Performance bound for Approximate Optimistic Policy Iteration

Bruno Scherrer, Christophe Thiery

Performance bound for Approximate Optimistic Policy Iteration

Christophe Thiery, Bruno Scherrer

We provide here a proof of the performance bound theorem published in Thiery and Scherrer (2010). This theorem applies to Least-Squares λ Policy Iteration and more generally approximate, optimistic Policy Iteration algorithms.

Theorem 1 (Performance bound for Approximate Optimistic Policy Iteration)

Let $(\lambda_n)_{n\geq 1}$ be a sequence of positive weights such that $\sum_{n\geq 1}\lambda_n = 1$. Let Q_0 be an arbitrary initialization. We consider an iterative algorithm that generates the sequence $(\pi_k, Q_k)_{k\geq 1}$ with

$$\pi_{k+1} \leftarrow \text{greedy}(Q_k),$$
$$Q_{k+1} \leftarrow \sum_{n\geq 1}\lambda_n (B_{\pi_{k+1}})^n Q_k + \epsilon_{k+1}.$$

ϵ_{k+1} is the approximation error made when estimating the next value function. Let ϵ be a uniform majoration of that error, i.e. for all k, $\|\epsilon_k\|_\infty \leq \epsilon$. Then

$$\limsup_{k\to\infty} \|Q^* - Q^{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\epsilon.$$

Proof

Notations and main idea of the proof We will use the following notations:

- $b_k = Q_k - B_{\pi_{k+1}} Q_k$ is the Bellman error,
- $d_k = Q^* - (Q_k - \epsilon_k)$ is the difference between the optimal value function and the Q_k iterate (before error),
- $s_k = Q_k - \epsilon_k - Q^{\pi_k}$ is the difference between the Q_k iterate (before error) and the (true) value of the policy π_k,
- $\beta = \sum_{n\geq 1}\lambda_n\gamma^n$ (note that $0 \leq \beta \leq \gamma$).

The distance between the value of the optimal policy and the value of the current policy can be formulated as

$$\begin{aligned}\|Q^* - Q^{\pi_k}\|_\infty &= \max(Q^* - Q^{\pi_k}) \\ &= \max(Q^* - Q_k + \epsilon_k + Q_k - \epsilon_k - Q^{\pi_k}) \\ &= \max(d_k + s_k) \\ &\leq \max d_k + \max s_k \end{aligned} \tag{1}$$

The idea of the proof is to compute upper bounds on d_k and s_k. As we will see, the bounds we will obtain will both depend on an upper on the Bellman error b_k, that we derive first.

An upper bound on the Bellman error b_k: As π_{k+1} is the greedy policy with respect to Q_k, we have $B_{\pi_k} Q_k \leq B_{\pi_{k+1}} Q_k$, which allows us to write

$$\begin{aligned} b_k &= Q_k - B_{\pi_{k+1}} Q_k \\ &= Q_k - B_{\pi_k} Q_k + B_{\pi_k} Q_k - B_{\pi_{k+1}} Q_k \\ &\leq Q_k - B_{\pi_k} Q_k \\ &= (Q_k - \epsilon_k + \epsilon_k) - B_{\pi_k}(Q_k - \epsilon_k + \epsilon_k) \\ &= (Q_k - \epsilon_k) - B_{\pi_k}(Q_k - \epsilon_k) + \epsilon_k - \gamma P_{\pi_k} \epsilon_k \\ &= \sum_{n\geq 1} \lambda_n \left[(B_{\pi_k})^n Q_{k-1}\right] - \sum_{n\geq 1} \lambda_n \left[(B_{\pi_k})^{n+1} Q_{k-1}\right] + (I - \gamma P_{\pi_k})\epsilon_k \\ &= \sum_{n\geq 1} \lambda_n \left[(B_{\pi_k})^n Q_{k-1}] - (B_{\pi_k})^{n+1} Q_{k-1}\right] + (I - \gamma P_{\pi_k})\epsilon_k \\ &= \sum_{n\geq 1} \lambda_n (\gamma P_{\pi_k})^n (Q_{k-1} - B_{\pi_k} Q_{k-1}) + (I - \gamma P_{\pi_k})\epsilon_k \\ &= \sum_{n\geq 1} \lambda_n (\gamma P_{\pi_k})^n b_{k-1} + (I - \gamma P_{\pi_k})\epsilon_k. \end{aligned}$$

By using the fact that P_{π_k} is a stochastic matrix, we have

$$\max b_k \leq \sum_{n\geq 1} \lambda_n \gamma^n \max b_{k-1} + (1+\gamma)\epsilon = \beta \max b_{k-1} + (1+\gamma)\epsilon.$$

We then deduce by induction that

$$\max b_k \leq \sum_{j=0}^{k-1} \beta^j (1+\gamma)\epsilon + \beta^k \max b_0 = \frac{1+\gamma}{1-\beta}\epsilon + O(\gamma^k). \tag{2}$$

An upper bound on d_k : Let us now consider the d_k term and its evolution.

$$\begin{aligned} d_{k+1} &= Q^* - (Q_{k+1} - \epsilon_{k+1}) \\ &= Q^* - \sum_{n\geq 1} \lambda_n (B_{\pi_{k+1}})^n Q_k \\ &= \sum_{n\geq 1} \lambda_n \left[Q^* - (B_{\pi_{k+1}})^n Q_k\right]. \end{aligned} \tag{3}$$

Since π_{k+1} is the greedy policy with respect to Q_k, we have $B_{\pi^*}Q_k \leq B_{\pi_{k+1}}Q_k$. Therefore

$$\begin{aligned}
& Q^* - (B_{\pi_{k+1}})^n Q_k \\
= \; & B_{\pi^*}Q^* - B_{\pi^*}Q_k + B_{\pi^*}Q_k - B_{\pi^{k+1}}Q_k + B_{\pi_{k+1}}Q_k - \\
& -(B_{\pi_{k+1}})^2 Q_k + (B_{\pi_{k+1}})^2 Q_k - \ldots + (B_{\pi_{k+1}})^{n-1} Q_k - (B_{\pi_{k+1}})^n Q_k \\
\leq \; & B_{\pi^*}Q^* - B_{\pi^*}Q_k + \gamma P_{\pi_{k+1}}(Q_k - B_{\pi_{k+1}}Q_k) + \\
& +(\gamma P_{\pi_{k+1}})^2 (Q_k - B_{\pi_{k+1}}Q_k) + \ldots + (\gamma P_{\pi_{k+1}})^{n-1}(Q_k - B_{\pi_{k+1}}Q_k) \\
= \; & \gamma P_{\pi^*}(Q^* - Q_k) + \\
& + \left[\gamma P_{\pi_{k+1}} + (\gamma P_{\pi_{k+1}})^2 + \ldots + (\gamma P_{\pi_{k+1}})^{n-1}\right](Q_k - B_{\pi_{k+1}}Q_k) \\
= \; & \gamma P_{\pi^*}(Q^* - (Q_k - \epsilon_k)) - \gamma P_{\pi^*}\epsilon_k + \\
& + \left[\gamma P_{\pi_{k+1}} + (\gamma P_{\pi_{k+1}})^2 + \ldots + (\gamma P_{\pi_{k+1}})^{n-1}\right](Q_k - B_{\pi_{k+1}}Q_k) \\
= \; & \gamma P_{\pi^*} d_k - \gamma P_{\pi^*}\epsilon_k + \left[\gamma P_{\pi_{k+1}} + (\gamma P_{\pi_{k+1}})^2 + \ldots + (\gamma P_{\pi_{k+1}})^{n-1}\right] b_k.
\end{aligned}$$

As P_{π^*} and $P_{\pi_{k+1}}$ are stochastic matrices, we deduce

$$\begin{aligned}
\max[Q^* - (B_{\pi_{k+1}})^n Q_k] \; & \leq \gamma \max d_k + \gamma\epsilon + (\gamma + \gamma^2 + \ldots + \gamma^{n-1}) \max b_k \\
& = \gamma \max d_k + \gamma\epsilon + \frac{\gamma - \gamma^n}{1-\gamma} \max b_k.
\end{aligned}$$

By using Equation 3, we obtain the following induction on $\max d_k$:

$$\max d_{k+1} \leq \gamma \max d_k + \gamma\epsilon + \sum_{n\geq 1} \lambda_n \left[\frac{\gamma - \gamma^n}{1-\gamma} \max b_k\right].$$

With the help of the Bellman error upper bound obtained earlier (Equation 2) we obtain

$$\begin{aligned}
\max d_{k+1} \; & \leq \gamma \max d_k + \gamma\epsilon + \sum_{n\geq 1} \lambda_n \left[\frac{\gamma - \gamma^n}{(1-\gamma)(1-\beta)}\right](1+\gamma)\epsilon + O(\gamma^k) \\
& = \gamma \max d_k + \gamma\epsilon + \frac{\gamma - \beta}{(1-\gamma)(1-\beta)}(1+\gamma)\epsilon + O(\gamma^k)
\end{aligned}$$

which gives, by taking the limit superior,

$$\limsup_{k\to\infty} \max d_k \leq \frac{\gamma}{1-\gamma}\epsilon + \left[\frac{\gamma - \beta}{(1-\gamma)^2(1-\beta)}\right](1+\gamma)\epsilon. \tag{4}$$

An upper bound on s_k : Let us now consider the s_k term from Equation 1:

$$\begin{aligned}
s_{k+1} \; & = Q_{k+1} - \epsilon_{k+1} - Q^{\pi_{k+1}} \\
& = \sum_{n\geq 1} \lambda_n \left[(B_{\pi_{k+1}})^n Q_k\right] - (B_{\pi_{k+1}})^\infty Q_k \\
& = \sum_{n\geq 1} \lambda_n \left[(B_{\pi_{k+1}})^n Q_k - (B_{\pi_{k+1}})^\infty Q_k\right].
\end{aligned} \tag{5}$$

It can be seen that

$$
\begin{aligned}
& (B_{\pi_{k+1}})^n Q_k - (B_{\pi_{k+1}})^\infty Q_k \\
= \; & (B_{\pi_{k+1}})^n Q_k - (B_{\pi_{k+1}})^{n+1} Q_k + (B_{\pi_{k+1}})^{n+1} Q_k - (B_{\pi_{k+1}})^{n+2} Q_k + \ldots \\
= \; & (\gamma P_{\pi_{k+1}})^n (Q_k - B_{\pi_{k+1}} Q_k) + (\gamma P_{\pi_{k+1}})^{n+1} (Q_k - B_{\pi_{k+1}} Q_k) + \ldots \\
= \; & (\gamma P_{\pi_{k+1}})^n [I + \gamma P_{\pi_{k+1}} + (\gamma P_{\pi_{k+1}})^2 + \ldots] b_k.
\end{aligned}
$$

As above, by using the stochasticity of $P_{\pi_{k+1}}$, we obtain

$$
\max[(B_{\pi_{k+1}})^n Q_k - (B_{\pi_{k+1}})^\infty Q_k] \leq \gamma^n (1 + \gamma + \gamma^2 + \ldots) \max b_k = \frac{\gamma^n}{1-\gamma} \max b_k.
$$

By using Equation 5, we obtain an upper bound on $\max s_{k+1}$:

$$
\max s_{k+1} \leq \frac{1}{1-\gamma} \left[\sum_{n \geq 1} \lambda_n \gamma^n \max b_k \right].
$$

With the help of the Bellman error upper bound (Equation 2) and by taking the limit superior, we have

$$
\limsup_{k \to \infty} \max s_k \leq \frac{1}{1-\gamma} \left(\sum_{m \geq 1} \lambda_n \gamma^n \frac{1+\gamma}{1-\beta} \epsilon \right) = \frac{\beta}{(1-\gamma)(1-\beta)} (1+\gamma)\epsilon. \tag{6}
$$

Conclusion of the proof Finally, let us get back to Equation 1 and use the upper bounds we just derived for d_k (Equation 4) and s_k (Equation 6):

$$
\begin{aligned}
\limsup_{k \to \infty} \|Q^* - Q^{\pi_k}\|_\infty & \leq \limsup_{k \to \infty} \max d_k + \limsup_{k \to \infty} \max s_k \\
& = \frac{\gamma}{1-\gamma} \epsilon + \left[\frac{\gamma - \beta}{(1-\gamma)^2 (1-\beta)} + \frac{\beta}{(1-\gamma)(1-\beta)} \right] (1+\gamma)\epsilon. \\
& = \frac{\gamma}{1-\gamma} \epsilon + \left[\frac{\gamma - \beta + (1-\gamma)\beta}{(1-\gamma)^2 (1-\beta)} \right] (1+\gamma)\epsilon. \\
& = \frac{\gamma}{1-\gamma} \epsilon + \left[\frac{\gamma}{(1-\gamma)^2} \right] (1+\gamma)\epsilon. \\
& = \frac{\gamma(1-\gamma) + \gamma(1+\gamma)}{(1-\gamma)^2} \epsilon \\
& = \frac{2\gamma}{(1-\gamma)^2} \epsilon. \quad \blacksquare
\end{aligned}
$$

References

Thiery, C. and B. Scherrer (2010). Least-squares λ policy iteration: Bias-variance trade-off in control problems. In *ICML'10: Proceedings of the 27th Annual International Conference on Machine Learning.*