Course Notes On Dynamic Optimization (Fall 2023)
Lecture 9B: The robustness benefits of online value iteration

Instructor: Daniel Russo

Email: djr2174@gsb.columbia.edu

Graduate Instructor: David Cheikhi
Email: d.cheikhi@columbia.edu

These notes are partly based of scribed notes from a previous edition of the class. I have done some follow up light editing, but there may be typos or errors.

Most RL algorithms proceed in an online manner. Rather than loop over states, as in classic DP algorithms, or sample states from a fixed distribution, as in our fitted VI algorithms, they learn from data from deploying policies in their environment. This raises lots of challenges (how do you explore?) and seems like a questionable constraint in practice when so many success stories use simulators.

So, *if you did have a simulator, is there something advantageous about trying to solve for effective policy in an "online manner."* In the special case of state-aggregated representations, we'll be able to answer that affirmatively.

These notes are largely inspired from the following references.

- Performance loss bounds for approximate value iteration with state aggregation. B Van Roy, 2006.
- Simple agent, complex environment: Efficient reinforcement learning with agent states. Shi Dong, Benjamin Van Roy, Zhengyuan Zhou, 2022.

However, in my attempt to understand (and simplify) things, the proofs, formulation, etc turned out to be quite different. Any errors are my own.

1 Error bound for fitted value iteration in the state-aggregated

Define the approximation error

$$\epsilon_* = \sup_{x \in \mathcal{X}} |J^*(x) - J^*(\phi(x))|.$$

This depends only on the cost-to-go function *under the optimal policy*, a huge improvement over ϵ_Θ in the previous notes.

But we still get an unfortunate amplification of error as the horizon grows.

Lemma 1 (Informal)**.** *Consider a discounted problem with discount factor α. Under approximate value iteration with state-aggregated representations*

$$\lim_{n\to\infty} \|J_{\mu_n} - J^*\|_\infty = O\left(\frac{\epsilon_*}{(1-\alpha)^2}\right)$$

It is possible to construct a problem in which this bound is tight and the quadratic scaling in the effective horizon $1/(1-\alpha)$ occurs.

2 State-aggregated Real time dynamic programming

We revisit RTDP in the state-aggregated case. As before, the algorithm requires the following primitives:

- The ability to simulate the state trajectory under a given policy.
- The ability to perform one-step lookahead at a specific state. For instance, we can compute $TJ(x)$.
  - It is worth emphasizing that many problems have large state spaces but sparse transitions. That is, it is easy to loop over all possible successor states from a given state but not to loop over all possible states. RTDP is especially natural in this case.

Algorithm 1 Real-Time Dynamic Programming (RTDP)

Require: $J_0 = (J_0(\bar{x}_1), \ldots, J_0(\bar{x}_m)) \in \mathbb{R}^m$, episode limit N, aggregation rule ϕ, error limit ϵ^*.

for Episode $n = 0, 1, 2, \ldots,$ **do**
  Reset $J_n^+ \leftarrow J_n$ and $k \leftarrow 0$
  Sample $x_0^n \sim d_0$
  while $x_k^n \neq \emptyset$ **do**
    Compute $J_n^+(\phi(x_k^n)) \leftarrow \min_{u\in U(x_k^n)} g(x_k^n, u_k^n) + \sum_{x'\in\mathcal{X}} p(x'|x_k^n, u_k^n) J_n(\phi(x')) - \epsilon^*$.
    Choose action $u_k^n \in \arg\min_{u\in U(x_k^n)} g(x_k^n, u_k^n) + \sum_{x'\in\mathcal{X}} p(x'|x_k^n, u_k^n) J_n(\phi(x'))$
    Apply u_k^n and observe next state x_{k+1}^n.
    Update period count: $k \leftarrow k+1$.
  end while
  Update $J_{n+1} \leftarrow J_n^+$
end for
return J_N

3 Regret bound under optimism

Right now, my proof requires the following condition to get the tightest bound. I know how to get rid of this condition in the discounted case, at the expense of worse bounds, but not yet in general.

Definition 1. We say the state transitions are ϕ-acyclic if, under any policy, for any $i \in [m]$, $\phi(x_k) = \bar{x}_i$ occurs at most once during an episode.

Proposition 1. *Suppose transitions are ϕ-acylcic; If ϕ-RTDP is applied with $J_0 \preceq J^*$, then,*

$$\underbrace{\mathbb{E}\left[\sum_{n=0}^{N-1}(J_{\mu_n}(x_0^n) - J^*(x_0^n))\right]}_{\textit{Cummulative Regret}} \leqslant \underbrace{\sum_{i=1}^{m}(J^*(\bar{x}_i) - J_0(\bar{x}_i))}_{\textit{Initialization error}} + \epsilon^* \times \mathbb{E}\left[\sum_{n=0}^{N-1}\tau^n\right]. \tag{1}$$

In particular,

$$\limsup_{N\to\infty} \underbrace{\frac{\mathbb{E}\left[\sum_{n=0}^{N-1}(J_{\mu^n}(x_0^n) - J^*(x_0^n))\right]}{\mathbb{E}\left[\sum_{n=0}^{N-1}\tau^n\right]}}_{\textit{average regret per-period}} \leqslant \epsilon^*.$$

3.1 First step: some rewriting

Lemma 2. *For any n and $k \leqslant \tau^n - 1$, $J_{n+1}(x_k^n) = TJ_n(x_k^n) - \epsilon^*$.*

Proof. State aggregation implies $J_{n+1}(x) = J_{n+1}(\phi(x))$. Because transitions are ϕ-acyclic, once J_n^+ is updated at some cluster its value is never changed again within an episode. Therefore,

$$J_{n+1}(x_k^n) = J_n^+(\phi(x_k^n)).$$

Now

$$\begin{aligned} J_n^+(\phi(x_k^n)) &= \min_{u\in U(x_k^n)} g(x_k^n, u_k^n) + \sum_{x'\in\mathcal{X}} p(x'|x_k^n, u_k^n) J_n(\phi(x')) - \epsilon^* \\ &= \min_{u\in U(x_k^n)} g(x_k^n, u_k^n) + \sum_{x'\in\mathcal{X}} p(x'|x_k^n, u_k^n) J_n(x') - \epsilon^* \quad \text{(Since } J_n \text{ is state-aggregated)} \\ &= TJ_n(x_k^n) - \epsilon^*. \end{aligned}$$

□

3.2 Second analysis step: preservation of optimism

The value V_0 in the initial iterate of the RTDP-ϕ is optimistic in the sense of $J_0 \preceq J^*$. We will show that this optimism is preserved throughout the iterations of RTDP-ϕ or, equivalently,

Lemma 3 (Preservation of Optimism)**.** *Under ϕ-RTDP, if $J_0 \preceq J^*$ then with probability 1,*

$$J_n \preceq J^*, \textit{ for } n = 0, \ldots, N-1. \tag{2}$$

Proof. Suppose that $J_n \preceq J^*$ for some $n \in \{0, 1, \ldots, N-2\}$. Then,

$$
\begin{aligned}
J_n^+(\phi(x_k^n)) &= \min_{u \in U(x_k^n)} g(x_k^n, u_k^n) + \sum_{x' \in \mathcal{X}} p(x'|x_k^n, u_k^n) J_n(\phi(x')) - \epsilon^* \\
&= \min_{u \in U(x_k^n)} g(x_k^n, u_k^n) + \sum_{x' \in \mathcal{X}} p(x'|x_k^n, u_k^n) J_n(x') - \epsilon^* \quad \text{(Since } J_n \text{ is state-aggregated)} \\
&= TJ_n(x_k^n) - \epsilon^* \\
&\leqslant TJ^*(x_k^n) - \epsilon^* \qquad \text{(By monotinicty and optimism of } J_n) \\
&= J^*(x_k^n) - \epsilon^* \qquad \text{(Since } J^* \text{ is a fixed point of T)} \\
&\leqslant J^*(x_k^n) + \epsilon^* - \epsilon^* \qquad \text{(Definition of approximation error)} \\
&= J^*(x_k^n).
\end{aligned}
$$

□

3.3 Second analysis step: relating regret to the Bellman gap at visited states

This lemma is copy pasted from last week's notes.

Lemma 4. *Let $J \in \mathbb{R}^{\mathcal{X}}$ satisfy $J \preceq J^*$ and the stationary policy $\mu \in G(J)$ be greedy with respect to J. Then*

$$
J^\mu(x) - J^*(x) \leqslant \mathbb{E}\left[\sum_{k=0}^{\tau} \left(TJ(x_k) - J(x_k)\right) \mid x_0 = x\right].
$$

Proof. Write $P_\mu \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ for the sub-stochastic transition matrix under μ, $P_\mu(x, x') = p(x'|x, \mu(x))$. If $g_\mu \in \mathbb{R}^{|\mathcal{X}|}$, then $T_\mu J = g_\mu + P_\mu J$. We find,

$$
\begin{aligned}
J^\mu - J^* &\preceq J^\mu - J && \text{(by optimism)} \\
&= T_\mu J^\mu - J && \text{(since } J_\mu \text{ is a fixed point)} \\
&= \left(T_\mu J^\mu - T_\mu J\right) + \left(T_\mu J - J\right) \\
&= \left(T_\mu J^\mu - T_\mu J\right) + \left(TJ - J\right) && \text{(since } \mu \in G(J)) \\
&= \left(g_\mu + P_\mu J^\mu - g_\mu - P_\mu J\right) \\
&= P_\mu \left(J^\mu - J\right) + \left(TJ - J\right) \\
&= \cdots \\
&= \sum_{k=0}^{\infty} P_\mu^k \left(TJ - J\right) && \text{(by iterating the recursion).}
\end{aligned}
$$

For any x, $\left(\sum_{k=0}^{\infty} P_\mu^k \left(TJ - J\right)\right)(x) = \mathbb{E}\left[\sum_{k=0}^{\tau} \left(TJ(x_k) - J(x_k)\right) \mid x_0 = x\right]$, completing the proof. □

3.4 Completing the proof

When reading this proof, it is helpful to imagine that

$$
J_0 \preceq J_1 \preceq \cdots \preceq J_n \preceq \cdots \preceq J^*.
$$

The main idea in the proof is to relate regret in episode n to the overall reduction in optimism $\sum_{x\in\mathcal{X}}(J_{n+1}(x) - J_n(x))$

Proof. By Lemma 2,

$$J_{n+1}(x_k^n) = TJ_n(x_k^n) - \epsilon^*.$$

Using this together with Lemma 4 and optimism, we find that for any fixed n,

$$\begin{aligned}
\underbrace{J^{\mu_n}(x_0^n) - J^*(x_0^n)}_{\text{Regret}} &\leqslant \mathbb{E}\left[\sum_{k=0}^{\tau_n}(TJ_n(x_k^n) - J_n(x_k^n)) \mid x_0^n, J_n\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\tau_n}(J_{n+1}(x_k^n) - J_n(x_k^n) + \epsilon^*) \mid x_0^n, J_n\right] \\
&\overset{(*)}{=} \mathbb{E}\left[\sum_{k=0}^{\tau_n}(J_{n+1}(\phi(x_k^n)) - J_n(\phi(x_k^n)) + \epsilon^*) \mid x_0^n, J_n\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{m}(J_{n+1}(\bar{x}_i) - J_n(\bar{x}_i)) \times T_i^n + \tau^n\epsilon^* \mid x_0^n, J_n\right] \quad \text{where} \quad T_i^n := \sum_{k=0}^{\tau^n-1}\mathbb{1}(\phi(x_k^n) = i) \\
&\overset{(*)}{=} \mathbb{E}\left[\sum_{i=1}^{m}(J_{n+1}(\bar{x}_i) - J_n(\bar{x}_i)) + \tau^n\epsilon^* \mid x_0^n, J_n\right].
\end{aligned}$$

The key step is equality (*), which uses that $T_i^n \in \{0,1\}$ since the transitions are ϕ-acyclic and (by the definition of the algorithm), estimated cost-to-go is not updated for a cluster when $T_i^n = 0$.

Summing over n, using the law of iterated expectations, and simplifying a telescoping sum yields

$$\mathbb{E}\left[J^{\mu_n}(x_0^n) - J^*(x_0^n)\right] \leqslant \mathbb{E}\left[\sum_{i=1}^{k}(J_N(\bar{x}_i) - J_0(\bar{x}_i))\right] + \epsilon_*\mathbb{E}\left[\sum_{n=0}^{N-1}\tau^n\right].$$

The last step uses optimism, i.e. $J_N \preceq J^*$ to rewrite the bound as

$$\mathbb{E}\left[J^{\mu_n}(x_0^n) - J^*(x_0^n)\right] \leqslant \sum_{i=1}^{k}(J^*(\bar{x}_i) - J_0(\bar{x}_i)) + \epsilon_*\mathbb{E}\left[\sum_{n=0}^{N-1}\tau^n\right].$$

□