left(|\Phi|^\beta T^{1-\beta}\right)$$

Compact policy set, Φ

If

1. $J_{\mathcal{M}}: \Phi \to [0,1]$ is L_J-Lipschitz,
2. The algorithm for finite policy set is run with Φ_ϵ
3. $\epsilon = T^{-\frac{\beta}{d^\Phi \beta+1}}$, where d^Φ is the dimension Φ,

Then,

$$R_\Phi(T;\psi) \le \tilde{O}\left(T^{\frac{(d^\Phi-1)\beta+1}{d^\Phi\beta+1}}\right)$$

- For example, $\beta = 0.5 \Rightarrow R_\Phi(T;\psi) \le \tilde{O}\left(T^{\frac{d^\Phi+1}{d^\Phi+2}}\right)$

Zooming in Policy Space

<u>Policy Zooming for RL</u>

Inputs: Horizon T, A policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\ \}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k + 1$

    $\Phi_{\mathrm{t}} = \Phi_{t-1}$

    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B\big(\phi; c_t(\phi)\big)$:

      $\Phi_{\mathrm{t}} \leftarrow \Phi_t \cup \{\phi'\}$

    $\phi_k \in \operatorname{argmax}_{\phi \in \Phi_t} Index_t(\phi)$

    $H = \max\{1, N_t(\phi_k)\}$

  $h \leftarrow h + 1$

  Play $a_t = \phi_k(s_t)$

Zooming in Policy Space

<u>Policy Zooming for RL</u>

Inputs: Horizon T, A policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\ \}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k + 1$

    $\Phi_{\mathrm{t}} = \Phi_{t-1}$

    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B\big(\phi; c_t(\phi)\big)$:

      $\Phi_{\mathrm{t}} \leftarrow \Phi_t \cup \{\phi'\}$

    $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t}\, Index_t(\phi)$

    $H = \max\{1, N_t(\phi_k)\}$

  $h \leftarrow h + 1$

  Play $a_t = \phi_k(s_t)$

Model-free

Model-based

Zooming in Policy Space

<u>Policy Zooming for RL (Model-free)</u>

Inputs: Horizon T, A policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\ \}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

    if $h = H_k$:

        $h = 0, k \leftarrow k + 1$

        $\Phi_{\mathrm{t}} = \Phi_{t-1}$

        while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:

            $\Phi_{\mathrm{t}} \leftarrow \Phi_t \cup \{\phi'\}$

        $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t} Index_t(\phi)$

        $H = \max\{1, N_t(\phi_k)\}$

    $h \leftarrow h + 1$

    Play $a_t = \phi_k(s_t)$

$$Index_t(\phi) = \bar{r}_t(\phi) + c_t(\phi)$$

where

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N}_t(\phi)} \sum_t r_t \mathbb{I}(\phi_t = \phi),$$

$$c_t(\phi) = const \cdot N_t(\phi)^{-\frac{1}{2}}, \text{ and}$$

$$N_t(\phi) = \sum_t \mathbb{I}(\phi_t = \phi)$$

Zooming in Policy Space

Policy Zooming for RL (Model-free)

Inputs: Horizon T, A policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\ \}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k + 1$

    $\Phi_\mathrm{t} = \Phi_{t-1}$

    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:

      $\Phi_\mathrm{t} \leftarrow \Phi_t \cup \{\phi'\}$

    $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t} Index_t(\phi)$

    $H = \max\{1, N_t(\phi_k)\}$

  $h \leftarrow h + 1$

  Play $a_t = \phi_k(s_t)$

$$Index_t(\phi) = \bar{r}_t(\phi) + c_t(\phi)$$

where

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N}_\mathrm{t}(\phi)} \sum_t r_t \mathbb{I}(\phi_t = \phi)\,,$$

$$c_t(\phi) = const \cdot N_t(\phi)^{-\frac{1}{2}}, \text{ and}$$

$$N_t(\phi) = \sum_t \mathbb{I}(\phi_t = \phi)$$

Zooming dimension:

$$d_z^\Phi = \inf\{d > 0 \colon N_\gamma(\Phi_\gamma) \leq c_z \gamma^{-d}\ \forall \gamma > 0\}$$

where Φ_γ is the set of $[\gamma, 2\gamma)$-suboptimal policies.

Zooming in Policy Space

Policy Zooming for RL (Model-free)

Inputs: Horizon T, A policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\ \}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k + 1$

    $\Phi_t = \Phi_{t-1}$

    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B\big(\phi; c_t(\phi)\big)$:

      $\Phi_t \leftarrow \Phi_t \cup \{\phi'\}$

    $\phi_k \in \text{argmax}_{\phi \in \Phi_t} Index_t(\phi)$

    $H = \max\{1, N_t(\phi_k)\}$

  $h \leftarrow h + 1$

  Play $a_t = \phi_k(s_t)$

$$Index_t(\phi) = \bar{r}_t(\phi) + c_t(\phi)$$

where

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N}_t(\phi)} \sum_t r_t \mathbb{I}(\phi_t = \phi),$$

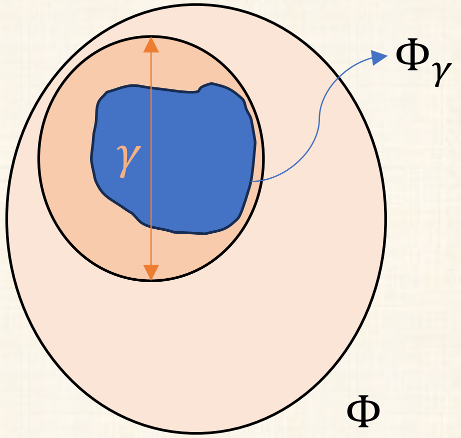
$$c_t(\phi) = const \cdot N_t(\phi)^{-\frac{1}{2}}, \text{ and}$$

$$N_t(\phi) = \sum_t \mathbb{I}(\phi_t = \phi)$$

Zooming dimension:

$$d_z^{\Phi} = \inf\{d > 0 : N_\gamma(\Phi_\gamma) \le c_z \gamma^{-d}\ \forall \gamma > 0\}$$

where Φ_γ is the set of $[\gamma, 2\gamma)$-suboptimal policies.



Zooming in Policy Space

Policy Zooming for RL (Model-free)

Inputs: Horizon T, A policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\,\}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

    if $h = H_k$:

        $h = 0, k \leftarrow k + 1$

        $\Phi_{\mathrm{t}} = \Phi_{t-1}$

        while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:

            $\Phi_{\mathrm{t}} \leftarrow \Phi_t \cup \{\phi'\}$

        $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t} Index_t(\phi)$

        $H = \max\{1, N_t(\phi_k)\}$

    $h \leftarrow h + 1$

    Play $a_t = \phi_k(s_t)$

$$Index_t(\phi) = \bar{r}_t(\phi) + c_t(\phi)$$

where

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N}_\mathrm{t}(\phi)} \sum_t r_t \mathbb{I}(\phi_t = \phi),$$

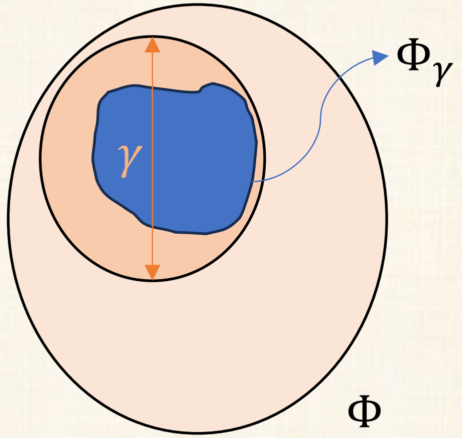
$$c_t(\phi) = const \cdot N_t(\phi)^{-\frac{1}{2}}, \text{ and}$$

$$N_t(\phi) = \sum_t \mathbb{I}(\phi_t = \phi)$$

Zooming dimension:

$$d_z^\Phi = \inf\{d > 0 : N_\gamma(\Phi_\gamma) \le c_z \gamma^{-d} \ \forall \gamma > 0\}$$

where Φ_γ is the set of $[\gamma, 2\gamma)$-suboptimal policies.



In this case, $d_z^\Phi = 1$, but $d^\Phi = 2$.

Zooming in Policy Space

<u>Policy Zooming for RL (Model-free)</u>

Inputs: Horizon T, A policy class Φ
Initialize: Set of active policies $\Phi_0 \leftarrow \{\,\}$, $k=0$, $h=0$, $H_0=0$
for $t=1$ to T:
    if $h=H_k$:
        $h=0, k \leftarrow k+1$
        $\Phi_{\mathrm{t}} = \Phi_{t-1}$
        while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B\big(\phi; c_t(\phi)\big)$:
            $\Phi_{\mathrm{t}} \leftarrow \Phi_t \cup \{\phi'\}$
        $\phi_k \in \mathrm{argmax}_{\phi\in\Phi_t}\, Index_t(\phi)$
        $H = \max\{1, N_t(\phi_k)\}$
    $h \leftarrow h+1$
    Play $a_t = \phi_k(s_t)$

$$Index_t(\phi) = \bar{r}_t(\phi) + c_t(\phi)$$

where

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N_t}(\phi)} \sum_t r_t \mathbb{I}(\phi_t = \phi),$$

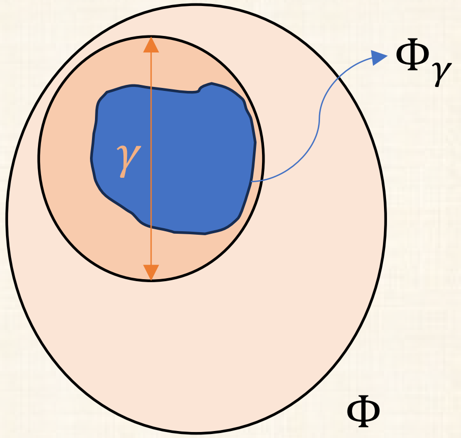
$$c_t(\phi) = const \cdot N_t(\phi)^{-\frac{1}{2}}, \text{ and}$$

$$N_t(\phi) = \sum_t \mathbb{I}(\phi_t = \phi)$$

<u>Zooming dimension:</u>

$$d_z^\Phi = \inf\{d > 0 : N_\gamma(\Phi_\gamma) \le c_z \gamma^{-d}\ \forall \gamma > 0\}$$

where Φ_γ is the set of $[\gamma, 2\gamma)$-suboptimal policies.



In this case, $d_z^\Phi = 1$, but $d^\Phi = 2$.

<u>Theroem:</u> If $\mathcal{M}$ satisfies Assumption 1, 2 and 3, then with high probability

$$R_\Phi(T; PZRLMF) \le \tilde{O}\left(T^{\frac{d_z^\Phi+1}{d_z^\Phi+2}}\right)$$

Zooming in Policy Space

Policy Zooming for RL (Model-free)

Inputs: Horizon T, A policy class Φ
Initialize: Set of active policies $\Phi_0 \leftarrow \{\,\}$, $k = 0$, $h = 0$, $H_0 = 0$
for $t = 1$ to T:
  if $h = H_k$:
    $h = 0, k \leftarrow k + 1$
    $\Phi_{\mathrm{t}} = \Phi_{t-1}$
    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:
      $\Phi_{\mathrm{t}} \leftarrow \Phi_t \cup \{\phi'\}$
    $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t} Index_t(\phi)$
    $H = \max\{1, N_t(\phi_k)\}$
  $h \leftarrow h + 1$
  Play $a_t = \phi_k(s_t)$

$$Index_t(\phi) = \bar{r}_t(\phi) + c_t(\phi)$$

where

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N_t}(\phi)} \sum_t r_t \mathbb{I}(\phi_t = \phi),$$

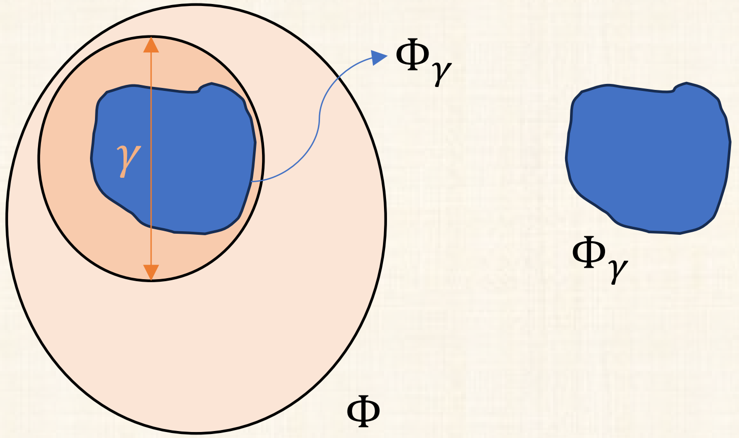
$$c_t(\phi) = const \cdot N_t(\phi)^{-\frac{1}{2}}, \text{and}$$

$$N_t(\phi) = \sum_t \mathbb{I}(\phi_t = \phi)$$

Zooming dimension:

$$d_z^{\Phi} = \inf\{d > 0: N_\gamma(\Phi_\gamma) \le c_z \gamma^{-d} \ \forall \gamma > 0\}$$

where Φ_γ is the set of $[\gamma, 2\gamma)$-suboptimal policies.



In this case, $d_z^{\Phi} = 1$, but $d^{\Phi} = 2$.

Theroem: If $\mathcal{M}$ satisfies Assumption 1, 2 and 3, then with high probability

$$R_\Phi(T; PZRLMF) \le \tilde{O}\left(T^{\frac{d_z^\Phi + 1}{d_z^\Phi + 2}}\right)$$

Zooming in Policy Space

Policy Zooming for RL (Model-free)

Inputs: Horizon T, A policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\ \}$, $k=0$, $h=0$, $H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k+1$

    $\Phi_t = \Phi_{t-1}$

    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:

      $\Phi_t \leftarrow \Phi_t \cup \{\phi'\}$

    $\phi_k \in \operatorname{argmax}_{\phi \in \Phi_t} Index_t(\phi)$

    $H = \max\{1, N_t(\phi_k)\}$

  $h \leftarrow h+1$

  Play $a_t = \phi_k(s_t)$

$$Index_t(\phi) = \bar{r}_t(\phi) + c_t(\phi)$$

where

$$\bar{r}_t(\phi) = \frac{1}{N_t(\phi)} \sum_t r_t \mathbb{I}(\phi_t = \phi),$$

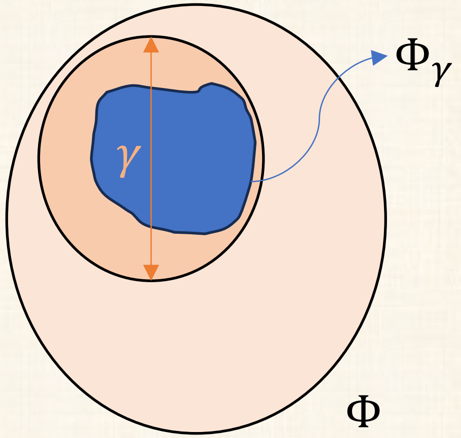
$$c_t(\phi) = const \cdot N_t(\phi)^{-\frac{1}{2}}, \text{ and}$$

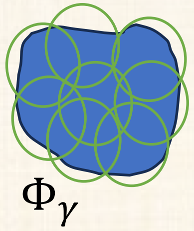
$$N_t(\phi) = \sum_t \mathbb{I}(\phi_t = \phi)$$

Zooming dimension:

$$d_z^\Phi = \inf\{d > 0: N_\gamma(\Phi_\gamma) \le c_z \gamma^{-d}\ \forall \gamma > 0\}$$

where Φ_γ is the set of $[\gamma, 2\gamma)$-suboptimal policies.





In this case, $d_z^\Phi = 1$, but $d^\Phi = 2$.

Theroem: If $\mathcal{M}$ satisfies Assumption 1, 2 and 3, then with high probability

$$R_\Phi(T; PZRLMF) \le \tilde{O}\left(T^{\frac{d_z^\Phi + 1}{d_z^\Phi + 2}}\right)$$

Zooming in Policy Space

Policy Zooming for RL (Model-free)

Inputs: Horizon T, A policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\,\}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:
- **if** $h = H_k$:
  - $h = 0, k \leftarrow k + 1$
  - $\Phi_t = \Phi_{t-1}$
  - **while** $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:
    - $\Phi_t \leftarrow \Phi_t \cup \{\phi'\}$
  - $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t}\, Index_t(\phi)$
  - $H = \max\{1, N_t(\phi_k)\}$
- $h \leftarrow h + 1$
- Play $a_t = \phi_k(s_t)$

$$Index_t(\phi) = \bar{r}_t(\phi) + c_t(\phi)$$

where

$$\bar{r}_t(\phi) = \frac{1}{\mathrm{N_t}(\phi)} \sum_t r_t \mathbb{I}(\phi_t = \phi),$$

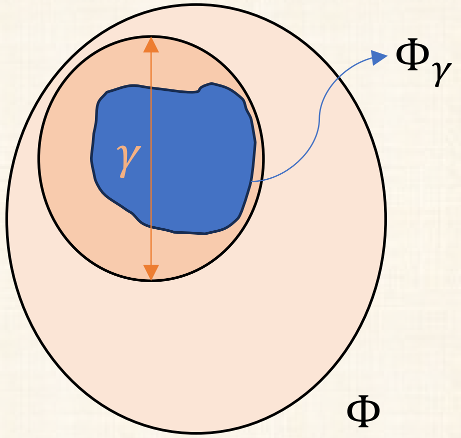
$$c_t(\phi) = const \cdot N_t(\phi)^{-\frac{1}{2}}, \text{and}$$

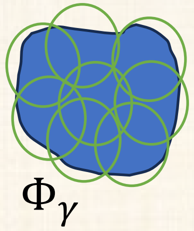
$$N_t(\phi) = \sum_t \mathbb{I}(\phi_t = \phi)$$

Zooming dimension:

$$d_z^\Phi = \inf\{d > 0 : N_\gamma(\Phi_\gamma) \le c_z \gamma^{-d}\ \forall \gamma > 0\}$$

where Φ_γ is the set of $[\gamma, 2\gamma)$-suboptimal policies.





#active policies in $\Phi_\gamma \le \#\ \gamma$-covering balls

In this case, $d_z^\Phi = 1$, but $d^\Phi = 2$.

Theroem: If $\mathcal{M}$ satisfies Assumption 1, 2 and 3, then with high probability

$$R_\Phi(T; PZRLMF) \le \tilde{O}\left(T^{\frac{d_z^\Phi + 1}{d_z^\Phi + 2}}\right)$$

Zooming in Policy Space

<u>Policy Zooming for RL (Model-based)</u>

Inputs: Horizon T, A Lipschitz policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\ \}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k + 1$

    $\Phi_{\mathrm{t}} = \Phi_{t-1}$

    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B\big(\phi; c_t(\phi)\big)$:

      $\Phi_{\mathrm{t}} \leftarrow \Phi_t \cup \{\phi'\}$

    $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t} Index_t(\phi)$

    $H = \max\{1, N_t(\phi_k)\}$

  $h \leftarrow h + 1$

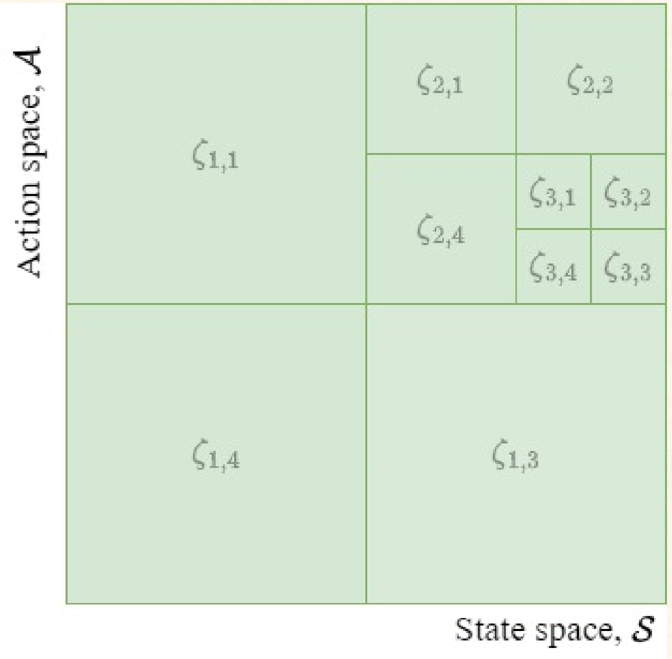
  Play $a_t = \phi_k(s_t)$

Zooming in Policy Space

<u>Policy Zooming for RL (Model-based)</u>

Inputs: Horizon T, A Lipschitz policy class Φ
Initialize: Set of active policies $\Phi_0 \leftarrow \{\,\}$, $k = 0$, $h = 0$, $H_0 = 0$
for $t = 1$ to T:
  if $h = H_k$:
    $h = 0, k \leftarrow k + 1$
    $\Phi_t = \Phi_{t-1}$
    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:
      $\Phi_t \leftarrow \Phi_t \cup \{\phi'\}$
    $\phi_k \in \operatorname{argmax}_{\phi \in \Phi_t} Index_t(\phi)$
    $H = \max\{1, N_t(\phi_k)\}$
  $h \leftarrow h + 1$
  Play $a_t = \phi_k(s_t)$

1. Adaptively discretized partition of $\mathcal{S} \times \mathcal{A}$:
   ζ is active at time $t \Leftrightarrow N_t(\zeta) \geq diam(\zeta)^{-(d_\mathcal{S}+2)}$ but $< 2^{d_\mathcal{S}+2} diam(\zeta)^{-(d_\mathcal{S}+2)}$



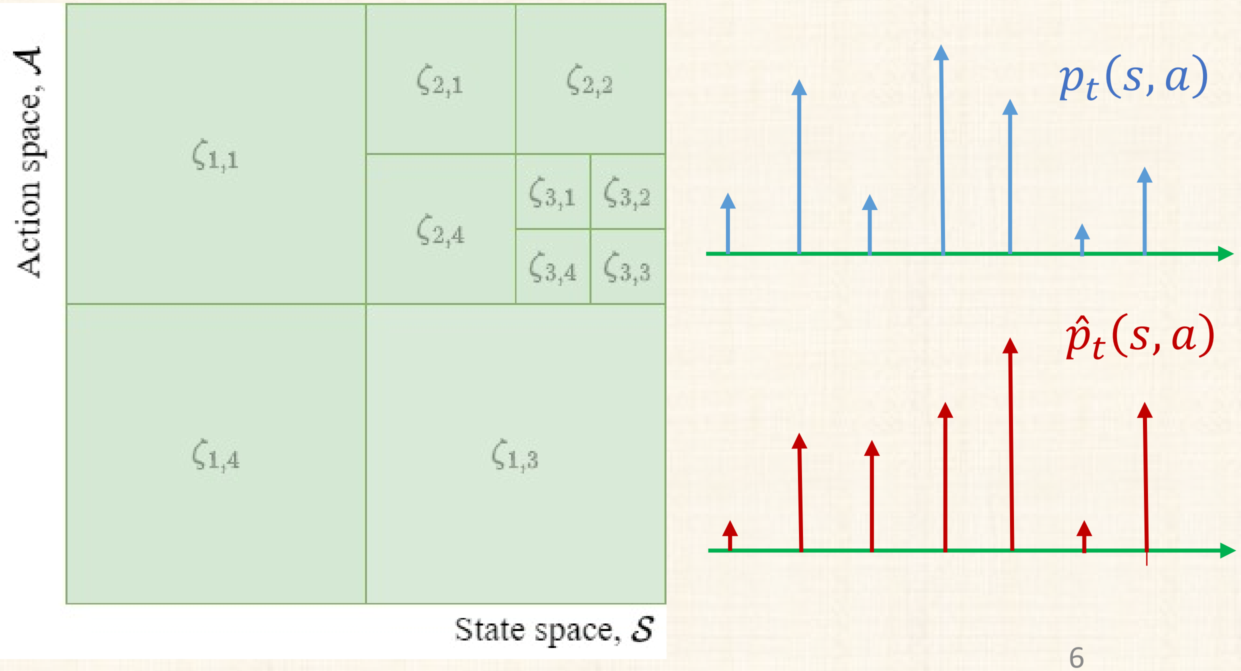
Zooming in Policy Space

<u>Policy Zooming for RL (Model-based)</u>

Inputs: Horizon T, A Lipschitz policy class Φ
Initialize: Set of active policies $\Phi_0 \leftarrow \{\,\}$, $k = 0$, $h = 0$, $H_0 = 0$
for $t = 1$ to T:
  if $h = H_k$:
    $h = 0, k \leftarrow k + 1$
    $\Phi_t = \Phi_{t-1}$
    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:
      $\Phi_t \leftarrow \Phi_t \cup \{\phi'\}$
    $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t} Index_t(\phi)$
    $H = \max\{1, N_t(\phi_k)\}$
  $h \leftarrow h + 1$
  Play $a_t = \phi_k(s_t)$

1. Adaptively discretized partition of $\mathcal{S} \times \mathcal{A}$:
   ζ is active at time $t \Leftrightarrow N_t(\zeta) \geq diam(\zeta)^{-(d_{\mathcal{S}}+2)}$ but $< 2^{d_{\mathcal{S}}+2} diam(\zeta)^{-(d_{\mathcal{S}}+2)}$
2. Confidence set for p:
   - Z_t: Representative point of cells

   $\mathcal{C}_t = \{\theta : \|\theta(s,a) - \hat{p}_t(s,a)\|_{TV} \leq \mathrm{diam}(\zeta_{s,a})\ \forall (s,a) \in Z_t\}$
   <u>**Lemma:**</u> $p_t \in \mathcal{C}_t\ \forall\, t \in [T]$ with high probability.



Zooming in Policy Space

<u>Policy Zooming for RL (Model-based)</u>

Inputs: Horizon T, A Lipschitz policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\,\}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k + 1$

    $\Phi_t = \Phi_{t-1}$

    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:

      $\Phi_t \leftarrow \Phi_t \cup \{\phi'\}$

    $\phi_k \in \text{argmax}_{\phi \in \Phi_t} Index_t(\phi)$

    $H = \max\{1, N_t(\phi_k)\}$

  $h \leftarrow h + 1$

  Play $a_t = \phi_k(s_t)$

1. Adaptively discretized partition of $\mathcal{S} \times \mathcal{A}$:
   ζ is active at time $t \Leftrightarrow N_t(\zeta) \geq diam(\zeta)^{-(d_{\mathcal{S}}+2)}$ but $< 2^{d_{\mathcal{S}}+2} diam(\zeta)^{-(d_{\mathcal{S}}+2)}$
2. Confidence set for p:
   - Z_t: Representative point of cells

   $\mathcal{C}_t = \left\{\theta: \|\theta(s,a) - \hat{p}_t(s,a)\|_{TV} \leq \text{diam}\left(\zeta_{s,a}\right) \forall (s,a) \in Z_t\right\}$

   <u>Lemma:</u> $p_t \in \mathcal{C}_t \; \forall\, t \in [T]$ with high probability.
3. Index of policy ϕ:
   $\bar{V}_0^{\phi}(s) = 0$

   $$\bar{V}_{i+1}^{\phi}(s) = r^{+}\left(\zeta_{s,\phi(s)}\right) + \max_{\theta \in \mathcal{C}_t} \sum \theta\left(q\left(\zeta_{s,\phi(s)}\right), s'\right) \bar{V}_i^{\phi}(s')$$

   where $r^{+} = r + const \cdot diam$,
   $q(\cdot)$ denotes quantized point

   $$Index_t(\phi) := \lim_{i \to \infty} \frac{1}{i} \bar{V}_i^{\phi}(s)$$

Zooming in Policy Space

Policy Zooming for RL (Model-based)

Inputs: Horizon T, A Lipschitz policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\,\}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k + 1$

    $\Phi_t = \Phi_{t-1}$

    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:

      $\Phi_t \leftarrow \Phi_t \cup \{\phi'\}$

    $\phi_k \in \text{argmax}_{\phi \in \Phi_t} Index_t(\phi)$

    $H = \max\{1, N_t(\phi_k)\}$

  $h \leftarrow h + 1$

  Play $a_t = \phi_k(s_t)$

Assumptions (Contd.)

4. Lower bound on stationary measure: $\exists \kappa > 0$ such that for every ϕ, $\mu_{\phi,p}^{(\infty)} \geq \kappa \cdot \lambda$, where λ is the Lebesgue measure on $\mathcal{S}$.
5. Partial derivatives of transition densities are bounded.

Zooming in Policy Space

Policy Zooming for RL (Model-based)

Inputs: Horizon T, A Lipschitz policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\ \}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k + 1$

    $\Phi_{\mathrm{t}} = \Phi_{t-1}$

    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B\big(\phi; c_t(\phi)\big)$:

      $\Phi_{\mathrm{t}} \leftarrow \Phi_t \cup \{\phi'\}$

    $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t}\, Index_t(\phi)$

    $H = \max\{1, N_t(\phi_k)\}$

  $h \leftarrow h + 1$

  Play $a_t = \phi_k(s_t)$

Assumptions (Contd.)

4. Lower bound on stationary measure: $\exists \underline{\kappa} > 0$ such that for every ϕ, $\mu_{\phi,p}^{(\infty)} \geq \underline{\kappa} \cdot \lambda$, where λ is the Lebesgue measure on $\mathcal{S}$.
5. Partial derivatives of transition densities are bounded.

Lemma: Under Assumption 1 and 5, with high probability

$$J_{\mathcal{M}}(\phi) \leq Index_t(\phi), \forall t = 1,2, \dots, T.$$

Zooming in Policy Space

Policy Zooming for RL (Model-based)

Inputs: Horizon T, A Lipschitz policy class Φ

Initialize: Set of active policies $\Phi_0 \leftarrow \{\,\}$, $k = 0$, $h = 0$, $H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k + 1$

    $\Phi_t = \Phi_{t-1}$

    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:

      $\Phi_t \leftarrow \Phi_t \cup \{\phi'\}$

    $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t} Index_t(\phi)$

    $H = \max\{1, N_t(\phi_k)\}$

  $h \leftarrow h + 1$

  Play $a_t = \phi_k(s_t)$

Assumptions (Contd.)

4. Lower bound on stationary measure: $\exists \kappa > 0$ such that for every ϕ, $\mu_{\phi,p}^{(\infty)} \geq \kappa \cdot \lambda$, where λ is the Lebesgue measure on $\mathcal{S}$.
5. Partial derivatives of transition densities are bounded.

Lemma: Under Assumption 1 and 5, with high probability

$$J_{\mathcal{M}}(\phi) \leq Index_t(\phi), \forall t = 1,2, \ldots, T.$$

Lemma: Under Assumption 1, 2, 3, 4 and 5, with high probability

$$Index_t(\phi)$$
$$\leq J_{\mathcal{M}}(\phi) + const \cdot \int diam\big(\zeta_{s,\phi(s)}\big) d\mu_{\phi,p}^{(\infty)}(s),$$
$$\forall t = 1,2, \ldots, T.$$

Zooming in Policy Space

Policy Zooming for RL (Model-based)

Inputs: Horizon T, A Lipschitz policy class Φ
Initialize: Set of active policies $\Phi_0 \leftarrow \{\ \}$, $k = 0$, $h = 0$, $H_0 = 0$
for $t = 1$ to T:
  if $h = H_k$:
    $h = 0, k \leftarrow k + 1$
    $\Phi_t = \Phi_{t-1}$
    while $\exists \phi' \in \Phi$ such that $\phi' \in \cup_{\phi \in \Phi_t} B(\phi; c_t(\phi))$:
      $\Phi_t \leftarrow \Phi_t \cup \{\phi'\}$
    $\phi_k \in \mathrm{argmax}_{\phi \in \Phi_t}\, Index_t(\phi)$
    $H = \max\{1, N_t(\phi_k)\}$
  $h \leftarrow h + 1$
  Play $a_t = \phi_k(s_t)$

Assumptions (Contd.)

4. Lower bound on stationary measure: $\exists \kappa > 0$ such that for every ϕ, $\mu_{\phi,p}^{(\infty)} \geq \kappa \cdot \lambda$, where λ is the Lebesgue measure on $\mathcal{S}$.
5. Partial derivatives of transition densities are bounded.

Lemma: Under Assumption 1 and 5, with high probability

$$J_{\mathcal{M}}(\phi) \leq Index_t(\phi), \forall t = 1,2, \ldots, T.$$

Lemma: Under Assumption 1, 2, 3, 4 and 5, with high probability

$$Index_t(\phi) \leq J_{\mathcal{M}}(\phi) + const \cdot \int diam\left(\zeta_{s,\phi(s)}\right) d\mu_{\phi,p}^{(\infty)}(s), \quad \forall t = 1,2, \ldots, T.$$

Theorem: If $\mathcal{M}$ satisfies Assumption 1, 2, 3, 4 and 5, then with high probability

$$R_\Phi(T; PZRLMB) \leq \tilde{O}\left(T^{\frac{2d_{\mathcal{S}}+d_z^\Phi+2}{2d_{\mathcal{S}}+d_z^\Phi+3}}\right)$$

Zooming in Policy Space

Two Special Cases:

Parameterized policy space:

- $\phi(\cdot; w)$ – policy parameterized by w
- $w \in W \subset \mathbb{R}^{d_w}$
- $\|\phi(\cdot; w_1) - \phi(\cdot; w_2)\|_\nu \leq \|w_1 - w_2\|$

Then

$$d_z^\Phi \leq d_w$$

Zooming in Policy Space

Two Special Cases:

Parameterized policy space:

- $\phi(\cdot; w)$ – policy parameterized by w
- $w \in W \subset \mathbb{R}^{d_w}$
- $\|\phi(\cdot; w_1) - \phi(\cdot; w_2)\|_\nu \leq \|w_1 - w_2\|$

Then

$$d_z^\Phi \leq d_w$$

Curvature condition:

If there is a unique maximum of $J_\mathcal{M}(\phi(;\cdot)): W \to [0,1]$ at $w^\star$, and

$$J_\mathcal{M}(\phi(\cdot; w^\star)) - J_\mathcal{M}(\phi(\cdot; w)) \geq K_w \|w - w^\star\|, \forall w \in W.$$

Then

$$d_z^\Phi = 0$$

Zooming in Policy Space

Two Special Cases:

Parameterized policy space:
- $\phi(\cdot; w)$ – policy parameterized by w
- $w \in W \subset \mathbb{R}^{d_w}$
- $\|\phi(\cdot; w_1) - \phi(\cdot; w_2)\|_\nu \leq \|w_1 - w_2\|$

Then
$$d_z^\Phi \leq d_w$$

Curvature condition:

If there is a unique maximum of $J_\mathcal{M}(\phi(;\cdot)): W \to [0,1]$ at $w^\star$, and
$$J_\mathcal{M}(\phi(\cdot; w^\star)) - J_\mathcal{M}(\phi(\cdot; w)) \geq K_w \|w - w^\star\|, \forall w \in W.$$
Then
$$d_z^\Phi = 0$$

Drawbacks:

- Dimension of Φ could be huge
- Knowledge of low-complexity Φ may not be available
- Computationally heavy

Zooming in State-Action Space

Zooming in State-Action Space

<u>Zooming for RL (ZoRL)</u>
Inputs: Horizon T
Initialize: Set of $\mathcal{P}_0 = \mathcal{S} \times \mathcal{A}, k = 0, h = 0, H_0 = 0$
for $t = 1$ to T:
    if $h = H_k$:
        $h = 0, k \leftarrow k + 1$
        Update partition, $\mathcal{P}_t$
        Construct extended MDP, $\mathcal{M}_t^+$
        $\phi_k \leftarrow \mathrm{EVI}\left(\mathcal{M}_t^+, 1/\sqrt{T}\right)$
        $d_k \leftarrow EPE\left(\mathcal{M}_t^{d,+}, \phi_k, 1/\sqrt{T}\right)$
        $H_k = const \cdot d_k^{-2(d_\mathcal{S}+1)}$
    $h \leftarrow h + 1$
    Play $a_t = \phi_k(s_t)$

Zooming in State-Action Space

<u>Zooming for RL (ZoRL)</u>
Inputs: Horizon T
Initialize: Set of $\mathcal{P}_0 = \mathcal{S} \times \mathcal{A}, k = 0, h = 0, H_0 = 0$
for $t = 1$ to T:
  if $h = H_k$:
    $h = 0, k \leftarrow k + 1$
    Update partition, $\mathcal{P}_t$
    Construct extended MDP, $\mathcal{M}_t^+$
    $\phi_k \leftarrow \mathrm{EVI}\left(\mathcal{M}_t^+, 1/\sqrt{T}\right)$
    $d_k \leftarrow EPE\left(\mathcal{M}_t^{d,+}, \phi_k, 1/\sqrt{T}\right)$
    $H_k = const \cdot d_k^{-2(d_{\mathcal{S}}+1)}$
  $h \leftarrow h + 1$
  Play $a_t = \phi_k(s_t)$

- $\mathcal{M}_t^+ = (S_t, A_t, \mathcal{C}_t, r_t^+)$ where
  $r_t^+(s,a) = r(s,a) + L_r diam\left(\zeta_{s,a}^t\right)$
- $\mathcal{M}_t^{d,+} = (S_t, A_t, \mathcal{C}_t, d_t)$ where
  $d_t(s,a) = diam\left(\zeta_{s,a}^t\right)$

Zooming in State-Action Space

<u>Zooming for RL (ZoRL)</u>
Inputs: Horizon T
Initialize: Set of $\mathcal{P}_0 = \mathcal{S} \times \mathcal{A}$, $k = 0$, $h = 0$, $H_0 = 0$
for $t = 1$ to T:
  if $h = H_k$:
    $h = 0$, $k \leftarrow k + 1$
    Update partition, $\mathcal{P}_t$
    Construct extended MDP, $\mathcal{M}_t^+$
    $\phi_k \leftarrow \mathrm{EVI}\left(\mathcal{M}_t^+, 1/\sqrt{T}\right)$
    $d_k \leftarrow EPE\left(\mathcal{M}_t^{d,+}, \phi_k, 1/\sqrt{T}\right)$
    $H_k = const \cdot d_k^{-2(d_\mathcal{S}+1)}$
  $h \leftarrow h + 1$
  Play $a_t = \phi_k(s_t)$

- $\mathcal{M}_t^+ = (S_t, A_t, \mathcal{C}_t, r_t^+)$ where
  $r_t^+(s,a) = r(s,a) + L_r diam\left(\zeta_{s,a}^t\right)$
- $\mathcal{M}_t^{d,+} = (S_t, A_t, \mathcal{C}_t, d_t)$ where
  $d_t(s,a) = diam\left(\zeta_{s,a}^t\right)$

Consider the average reward optimality equation (AROE) of $\mathcal{M}$

$$J + h(s) = \max_{a \in \mathcal{A}}\left\{r(s,a) + \int p(s,a,ds')h(s')\right\}$$

Zooming in State-Action Space

Zooming for RL (ZoRL)

Inputs: Horizon T

Initialize: Set of $\mathcal{P}_0 = \mathcal{S} \times \mathcal{A}, k = 0, h = 0, H_0 = 0$

for $t = 1$ to T:

  if $h = H_k$:

    $h = 0, k \leftarrow k + 1$

    Update partition, $\mathcal{P}_t$

    Construct extended MDP, $\mathcal{M}_t^+$

    $\phi_k \leftarrow \mathrm{EVI}\left(\mathcal{M}_t^+, 1/\sqrt{T}\right)$

    $d_k \leftarrow EPE\left(\mathcal{M}_t^{d,+}, \phi_k, 1/\sqrt{T}\right)$

    $H_k = const \cdot d_k^{-2(d_{\mathcal{S}}+1)}$

  $h \leftarrow h + 1$

  Play $a_t = \phi_k(s_t)$

- $\mathcal{M}_t^+ = (S_t, A_t, \mathcal{C}_t, r_t^+)$ where
  $r_t^+(s,a) = r(s,a) + L_r diam\left(\zeta_{s,a}^t\right)$
- $\mathcal{M}_t^{d,+} = (S_t, A_t, \mathcal{C}_t, d_t)$ where
  $d_t(s,a) = diam\left(\zeta_{s,a}^t\right)$

Consider the average reward optimality equation (AROE) of $\mathcal{M}$

$$J + h(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \int p(s,a,ds')h(s') \right\}$$

$\star\ \exists h_{\mathcal{M}} : \mathcal{S} \to \mathbb{R}$ s.t.

$$J_{\mathcal{M}}^{\star} + h_{\mathcal{M}}(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \int p(s,a,ds')h_{\mathcal{M}}(s') \right\}$$

Zooming in State-Action Space

<u>Zooming for RL (ZoRL)</u>
Inputs: Horizon T
Initialize: Set of $\mathcal{P}_0 = \mathcal{S} \times \mathcal{A}, k = 0, h = 0, H_0 = 0$
for $t = 1$ to T:
  if $h = H_k$:
    $h = 0, k \leftarrow k + 1$
    Update partition, $\mathcal{P}_t$
    Construct extended MDP, $\mathcal{M}_t^+$
    $\phi_k \leftarrow \mathrm{EVI}(\mathcal{M}_t^+, 1/\sqrt{T})$
    $d_k \leftarrow EPE(\mathcal{M}_t^{d,+}, \phi_k, 1/\sqrt{T})$
    $H_k = const \cdot d_k^{-2(d_\mathcal{S}+1)}$
  $h \leftarrow h + 1$
  Play $a_t = \phi_k(s_t)$

- $\mathcal{M}_t^+ = (S_t, A_t, \mathcal{C}_t, r_t^+)$ where
  $r_t^+(s,a) = r(s,a) + L_r diam(\zeta_{s,a}^t)$
- $\mathcal{M}_t^{d,+} = (S_t, A_t, \mathcal{C}_t, d_t)$ where
  $d_t(s,a) = diam(\zeta_{s,a}^t)$

Consider the average reward optimality equation (AROE) of $\mathcal{M}$

$$J + h(s) = \max_{a\in\mathcal{A}}\{r(s,a) + \int p(s,a,ds')h(s')\}$$

⋆ $\exists h_\mathcal{M}: \mathcal{S} \rightarrow \mathbb{R}$ s.t.

$$J_\mathcal{M}^\star + h_\mathcal{M}(s) = \max_{a\in\mathcal{A}}\{r(s,a) + \int p(s,a,ds')h_\mathcal{M}(s')\}$$

⋆ $gap(s,a) \coloneqq J_\mathcal{M}^\star + h_\mathcal{M}(s) - \left(r(s,a) + \int p(s,a,ds')h_\mathcal{M}(s')\right)$

Zooming in State-Action Space

<u>Zooming for RL (ZoRL)</u>
Inputs: Horizon T
Initialize: Set of $\mathcal{P}_0 = \mathcal{S} \times \mathcal{A}, k = 0, h = 0, H_0 = 0$
for $t = 1$ to T:
  if $h = H_k$:
    $h = 0, k \leftarrow k + 1$
    Update partition, $\mathcal{P}_t$
    Construct extended MDP, $\mathcal{M}_t^+$
    $\phi_k \leftarrow \mathrm{EVI}\left(\mathcal{M}_t^+, 1/\sqrt{T}\right)$
    $d_k \leftarrow EPE\left(\mathcal{M}_t^{d,+}, \phi_k, 1/\sqrt{T}\right)$
    $H_k = const \cdot d_k^{-2(d_{\mathcal{S}}+1)}$
  $h \leftarrow h + 1$
  Play $a_t = \phi_k(s_t)$

- $\mathcal{M}_t^+ = (S_t, A_t, \mathcal{C}_t, r_t^+)$ where
  $r_t^+(s,a) = r(s,a) + L_r diam\left(\zeta_{s,a}^t\right)$
- $\mathcal{M}_t^{d,+} = (S_t, A_t, \mathcal{C}_t, d_t)$ where
  $d_t(s,a) = diam\left(\zeta_{s,a}^t\right)$

Consider the average reward optimality equation (AROE) of $\mathcal{M}$

$$J + h(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \int p(s,a,ds')h(s') \right\}$$

⋆ $\exists h_{\mathcal{M}}: \mathcal{S} \to \mathbb{R}$ s.t.

$$J_{\mathcal{M}}^\star + h_{\mathcal{M}}(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \int p(s,a,ds')h_{\mathcal{M}}(s') \right\}$$

⋆ $gap(s,a) \coloneqq J_{\mathcal{M}}^\star + h_{\mathcal{M}}(s) - \left(r(s,a) + \int p(s,a,ds')h_{\mathcal{M}}(s')\right)$

<u>Zooming dimension:</u>

$$d_z = \inf\left\{d > 0 : N_\gamma\left(\mathrm{Z}_\gamma\right) \leq c_z \gamma^{-d} \ \forall \gamma > 0\right\}$$

$where\ \mathrm{Z}_\gamma = \{(s,a) \mid gap(s,a) \leq \gamma\}.$

Zooming in State-Action Space

<u>Zooming for RL (ZoRL)</u>
Inputs: Horizon T
Initialize: Set of $\mathcal{P}_0 = \mathcal{S} \times \mathcal{A}$, $k = 0$, $h = 0$, $H_0 = 0$
for $t = 1$ to T:
  if $h = H_k$:
    $h = 0, k \leftarrow k + 1$
    Update partition, $\mathcal{P}_t$
    Construct extended MDP, $\mathcal{M}_t^+$
    $\phi_k \leftarrow \mathrm{EVI}(\mathcal{M}_t^+, 1/\sqrt{T})$
    $d_k \leftarrow EPE(\mathcal{M}_t^{d,+}, \phi_k, 1/\sqrt{T})$
    $H_k = const \cdot d_k^{-2(d_{\mathcal{S}}+1)}$
  $h \leftarrow h + 1$
  Play $a_t = \phi_k(s_t)$

- $\mathcal{M}_t^+ = (S_t, A_t, \mathcal{C}_t, r_t^+)$ where
  $r_t^+(s,a) = r(s,a) + L_r diam(\zeta_{s,a}^t)$
- $\mathcal{M}_t^{d,+} = (S_t, A_t, \mathcal{C}_t, d_t)$ where
  $d_t(s,a) = diam(\zeta_{s,a}^t)$

Consider the average reward optimality equation (AROE) of $\mathcal{M}$

$$J + h(s) = \max_{a \in \mathcal{A}} \{ r(s,a) + \int p(s,a,ds') h(s') \}$$

$\star$ $\exists h_{\mathcal{M}}: \mathcal{S} \to \mathbb{R}$ s.t.

$$J_{\mathcal{M}}^\star + h_{\mathcal{M}}(s) = \max_{a \in \mathcal{A}} \{ r(s,a) + \int p(s,a,ds') h_{\mathcal{M}}(s') \}$$

$\star$ $gap(s,a) \coloneqq J_{\mathcal{M}}^\star + h_{\mathcal{M}}(s) - \left(r(s,a) + \int p(s,a,ds') h_{\mathcal{M}}(s') \right)$

<u>Zooming dimension:</u>

$$d_z = \inf\{ d > 0 : N_\gamma(\mathrm{Z}_\gamma) \le c_z \gamma^{-d} \ \forall \gamma > 0 \}$$

where $\mathrm{Z}_\gamma = \{ (s,a) \mid gap(s,a) \le \gamma \}$.

$\star$ $d_z \le d_{\mathcal{S}} + d_{\mathcal{A}}$

Zooming in State-Action Space

Zooming for RL (ZoRL)
Inputs: Horizon T
Initialize: Set of $\mathcal{P}_0 = \mathcal{S} \times \mathcal{A}, k = 0, h = 0, H_0 = 0$
for $t = 1$ to T:
  if $h = H_k$:
    $h = 0, k \leftarrow k + 1$
    Update partition, $\mathcal{P}_t$
    Construct extended MDP, $\mathcal{M}_t^+$
    $\phi_k \leftarrow \mathrm{EVI}\left(\mathcal{M}_t^+, 1/\sqrt{T}\right)$
    $d_k \leftarrow EPE\left(\mathcal{M}_t^{d,+}, \phi_k, 1/\sqrt{T}\right)$
    $H_k = const \cdot d_k^{-2(d_{\mathcal{S}}+1)}$
  $h \leftarrow h + 1$
  Play $a_t = \phi_k(s_t)$

- $\mathcal{M}_t^+ = (S_t, A_t, \mathcal{C}_t, r_t^+)$ where
  $r_t^+(s,a) = r(s,a) + L_r diam\left(\zeta_{s,a}^t\right)$
- $\mathcal{M}_t^{d,+} = (S_t, A_t, \mathcal{C}_t, d_t)$ where
  $d_t(s,a) = diam\left(\zeta_{s,a}^t\right)$

Consider the average reward optimality equation (AROE) of $\mathcal{M}$

$$J + h(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \int p(s,a,ds')h(s') \right\}$$

$\star$ $\exists h_{\mathcal{M}}: \mathcal{S} \to \mathbb{R}$ s.t.

$$J_{\mathcal{M}}^\star + h_{\mathcal{M}}(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \int p(s,a,ds')h_{\mathcal{M}}(s') \right\}$$

$\star$ $gap(s,a) \coloneqq J_{\mathcal{M}}^\star + h_{\mathcal{M}}(s) - \left(r(s,a) + \int p(s,a,ds')h_{\mathcal{M}}(s')\right)$

Zooming dimension:

$$d_z = \inf\left\{d > 0 : N_\gamma\left(\mathrm{Z}_\gamma\right) \le c_z \gamma^{-d}\ \forall \gamma > 0\right\}$$

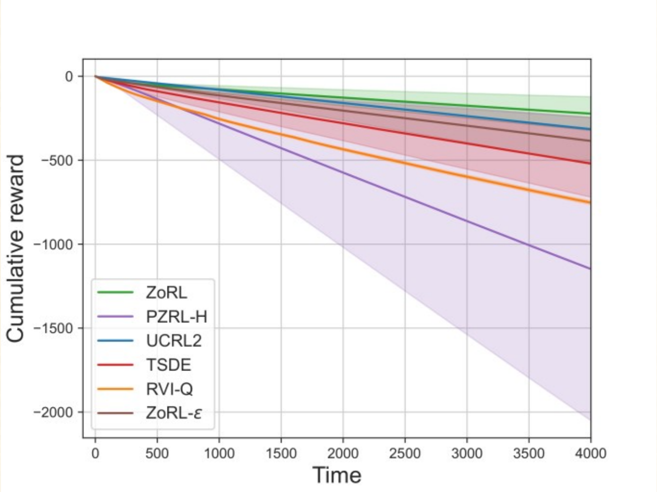
$where\ \mathrm{Z}_\gamma = \{(s,a) \mid gap(s,a) \le \gamma\}$.

$\star$ $d_z \le d_{\mathcal{S}} + d_{\mathcal{A}}$

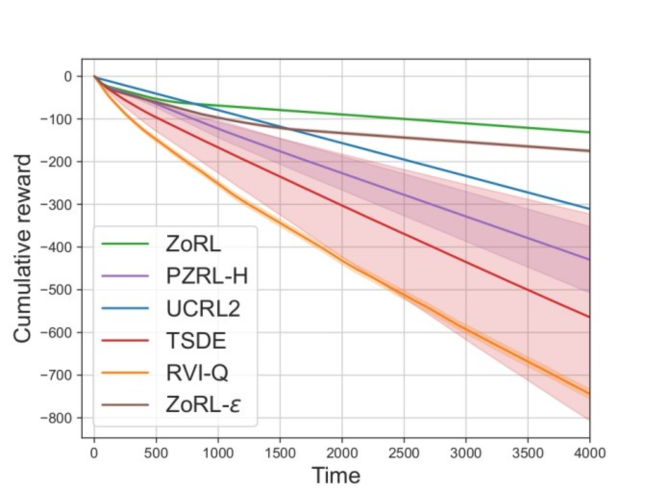
Theorem: If $\mathcal{M}$ satisfies Assumption 1, 2, 3, 4 and 5, then with high probability

$$R(T; ZoRL) \le \tilde{O}\left(T^{\frac{2d_{\mathcal{S}}+d_z+2}{2d_{\mathcal{S}}+d_z+3}}\right)$$

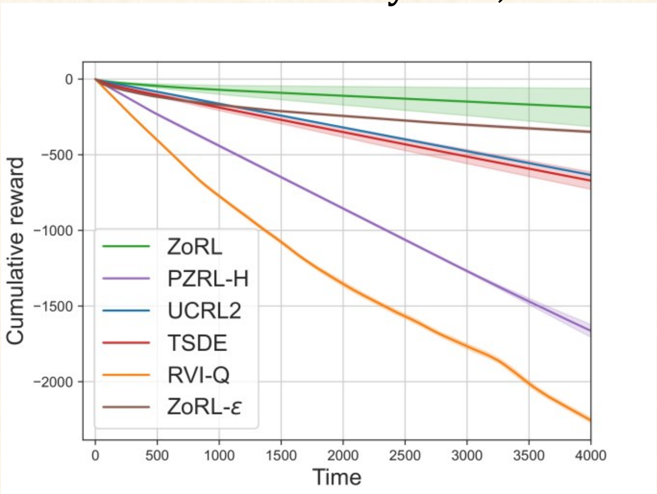
Zooming in State-Action Space



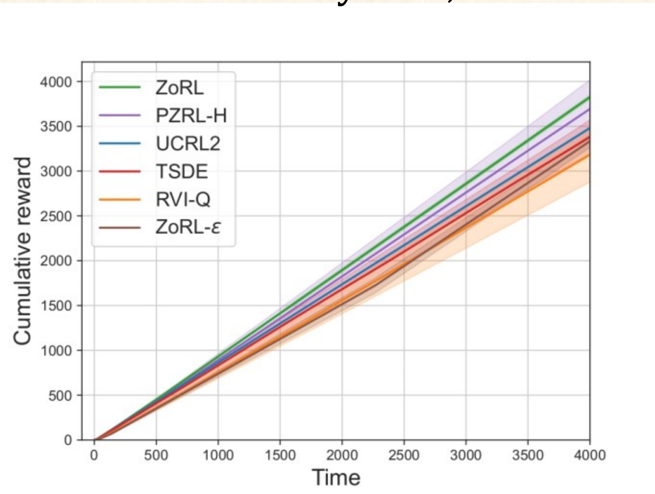
Truncated Linear System, 2x2



Nonlinear System, 2x2

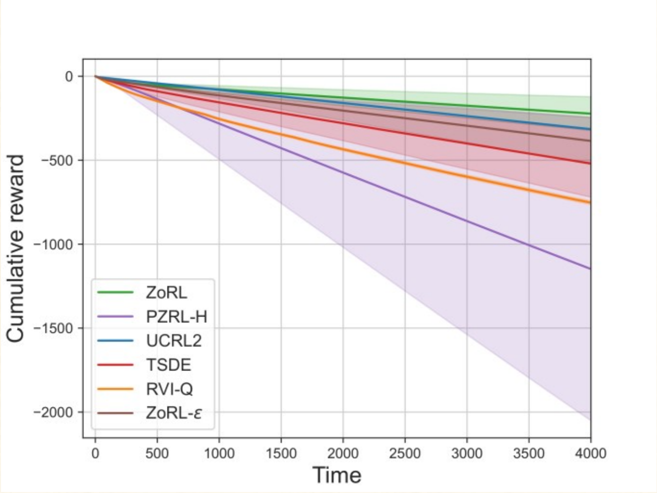


Truncated Linear System, 2x4

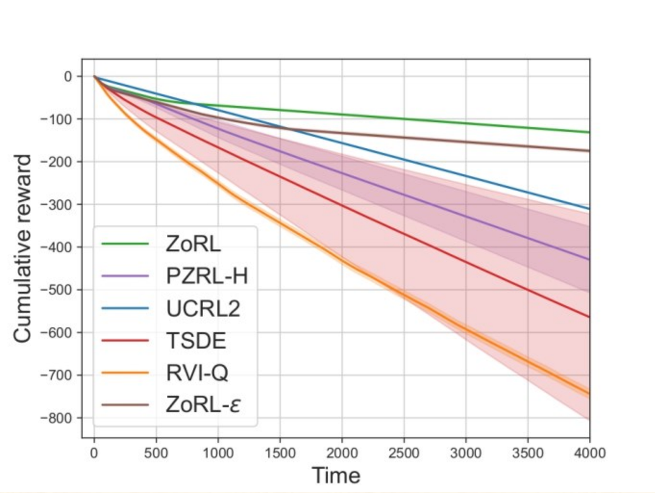


Continuous RiverSwim

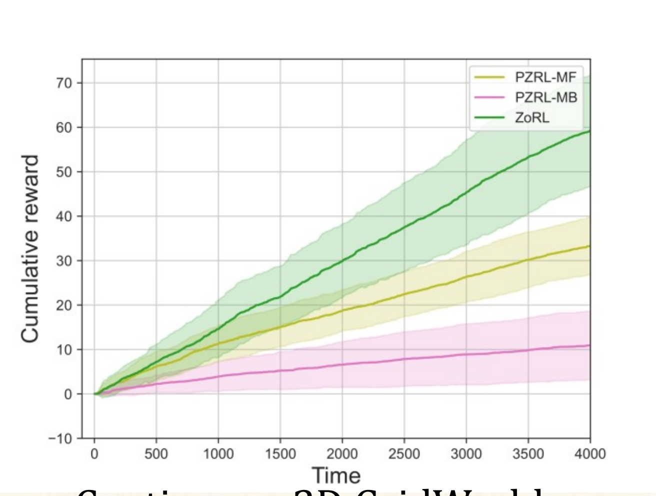
Zooming in State-Action Space



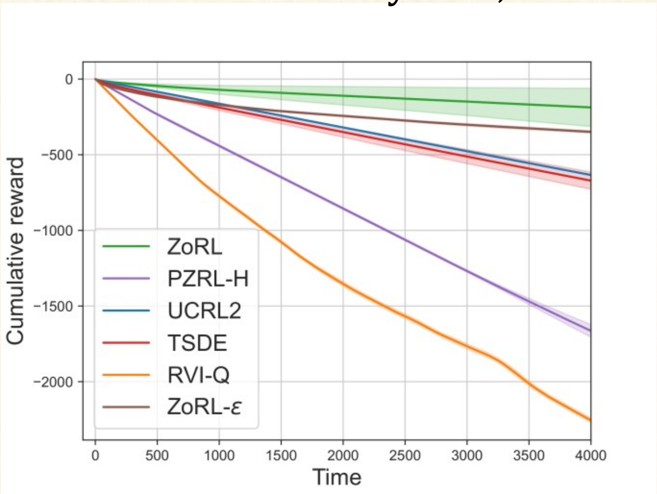
Truncated Linear System, 2x2



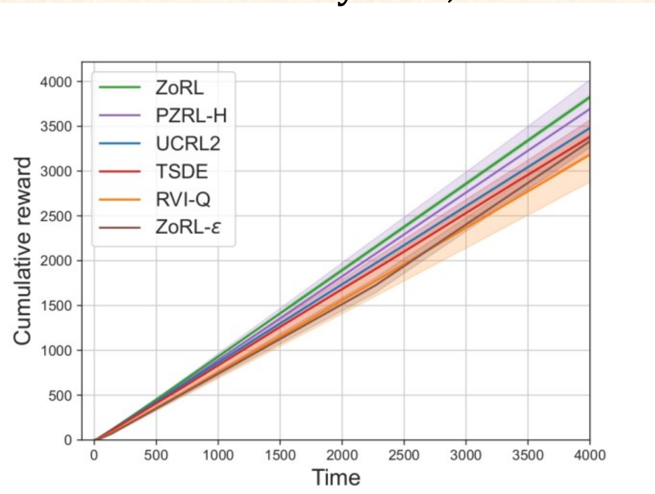
Nonlinear System, 2x2



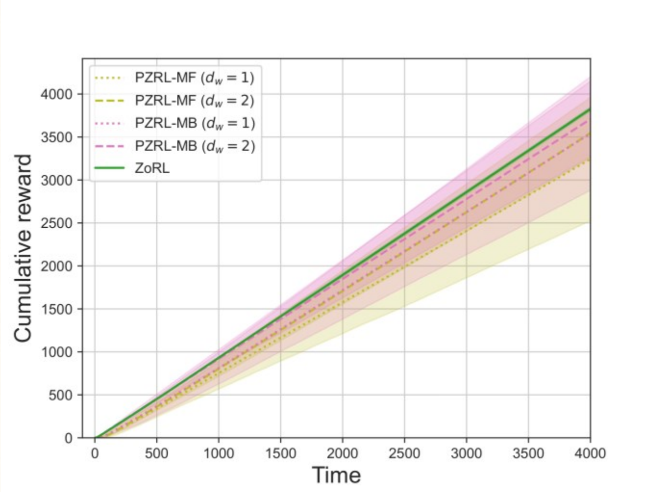
Continuous 2D GridWorld



Truncated Linear System, 2x4



Continuous RiverSwim



Continuous RiverSwim

Zooming in State-Action Space

ZoRL

I. is statistically efficient
II. covers a broad class of MDPs
III. is computationally efficient

Zooming in State-Action Space

ZoRL

I. is statistically efficient
II. covers a broad class of MDPs
III. is computationally efficient

Future directions:

I. Generalization of zooming idea beyond Lipschitz assumption
II. Relaxation of other assumptions such as ergodicity
III. Exploring connection with existing complexity measures

Zooming in State-Action Space

ZoRL

I. is statistically efficient
II. covers a broad class of MDPs
III. is computationally efficient

Future directions:

I. Generalization of zooming idea beyond Lipschitz assumption
II. Relaxation of other assumptions such as ergodicity
III. Exploring connection with existing complexity measures

Preprints:

1. Kar, Avik, and Rahul Singh. "Provably Adaptive Average Reward Reinforcement Learning for Metric Spaces." arXiv preprint arXiv:2410.19919 (2024).
2. Kar, Avik, and Rahul Singh. "Policy Zooming: Adaptive Discretization-based Infinite-Horizon Average-Reward Reinforcement Learning." arXiv preprint arXiv:2405.18793 (2024).

Thank you!