3) Model-Based Reinforcement Learning

Melih Kandemir

University of Southern Denmark
Department of Mathematics and Computer Science (IMADA)

Some concepts

- **Model-based RL:** Approximate $\widehat{M} = \langle \mathcal{S}, \mathcal{A}, \widehat{r}, \widehat{P} \rangle$ from data where $\widehat{r}$ is a reward estimate and $\widehat{P}$ is a transition probability estimate.
- **Model-free RL:** Approximate $\widehat{V}$
- **Offline RL:** Learning from observations collected beforehand.
- **Online RL:** Learning while in action.
- **On-policy RL:** Online RL by acting based on the policy being learned.
- **Off-policy RL:** Online RL by acting based on an exploration (behavior) policy.

Effective horizon of discounted return

Theorem

Given a discount factor γ, the discounted return in the first $T = \frac{1}{1-\gamma} \log \frac{\varepsilon(1-\gamma)}{R_{\max}}$ time steps, is within ε of the total discounted return.

Proof. Recall that the rewards are $r_t \in [0, R_{\max}]$. Fix an infinite sequence of rewards $(r_0, \ldots, r_t, \ldots)$. We would like to consider the following difference:

$$\sum_{t=0}^{\infty} r_t \gamma^t - \sum_{t=0}^{T-1} r_t \gamma^t = \sum_{t=T}^{\infty} r_t \leq \frac{\gamma^T}{1-\gamma} R_{max}$$

We want this difference to be bounded by ε, hence $\frac{\gamma^T}{1-\gamma} R_{\max} \leq \varepsilon$. This is equivalent to $T \log(1/\gamma) \leq \log R_{\max} - \log(\epsilon(1-\gamma))$. Since $\log(1+x) \leq x$, we can bound $\log(1/\gamma) = \log(1 + \frac{1-\gamma}{\gamma}) \leq \frac{1-\gamma}{\gamma}$. Since $\gamma < 1$, we have that $\frac{\gamma}{1-\gamma} \leq \frac{1}{1-\gamma}$ and hence it is sufficient to have $T \geq \frac{1}{1-\gamma} \log \frac{R_{\max}}{\varepsilon(1-\gamma)}$ $\square$

Concentration of a single estimator

Theorem (Chernoff-Hoeffding)

Let $R_1, \ldots, R_m$ *be* m *i.i.d. samples of a random variable* $R \in [0, 1]$. *Let* $\mu = E[R]$ *and* $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} R_i$. *For any* $\varepsilon \in (0, 1)$ *we have,*

$$P(\hat{\mu} - \mu \geq \varepsilon) \leq e^{-2\varepsilon^2 m}$$

Setting $e^{-2\varepsilon^2 m} \leq \delta$ and solving for m yields the result below.

Corollary

Let $R_1, \ldots, R_m$ *be* m *i.i.d. samples of a random variable* $R \in [0, 1]$. *Let* $\mu = E[R]$ *and* $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} R_i$. *Fix* $\varepsilon, \delta > 0$. *Then, for* $m \geq \frac{1}{2\varepsilon^2}\log(1/\delta)$, *with probability at least* $1 - \delta$, *we have that* $\hat{\mu} - \mu \leq \varepsilon$.

Simultaneous concentration of K estimators

Now consider the case where we have true means $\mu_1, \ldots, \mu_K$ of K random variables in the interval $[0, R_{\max}]$ and their corresponding empirical means $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$. We are interested to bound the probability of the unwanted event that at least one of the empirical means are more erroneous than we can tolerate

$$
\begin{aligned}
P(\exists j \text{ s.t. } \widehat{\mu}_j - \mu_j \geq \varepsilon) &= P((\widehat{\mu}_1 - \mu_1) \geq \varepsilon \cup (\widehat{\mu}_2 - \mu_2) \geq \varepsilon \cup \\
&\qquad \ldots \cup (\widehat{\mu}_K - \mu_K) \geq \varepsilon) \\
&\leq \sum_{j=1}^{K} P\Big((\widehat{\mu}_j - \mu_j)/R_{\max} \geq \varepsilon/R_{\max}\Big) \\
&\leq \sum_{j=1}^{K} e^{-2\varepsilon^2 m/R_{\max}^2} = K e^{-2\varepsilon^2 m/R_{\max}^2}
\end{aligned}
$$

Remark: We here assume that *each* of the K random variables is observed m times.

Probably Approximately Correct (PAC) analysis

If we set $Ke^{-2\varepsilon^2 m/R_{\max}^2} \leq \delta$ for some $\delta \in [0, 1]$ and solve for m, we get

$$m \geq \frac{R_{\max}^2}{2\varepsilon^2} \log(K/\delta).$$

The r.h.s. gives a lower bound on the number of samples required to reduce the probability of an approximation error of at most ε below δ. Leaving ε alone on one side of the inequality and plugging the related statement into the original expression we also get

$$P\left(\forall j : \widehat{\mu}_j - \mu_j \leq \sqrt{\frac{R_{\max}^2}{2m} \log(K/\delta)}\right) \geq 1 - \delta.$$

This statement says that it is highly **probable** that $\widehat{\mu}_j$'s are **approximately correct** estimates of μ_j's. Hence the name.

Concentration of an empirical distribution

Let p denote the vector of probability masses of a categorical distribution with d categories and $\widehat{p}$ be its empirical estimate. We would like to find the concentration of $\|p - \widehat{p}\|_1$. Consider the fact that

$$\|a\|_1 = \max_{u \in \{-1,+1\}^d} u^\top a.$$

As there exist 2^d possible u instances, we have

$$P\left(\forall u : u^\top(\widehat{p} - p) \leq \sqrt{\frac{R_{\max}^2}{2m}\log(2^d/\delta)}\right) \geq 1 - \delta$$

which implies

$$P\left(\|\widehat{p} - p\|_1 \leq \sqrt{\frac{R_{\max}^2 d}{2m}\log(2/\delta)}\right) \geq 1 - \delta.$$

Concentration of K empirical distributions

Consider the case where we are interested in bounding K probability distributions $(p_j)_{j=0}^{K-1}$ simultaneously after observing m samples from each, making $N = mK$ observations in total. Plugging the values into the results developed earlier, we get

$$P\left(\forall j \in [K] : \|\widehat{p}_j - p_j\|_1 \leq \sqrt{\frac{R_{\max}^2 dK}{2N} \log(2K/\delta)}\right) \geq 1 - \delta.$$

The empirical MDP

Given an i.i.d. set of tuples $D = \{(s, a, r_i, s_i') : 1 \leq i \leq m\}$ for a given (s, a), estimate the empirical transition probability

$$\widehat{P}(s'|s,a) = \frac{\sum_{j=1}^{m} \mathbb{I}(s_j = s, a_j = a, s_j' = s')}{\sum_{j=1}^{m} \mathbb{I}(s_j = s, a_j = a)}$$

and the empirical reward

$$\widehat{r}(s,a) = \frac{\sum_{j=1}^{m} r_j \mathbb{I}\left(s_j = s, a_j = a\right)}{\sum_{j=1}^{m} \mathbb{I}(s_j = s, a_j = a)}.$$

Denote the below tuple as the **empirical MDP**

$$\widehat{M} = \langle \mathcal{S}, \mathcal{A}, \widehat{P}, p_0, \widehat{r} \rangle$$

which is an empirical estimate of the true MDP

$$M = \langle \mathcal{S}, \mathcal{A}, P, p_0, r \rangle.$$

True value and estimated value

Assuming access to the true P (temporarily), define the **estimated value function** as

$$\widehat{V}_T^{\pi}(s_0) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{T} \widehat{r}_t(s_t, a_t)\right].$$

We would like to know how much the estimation resembles the true quantity

$$|V_T^{\pi}(s_0) - \widehat{V}_T^{\pi}(s_0)| = \left|\mathbb{E}^{\pi}\left[\sum_{t=0}^{T} r_t(s_t, a_t)\right] - \mathbb{E}^{\pi}\left[\sum_{t=0}^{T} \widehat{r}_t(s_t, a_t)\right]\right|$$

Remark: The term *empirical value function* is saved for later use.

Propagation of reward estimation error to the value

Theorem

Assume that for every (s,a) *and* t *we have* $|r_t(s,a) - \hat{r}_t(s,a)| \leq \varepsilon$. *Then, for any policy* $\pi \in \Pi_{MS}$ *we have* $|V_T^\pi(s_0) - \widehat{V}_T^\pi(s_0)| \leq \varepsilon(T+1)$.

Proof.

$$|V_T^\pi(s_0) - \widehat{V}_T^\pi(s_0)| = \left| \mathbb{E}^\pi \left[\sum_{t=0}^{T} r_t(s_t, a_t) - \sum_{t=0}^{T} \hat{r}_t(s_t, a_t) \right] \right|$$

$$\leq \mathbb{E}^\pi \left[\left| \sum_{t=0}^{T} r_t(s_t, a_t) - \sum_{t=0}^{T} \hat{r}_t(s_t, a_t) \right| \right] \qquad \text{Jensen's ineq. and } |\cdot| \text{ convex}$$

$$\leq \mathbb{E}^\pi \left[\sum_{t=0}^{T} |r_t(s_t, a_t) - \hat{r}_t(s_t, a_t)| \right] \qquad \text{Triangle ineq. and } \mathbb{E} \text{ monotone}$$

$$= \varepsilon(T+1) \quad \square$$

Remark: For the non-episodic setting, replace $T+1$ by $1/(1-\gamma)$.

What we know and what we want

When we have an observation set D, our epistemic situation is as below.

		Value	
		True	**Approx**
Optimal policy	**True**	V^{π_*} Wanted!	$\widehat{V}^{\pi_*}$ Unknown
	Approx	$V^{\widehat{\pi}_*}$ Unknown	$\widehat{V}^{\widehat{\pi}_*}$ Known

We further know the following

- $|V^{\pi_*} - \widehat{V}^{\pi_*}| \leq \varepsilon(T+1)$
- $|V^{\widehat{\pi}_*} - \widehat{V}^{\widehat{\pi}_*}| \leq \varepsilon(T+1)$
- $V^{\pi_*} \geq V^{\widehat{\pi}_*}$
- $\widehat{V}^{\widehat{\pi}_*} \geq \widehat{V}^{\pi_*}$

Propagation of error to the value of the optimal policy

Theorem

Assume that for every (s,a) *and* t *we have* $|r_t(s,a) - \hat{r}_t(s,a)| \leq \varepsilon$. *Then,*

$$V_T^{\pi_*}(s_0) - V_T^{\hat{\pi}_*}(s_0) \leq 2\varepsilon(T+1).$$

Proof.

$$\begin{aligned}
V_T^{\pi_*}(s_0) - V_T^{\hat{\pi}_*}(s_0) &= V_T^{\pi_*}(s_0) - \widehat{V}_T^{\pi_*}(s_0) + \widehat{V}_T^{\pi_*}(s_0) - V_T^{\hat{\pi}_*}(s_0) \\
&\leq \varepsilon(T+1) + \widehat{V}_T^{\pi_*}(s_0) - V_T^{\hat{\pi}_*}(s_0) \\
&\leq \varepsilon(T+1) + \widehat{V}_T^{\hat{\pi}_*}(s_0) - V_T^{\hat{\pi}_*}(s_0) \\
&\leq \varepsilon(T+1) + \varepsilon(T+1) \\
&= 2\varepsilon(T+1) \qquad \square
\end{aligned}$$

In the non-episodic setup we arrive at a famous result:

$$V_T^{\pi_*}(s_0) - V_T^{\hat{\pi}_*}(s_0) \leq \frac{2\varepsilon}{1-\gamma}.$$

Some useful inequalities

For vectors $a, b \in \mathbb{R}^d$, we have

$$\begin{aligned}
\|a^\top b\|_\infty &= \max_i \{|a_i b_i|\} \\
&\leq \max_i \{|a_i| \cdot |b_i|\} \\
&\leq \max_i \{|a_i| \cdot |b_i|\} \\
&\leq \max_i \left\{ |a_i| \cdot \max_j \{|b_j|\} \right\} \\
&= \|a\|_\infty \cdot \|b\|_\infty \\
&\leq \max_i \left\{ |a_i| \sum_j |b_j| \right\} \\
&= \|a\|_\infty \cdot \|b\|_1 \\
&\leq \|a\|_1 \cdot \|b\|_1.
\end{aligned}$$

Summary: $\|a^\top b\|_\infty \leq \|a\|_\infty \cdot \|b\|_\infty \leq \|a\|_1 \cdot \|b\|_1$

Some useful inequalities cont'd

Define norm on matrices $\|\Delta\|_{\infty,1} = \max_i \sum_j |\Delta_{ij}|$. For a matrix Δ and a vector a we have

$$
\begin{aligned}
\|\Delta a\|_\infty &= \max_i \left\{ \left| \sum_j \Delta_{ij} a_j \right| \right\} \\
&\leq \max_i \left\{ \sum_j |\Delta_{ij} a_j| \right\} \\
&\leq \max_i \left\{ \sum_j |\Delta_{ij}| \cdot |a_j| \right\} \\
&\leq \max_i \left\{ \sum_j |\Delta_{ij}| \sum_k |a_k| \right\} \\
&= \|\Delta\|_{\infty,1} \|a\|_1.
\end{aligned}
$$

Propagation of transition probability error to marginals

Theorem

Assume that $\|P_1 - P_2\|_{\infty,1} \leq \varepsilon$. Let p_1^t and p_2^t be the distributions over states after trajectories of length t of P_1 and P_2, respectively. Then $\|p_1^t - p_2^t\|_1 \leq \varepsilon t$.

Proof. Let p_0 be the distribution of the start state. Then $p_1^t = p_0^\top P_1^t$ and $p_2^t = p_0^\top P_2^t$. Proof by induction on t. For $t = 0$ we have $p_1^0 = p_2^0 = p_0$. Let $z^t = p_1^t - p_2^t$ and assume $\|z^{t-1}\|_1 \leq \varepsilon(t-1)$,

$$
\begin{aligned}
\|p_1^t - p_2^t\|_1 &= \|p_0^\top P_1^t - p_0^\top P_2^t\|_1 \\
&= \|p_1^{t-1} P_1 - (p_1^{t-1} - z^{t-1}) P_2\|_1 \\
&\leq \|p_1^{t-1}(P_1 - P_2)\|_1 + \|z^{t-1} P_2\|_1 \\
&\leq \underbrace{\|p_1^{t-1}\|_1}_{\to 1} \cdot \underbrace{\|P_1 - P_2\|_{\infty,1}}_{\leq \varepsilon} + \underbrace{\|z^{t-1}\|_1}_{\leq \varepsilon(t-1)} \cdot \underbrace{\|P_2\|_{\infty,1}}_{\to 1} \\
&\leq \varepsilon + \varepsilon(t-1) = \varepsilon t \quad \square
\end{aligned}
$$

Simulation lemma

Define $\widehat{M}$ as ε-approximate for M if $\|P - \widehat{P}\|_{\infty,1} \leq \varepsilon$ and $\|r_t^\pi - \widehat{r}_t^\pi\|_\infty \leq \varepsilon$ for all π and t.

Lemma (Simulation lemma)

For an ε-approximate $\widehat{M}$, the following holds

$$\|V_T^\pi - \widehat{V}_T^\pi\|_\infty \leq \frac{(R_{max}T^2 + (R_{max} + 2)T)\varepsilon}{2}$$

Proof. Define $\Delta_t := r_t - \widehat{r}_t$ and $P_\pi^t := (P_\pi)^t$, then

$$\begin{aligned}
\|V_T^\pi - \widehat{V}_T^\pi\|_\infty &= \left\| \sum_t p_0^\top P_\pi^t r_t - \sum_t p_0^\top \widehat{P}_\pi^t \widehat{r}_t \right\|_\infty \\
&= \left\| \sum_t p_0^\top P_\pi^t r_t - p_0^\top \widehat{P}_\pi^t (r_t - \Delta_t) \right\|_\infty \\
&\leq \sum_t \|p_0^\top (P_\pi^t - \widehat{P}_\pi^t) r_t + p_0^\top \widehat{P}_\pi^t \Delta_t\|_\infty
\end{aligned}$$

Simulation lemma cont'd

$$
\begin{aligned}
\|V_T^\pi - \widehat{V}_T^\pi\|_\infty &\leq \sum_t \|p_0^\top (P_\pi^t - \widehat{P}_\pi^t) r_t\|_\infty + \|p_0^\top \widehat{P}_\pi^t \Delta_t\|_\infty \\
&\leq \sum_t \|p_0^\top (P_\pi^t - \widehat{P}_\pi^t) r_t\|_\infty + \cancelto{1}{\|p_0\|_1} \cdot \|\widehat{P}_\pi^t \Delta_t\|_\infty \\
&\leq \sum_t \underbrace{\|p_0^\top (P_\pi^t - \widehat{P}_\pi^t)\|_\infty}_{\leq \varepsilon t} \underbrace{\|r_t\|_\infty}_{\leq R_{\max}} + \cancelto{1}{\|\widehat{P}_\pi^t\|_{\infty,1}} \underbrace{\|\Delta_t\|_\infty}_{\leq \varepsilon} \\
&\leq \sum_t (R_{\max} \varepsilon t + \varepsilon) \\
&= R_{\max} \varepsilon (T(T+1))/2 + \varepsilon T.
\end{aligned}
$$

Arranging the terms yields the result $\square$

Extended simulation lemma

We can extend the previous result to an error bound on the optimal policy.

Corollary

For an ε-approximate $\widehat{M}$, the following holds

$$\|V_T^{\pi^*} - V_T^{\widehat{\pi}^*}\|_\infty \leq (R_{max}T^2 + (R_{max}+2)T)\varepsilon$$

Proof. Denote $\xi := {}^{(R_{\max}T^2 + (R_{\max}+2)T)\varepsilon}/_{2}$

$$\begin{aligned}
\|V_T^{\pi^*} - V_T^{\widehat{\pi}^*}\|_\infty &= \|V_T^{\pi^*} - \widehat{V}_T^{\pi^*} + \widehat{V}_T^{\pi^*} - V_T^{\widehat{\pi}^*}\|_\infty \\
&\leq \|V_T^{\pi^*} - \widehat{V}_T^{\pi^*}\|_\infty + \|\widehat{V}_T^{\pi^*} - V_T^{\widehat{\pi}^*}\|_\infty \\
&\leq \underbrace{\|V_T^{\pi^*} - \widehat{V}_T^{\pi^*}\|_\infty}_{\leq \xi} + \underbrace{\|\widehat{V}_T^{\widehat{\pi}^*} - V_T^{\widehat{\pi}^*}\|_\infty}_{\leq \xi} \\
&\leq 2\xi = (R_{\max}T^2 + (R_{\max}+2)T)\varepsilon \quad \square
\end{aligned}$$

The last step exploits the fact that the simulation lemma applies to all policies simultaneously.

PAC analysis of model error

The conditions to get a ε-approximate empirical MDP with high probability can be attained from the PAC bound we developed earlier by setting $d := |\mathcal{S}|$ and $K = |\mathcal{S}||\mathcal{A}|$:

$$\varepsilon := \sqrt{\frac{R_{\max}^2 |\mathcal{S}|^2 |\mathcal{A}|}{2N} \log(|\mathcal{S}||\mathcal{A}|/\delta)}$$

leading to the PAC statement below

$$P\Bigg(\|V_T^{\pi^*} - V_T^{\widehat{\pi}^*}\|_\infty \leq (R_{\max}^2 (T^2 + T) + 2R_{\max}T) \sqrt{\frac{|\mathcal{S}|^2 |\mathcal{A}|}{2N} \log(|\mathcal{S}||\mathcal{A}|/\delta)} \Bigg) \geq 1 - \delta$$

which implies a sample complexity of $\mathcal{O}(\mathcal{S}|^2 |\mathcal{A}| T^4 \log(|\mathcal{S}||\mathcal{A}|))$. Suppressing the logarithmic dependencies, we get $\widetilde{\mathcal{O}}(|\mathcal{S}|^2 |\mathcal{A}| T^4)$.

Offline MBRL: Approximate Value Iteration

Collect m samples from each (s, a) pair and compute $\widehat{p}$

$V_0 := (V_0(s))$ for some $s \in \mathcal{S}$.

for $n = 0, 1, 2, \ldots$ **do**

$\quad V_{n+1}(s) := \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \widehat{p}(s'|s,a) V_n(s') \right\}, \quad \forall s \in \mathcal{S}$

end for

Offline because the first step needs to be managed by a special **behavior policy** or a **simulator**.

Online MBRL

$n(s,a) \leftarrow 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
$n(s,a,s') \leftarrow 0$ for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$
$K \leftarrow \{(s,a) : s \in \mathcal{S}, a \in \mathcal{A}, n(s,a) = m\}$
repeat
    $\widehat{M} \leftarrow$ ConstructMDP(K)
    $\widehat{\pi} \leftarrow$ DPSolve$(\widehat{M})$
    Collect an episode $s_1, a_1, \ldots, s_T, a_T$ using policy $\widehat{\pi}$
    for $t = 1$ to $T - 1$ **do**
        if $n(s_t, a_t) < m$ **then**
            $n(s_t, a_t) \leftarrow n(s_t, a_t) + 1$
            $n(s_t, a_t, s_{t+1}) \leftarrow n(s_t, a_t, s_{t+1}) + 1$
        end if
    end for
until a stopping criterion is satisfied

Explicit Explore-Exploit (E^3) algorithm

procedure CONSTRUCTMDP(K)
    for each $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$ **do**

$$\widehat{P}(s'|s,a) = \begin{cases} \frac{n(s,a,s')}{n(s,a)}, & \text{if } (s,a) \in K \\ \mathbb{I}[s' = s], & \text{otherwise} \end{cases}$$

$$\widehat{r}(s,a) = \begin{cases} 0, & \text{if } (s,a) \in K \\ 1, & \text{otherwise} \end{cases}$$

    end for
end procedure

R-MAX algorithm

Builds on the *Optimism in the Face of Uncertainty (OFU)* principle.

procedure CONSTRUCTMDP(K)

  for each $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$ **do**

$$\widehat{P}(s'|s,a) = \begin{cases} \frac{n(s,a,s')}{n(s,a)}, & \text{if } (s,a) \in K \\ \mathbb{I}[s' = s], & \text{otherwise} \end{cases}$$

$$\widehat{r}(s,a) = \begin{cases} r(s,a), & \text{if } (s,a) \in K \\ R_{max}, & \text{otherwise} \end{cases}$$

  end for

end procedure