Error Bounds for Approximate Dynamic Programming

JONATHAN COLAÇO CARR
McGill University
May 3, 2022.

This work highlights some important error bounds on the performance loss in Approximate Dynammic Programming (ADP). We first review some classic error bounds on the supremum (L_∞) norm of the performance loss before discussing the more recent, sharper bounds obtained with $L_{p<\infty}$ norms. We also discuss how the $L_{p<\infty}$ bounds can be used to analyze the error caused by differences in sampling and future-state distributions.

1. Introduction

The general setting of this paper is the *finite-action discounted MDP*. This is a 5-tuple $M = (\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is a measurable state space, $\mathcal{A}$ is a finite set of a actions, P is the probability transition kernel, $\mathcal{R}$ is the reward kernel, and $0 \leq \gamma < 1$ is the discount factor. We denote $r(s,a) = \mathbb{E}[\mathcal{R}(\cdot|s,a)]$. We 'act' in an MDP by following a *policy*[1] $\pi : \mathcal{S} \to \mathcal{A}$. Our goal is to find a policy which maximizes the discounted sum of rewards received after each action.

One way to find the best policy is to keep track of a *value function* $V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t R_t | S_0 = s\right]$ or an *action-value* function $Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t R_t | S_0 = s, A_0 = a\right]$. For a discounted MDP, the *optimal value* and *action-value* functions are defined by $V^\star(s) = \sup_\pi V^\pi(s)$ and $Q^\star(s,a) = \sup_\pi(s,a)$. A policy $\pi^\star$ is *optimal* if it achieves the best value in every state. One of the key principles of dynamic programming is that greedy policies with respect to $Q^\star$ or $V^\star$ are optimal [2]. For a policy π, we define the Bellman operators acting on V and Q by

$$(T^\pi V)(s) = r(s,\pi(s)) + \gamma \int V^\pi(s')P(ds'|s,a)$$

$$(T^\pi Q)(s,a) = r(s,a) + \gamma \int Q^\pi(s',\pi(s'))P(ds'|s,a).$$

Similarly, we define the Bellman optimality operators as

$$(T^\star V)(s) = \max_a \left\{ r(s,a) + \gamma \int V(s')P(ds'|s,a) \right\}$$

$$(T^\star Q)(s,a) = r(s,a) + \gamma \int \max_{a'} Q(s',a')P(ds'|s,a).$$

The fixed point of the Bellman operators are optimal value and action-value functions [2]. However, often the full MDP is unknown or too large to compute to solve directly. In practice this means we resort to finding approximate value functions and policies with iterative algorithms.

Approximate dynamic programming (ADP) [3] provides a framework to analyze the behaviour of these approximate solutions. It studies the influence of an *approximation operator* $\mathscr{A}$ on our value function estimates V_k. This gives rise to the Approximate Value Iteration (AVI) algorithm shown in Algorithm 1.

[1] For this paper we focus on deterministic policies. The more general definition of a policy is a mapping from $\mathcal{S}$ to probability measures over $\mathcal{A}$.

In the exact same way, we can describe approximate value iteration for action-value functions, applying updates $Q_{k+1} = \mathscr{A}T^\star Q_k$, and acting greedily with respect to Q_{k+1}. This is used in Approximate Policy Iteration (API) to find near-optimal policies, as shown in Algorithm 2.

Algorithm 1 Approximate Value Iteration (AVI)

```
Start with $v_0$
Update values: $v_{k+1} = \mathscr{A}T^\star v_k$                ▷ $v_{k+1} \approx T^\star v_k$
Return control policy $\pi_{k+1} = Greedy(v_{k+1})$
```

Algorithm 2 Approximate Policy Iteration (API)

```
Start with $\pi_0$
loop
    Policy Evaluation: $q_i = \mathscr{A} q_{\pi_i}$                ▷ $q_i \approx q_{\pi_i}$
    Greedy Improvement: $\pi_{i+1} = \text{argmax}_a\, q_i(s,a)$
```

In general, these sequences of approximate value and action-value functions do not converge to the optimal fixed points [3]. However, we are still able to bound the error of these approximations after a fixed number of iterations. The rest of this work is devoted to reviewing such error bounds for AVI and API.

2. L_∞ Error Bounds

This section reviews some of the famous results attributed to Bertsekas and Tsitsiklis [3]. They focus mainly on bounding the error of the approximate solutions with the L_∞ norm, $\|V(s)\|_\infty := \sup_{s \in \mathcal{S}} V(s)$. The goal of this section is to understand how to bound errors in the L_∞ setting before moving to the more complex $L_{p<\infty}$ case. The first result bounds the performance loss in AVI by an the approximation error incurred at each iteration as well as an initial error term.

Theorem 2.1 (Bertsekas & Tsitsiklis, 1996). *Given an MDP, let Q_k be the value function returned by AVI after k steps and let π_k be its corresponding greedy policy. Then*

$$\|Q^\star - Q_{\pi_n}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq k \leq n} \|T^\star q_k - \mathscr{A}T^\star Q_k\|_\infty + \frac{2\gamma^{n+1}}{1-\gamma}\varepsilon_0,$$

where $\varepsilon_0 = \|Q^\star - Q_0\|_\infty$ and $T^\star$ is the optimal Bellman operator associated with the MDP.

The proof of Theorem 2.1 is provided in Appendix A. The next theorem shows a similar result for approximate policy iteration, with the proof shown in Appendix B.

Theorem 2.2 (Bertsekas & Tsitsiklis, 1996)**.** *Given an MDP, let Q_k and π_k be the value function and policy acheived by API at iteration k. Then*

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

Note again that the error in the approximate solution will depend on the approximation error at iteration k, $\|Q_{\pi_k} - Q_k\|_\infty$.

The above two results quantify error in terms of the L_∞ norm. Compared to L_p norms, the L_∞ norm is quite conservative; it is possible that a learned value function might have a large L_∞ norm but a small L_p norm. This motivates the search for error bounds in terms of L_p norm, as discussed in the next section.

3. L_p Error Bounds for $p < \infty$

The goal of this section is to understand how to bound errors arising from AVI and API using L_p norms. While there are several papers on this subject [3, 7, 8, 5, 1], we focus on important results from Farahmand et. al [5].

3.1. *Background Math*

In order to address error bounds for L_p spaces, we require a few extra definitions from [5]. For our measurable state space $\mathcal{S}$, we define $\Delta(\mathcal{S})$ as the set of probability measures over the σ-algebra of $\mathcal{S}$. For a probability measure $\rho \in \Delta(\mathcal{S})$ and transition kernel P^π, we define $\rho P^\pi(ds') = \int P(ds'|s,\pi(s))d\rho(s)$ as the m-step-ahead probability distribution of states if the starting state distribution is ρ and we follow P^π for m steps. In this section, we also let $\|V\|_{p,\nu}$ be the $L^p(\nu)$ norm over measurable functions $V : \mathcal{S} \to \mathbb{R}$,

$$\|V\|_{p,\nu} := \int_{\mathcal{S}} |V(s)|^p d\nu(s).$$

Similarly, for a measurable function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define

$$\|Q\|_{p,\nu} = \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} |Q(s,a)|^p d\nu(s).$$

The performance loss $\|Q^\star - Q^{\pi_k}\|_{p,\rho}$ will partially depend on the difference between the sampling distribution ν and the future state distribution $\rho P^{\pi_1} P^{\pi_2} \cdots$. In Farahmand et. al's work (among others [7, 8]), this difference is encapsulated by the following *concentrability coefficients*.

Definition 3.1 (Expected Concentrability of the Future-State Distribution)**.** Given $\rho, \nu \in \Delta(\mathcal{S})$, $\nu \ll \lambda$[2], $m \geq 0$, and an arbitrary sequence of stationary policies $\{\pi_m\}_{m\geq 1}$, let $\rho P^{\pi_1} P^{\pi_2} \cdots P^{\pi_m} \in \Delta(\mathcal{S})$ denote the future-state distribution obtained when the first state is distributed according to ρ and then we follow the sequence of policies $\{\pi_k\}_{k=1}^m$.

[2] The notation here is to say that λ (the Lebesgue measure) dominates ν. I.e $\lambda(A) = 0 \Rightarrow \nu(A) = 0$.

a. For AVI results, we define the following concentrability coefficient,

$$c_{VI,\rho,\nu}(m_1,m_2;\pi) := \left(\mathbb{E}_{X\sim\nu}\left[\left|\frac{d\left(\rho(P^{\pi})^{m_1}(P^{\pi^\star})^{m_2}\right)}{d\nu}(X)\right|^2\right]\right)^{\frac{1}{2}}.$$

If the future-state distribution is not absolutely continuous with respect to ν, then we take $c_{VI,\rho,\nu}(m_1,m_2;\pi)=\infty$.

b. For API results, we define the following concentrability coefficients,

$$c_{PI_1,\rho,\nu}(m_1,m_2;\pi) := \left(\mathbb{E}_{X\sim\nu}\left[\left|\frac{d\left(\rho(P^{\pi^\star})^{m_1}(P^{\pi})^{m_2}\right)}{d\nu}(X)\right|^2\right]\right)^{\frac{1}{2}}$$

$$c_{PI_2,\rho,\nu}(m_1,m_2;\pi_1,\pi_2) := \left(\mathbb{E}_{X\sim\nu}\left[\left|\frac{d\left(\rho(P^{\pi^\star})^{m_1}(P^{\pi_1})^{m_2}P^{\pi_2}\right)}{d\nu}(X)\right|^2\right]\right)^{\frac{1}{2}}$$

$$c_{PI_3,\rho,\nu} := \left(\mathbb{E}_{X\sim\nu}\left[\left|\frac{d\left(\rho P^{\pi^\star}\right)}{d\nu}(X)\right|^2\right]\right)^{\frac{1}{2}}.$$

If the future state distribution is not absolutely continuous with respect to ν, then we take $c_{PI_1,\rho,\nu}=\infty$ (and similarly for the others).

These concentrability coefficients use the Radon-Nikodym derivative to measure the difference between the future state distribution and the sampling distribution. Farahmand et. al use these concentrability coefficients to show that the effect of the future-state distribution on the performance loss depends on the expectation of the squared Radon-Nikodym derivative, as reviewed Sections 3.2 and 3.3.

3.2. *AVI Error Bounds*

For AVI error bounds, we consider a sequence of functions $(V_k)_{k\geq 1}$ where V_0 is initialised arbitrarily and $V_{k+1}=\mathscr{A}T^\star V_k$, so that $V_{k+1}\approx T^\star V_k$. We define the approximation error at each iteration by $\varepsilon_k = T^\star V_k - V_{k+1}$. In this section, we are interested in studying how the approximation errors $\{\varepsilon_k\}_{k=0}^{K}$ relate to the performance loss $\|V^\star - V^{\pi_k}\|_{p,\rho}$ of the obtained policy π_K, which is greedy with respect to V_{K-1}. For ease of notation, we will also introduce

$$\alpha_k = \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}}, \quad 0\leq k<K.$$

Theorem 3.1 (Farahmand, 2010 - Error Propagation for AVI)**.** *Let $p\geq 1$ be a real number, K be a positive integer, and $V_{max}\leq \frac{R_{max}}{1-\gamma}$. Then for any sequence $\{V_k\}_{k=0}^{K-1}\subset B(\mathcal{S},V_{max})$ (the space of V_{max}-bounded measurable functions defined on $\mathcal{S}$) and the corresponding sequence $\{\varepsilon_k\}_{k=0}^{K-1}$ of AVI approximation* *errors, we have*

$$\|V^\star - V^{\pi_K}\|_{p,\rho} \le \frac{2\gamma}{(1-\gamma)^2}\left[\inf_{r\in[0,1]} C^{\frac{1}{2p}}_{VI,\rho,\nu}(K;r)\,\mathcal{E}^{\frac{1}{2p}}(\varepsilon_0,\dots,\varepsilon_{K-1};r) + \frac{2}{1-\gamma}\gamma^{\frac{K}{p}}R_{max}\right],$$

where

$$C_{VI,\rho,\nu}(K;r) = \left(\frac{1-\gamma}{2}\right)^2 \sup_{\pi'\in\Delta(\mathcal{A})} \sum_{k=0}^{K-1} \alpha_k^{2(1-r)} \Bigg(\sum_{m\ge 0}\gamma^m (c_{VI,\rho,\nu}(m,K-k;\pi') + c_{VI,\rho,\nu}(m+1,K-k-1;\pi'))\Bigg)^2,$$

and $\mathcal{E}(\varepsilon_0,\dots,\varepsilon_{K-1};r) = \sum_{k=0}^{K-1}\alpha_k^{2r}\|\varepsilon_k\|_{2p,\nu}$.

The interpretation of this theorem is that the performance loss after K iterations of AVI is bounded (up to some number $\frac{4\gamma}{(1-\gamma)^3}\gamma^{\frac{K}{p}}R_{\max}$) by the infimum of a product containing two terms. The first is a discounted sum of the concentrability coefficients (ie. measures of the difference between sampling and future-state distributions) from each iteration $0 \le k < K$. The second is a discounted sum of the errors $\varepsilon_k = T^\star V_k - V_{k+1}$ incurred at each iteration.

The full proof of Theorem 3.1 can be found in [5]. At a high level, the proof first obtains a bound on $|V^\star - V_k|$ for an arbitrary k. Using induction, we can then obtain a bound on $|V^\star - V^K|$ (accumulating the error at each of the $K-1$ steps). Next, we find a pointwise bound on $|V^\star - V^{\pi_K}|$, taking a supremum over sequences of policies. From there, we can apply a series of Hölder, Jensen, and Cauchy-Schwarz inequalities to recover the sum of concentrability coefficients in the theorem.

3.3. *API Error Bounds*

Similarly to AVI, we consider the API procedure and the sequence $\{Q_k\}_{k=0}^{K-1}$, where Q_k is the approximate action-value function for the greedy policy π_k. At each iteration, define the *Bellman Residual* (BR) and policy *Approximation Error* (AR) by

$$\varepsilon_k^{\text{BR}} = Q_k - T^{\pi_k}Q_k \tag{3.1}$$

$$\varepsilon_k^{\text{AE}} = Q_k - Q^{\pi_k}. \tag{3.2}$$

The following result studies how the error in our approximate policy relates to these errors under the $L_{2p}(\nu)$ norm.

Theorem 3.2 (Farahmand, 2010 - Error Propagation for API)**.** *Let $p \ge 1$ be a real number, K be a positive integer, and $Q_{max} \le \frac{R_{max}}{1-\gamma}$. Then for any sequence $\{Q_k\}_{k=0}^{K-1} \subset B(\mathcal{S}\times\mathcal{A}, Q_{max})$ (space of bounded measurable functions defined on $\mathcal{S}\times\mathcal{A}$) and the corresponding sequence $\{\varepsilon_k\}_{k=0}^{K-1}$ defined in (3.1) or (3.2), we have*

$$\|Q^\star - Q^{\pi_K}\|_{p,\rho} \le \frac{2\gamma}{(1-\gamma)^2}\left[\inf_{r\in[0,1]} C^{\frac{1}{2p}}_{PI(BR/AE),\rho,\nu}(K;r)\,\mathcal{E}^{\frac{1}{2p}}(\varepsilon_0,\dots,\varepsilon_{K-1};r) + \gamma^{\frac{K}{p}-1}R_{max}\right],$$

where $\mathcal{E}(\varepsilon_0,\dots,\varepsilon_{K-1};r) = \sum_{k=0}^{K-1}\alpha_k^{2r}\|\varepsilon_k\|_{2p,\nu}$

a. *If $\varepsilon_k = \varepsilon^{BR}$ for all $0 \le k < K$, then*

$$C_{PI(BR),\rho,\nu}(K;r) = \left(\frac{1-\gamma}{2}\right)^2 \sup_{\pi'_0,\dots,\pi'_K} \sum_{k=0}^{K-1} \alpha_k^{2(1-r)} \Bigg(\sum_{m\ge 0} (c_{PI_1,\rho,\nu}(K-k-1,m+1;\pi'_{k+1}) + c_{PI_1,\rho,\nu}(K-k,m;\pi'_k)) \Bigg)^2.$$

b. *If $\varepsilon_k = \varepsilon^{AE}$ for all $0 \le k < K$, then*

$$C_{PI(AE),\rho,\nu}(K;r,s) = \left(\frac{1-\gamma}{2}\right)^2 \sup_{\pi'_0,\dots,\pi'_K} \sum_{k=0}^{K-1} \alpha_k^{2(1-r)} \Bigg(\sum_{m\ge 0} \gamma^m c_{PI_1,\rho,\nu}(K-k-1,m+1;\pi'_{k+1}) + \sum_{m\ge 1} \gamma^m c_{PI_2,\rho,\nu}(K-k-1,m;\pi'_{k+1},\pi'_k) + c_{PI_3,\rho,\nu} \Bigg)^2.$$

The interpretation of Theorem 3.2 is the almost the same as that of Theorem 3.1: we can bound the error (up to some number $\frac{2\gamma}{(1-\gamma)^2}\gamma^{\frac{K}{P}-1}R_{\max}$) by the infimum of the product of two terms. The first is a discounted sum of concentrability coefficents and the second is a discounted sum of error terms (either Bellman residual or approximation errors). The full proof of Theorem 3.2 can be found in [5]; its proof is very similar to that of Theorem 3.1.

4. Discussion

Although less approachable at first glance, the L_p error bounds provide some serious advantages to their L_∞ counterparts. The first is the practical advantage of L_p norms over L_∞ norms. Not only do the L_p bounds provide more sharper estimates to the error; they also are more compatible with familiar L_p norms used in other fields machine learning [6, 9].

The second advantage of L_p results is that it provides a better understanding of how the future state distribution effects the performance loss. Theorems 3.1 and 3.2 relate the performance loss to the expectation of the squared Radon-Nikodym derivative between the future state distribution and the sampling distribution. Before Farahmand et. al's work, the key factor in performance loss was believed to be the supremum over this derivative, rather than an expectation [1, 8]. Intuitively these results show that if for some subset of the state space $\mathcal{S}' \subset \mathcal{S}$, the ratio $\frac{d(\rho(P^\pi)^m)}{d\nu}$ is large but the sampling probability $\nu(\mathcal{S}')$ is very small, the performance loss due to it is still small.

5. Conclusion

Both the L_∞ and L_p bounds presented in this paper provide valuable insight into the error propagation that arises in approximate dynamic programming. The L_p bounds on the performance loss are in-part due to concentrability coefficients that capture relationship between sampling and future-state distributions. A deeper dive into these coefficients could offer better bounds for ADP algorithms, which in-turn underlie some of the most important problems in reinforcement learning.

References

1. A. Antos, C. Szepesvári, and R. Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. volume 71, pages 574–588, 09 2006.
2. R. Bellman. *Dynamic Programming*. Dover Publications, 1957.
3. D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming.* Athena Scientific, Belmont, MA, 1996.
4. D. Borsa. Approximate dynamic programming.
5. A.-m. Farahmand, R. Munos, and C. Szepesvári. Error propagation for approximate policy and value iteration. pages 568–576, 01 2010.
6. M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018.
7. R. Munos. Error bounds for approximate policy iteration. In *ICML*, 2003.
8. R. Munos. Performance bounds in lp-norm for approximate value iteration. *SIAM J. Control. Optim.*, 46:541–561, 2007.
9. S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014.

A. Proof of Theorem 1

Proof (Adapted from [4]) Let $\varepsilon = \|Q^\star - Q_{\pi_n}\|_\infty \max_{0\le k<n} \|T^\star Q_k - \mathscr{A}T^\star Q_k\|_\infty$. Then for all $k < n$,

$$\begin{aligned} \|Q^\star - Q_{k+1}\|_\infty &\le \|Q^\star - T^\star Q_k\|_\infty + \|T^\star Q_k - Q_{k+1}\|_\infty \\ &\le \|T^\star Q^\star - T^\star Q_k\|_\infty + \varepsilon \\ &\le \gamma \|Q^\star Q_k\|_\infty + \varepsilon. \end{aligned}$$

So

$$\begin{aligned} \|Q^\star - Q_k\|_\infty &\le \gamma \|Q^\star - Q_{k-1}\|_\infty + \varepsilon \\ &\le \gamma(\gamma \|Q^\star - Q_{k-2}\|_\infty + \varepsilon) + \varepsilon \\ &\dots \\ &\le \gamma^k \|Q^\star - Q_0\|_\infty + \varepsilon(1 + \gamma + \cdots + \gamma^{K-1}) \\ &\le \gamma^k \|Q^\star - Q_0\|_\infty + \frac{1}{(1-\gamma)^\varepsilon}. \end{aligned} \tag{$\star$}$$

Next, recall from Assignment 2 that the performance of a greedy policy π_k based on Q_k is

$$\|Q^\star - Q_{\pi_k}\|_\infty \le \frac{2\gamma}{1-\gamma} \|Q^\star - Q_k\|_\infty. \tag{$\dagger$}$$

Combining ($\star$) and ($\dagger$) we get that

$$\|Q^\star - Q_{\pi_n}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2} \max_{0\le k<n} \|T^\star Q_k - \mathscr{A}T^\star Q_k\|_\infty + \frac{2\gamma^{n+1}}{(1-\gamma)} \|Q^\star - Q_0\|_\infty,$$

which is what we wanted to show. □

B. Proof of Theorem 2

Proof (Adapted from [4]). First, recall that the bellman expectation operator for policy π can be written as

$$T^{\pi}Q = R + \gamma P^{\pi}Q, \tag{$\star$}$$

where $R \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ is a vector enumerating all expected rewards $r(s,a)$. First we will introduce and bound the *gain* incurred after each iteration of the algorithm $gain_k := Q_{\pi_{k+1}} - Q_{\pi_k}$. Expanding the definition, we find that

$$\begin{aligned} gain_k =& Q_{\pi_{k+1}} - Q_{\pi_k} \\ =& T^{\pi_{k+1}}Q_{\pi_{k+1}} - T^{\pi_{k+1}}Q_{\pi_k} \\ &+ T^{\pi_{k+1}}Q_{\pi_k} - T^{\pi_{k+1}}Q_k \\ &+ T^{\pi_{k+1}}Q_k - T^{\pi_k}Q_k \\ &+ T^{\pi_k}Q_k - T^{\pi_k}Q_{\pi_k}. \end{aligned}$$

We can bound each line in the last equation using $(\star)$ to find that

$$\begin{aligned} gain_k \geq & \gamma P^{\pi_{k+1}} gain_k \\ &+ \gamma P^{\pi_{k+1}} e_k \\ &+ 0 \\ &+ \gamma P^{\pi_k} e_k \\ &\geq \gamma P^{\pi_{k+1}} gain_k + \gamma(P^{\pi_{k+1}} - P^{\pi_k})e_k, \end{aligned}$$

where $e_k = Q_{\pi_k} - Q_k$. Rearranging, we get a lower bound on how the policy improves at each iteration,

$$gain_k \geq \gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})e_k. \tag{$\dagger$}$$

We now turn our focus to the loss in performance at iteration k, $L_k := Q^{\star} - Q_{\pi_k}$. Observe that

$$\begin{aligned} L_{k+1} =& Q^{\star} - Q_{\pi_{k+1}} \\ =& T^{\pi^{\star}}Q_{\pi^{\star}} - T^{\pi_{k+1}}Q_{\pi_{k+1}} \\ =& T^{\pi^{\star}}Q_{\pi^{\star}} - T^{\pi^{\star}}Q_{\pi_k} \\ &+ T^{\pi^{\star}}Q_{\pi_k} - T^{\pi^{\star}}Q_k \\ &+ T^{\pi^{\star}}Q_k - T^{\pi_{k+1}}Q_k \\ &+ T^{\pi_{k+1}}Q_k - T^{\pi_{k+1}}Q_{\pi_k} \\ &+ T^{\pi_{k+1}}Q_{\pi_k} - T^{\pi_{k+1}}Q_{\pi_{k+1}}. \end{aligned}$$

Similarly to the gain, we can bound each line of the expanded equation to find that

$$L_{k+1} \leq \gamma P^{\pi^{\star}} L_k + \gamma(P^{\pi^{\star}} - P^{\pi_{k+1}})e_k - \gamma P^{\pi_{k+1}} gain_k.$$

Using the bound in $(\dagger)$ on the last equation, we find that

$$L_{k+1} \leq \gamma P^{\pi^{\star}} L_k + \gamma(P^{\pi^{\star}} - P^{\pi_{k+1}})e_k - \gamma P^{\pi_{k+1}} gain_k$$

$$
\begin{aligned}
&\leq \gamma P^{\pi^\star} L_k + \gamma(P^{\pi^\star} - P^{\pi_{k+1}})e_k - \gamma P^{\pi_{k+1}}(\gamma(I-\gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})e_k) \\
&\leq \gamma P^{\pi^\star} L_k + \gamma\left(P^{\pi^\star} + \gamma P^{\pi_{k+1}}(I-\gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k}) - P^{\pi_{k+1}}\right) e_k \\
&\leq \gamma P^{\pi^\star} L_k + \gamma\left(P^{\pi^\star} + P^{\pi_{k+1}}\left(I-\gamma P^{\pi_{k+1}}\right)^{-1}\left(I - \gamma P^{\pi_k}\right)\right) e_k.
\end{aligned}
$$

As $k \to \infty$,

$$
\limsup_{k\to\infty} L_k \leq \gamma(I-\gamma P^{\pi^\star}) \limsup_{k\to\infty} \left(P^{\pi^\star} + P^{\pi_{k+1}}(I-\gamma P^{\pi_{k+1}})^{-1}(I-\gamma P^{\pi_k})\right) e_k.
$$

Taking the L_∞ norm,

$$
\begin{aligned}
\limsup_{k\to\infty} \|L_k\|_\infty &\leq \frac{\gamma}{1-\gamma} \limsup_{k\to\infty} \left\| \left(P^{\pi^\star} + P^{\pi_{k+1}}(I-\gamma P^{\pi_{k+1}})^{-1}(I-\gamma P^{\pi_k})\right) \right\|_\infty \|e_k\|_\infty \\
&\leq \frac{\gamma}{1-\gamma}\left(\frac{1+\gamma}{1-\gamma} + 1\right) \limsup_{k\to\infty} \|e_k\|_\infty.
\end{aligned}
$$

In the last line we've used the fact that $\|P\|_\infty = 1$ when P is a row stochastic matrix. □