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
• 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
Markov Decision Process (MDP) task:

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

Expected performance of a policy $\pi$:

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

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

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

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

$$V^\star = T^\star V^\star := \max_\pi T^\pi V^\star$$
Policy evaluation algorithm:

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

- Proof of convergence to \( V^\pi \): contraction property of \( \mathcal{T}^\pi \) and the Banach fixed point theorem.
Policy evaluation algorithm:

\[ V_{n+1}(s) = T^\pi V_n(s) \]

- Proof of convergence to \( V^\pi \): contraction property of \( T^\pi \) and the Banach fixed point theorem.

Policy iteration algorithm:

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

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

$$V_n+1(s) = (1-\alpha)V_n(s) + \alpha(r + \gamma V_n(s')) \leftarrow \{ a \sim \pi(\cdot | s), r, s' \sim MDP \}$$

Proof of convergence: more involved due to sampling. Involves stochastic approximation theory.
In the Reinforcement Learning setting, we cannot evaluate $\mathcal{T}_\pi$ or $\mathcal{T}^\star$. Approximate them via *sampling*, e.g. TD(0) algorithm:

$$V_{n+1}(s) = (1-\alpha)V_n(s) + \alpha(r + \gamma V_n(s'))$$

$$\left\{ \begin{array}{l}
  a \sim \pi(\cdot|s) \\
  r, s' \sim \text{MDP}
\end{array} \right.$$  

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

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

- 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

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

The functions \( V_n \) obtained from sample-based algorithms are random variables. We study their distributions:
A Distributional Analysis

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

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

- $s_1 \sim MDP$
  - $r = 1, p = 1/2$
  - $V_0(s) = 0$
  - $V_1(s) = 1$

- $s \sim MDP$
  - $r = -1, p = 1/2$
  - $V_0(s) = 0$
  - $V_1(s) = -1$

- $s_2 \sim MDP$
  - $1 + \gamma$
  - $1 + \gamma + \gamma^2$

- \ldots
Does the sequence of distributions converge? To which limit?

\[ V_0(s) = 0 \]

\[ V_1(s) = 1 \]

\[ s_1 \sim \text{MDP} \]

\[ r = 1, p = 1/2 \]

\[ s \]

\[ s_2 \sim \text{MDP} \]

\[ r = -1, p = 1/2 \]

\[ V_1(s) = -1 \]
A Distributional Equation

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

- A similar equation can be written for any sampling-based algorithm
  - Monte Carlo
  - TD(\lambda)
  - Q-Learning
  - SARSA
  - Double Q-Learning
  - etc...
A Distributional Equation

\[ V_{n+1}(s) = (1-\alpha)V_n(s) + \alpha(R(s, A) + \gamma V_n(S')) \]

\[ \sim \left\{ \begin{array}{l}
A \sim \pi(\cdot|s) \\
R, S' \sim MDP
\end{array} \right. \]

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

- These equations define Markov chains over space of value functions
A Distributional Equation

\begin{align*}
TD(0): \quad V_{n+1}(s) & \overset{D}{=} (1-\alpha)V_n(s) + \alpha (R(s,A) + \gamma V_n(S')) \\
\end{align*}

- A similar equation can be written for any sampling-based algorithm
  - Monte Carlo
  - TD(\lambda)
  - 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

\[ TD(0) : \quad V_{n+1}(s) \overset{D}{=} (1-\alpha)V_n(s) + \alpha(R(s,A) + \gamma V_n(S')) \quad \left\{ \begin{array}{l} A \sim \pi(\cdot|s) \\ R, S' \sim \text{MDP} \end{array} \right. \]

- A similar equation can be written for any sampling-based algorithm
  → Monte Carlo
  → TD(\lambda)
  → 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
For any update rule and step-size, consider its Markov kernel $K$

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

Lift stochastic update rule to operator over distributions:

$$V_n \sim \mu_n$$

$$V_{n+1} \sim \mu_{n+1} = (\mu_n)K = (\mu_0)K^{n+1}.$$
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:

\[
d_{TV}(\mu, \nu) = \sup_{A} |\mu(A) - \nu(A)|
\]
• Measuring convergence of Markov chains requires a metric between probability distributions

• Common choice in the Markov chain literature is the Total Variation metric

\[
d_{TV}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|
\]

• Will not work for us!

\[
d_{TV}(\delta_0, \delta_{V_n}) = 1 \quad \forall n
\]
• We use the *Wasserstein* metric between probability distributions

\[
\mathcal{W}(\mu, \nu) = \inf_{X \sim \mu, Y \sim \nu} \mathbb{E} [\|X - Y\|_\infty]
\]
Wasserstein metric

- We use the Wasserstein metric between probability distributions

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

- 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

\[ \mu \xrightarrow{1/3} X = x_1 \xrightarrow{\text{C1}} X = x_2 \xrightarrow{\text{C2}} X = x_3 \]

\[ \nu \xrightarrow{\text{C3}} Y = y_1 \xrightarrow{1/3} Y = y_2 \xrightarrow{\text{C4}} Y = y_3 \]
Wasserstein metric

- We use the Wasserstein metric between probability distributions

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

- 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

![Diagram of Wasserstein metric with probability distributions]
• We use the Wasserstein metric between probability distributions

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

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

$$\mathcal{W}(\mu K, \nu K) \leq \left(1 - \alpha + \alpha \gamma\right)\mathcal{W}(\mu, \nu),$$
**Punchline:** For TD(0), the induced operator $K$ is **contractive** with respect to the Wasserstein metric

\[
\mathcal{W}(\mu K, \nu K) \leq (1 - \alpha + \alpha \gamma) \mathcal{W}(\mu, \nu),
\]

with $\gamma < 1$.

By Banach’s fixed point theorem the distributions $\mu K^n$ converge to a fixed point

\[
\psi = \psi K.
\]

This is exactly the property of a stationary distribution!
Analogies

We have:

Bellman operator $T^\pi$:
- Contraction with respect to $\|\cdot\|_\infty$ w/ factor $\gamma$
- Unique fixed point $V^\pi \in \mathbb{R}^{|S|}$
We have:

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

$\text{TD}(0)$ (with policy $\pi$, step-size $\alpha$):
- Contraction with respect to $\mathcal{W}_{\|\cdot\|_\infty}$ w/ factor $1 - \alpha + \alpha \gamma$
- Unique fixed point $\psi^{\pi,\alpha,\text{TD}(0)} \in \text{Dists}(\mathbb{R}^{|S|})$
For any stepsizes $\alpha \in (0, 1]$, the following algorithms are contractive:

- Monte Carlo Evaluation w/ factor $1 - \alpha$
- TD($\lambda$) 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.
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)): Consider $V_0 \sim \mu$, $W_0 \sim \nu$. Define a coupling of $V_1$ and $W_1$ as follows:

$$V_0(s) \quad V_1 \text{ samples } s_i \iff W_1 \text{ samples } s_i \quad W_0(s)$$

$$r_1 + \gamma V_0(s_1) \quad r_1 + \gamma W_0(s_1)$$

$$r_2 + \gamma V_0(s_2) \quad r_2 + \gamma W_0(s_2)$$

$r_1, s_1 \sim \text{MDP}$

$r_2, s_2 \sim \text{MDP}$
Proof (TD(0)) (continued)

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

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

$$\leq \gamma \mathbb{E} \left[ \| V_0 - W_0 \|_{\infty} \right]$$
Proof (TD(0)) (continued)

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

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

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

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

\[ f_{n+1} = (1 - \alpha) f_n + \alpha \hat{T}(f_n), \]

where \( \hat{T} \) is the target.
Evaluation setting

If the stochastic updates are, in expectation, a Bellman operator of $\pi$

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

then:
Evaluation setting

If the stochastic updates are, in expectation, a Bellman operator of $\pi$

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

then:

- The mean of the stationary distributions is the true value function ($V^\pi$ or $Q^\pi$)
If the stochastic updates are, in expectation, a Bellman operator of $\pi$ then:

- The mean of the stationary distributions is the true value function ($V^\pi$ or $Q^\pi$)
- 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 $\pi$

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

then:

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

$$\mathbb{P}_{f_\alpha \sim \text{stationary dist.}} \left\{ \min_i |f_\alpha(i) - f^\pi(i)| \geq \varepsilon \right\} \xrightarrow{\alpha \to 0} 0$$
If the stochastic updates are, in expectation, a Bellman *optimality* operator

\[ \mathbb{E}[\hat{T} f] = T^* f, \quad \forall f \]

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

\[ \mathbb{E}[\hat{T}f] = T^* f, \quad \forall f \]

then:

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

\[
Q_{n+1}(s, a) = (1 - \alpha) Q_n(s, a) + \alpha G^{\pi_n}(s, a), \\
\pi_{n+1} = \text{greedy } (Q_{n+1})
\]

where \( G^{\pi}(s, a) \) is a discounted return sampled from the MDP using \( \pi \).
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 $\alpha = 1$

\[
Q_{n+1}(s, a) = G^{\pi_n}(s, a), \\
\pi_{n+1} = \text{greedy}(Q_{n+1})
\]
Proof via greedy partitions

- Special case of $\alpha = 1$

\[
Q_{n+1}(s, a) = G^{\pi_n}(s, a), \\
\pi_{n+1} = \text{greedy} (Q_{n+1})
\]

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

• Special case of $\alpha = 1$

\[
Q_{n+1}(s, a) = G^{\pi_{n}}(s, a), \\
\pi_{n+1} = \text{greedy}(Q_{n+1})
\]

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

• This is a finite state Markov chain
Proof via greedy partitions

• Probabilistic policy improvement:

\[ P\{\text{sampling } G^{\pi_n} \text{ that has greedy}(G^{\pi_n}) = \text{greedy}(Q^{\pi_n})\} > 0 \]
• Probabilistic policy improvement:

\[ \mathbb{P}\{\text{sampling } G^\pi_n \text{ that has } \text{greedy}(G^\pi_n) = \text{greedy}(Q^\pi_n)\} > 0 \]

• Therefore every initial policy \( \pi \) can reach \( \pi^* \) with some probability...
Proof via greedy partitions

- Probabilistic policy improvement:

\[ \mathbb{P}\{\text{sampling } G^{\pi_n} \text{ that has } \text{greedy}(G^{\pi_n}) = \text{greedy}(Q^{\pi_n})\} > 0 \]

- Therefore every initial policy \( \pi \) can reach \( \pi^* \) with some probability...
- ...and \( \pi^* \) is a recurrent state
Proof via greedy partitions

- Probabilistic policy improvement:

\[ \mathbb{P}\{\text{sampling } G^{\pi_n} \text{ that has }\ \text{greedy}(G^{\pi_n}) = \text{greedy}(Q^{\pi_n})\} > 0 \]

- Therefore every initial policy \( \pi \) can reach \( \pi^* \) with some probability...
- ...and \( \pi^* \) 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
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:

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

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

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

• Decreasing step-sizes and/or online updates → Corresponds to time-dependent Markov chains → Applying a sequence of contractive kernels

\[ \mu \alpha_1 K \alpha_2 \cdots K \alpha_n \]

• Function approximation → Preliminary results for linear function approximation

• Optimistic Policy Iteration and other stochastic policy iteration methods (e.g. actor-critic methods)
Future work

- Decreasing step-sizes and/or online updates
  → Corresponds to time-dependent Markov chains
  → Applying a sequence of contractive kernels $\mu K_{\alpha_1} K_{\alpha_2} \cdots K_{\alpha_n}$

- Function approximation
  → Preliminary results for linear function approximation

- Optimistic Policy Iteration and other stochastic policy iteration methods (e.g. actor-critic methods)
Merci