Improving Exploration in Reinforcement Learning: Recent Theoretical Insights

ML IN REAL WORLD (CRITEO) – JUNE 28TH 2018

Alessandro Lazaric – Facebook AI Research (on leave from Inria)
Reinforcement learning is learning how to map situations to actions so as to maximize a numerical reward signal in an unknown and uncertain environment.

The learner is not told which actions to take but she must discover which actions yield the most reward by trying them (trial-and-error). In the most interesting and challenging cases, actions affect not only the immediate reward but also the next situation and all subsequent rewards (delayed reward).
Recent RL Successes

**ATARI**

*improved even further over the years*

**Game of GO**

Given the generality of the RL framework, we can expect these algorithms could be applied to a wide range of applications (e.g., recommendation, education, human-robot interaction)

*improved even further over the years*
Recent RL Successes

- "Mastering the Game of Go without Human Knowledge", Silver et al. (2017)

**RL most successful algorithms are sample inefficient (both in collecting and using samples)**

**We need better understanding of how to effectively explore an unknown environment and learn an optimal policy.**

- 4.9 million games
- 1 epoch = samples = 10 million frames
- Potential applications: robotics, personalized recommendation, human-computer interaction, ...
Outline

➢ Optimism-in-face-of-uncertainty principle
➢ Improving exploration with prior knowledge on the bias space
➢ Efficient exploration with misspecified states
➢ Conclusions
Exploration with Optimism-in-face-of-uncertainty

Relevant literature:
Markov Decision Process

An MDP is a tuple $M = \langle S, A, p, r \rangle$

- State space $S$ finite
- Action space $A$
- Transition probability $p(s'|s, a)$
- Reward function $r(s, a)$

Stationary Markov policy $\pi : S \rightarrow \Delta(A)$
Average Reward (undiscounted infinite horizon)

\[ g(M, \pi) = \lim_{n \to \infty} \mathbb{E}\left[ \frac{1}{n} \sum_{t=1}^{n} r_t \right] \]

\[ r_t = r(s_t, \pi(s_t)) \]

\[ s_{t+1} \sim p(\cdot | s_t, \pi(s_t)) \]

\[ g^* = \max_{\pi} g(M, \pi) \]

\[ \pi^* = \arg \max_{\pi} g(M, \pi) \]
The Learning Problem

- Set initial state $s_0$
- **While** (true)
  - Observe $s_t$
  - Execute action $a_t$
  - Observe $s_{t+1}, r_t$

Diagram:

- Agent
- Environment
- Critic

Step: $\langle s_t, a_t, s_{t+1}, r_t \rangle$

Trajectory: $\langle s_1, a_1, s_2, r_1, s_2, a_2, \ldots \rangle$
True Environment

No initial knowledge
Noisy observations
Estimation of the environment trajectory \( \langle s_1, a_1, s_2, r_1, s_2, a_2, \cdots \rangle \)

Estimated environment

\[
\hat{M}_t = \langle S, A, \hat{r}_t, \hat{p}_t \rangle
\]

\[
\hat{r}_t(s, a) = \frac{\hat{R}_t(s, a)}{N_t(s)}\quad \hat{p}_t(s'|s, a) = \frac{N_t(s, a, s')}{N_t(s, a)}
\]
Estimated environment

Both estimated rewards and dynamics may be inaccurate
Plausible environments

Estimated environment

\[ \hat{M}_t = \langle S, A, \hat{r}_t, \hat{p}_t \rangle \]

\[ \hat{r}_t(s, a) = \frac{R_t(s, a)}{N_t(s)} \]
\[ \hat{p}_t(s'|s, a) = \frac{N_t(s, a, s')}{N_t(s, a)} \]

Uncertainty

\[ |\tilde{r}(s, a) - \hat{r}_t(s, a)| \leq B_{r,t}(s, a) \]
\[ \|\tilde{p}(\cdot|s, a) - \hat{p}_t(\cdot|s, a)\|_1 \leq B_{p,t}(s, a) \]

Plausible environments

\[ \mathcal{M}_t = \{ \hat{M} = \langle S, A, \tilde{r}, \tilde{p} \rangle \} \]
Current estimate

Plausible environments

True Environment
Optimistic environment

Optimism is used in a growing number of deep RL methods to improve exploration

\[
(\tilde{\pi}_t, \tilde{M}_t) = \operatorname*{arg\,max}_{M \in \mathcal{M}_t} \max_{\pi} g(\pi, M)
\]

“Unifying Count-Based Exploration and Intrinsic Motivation”, Bellemare et al. (2016)
“The uncertainty Bellman equation and exploration”, Osband et al. (2018)
True Environment

Optimistic environment and policy
True Environment

Noisy observations
Noisy observations

True Environment
Better estimation of the environment (effective exploration), while attempting to collect high reward (effective exploitation)
Regret guarantees

\[ R_n = ng^* - \sum_{t=1}^{n} r_t \]
Upper-Confidence for RL (UCRL)

**Theorem (Jaksch et al., 2010)**

For any $n$ and any MDP with $S$ states, $A$ actions, and diameter $D$, with probability $1 - \delta$, UCRL suffers a cumulative regret

$$R_n = \tilde{O}(DS\sqrt{An})$$
Diameter of an MDP

\[ D = \max_{s, s' \in S} \left\{ \min_{\pi: S \rightarrow A} \left\{ \mathbb{E}^\pi[T(s, s')] \right\} \right\} \]
Limitations of UCRL: (1) Diameter

$$D = \max_{s,s' \in S} \left\{ \min_{\pi : S \rightarrow A} \left\{ \mathbb{E}^\pi \left[ T(s, s') \right] \right\} \right\}$$

longest shortest path

Longer paths should not necessarily correspond to large regret
Limitations of UCRL: (2) Misspecified states

An MDP is a tuple $M = \langle S, A, p, r \rangle$

- **State space** $S$
- **Action space** $A$
- **Transition probability** $p$
- **Reward function** $r$

Not necessarily all reachable

Very common in practice: we do not know in advance all reachable states

Optimism “favors” unknown states, but if they are unreachable, then it suffers **unbounded** regret.
Bias-span constrained exploration

Relevant literature:
Bias function

\[ h^\pi(s) = \lim_{n \to \infty} \mathbb{E} \left[ \sum_{t=1}^{n} r_t - g^\pi(s) \right] \]

difference between actual reward and asymptotic reward
Average Reward (undiscounted infinite horizon)

Each step has reward -1 and puddle has reward -3

$$h^{\pi^*}(s') = -36 - 3$$

$$h^{\pi^*}(s) = -5$$

Difference in “potential”

$$h^{\pi^*}(s) - h^{\pi^*}(s')$$ (under the optimal policy)
Optimal Bias-span

Assumption

\[
\max_{s \in S} h^{\pi^*}(s) - \min_{s \in S} h^{\pi^*}(s) = sp(h^{\pi^*}) \leq c
\]
Bias-span Constrained Optimism

\[ (\tilde{\pi}_t, \tilde{M}_t) = \arg \max_{M \in \mathcal{M}_t} \max_{\pi} g(\pi, M) \]

\[
\text{s.t. } \text{sp}(h(\pi, M)) \leq c
\]

Non-trivial optimization problem!

only “reasonable” MDPs are considered

only “reasonable” policies without too big potentials are considered
Solving an MPD

\[ \pi^*(M) = \arg \max_{\pi} g(\pi, M) \]

Value iteration

\[ v_0(s) = 0 \]

\[ v_{n+1}(s) = \max_a \left( r(s, a) + \sum_{s'} p(s'|s, a) v_n(s') \right) \]

\[ \pi_{n+1}(s) = \arg \max_a \left( r(s, a) + \sum_{s'} p(s'|s, a) v_{n+1}(s') \right) \]
Solving a constrained MPD

In general:
- no convergence,
- even when convergent not associated to a policy

\[
\pi^*(M) = \arg\max_{\pi} g(\pi, M) \quad \text{s.t.} \quad \text{sp}(h(\pi, M)) \leq c
\]

(span-constrained) value iteration

\[
v_0(s) = 0 \\
v_{n+1/2}(s) = \max_a \left( r(s, a) + \sum_{s'} p(s'|s, a) v_n(s') \right) \\
v_{n+1} = \text{trunc}_c(v_{n+1/2})
\]
Bias-span Constrained Optimism

\[
(\tilde{\pi}_t, \tilde{M}_t) = \arg \max_{M \in \mathcal{M}_t} \max_{\pi} g(\pi, M) \\
\text{s.t. } \text{sp}(h(\pi, M)) \leq c
\]

Plausible MDPs

\[
\mathcal{M}_t = \{ \tilde{M} = \langle S, A, \tilde{r}, \tilde{p} \rangle \} \\
|\tilde{r}(s, a) - \tilde{r}_t(s, a)| \leq B_{r,t}(s, a) \\
\|\tilde{p}(\cdot | s, a) - \tilde{p}_t(\cdot | s, a)\|_1 \leq B_{p,t}(s, a)
\]

include \(\tilde{r}(s, a) = 0\)

allow non-zero transitions to an arbitrary \(\tilde{p}(s | s_a) \geq \eta\)
Bias-span Constrained Optimism

\[
(\tilde{\pi}_t, \tilde{M}_t) = \arg \max_{M \in \mathcal{M}_t^+} \max_{\pi} g(\pi, M) \\
\text{s.t. } \sp(h(\pi, M)) \leq c
\]

(span-constrained) “extended” value iteration

\[
v_0(s) = 0 \\
v_{n+1/2}(s) = \max_a \left( \max_{\tilde{r} \in \mathcal{R}^+\ell} \tilde{r}(s, a) + \max_{\tilde{p} \in \mathcal{P}^+\ell} \sum_{s'} \tilde{p}(s'|s, a)v_n(s') \right) \\
v_{n+1} = \text{trunc}_c(v_{n+1/2})
\]
Span-constrained Optimization

Theorem *(Fruit, Pirotta, Ortner, L, 2018)*

The span-constrained extended value iteration

- **Converges**
- Returns a span-constrained (stochastic) **policy**
- Solves the original constrained optimization problem up to an **additive error** $\eta \epsilon$
Span-constrained Learning (SCAL)

**Theorem (Fruit, Pirotta, Ortner, L, 2018)**

For any $n$ and any MDP with $S$ states, $A$ actions, and bias span upper-bounded by $c$, with probability $1 - \delta$, SCAL suffers a cumulative regret

$$R_n = \tilde{O}(cS\sqrt{An})$$
A “complex” navigation problem

\[
S = 360, \ A = 8
\]

\[
D = 250, \ sp(h^*) \approx 3.28
\]
A “complex” navigation problem
A “complex” navigation problem
A “complex” navigation problem

![Graph showing cumulative regret over duration for SCAL and UCRL with different c values.](image)
Learning with Misspecified State Spaces

Relevant literature:
No assumption about whether all the states in $S$ are actually reachable.

No assumption on the bias span.
Misspecified State Space
Misspecified State Space

If \( \bar{s} \) exists, then we can discover it and learn the optimal policy.

If \( \bar{s} \) does not exist, then we suffer linear regret (\( D=\infty \)).

Optimism

If \( \bar{s} \) exists, then we can discover it and learn the optimal policy.
Misspecified State Space

"Greedy" approach
If $\mathcal{S}$ does not exist, then we learn the optimal policy on reachable states.

If $\overline{\mathcal{S}}$ does exist, then we suffer linear regret.
Truncated Plausible MDPs

Estimated transition probability

\[ \hat{p}_t(s'|s, a) \]

Uncertainty

\[ |\tilde{p}(s'|s, a) - \hat{p}_t(s'|s, a)| \leq B_{p,t}(s, a, s') \]

Largest plausible transition probability to \( \bar{s} \)

\[ \hat{p}_t(\bar{s}|s, a) + B_{p,t}(s, a, \bar{s}) \leq \rho_t \Rightarrow \tilde{p}_t(\bar{s}|s, a) = 0 \]
Truncated Plausible MDPs

\[
(\widehat{\pi}_t, \widehat{M}_t) = \arg \max_{M \in \mathcal{M}_t^T} \max_{\pi} g(\pi, M)
\]

\[S_t^c \quad \text{set of states observed so far}\]
\[\rho_t \approx \sqrt{1/t} \quad \text{decreasing threshold}\]
\[\forall s \in S_t^c, \bar{s} \notin S_t^c, \quad \text{if } \widehat{p}_t(\bar{s}|s, a) + B_{p,t}(s, a, \bar{s}) \leq \rho_t \Rightarrow \widehat{p}(\bar{s}|s, a) = 0\]
Truncated UCRL (TUCRL)

Theorem (Fruit, Pirotta, L, 2018)

For any $n$ and any MDP with $S$ states, $A$ actions, and diameter of the well-specified states $D_{\text{comm}}$ (the “true” diameter is $\infty$), with probability $1-\delta$, TUCRL suffers a cumulative regret

$$R_n = \tilde{O}(D_{\text{comm}} S \sqrt{A n})$$
The Taxi Navigation Problem

- 500 states defined (all possible combinations of passenger, taxi, and destination positions)
- Only 400 states are actually reachable
The Taxi Navigation Problem \[\text{[the higher the worse]}\]

Misspecified states

- **UCRL**: linear regret
- **SCAL**: Prior knowledge helps
- **TUCRL**: even without prior knowledge, it can still learn effectively
Conclusion
Conclusion

➢ Effective exploration is critical to apply RL in sample-expensive applications
➢ Optimistic exploration could be inefficient in “large” problems
➢ Prior knowledge on the range of the bias function helps avoiding “useless” exploration
➢ Misspecified states can be effectively managed

*Integrate these findings into efficient deep RL approaches (e.g., model-based, policy gradient, value-based)*
Thanks!

Questions?