POLITEX: Regret Bounds for Policy Iteration Using Expert Prediction

Yasin Abbasi-Yadkori, Peter L. Bartlett, Kush Bhatia, Nevena Lazić, Csaba Szepesvári, Gellért Weisz

Summary

- Setting: average-cost RL with discrete actions and value function approximation
- POLITEX: *softened* and *averaged* policy iteration.
- If the value function error after τ steps satisfies

$$\|Q_\pi - \widehat{Q}_\pi\|_\nu \le \varepsilon(\tau) = \varepsilon_0 + O(\sqrt{1/\tau})$$

where ε_0 is the approximation error, and ν is the stationary state-action distribution, then the regret of POLITEX in uniformly mixing MDPs is of the order

$$\mathfrak{R}_T = \widetilde{O}(T^{3/4} + \varepsilon_0 T)\,.$$

- Regret bound does not scale in the size of the MDP, does not depend on the "concentrability coefficient", easy to implement (no confidence bounds required).

POLITEX algorithm

Input: phase length $\tau > 0$, initial state x_0
Set $\widehat{Q}_0(x,a) = 0 \;\; \forall x, a$
for $i := 1, 2, \ldots,$ **do**

Policy iteration: $\pi_i(\cdot|x) = \underset{u\in\Delta}{\operatorname{argmin}} \langle u, \widehat{Q}_{i-1}(x,\cdot)\rangle$

$$\text{POLITEX}: \quad \pi_i(\cdot|x) = \underset{u\in\Delta}{\operatorname{argmin}} \langle u, \sum_{j=0}^{i-1} \widehat{Q}_j(x,\cdot)\rangle - \eta^{-1}\mathcal{H}(u)$$

$$\propto \exp\left(-\eta \sum_{j=0}^{i-1} \widehat{Q}_j(x,\cdot)\right)$$

Execute π_i for τ time steps and collect dataset $\mathcal{Z}_i$
Estimate $\widehat{Q}_i$ from $\mathcal{Z}_1, \ldots, \mathcal{Z}_i, \pi_1, \ldots, \pi_i$
end for

Different value estimation methods are possible. For non-linear Q-functions, one can maintain only the most recent n estimates.

POLITEX analysis

Assumptions:

- A1 (*Unichain*). MDP states form a single recurrent class.
- A2 (*Uniform mixing*). $\sup_\pi \|(\nu_\pi - \nu)^\top H_\pi\|_1 \le \exp(-\kappa^{-1})\|\nu_\pi - \nu\|_1$, where H_π is the transition probability matrix for (s,a) pairs under π, and ν_π is the stationary state-action distribution.

Let $\lambda_\pi = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T c(x_t, \pi(x_t))$ be the average cost of a policy.
Regret decomposes as

$$\mathfrak{R}_T = \sum_{t=1}^T c(x_t,a_t) - c(x_t^*, a_t^*) = V_T + \overline{\mathfrak{R}}_T + W_T$$

$$V_T = \sum_{t=1}^T c_t - \lambda_{\pi_{(t)}} \quad \overline{\mathfrak{R}}_T = \sum_{t=1}^T \lambda_{\pi_{(t)}} - \lambda_{\pi^*} \quad W_T = \sum_{t=1}^T \lambda_{\pi^*} - c_t^*$$

- V_T and W_T are the differences between the instantaneous and average costs, respectively scale as $\kappa T^{3/4}$ and $\kappa\sqrt{T}$ w.h.p.
- $\overline{\mathfrak{R}}_T$ is the *pseudo-regret*, bounded using the regret bound of the Exponentially Weighted Average (EWA) forecaster and the value error bound.

Least squares policy estimation (LSPE)

Bounding $\overline{\mathfrak{R}}_T$ requires $\widehat{Q}_i(x,a) \in [b, b + Q_{\max}]$ and bounded error

$$\|Q_{\pi_i} - \widehat{Q}_i\|_{\nu^*},\; \|Q_{\pi_i} - \widehat{Q}_i\|_{\mu^*\otimes\pi_i} \le \varepsilon(\tau) = \varepsilon_0 + O(\sqrt{1/\tau})\,.$$

LSPE:

- Linear value function approximation $\widehat{Q}_\pi = \Psi w_\pi$
- Obtains a simulation-based solution to the projected Bellman equation $\Psi w = \Pi_\pi(c - \lambda\mathbf{1} + H\Psi w)$

Under the additional assumptions below, w.h.p. LSPE satisfies the error bound, and $Q_{\max}$ is bounded.

- A3 (*Features*). Columns of $[\Psi\ \mathbf{1}]$ are linearly independent, and features are bounded.
- A4 (*Feature excitation*). For any π, $\lambda_{\min}(\Psi^\top \mathrm{diag}(\nu_\pi)\Psi) \ge \sigma > 0$.

Experiments

POLITEX + LSPE on Queueing problems

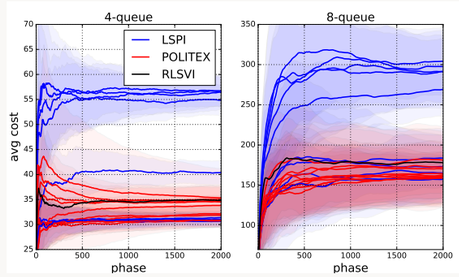

Figure: Average cost at the end of each phase for the 4-queue and 8-queue environments (mean and std of 50 runs), for different η.

POLITEX + neural networks on Atari

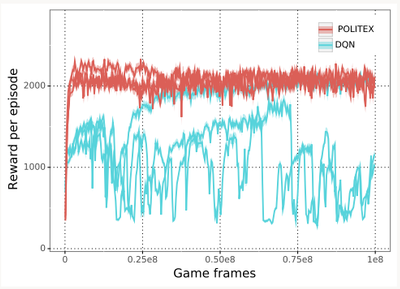

Figure: Ms Pacman game scores obtained by the agents at the end of each game, using runs with different random seeds.

Related work

Y. Abbasi-Yadkori, N. Lazić, and C. Szepesvári. "Regret bounds for model-free LQ control." AISTATS (2019).
E. Even-Dar, S. M. Kakade, and Y. Mansour. "Online MDPs." Mathematics of Operations Research 34.3 (2009).
H. Yu and D. P. Bertsekas. "Convergence results for some temporal difference methods based on least squares." IEEE Transactions on Automatic Control 54.7 (2009).