Reinforcement Learning with Imperfect Transition Predictions: A Bellman–Jensen Approach

Chenbei Lu[1], Zaiwei Chen[2], Tongxin Li[3], Chenye Wu[3], Adam Wierman[4]

[1] Cornell University, [2] Purdue University, [3] The Chinese University of Hong Kong, Shenzhen, [4] Caltech

The Thirty-Ninth Annual Conference on Neural Information Processing Systems (Spotlight)

Motivation and Challenges

Predictions

- Traditional RL is based on MDP with one-step transition model $P(s'|s,a)$
- Many real systems provide K-step forecasts on the future
- There is no theory or algorithms for incorporating predictions into MDPs
  - Curse of Dimensionality: predictions geometrically increase state-action space
  - Theoretical Limits with Multi-step Predictions: classical MDP/optimality theory focuses on one-step transitions

Fig.1: Weather forecast

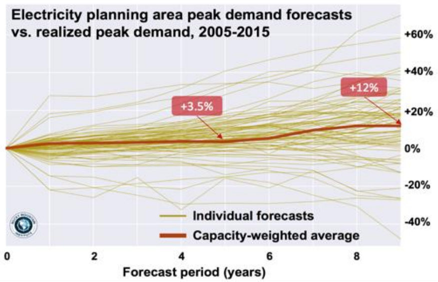

Fig.2: Power system load forecasts

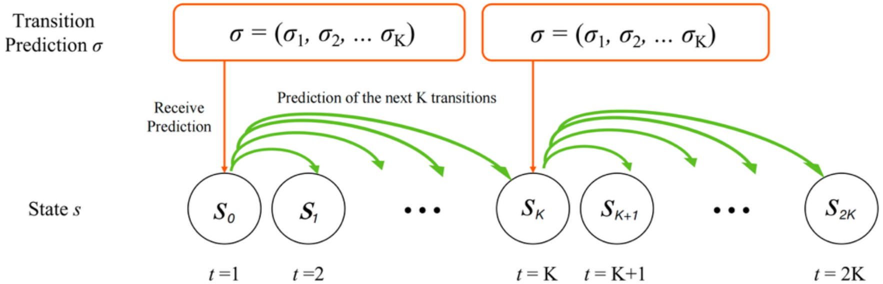

Fig. 3: MDP with Predictions

Key Idea I: Bayesian Value Theory

Compress Predictions via Bayesian Value

- Markov property helps reduce the dimensionality for decision making
- Estimate Bayesian value function (low dimensional) over the distribution of prediction $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_K)$

$$V^{\text{Bayes},\pi}_{K,\mathcal{A}^-,\varepsilon}(s) := \mathbb{E}_{\boldsymbol{\sigma}}\left[\mathbb{E}_{\pi}\left[\sum\nolimits_{t=0}^{\infty}\gamma^t r(s_t, a_t)\,\middle|\, s_0 = s, \boldsymbol{\sigma}_0 = \boldsymbol{\sigma}\right]\right].$$

Bellman Optimality & Optimal Policy

- Key result: Bayesian value function + predictions + partial transition model for $a \in \mathcal{A} \setminus \mathcal{A}^-$ is enough for optimal decision making:

Corollary 3.2 (Optimal Policy with Bayesian Value Function and Transition Predictions). *The optimal policy $\pi^*(\cdot \mid s, \boldsymbol{\sigma})$ with K-step transition predictions $\boldsymbol{\sigma}$ satisfies:*

$$\{\boldsymbol{a} \in \mathcal{A}^K \,|\, \pi^*(\boldsymbol{a} \mid s, \boldsymbol{\sigma}) > 0\} \subseteq \arg\max_{\boldsymbol{a}\in\mathcal{A}^K} \left(\sum_{t=0}^{K-1}\gamma^t\left(\sum_{s_t} P(s_t|s, \boldsymbol{a}_{0:t-1}, \boldsymbol{\sigma}_{1:t}) r(s_t, a_t)\right)\right.$$
$$\left. + \gamma^K \sum\nolimits_{s_K} P(s_K|s, \boldsymbol{a}, \boldsymbol{\sigma}) V^{\text{Bayes},*}_{K,\mathcal{A}^-,\boldsymbol{\varepsilon}}(s_K)\right), \forall\, s \in \mathcal{S}, \boldsymbol{\sigma} \in \mathcal{Q}_K. \quad (6)$$

Key Idea II: The Bellman-Jensen Gap

Theoretical Understanding

- **Key idea**: the value of predictions comes from the nested Jensen gap induced by infinite operator reordering of max-over-$\mathbb{E}$ operations. We call it the **Bellman-Jensen Gap.**
- Understanding the Bellman-Jensen Gap using Bellman expansion:
  - No prediction:

$$V^*_{\mathrm{MDP}}(s_0) = \max_{a_0}\left[r(s_0,a_0) + \gamma \mathbb{E}_{\sigma_1^*}\left[\max_{a_1}\left[r(s_1,a_1) + \gamma \mathbb{E}_{\sigma_2^*}\left[\max_{a_2}[r(s_2,a_2) + \cdots]\right]\right]\right]\right],$$

  - One-step prediction:

$$V^{\mathrm{Bayes},*}_{K=1,\mathcal{A},\mathbf{0}}(s_0) = \mathbb{E}_{\sigma_1^*}\left[\max_{a_0}\left[r(s_0,a_0) + \gamma \mathbb{E}_{\sigma_2^*}\left[\max_{a_1}[r(s_1,a_1) + \cdots]\right]\right]\right].$$

  - Infinite-step prediction:

$$\begin{aligned} V^{\mathrm{Bayes},*}_{\mathrm{off}}(s_0) &= \mathbb{E}_{\boldsymbol{\sigma}_1^*}\mathbb{E}_{\boldsymbol{\sigma}_2^*}\cdots\left[\max_{a_0}\left[r(s_0,a_0) + \gamma\max_{a_1}\left[r(s_1,a_1) + \gamma\max_{a_2}[r(s_2,a_2) + \cdots]\right]\right]\right] \\ &= \lim_{k\to\infty}\mathbb{E}_{\boldsymbol{\sigma}^*_{1:k}}\left[\max_{\boldsymbol{a}_{0:k-1}}\left[\sum_{t=0}^{k-1}\gamma^t r(s_t,a_t)\right]\right], \end{aligned}$$

Value of Predictions

Predictions

Theorem 4.1 (Bellman-Jensen Performance Bound). *Given any prediction with horizon $K \geq 1$, predictable action set $\mathcal{A}^- \subseteq \mathcal{A}$ and prediction errors $\boldsymbol{\varepsilon}$, the performance gap between the prediction-aware policy and the offline optimal policy satisfies:*

$$\max_{s \in \mathcal{S}} \left(V_{\text{off}}^{\text{Bayes},*}(s) - V_{K,\mathcal{A}^-,\boldsymbol{\varepsilon}}^{\text{Bayes},*}(s) \right) \leq \underbrace{\frac{C_1 \gamma^K \sqrt{K \log |\mathcal{A}|}}{(1-\gamma)^{\frac{6}{5}}(1-\gamma^{2K})}}_{A_1\text{:loss due to finite prediction window}} + \underbrace{\sum_{j=1}^{K} \frac{\gamma^j}{(1-\gamma)(1-\gamma^K)} \epsilon_j}_{A_2\text{:loss due to prediction error}}$$
$$+ \underbrace{C_2 \sum_{t=1}^{\infty} \gamma^t \sqrt{\log(|\mathcal{A}|^{t+1} - |\mathcal{A}^-|^{t+1} + 1)\theta_{\max}^2}}_{A_3\text{:loss due to partial action predictability}},$$

- A1: With larger prediction window *K*, the policy performance approaches the offline optimal policy geometrically fast.
- A2: The impact of predictions made later on decision-making efficiency drops exponentially.
- A3: A larger value function variance and a smaller predictive action set reduces the performance of control policy.

BOLA Algorithm

Algorithm Design

- **Offline** estimate the Bayesian value function + **Online** adapt to the high-dimensional predictions

$$\{\boldsymbol{a} \in \mathcal{A}^K \mid \pi^*(\boldsymbol{a} \mid s, \boldsymbol{\sigma}) > 0\} \subseteq \arg\max_{\boldsymbol{a} \in \mathcal{A}^K} \left(\sum_{t=0}^{K-1} \gamma^t \left(\sum_{s_t} P(s_t | s, \boldsymbol{a}_{0:t-1}, \boldsymbol{\sigma}_{1:t}) r(s_t, a_t) \right) \right.$$
$$\left. + \gamma^K \sum\nolimits_{s_K} P(s_K | s, \boldsymbol{a}, \boldsymbol{\sigma}) V^{\text{Bayes},*}_{K, \mathcal{A}^-, \boldsymbol{\varepsilon}}(s_K) \right), \forall\, s \in \mathcal{S}, \boldsymbol{\sigma} \in \mathcal{Q}_K.$$

- Avoid exponential complexity on calculating the optimal policy offline

Theoretical Guarantees

- The sample complexity of prediction-augmented MDP is lower than vanilla MDP
- With longer prediction window K, larger predictive action set $\mathcal{A}^-$, the sample complexity reduces
- Idea case: with infinite and comprehensive predictions, sample complexity reduces to zero
- Intuition: prediction provides additional information, which reduces the sample requirement

Thanks for Listening!

Welcome to our poster session:

Dec. 4, 11 a.m. PST — 2 p.m. PST, Exhibit Hall C,D,E

Chenbei Lu

Cornell University