A Distributional Analysis of Sampling-Based Reinforcement Learning Algorithms

Philip Amortila

Joint work w/: Doina Precup, Prakash Panangaden, Marc G. Bellemare, Nan Jiang

McGill University, Google Brain, UIUC

April 10th 2020

Spoilers

- Mathematical tool to study stochastic RL algorithms
- Analysis is much easier (generalization of bread-and-butter proof techniques)
- Direct tie-in to practical applications
- Progress towards open questions about convergence of difficult algorithms

Dynamic Programming 101

Markov Decision Process (MDP) task:

- Given an MDP, find the policy which maximizes lifetime returns

Expected performance of a policy π:

$$V^{\pi}(s) = \mathbb{E}_{\mathtt{MDP}}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

Value function is the fixed point of the $\mathcal{T}^{\pi}$:

$$V^{\pi} = \mathcal{T}^{\pi} V^{\pi} := R^{\pi} + \gamma P^{\pi} V^{\pi}$$

Value function of optimal policy $\pi^{\star}$ is the fixed point of $\mathcal{T}^{\star}$:

$$V^{\star} = \mathcal{T}^{\star} V^{\star} := \max_{\pi} \mathcal{T}^{\pi} V^{\star}$$

Dynamic Programming 102

Policy evaluation algorithm:

$$V_{n+1}(s) = \mathcal{T}^{\pi} V_n(s)$$

- Proof of convergence to V^{π}: contraction property of $\mathcal{T}^{\pi}$ and the Banach fixed point theorem.

Dynamic Programming 102

Policy evaluation algorithm:

$$V_{n+1}(s) = \mathcal{T}^{\pi} V_n(s)$$

- Proof of convergence to V^{π}: contraction property of $\mathcal{T}^{\pi}$ and the Banach fixed point theorem.

Policy iteration algorithm:

$$\begin{cases} \text{evaluate } V^{\pi_n} \\ \text{set } \pi_{n+1} = \text{greedy}(V_n^{\pi}) \end{cases}$$

- Proof of convergence to $\pi^{\star}$: monotonicity property of $\mathcal{T}^{\pi}$.

RL Proofs are Hard

In the Reinforcement Learning setting, we cannot evaluate $\mathcal{T}^\pi$ or $\mathcal{T}^\star$.

RL Proofs are Hard

In the Reinforcement Learning setting, we cannot evaluate $\mathcal{T}^\pi$ or $\mathcal{T}^\star$. Approximate them via *sampling*, e.g. TD(0) algorithm:

$$V_{n+1}(s) = (1-\alpha)V_n(s)+\alpha(r+\gamma V_n(s')) \quad \leftarrow \begin{cases} a \sim \pi(\cdot|s) \\ r, s' \sim \text{MDP} \end{cases}$$

- Proof of convergence: more involved due to sampling. Involves stochastic approximation theory.

Constant step-sizes

$$\texttt{TD(0)}: \quad V_{n+1}(s) = (1-\alpha)V_n(s) + \alpha(r + \gamma V_n(s')) \leftarrow \begin{cases} a \sim \pi(\cdot|s) \\ r, s' \sim \text{MDP} \end{cases}$$

- For constant step-sizes, the estimates will not converge to a single point estimate in general.
- Does there exist a limiting behaviour of the algorithm that is stationary?
    - Running another iteration of the algorithm keeps this larger behaviour unchanged.

A Distributional Analysis

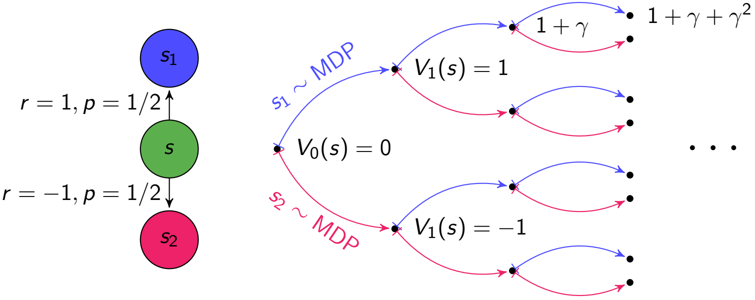
$$\texttt{TD(0)}: \quad V_{n+1}(s) = (1-\alpha)V_n(s) + \alpha(r + \gamma V_n(s')) \;\leftarrow \begin{cases} a \sim \pi(\cdot|s) \\ r, s' \sim \text{MDP} \end{cases}$$

The functions V_n obtained from sample-based algorithms are *random variables*. We study their distributions:

A Distributional Analysis

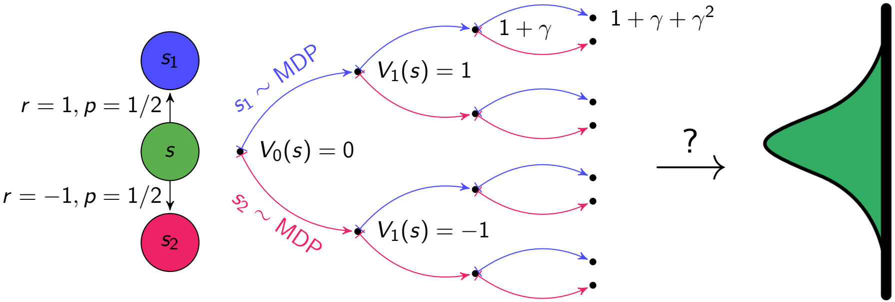
$$\texttt{TD(0)}: \quad V_{n+1}(s) = (1-\alpha)V_n(s) + \alpha(r + \gamma V_n(s')) \leftarrow \begin{cases} a \sim \pi(\cdot|s) \\ r, s' \sim \text{MDP} \end{cases}$$

The functions V_n obtained from sample-based algorithms are *random variables*. We study their distributions:



Convergence of distributions

Does the sequence of distributions converge? To which limit?



A Distributional Equation

$$\texttt{TD(0)}: \; V_{n+1}(s) \overset{D}{=} (1-\alpha)V_n(s) + \alpha(R(s,A) + \gamma V_n(S')) \quad \leftarrow \begin{cases} A \sim \pi(\cdot|s) \\ R, S' \sim \text{MDP} \end{cases}$$

- A similar equation can be written for any sampling-based algorithm
  - → Monte Carlo
  - → TD(λ)
  - → Q-Learning
  - → SARSA
  - → Double Q-Learning
  - → etc...

A Distributional Equation

$$\texttt{TD(0)}: \; V_{n+1}(s) \overset{D}{=} (1-\alpha)V_n(s) + \alpha(R(s,A) + \gamma V_n(S')) \quad \leftarrow \begin{cases} A \sim \pi(\cdot|s) \\ R, S' \sim \text{MDP} \end{cases}$$

- A similar equation can be written for any sampling-based algorithm
  - $\rightarrow$ Monte Carlo
  - $\rightarrow$ TD(λ)
  - $\rightarrow$ Q-Learning
  - $\rightarrow$ SARSA
  - $\rightarrow$ Double Q-Learning
  - $\rightarrow$ etc...
- These equations define Markov chains over space of value functions

A Distributional Equation

$$\texttt{TD(0)}: \; V_{n+1}(s) \overset{D}{=} (1-\alpha)V_n(s) + \alpha(R(s,A) + \gamma V_n(S')) \quad \leftarrow \begin{cases} A \sim \pi(\cdot|s) \\ R, S' \sim \text{MDP} \end{cases}$$

- A similar equation can be written for any sampling-based algorithm
  - → Monte Carlo
  - → TD(λ)
  - → Q-Learning
  - → SARSA
  - → Double Q-Learning
  - → etc...
- These equations define Markov chains over space of value functions
- Study this question for the case of constant step-sizes and synchronous updates.
  - → Markov chains are homogeneous
- Inspired by Dieuleveut, Durmus, Bach (2017)

A Distributional Equation

$$\texttt{TD(0)}: \; V_{n+1}(s) \overset{D}{=} (1-\alpha)V_n(s) + \alpha(R(s,A) + \gamma V_n(S')) \quad \leftarrow \begin{cases} A \sim \pi(\cdot|s) \\ R, S' \sim \text{MDP} \end{cases}$$

- A similar equation can be written for any sampling-based algorithm
  - → Monte Carlo
  - → TD(λ)
  - → Q-Learning
  - → SARSA
  - → Double Q-Learning
  - → etc...
- These equations define Markov chains over space of value functions
- Study this question for the case of constant step-sizes and synchronous updates.
  - → Markov chains are homogeneous
- Inspired by Dieuleveut, Durmus, Bach (2017)
- Special case: TD(0) with $\alpha = 1$ is the distributional RL operator

Operator between distributions

For any update rule and step-size, consider its Markov kernel K

$$K(V_n, \mathcal{B}) = \mathbb{P}\{V_{n+1} \in \mathcal{B} \mid V_n\}, \ \mathcal{B} \in \texttt{Borel}(\mathbb{R}^n)$$

Lift stochastic update rule to operator over distributions:

$$\begin{aligned} V_n &\sim \mu_n \\ V_{n+1} &\sim \mu_{n+1} = (\mu_n)K = (\mu_0)K^{n+1}. \end{aligned}$$

Convergence of stochastic processes

- Measuring convergence of Markov chains requires a metric between probability distributions

Convergence of stochastic processes

- Measuring convergence of Markov chains requires a metric between probability distributions
- Common choice in the Markov chain literature is the Total Variation metric

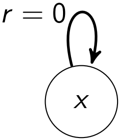
$$d_{\mathtt{TV}}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$$

Convergence of stochastic processes

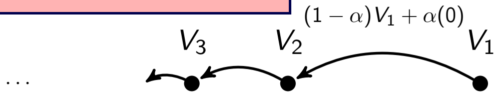
- Measuring convergence of Markov chains requires a metric between probability distributions
- Common choice in the Markov chain literature is the Total Variation metric

$$d_{\mathtt{TV}}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$$

- Will not work for us!



$$d_{\mathtt{TV}}(\delta_0, \delta_{V_n}) = 1 \quad \forall n$$



Wasserstein metric

- We use the *Wasserstein* metric between probability distributions

$$\mathcal{W}(\mu, \nu) = \inf_{\substack{X \sim \mu \\ Y \sim \nu}} \mathbb{E}\left[\|X - Y\|_\infty\right]$$

Wasserstein metric

- We use the *Wasserstein* metric between probability distributions

$$\mathcal{W}(\mu, \nu) = \inf_{\substack{X \sim \mu \\ Y \sim \nu}} \mathbb{E}\left[\|X - Y\|_\infty\right]$$

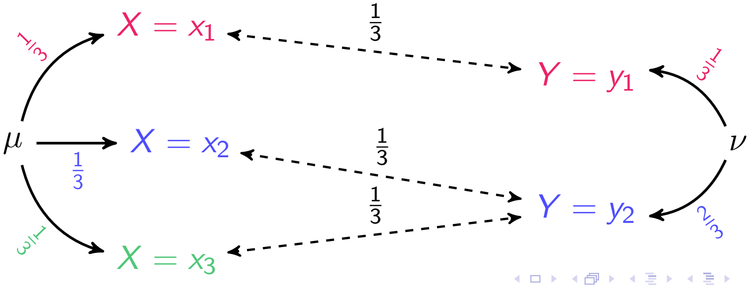
- Our choice of cost function: $\|\cdot\|_\infty$
- Minimization over couplings: pairs of random variables (X, Y) such that $X \sim \mu, Y \sim \nu$ marginally

Wasserstein metric

- We use the *Wasserstein* metric between probability distributions

$$\mathcal{W}(\mu, \nu) = \inf_{\substack{X \sim \mu \\ Y \sim \nu}} \mathbb{E}\left[\|X - Y\|_\infty\right]$$

- Our choice of cost function: $\|\cdot\|_\infty$
- Minimization over couplings: pairs of random variables (X, Y) such that $X \sim \mu, Y \sim \nu$ marginally

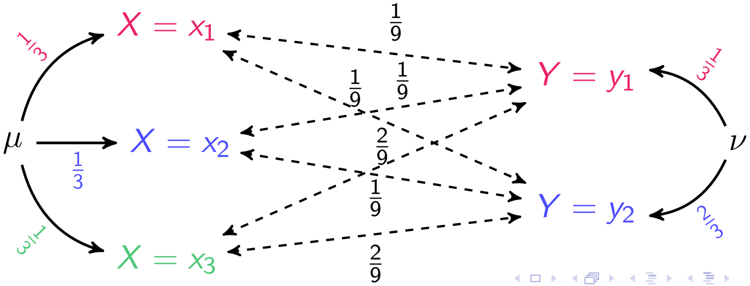


Wasserstein metric

- We use the *Wasserstein* metric between probability distributions

$$\mathcal{W}(\mu, \nu) = \inf_{\substack{X \sim \mu \\ Y \sim \nu}} \mathbb{E}\left[\|X - Y\|_\infty\right]$$

- Our choice of cost function: $\|\cdot\|_\infty$
- Minimization over couplings: pairs of random variables (X, Y) such that $X \sim \mu, Y \sim \nu$ marginally



Contraction in the space of distributions on functions

Punchline: For TD(0), the induced operator K is contractive with respect to the Wasserstein metric

$$\mathcal{W}(\mu K, \nu K) \leq \underbrace{(1 - \alpha + \alpha\gamma)}_{<1} \mathcal{W}(\mu, \nu),$$

Contraction in the space of distributions on functions

Punchline: For TD(0), the induced operator K is contractive with respect to the Wasserstein metric

$$\mathcal{W}(\mu K, \nu K) \leq \underbrace{(1 - \alpha + \alpha\gamma)}_{<1} \mathcal{W}(\mu, \nu),$$

By Banach's fixed point theorem the distributions μK^n converge to a fixed point

$$\psi = \psi K.$$

This is exactly the property of a stationary distribution!

Analogies

We have:

Bellman operator $\mathcal{T}^\pi$:

- Contraction with respect to $\|\cdot\|_\infty$ w/ factor γ
- Unique fixed point $V^\pi \in \mathbb{R}^{|\mathcal{S}|}$

Analogies

We have:

Bellman operator $\mathcal{T}^\pi$:

- Contraction with respect to $\|\cdot\|_\infty$ w/ factor γ
- Unique fixed point $V^\pi \in \mathbb{R}^{|\mathcal{S}|}$

TD(0) (with policy π, step-size α):

- contraction with respect to $\mathcal{W}_{\|\cdot\|_\infty}$ w/ factor $1 - \alpha + \alpha\gamma$
- Unique fixed point $\psi^{\pi,\alpha,\texttt{TD(0)}} \in \texttt{Dists}(\mathbb{R}^{|\mathcal{S}|})$

Contractive Algorithms

For any stepsizes $\alpha \in (0, 1]$, the following algorithms are contractive:

- Monte Carlo Evaluation w/ factor $1 - \alpha$
- TD(λ) w/ factor $1 - \alpha + \alpha\gamma\frac{1-\lambda}{1-\lambda\gamma}$
- SARSA & Expected SARSA w/ factor $1 - \alpha + \alpha\gamma$
- Q-Learning w/ factor $1 - \alpha + \alpha\gamma$
- Double Q-Learning w/ factor $\frac{1}{2}(2 - \alpha + \alpha\gamma)$

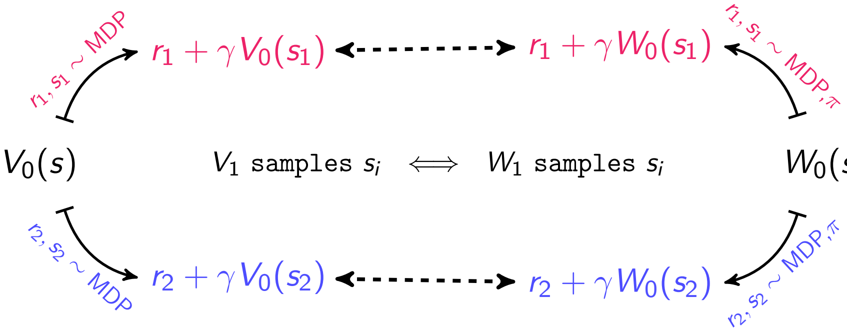
The same proof technique extends to all the above algorithms.

Contractive Algorithms

Proof (TD(0)): Consider $V_0 \sim \mu, W_0 \sim \nu$. Define a coupling of V_1 and W_1 as follows:

Contractive Algorithms

Proof (TD(0)): Consider $V_0 \sim \mu, W_0 \sim \nu$. Define a coupling of V_1 and W_1 as follows:



Proof (TD(0)) (continued)

Proof (TD(0)): Distance between the targets (under the coupling):

$$\mathbb{E}_{\texttt{coupling}}\left[\max_s |r + \gamma V_0(s') - r - \gamma W_0(s')|\right] = \gamma \mathbb{E}\left[\max_s |V_0(s') - W_0(s')|\right]$$
$$\leq \gamma \mathbb{E}\left[\|V_0 - W_0\|_\infty\right]$$

Proof (TD(0)) (continued)

Proof (TD(0)): Distance between the targets (under the coupling):

$$\mathbb{E}_{\texttt{coupling}}\left[\max_s |r + \gamma V_0(s') - r - \gamma W_0(s')|\right] = \gamma \mathbb{E}\left[\max_s |V_0(s') - W_0(s')|\right]$$
$$\leq \gamma \mathbb{E}\left[\|V_0 - W_0\|_\infty\right]$$

Upper bound $\mathcal{W}(\mu K, \nu K)$ by the coupling:

$$\mathcal{W}(\mu K, \nu K) \leq (1-\alpha)\mathcal{W}(\mu, \nu) + \alpha\gamma\mathbb{E}\left[\|V_0 - W_0\|_\infty\right]$$
$$= \underbrace{(1 - \alpha + \alpha\gamma)}_{<1} \mathcal{W}(\mu, \nu)$$

Stationary distributions

Q: If an algorithm with constant step-sizes converges, what is its stationary distribution?

Q: If an algorithm with constant step-sizes converges, what is its stationary distribution?

Suppose an algorithm has the form

$$f_{n+1} = (1-\alpha) f_n + \alpha \underbrace{\hat{\mathcal{T}}(f_n)}_{\texttt{target}},$$

Evaluation setting

If the stochastic updates are, in expectation, a Bellman operator of π

$$\mathbb{E}_{\texttt{sampling}}[\hat{\mathcal{T}} f] = \mathcal{T}^{\pi} f, \quad \forall f$$

then:

Evaluation setting

If the stochastic updates are, in expectation, a Bellman operator of π

$$\mathbb{E}_{\texttt{sampling}}[\hat{\mathcal{T}} f] = \mathcal{T}^{\pi} f, \quad \forall f$$

then:

- The mean of the stationary distributions is the true value function (V^{π} or Q^{π})

Evaluation setting

If the stochastic updates are, in expectation, a Bellman operator of π

$$\mathbb{E}_{\texttt{sampling}}[\hat{\mathcal{T}} f] = \mathcal{T}^\pi f, \quad \forall f$$

then:

- The mean of the stationary distributions is the true value function (V^π or Q^π)
- The covariance is linear in the step-size and the covariance of $\hat{\mathcal{T}} f - \mathcal{T}^\pi f$

Evaluation setting

If the stochastic updates are, in expectation, a Bellman operator of π

$$\mathbb{E}_{\texttt{sampling}}[\hat{\mathcal{T}} f] = \mathcal{T}^\pi f, \quad \forall f$$

then:

- The mean of the stationary distributions is the true value function (V^π or Q^π)
- The covariance is linear in the step-size and the covariance of $\hat{\mathcal{T}} f - \mathcal{T}^\pi f$
- The distributions concentrate around these means:

$$\mathbb{P}_{f_\alpha \sim \texttt{stationary dist.}} \left\{ \min_i |f_\alpha(i) - f^\pi(i)| \geq \varepsilon \right\} \xrightarrow{\alpha \to 0} 0$$

Control setting

If the stochastic updates are, in expectation, a Bellman *optimality* operator

$$\mathbb{E}[\hat{\mathcal{T}} f] = \mathcal{T}^{\star} f, \quad \forall f$$

then:

Control setting

If the stochastic updates are, in expectation, a Bellman *optimality* operator

$$\mathbb{E}[\hat{\mathcal{T}} f] = \mathcal{T}^{\star} f, \quad \forall f$$

then:

- Mean of the stationary distribution *overestimates* the true value function ($V^{\star}$ or $Q^{\star}$)

A non-contractive example: Optimistic Policy Iteration

- Algorithms previously seen were sampling analogues of contractive mappings.
- What about stochastic analogues of policy improvement algorithms?

A non-contractive example: Optimistic Policy Iteration

- Algorithms previously seen were sampling analogues of contractive mappings.
- What about stochastic analogues of policy improvement algorithms?

We study the Optimistic Policy Iteration (OPI) algorithm

$$Q_{n+1}(s,a) = (1-\alpha)Q_n(s,a) + \alpha \mathcal{G}^{\pi_n}(s,a),$$
$$\pi_{n+1} = \text{greedy}\left(Q_{n+1}\right)$$

where $\mathcal{G}^{\pi}(s,a)$ is a discounted return sampled from the MDP using π.

Optimistic Policy Iteration

- The analysis of this method is not straightforward with typical stochastic approximation techniques
- Convergence known only in limited cases (Robbins-Monro step-sizes and sampling conditions)
- Contraction does not hold for classic policy iteration or its sampling-based variant
- Simple coupling argument ruled out: different functions have different sampling distributions

Proof via greedy partitions

- Special case of $\alpha = 1$

$$\begin{aligned} Q_{n+1}(s,a) &= \mathcal{G}^{\pi_n}(s,a), \\ \pi_{n+1} &= \text{greedy}\left(Q_{n+1}\right) \end{aligned}$$

Proof via greedy partitions

- Special case of $\alpha = 1$

$$\begin{aligned} Q_{n+1}(s,a) &= \mathcal{G}^{\pi_n}(s,a), \\ \pi_{n+1} &= \text{greedy}\left(Q_{n+1}\right) \end{aligned}$$

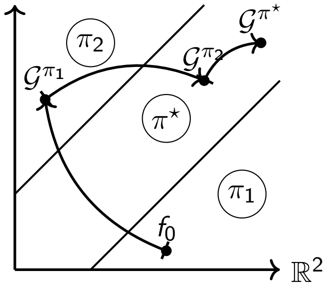
- Here the algorithm is Markovian over the *greedy partition* of $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$

Proof via greedy partitions

- Special case of $\alpha = 1$

$$Q_{n+1}(s,a) = \mathcal{G}^{\pi_n}(s,a), \\ \pi_{n+1} = \text{greedy}\left(Q_{n+1}\right)$$

- Here the algorithm is Markovian over the *greedy partition* of $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$
- This is a finite state Markov chain



Proof via greedy partitions

- Probabilistic policy improvement:

$$\mathbb{P}\left\{\texttt{sampling } \mathcal{G}^{\pi_n} \texttt{ that has } \operatorname{greedy}(\mathcal{G}^{\pi_n}) = \operatorname{greedy}(Q^{\pi_n})\right\} > 0$$

Proof via greedy partitions

- Probabilistic policy improvement:

$$\mathbb{P}\left\{\texttt{sampling } \mathcal{G}^{\pi_n} \texttt{ that has } \text{greedy}(\mathcal{G}^{\pi_n}) = \text{greedy}(Q^{\pi_n})\right\} > 0$$

- Therefore every initial policy π can reach $\pi^\star$ with some probability...

Proof via greedy partitions

- Probabilistic policy improvement:

$$\mathbb{P}\left\{\texttt{sampling } \mathcal{G}^{\pi_n} \texttt{ that has } \text{greedy}(\mathcal{G}^{\pi_n}) = \text{greedy}(Q^{\pi_n})\right\} > 0$$

- Therefore every initial policy π can reach $\pi^\star$ with some probability...
- ...and $\pi^\star$ is a recurrent state

Proof via greedy partitions

- Probabilistic policy improvement:

$$\mathbb{P}\left\{\texttt{sampling } \mathcal{G}^{\pi_n} \texttt{ that has } \mathrm{greedy}(\mathcal{G}^{\pi_n}) = \mathrm{greedy}(Q^{\pi_n})\right\} > 0$$

- Therefore every initial policy π can reach $\pi^\star$ with some probability...
- ...and $\pi^\star$ is a recurrent state
- So the Markov chain is ergodic and converges to a stationary distribution over policies!

Difficulty of the $\alpha < 1$ case

- The analysis does not quite extend to the general case of $\alpha < 1$

Difficulty of the $\alpha < 1$ case

- The analysis does not quite extend to the general case of $\alpha < 1$
- No longer Markovian over policies, now on the continuous space of value functions

Difficulty of the $\alpha < 1$ case

- The analysis does not quite extend to the general case of $\alpha < 1$
- No longer Markovian over policies, now on the continuous space of value functions
- Continuous space ergodic theorems require "smoothness" properties
  - Not satisfied by this algorithm
  - Discontinuous at the boundary between greedy partitions

Boltzmann OPI

For $\alpha < 1$, can establish convergence for a variant that uses Boltzmann (softmax) policies

$$\pi_{f,\beta}(a|s) = \frac{\exp(\beta f(s,a))}{\sum_a \exp(\beta f(s,a))}, \quad \beta > 0$$

This system is *Lyapunov stable* with respect to Wasserstein metric:

$$\lim_{\nu \to \mu} \sup_{n \geq 0} \mathcal{W}(\nu K^n, \mu K^n) = 0$$

$\rightarrow$ (via. another simple coupling argument)

Establishes convergence when combined with reachability and aperiodicity of $\pi^\star$, as before.

Future work

Future work

- Decreasing step-sizes and/or online updates
  - $\rightarrow$ Corresponds to time-dependent Markov chains
  - $\rightarrow$ Applying a sequence of contractive kernels $\mu K_{\alpha_1} K_{\alpha_2} \cdots K_{\alpha_n}$
- Function approximation
  - $\rightarrow$ Preliminary results for linear function approximation
- Optimistic Policy Iteration and other stochastic policy iteration methods (e.g. actor-critic methods)

Merci