ON GLOBAL OPTIMALITY GUARANTEE FOR ADVANTAGE-WEIGHTED REGRESSION

Changzhi Yan[*]

ABSTRACT

In this work, we develop a global optimality guarantee for the advantage-weighted regression (AWR) algorithm in the tabular setting. Based on the concept of reduction to supervised learning, AWR is an iterative reinforcement learning algorithm where the policy is updated using standard regression. Our goal is to derive an upper bound on the sub-optimality of the policy output by AWR. Focusing on the expected improvement of successive iterates of policy, we motivate the AWR algorithm as an approximate optimization to a constrained policy search problem, where the approximation can be decomposed into two steps. In the first step, we formulate a target policy update rule based on the optimal solution to the problem as an intermediate result. We show that this intermediate policy update rule not only ensures monotonic policy improvements but also enjoys a convergence rate of $O(1/K)$ for near-optimal policies. By analyzing the second step, we derive the final update rule and provide a sub-optimality upper bound for it, where the upper bound has no dependence on the size of the state and action space, thus enabling the near optimality of the output policy and the extension to the case of continuous state action space. We verify that this update rule has a crucial and desirable property for ensuring near-optimal output policies, and under certain conditions, global optimality can be guaranteed. These results provide a theoretical understanding of the iterative regression-based reinforcement learning algorithms and provide intuitions for how to obtain more effective policies.

1 Introduction

The development of methodologies of *reduction to supervised learning* never ceases in the reinforcement learning (RL) literature, since it allows us to extend the well-established guarantees of supervised learning methods to RL algorithms. Arguably the most basic setting in supervised learning is agnostic learning, in which we are trying to find the best classifier or hypothesis in a given class. In the context of RL, agnostic learning aims to discover the sample complexity of finding an optimal policy that maximizes the expected return, where each policy is associated with a function in the hypothesis class (e.g., the hypothesis class itself is a policy class). Under what conditions will agnostic learning in RL have analogous results to that of supervised learning? Surprisingly, it turns out that this is a fundamentally hard problem. Kearns et al. [1999] first introduced that, without making further assumptions, even though we can avoid dependence on the size of the state space, agnostic learning is not possible in RL, unless we are willing to pay a sample complexity exponential in the problem horizon. If we make stronger assumptions about the hypothesis class (such as the Bellman completeness in fitted-Q methods), or we develop distribution-dependent results (such as the concentrability coefficient in fitted-Q methods and distribution mismatch coefficient in policy gradient methods), we stand a chance to circumvent the statistical hardness results. In this work, our analysis aligns with these observations, thus introducing the advantages and drawbacks of our results: first, we don't make additional assumptions about the function class (in our case it's policy class), the realizability is the only concern, and our final result is distribution-dependent; second, even though the sub-optimality upper bound of AWR has no dependence on the size of the state and action space and can be reduced under some conditions, we fail to establish the convergence guarantee for this algorithm.

Another key concept AWR is based on is *incremental policy updates*, which can be viewed as making small incremental updates to the policy by forcing that the inducing state distribution from the new policy is not far away from that of the current policy. In the literature, both CPI (Kakade and Langford [2002]) and TRPO (Schulman et al. [2015])

[*] This is an individual work. No affiliation exists. (In this work, all the expressions "we" actually mean "I".)

are derived from this notion. CPI achieves this by forming the new policy as a combination of the current policy and a local greedy policy, and with the careful choice of the combination weight, the state distribution mismatch between successive iterates of policy won't be large so that CPI guarantees to achieve monotonic policy improvements. TRPO forces the new policy to be close to the current policy by explicitly imposing a KL constraint in the optimization procedure and directly maximizes the (approximate) performance improvement of the new policy. Borrowing ideas from TRPO, AWR formulates a similar optimization problem, with a subtle difference in the problem approximation scheme. Although AWR tries to maximize the improvement of successive iterates of policy, like TRPO does, we'll show that AWR doesn't guarantee monotonic performance improvements. Note that TRPO gives an equivalent update procedure to the Natural Policy Gradient (NPG, Kakade [2001]), thus here the connection between AWR and NPG can be anticipated. Specifically, in our derivation, we motivate AWR as an approximate optimization of a constrained policy search problem, the optimal solution of which exactly recovers the NPG update rule for the softmax parameterization. Besides, the original TRPO analysis provides performance guarantees, largely relying on a reduction to the CPI guarantees, and so does this work.

AWR comes from the fusion of the concepts mentioned above, which makes it an iterative regression-based RL algorithm. The remainder of the paper is organized as follows. In section 2, we introduce notations and revisit the original formulation of advantage-weighted regression. Note that AWR is an off-policy method, and we restrict ourselves to the on-policy case for the convenience of analysis. In section 3, we first formulate a constrained policy search problem that aims to maximize the expected return. With a two-step approximation to its optimization, we derive a variant algorithm (of the original AWR) that enables global optimality (while the original AWR doesn't). Then, we analyze each approximation scheme and develop global optimality guarantees for the results in each approximation step individually. Finally, section 4 gives our conclusions.

2 Preliminaries

MDP Consider an infinite-horizon discounted Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$, specified by its state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ is the space of probability distributions over $\mathcal{S}$ (i.e., the probability simplex), reward function $r : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$, discount factor $\gamma \in [0, 1)$, and initial state distribution $\mu \in \Delta(\mathcal{S})$. For the convenience of analysis, we focus on the case where rewards are deterministic, and $\mathcal{S}$ and $\mathcal{A}$ are discrete and finite.[2] Given a (stationary and stochastic) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the interaction protocol of the agent with the environment can described as the following generative process: $s_0 \sim \mu$, $a_t \sim \pi(\cdot|s_t)$, $r_t = r(s_t, a_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$, $\forall t \geq 0$, which also induces a distribution of trajectories. The value function of state s under policy π is defined as the expected discounted return:

$$V^\pi(s) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \Big| s_0 = s\right], \tag{1}$$

where $\mathbb{E}_\pi[\cdot]$ refers to the distribution of trajectories under policy π. The state-action value (or Q-value) function of policy π is given as $Q^\pi(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]$. Let $A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s)$ be the advantage function of π. Denote $V^\pi(\mu) := \mathbb{E}_{s\sim\mu}[V^\pi(s)]$, and $A^\pi(s, \pi') := \mathbb{E}_{a\sim\pi'(\cdot|s)}[A^\pi(s, a)]$. Since the reward is bounded in $[0, R_{\max}]$, the value functions $V^\pi(s)$ and $Q^\pi(s, a)$ can both be bounded between 0 and $V_{\max} := \frac{R_{\max}}{1-\gamma}$. The (normalized) discounted state distribution induced by policy π is defined as:

$$d^\pi_\mu(s) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s \,|\, \pi, \mu), \tag{2}$$

where $\Pr(s_t \,|\, \pi, \mu)$ represents the state marginal distribution at time step t, induced by policy π and starting from $s_0 \sim \mu$. Given the initial state distribution μ, there exists an optimal policy $\pi^\star$ such that $V^{\pi^\star}(\mu) = \max_\pi V^\pi(\mu)$. We denote $V^\star(\mu) := V^{\pi^\star}(\mu)$ for conciseness.

AWR Peng et al. [2019] first proposed advantage-weighted regression (AWR), an iterative RL algorithm based on the concept of reduction to supervised learning, each iteration of which consists of two regression steps: one to regress onto cumulative rewards for an "advantage" function, and another to regress onto the advantage-weighted actions for the updated policy (thus has its name "advantage-weighted regression"). Although the beauty of AWR lies in its ability to incorporate off-policy data, for clarity and simplicity, we only consider the on-policy setting where the sampling policy is the output policy from the previous iteration. The purpose of this concern is two-fold: first, the

[2] The introduced results in this work can be easily adapted to the continuous case.

term "advantage" will have a consistent meaning with that in the literature, and second, it will be clearer to reveal the iterative nature of this algorithm. In each iteration, the AWR update rule is given as follows:

$$\pi_{k+1} = \arg\max_{\pi} \mathbb{E}_{s\sim d_\mu^{\pi_k}} \mathbb{E}_{a\sim\pi_k(\cdot|s)} \left[\exp\left(\frac{1}{\beta} A^{\pi_k}(s,a)\right) \log \pi(a|s) \right]. \tag{3}$$

The AWR update can be interpreted as solving a maximum likelihood problem that fits a new policy π_{k+1} to samples collected under the current policy π_k, where the likelihood of each action is weighted by the exponentiated advantage for that action, with a temperature parameter $\beta > 0$.

3 Global Optimality Guarantees

In this work, we distinguish two kinds of initial state distributions: one is the "performance measure", denoted as ρ, under which the performance of the learned policy is evaluated; and the other is the "optimization measure", denoted as μ, under which the policy is updated. While we are interested in the good performance under ρ, we will see how it will be helpful to optimize a policy under a different measure μ. Also, in this section, we stick with the convention of the "μ-restart" setting, where we have a restart distribution $s_0 \sim \mu$ in each episode.

Suppose the AWR algorithm eventually outputs a termination policy π_K. This work aims to show that π_K is near-optimal by deriving an upper bound on the sub-optimality $V^\star(\rho) - V^{\pi_K}(\rho)$ with dependence on some MDP-specific quantities[3], and under some conditions, the sub-optimality upper bound can be reduced. In the following subsections, we first derive a variant algorithm of AWR and then provide analyses of its global optimality guarantees. This variant differs from the original algorithm (Peng et al. [2019]) in a subtle but important way. As we will see, our variant algorithm has a crucial property that enables global optimality which the original AWR doesn't hold.

3.1 Derivation of the Variant AWR

In this subsection, we motivate the AWR algorithm as an approximate optimization of a constrained policy search problem. By decomposing the approximation into two steps and analyzing each approximation scheme individually, we develop a variant algorithm that enables global optimality. Overall, we want to find a policy π_{k+1} that maximizes the *expected improvement* $V^{\pi_{k+1}}(\mu) - V^{\pi_k}(\mu)$ over a sampling policy π_k which is given as the output from the previous iteration. According to the performance difference lemma[4], the expected improvement can be expressed as:

$$V^{\pi_{k+1}}(\mu) - V^{\pi_k}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi_{k+1}}} \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)} [A^{\pi_k}(s,a)]. \tag{4}$$

However, at the current iteration, we don't know the state distribution of π_{k+1}, and all we have access to is $d_\mu^{\pi_k}$. Following Schulman et al. [2015], we force π_{k+1} to be close to π_k by explicitly imposing KL constraints in the optimization procedure in a sense that $\mathbb{E}_{s\sim d_\mu^{\pi_k}} \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)} [A^{\pi_k}(s,a)]$ well approximates $\mathbb{E}_{s\sim d_\mu^{\pi_{k+1}}} \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)} [A^{\pi_k}(s,a)]$ so that we can circumvent the need to collect samples from π_{k+1}. Thus, we can formulate the following *constrained policy search problem*:

$$\max_{\pi} \; \mathbb{E}_{s\sim d_\mu^{\pi_k}} \mathbb{E}_{a\sim\pi(\cdot|s)} [A^{\pi_k}(s,a)] \tag{5}$$

$$\text{s.t. } D_{\mathrm{KL}}\big(\pi(\cdot|s) \big\| \pi_k(\cdot|s)\big) \le \delta, \;\; \forall s \tag{6}$$

$$\sum_a \pi(a|s) = 1, \;\; \forall s. \tag{7}$$

Denote its objective function by $\mathbb{A}_k(\pi) := \mathbb{E}_{s\sim d_\mu^{\pi_k}} \mathbb{E}_{a\sim\pi(\cdot|s)} [A^{\pi_k}(s,a)]$, which can be understood as the approximate expected improvement of π over the current policy π_k. Since it's impractical to solve this problem due to the large number of KL constraints, we use a heuristic approximation considering the average KL divergence.

Approximation I Enforcing KL constraints only in expectation $\mathbb{E}_{s\sim d_\mu^{\pi_k}} \left[D_{\mathrm{KL}}\big(\pi(\cdot|s) \big\| \pi_k(\cdot|s)\big) \right]$, and further relaxing the hard KL constraint by converting it into a soft constraint with coefficient β. This gives the following optimization problem:

$$\max_{\pi} \; \mathbb{E}_{s\sim d_\mu^{\pi_k}} \mathbb{E}_{a\sim\pi(\cdot|s)} [A^{\pi_k}(s,a)] + \beta\Big(\delta - \mathbb{E}_{s\sim d_\mu^{\pi_k}} \left[D_{\mathrm{KL}}\big(\pi(\cdot|s) \big\| \pi_k(\cdot|s)\big) \right]\Big) \tag{8}$$

$$\text{s.t. } \sum_a \pi(a|s) = 1, \;\; \forall s. \tag{9}$$

[3] Such as the effective horizon $1/(1-\gamma)$, the scale of the problem $V_{\max}$, and the initial state distribution ρ (or μ), preferably with no dependence on the size of state space $|\mathcal{S}|$ and action space $|\mathcal{A}|$.

[4] Since this lemma is quite standard in the literature, I directly invoke it without providing proof in this work.

It has a closed-form solution:

$$\pi_{k+1}(a|s) := \pi^\star(a|s) = \frac{1}{Z_k(s)}\pi_k(a|s)\exp\left(\frac{1}{\beta}A^{\pi_k}(s,a)\right), \tag{10}$$

with $Z_k(s)$ being the partition function. Note by Eq.(8) that β is on the same scale as V_{max}. As a remark, in the above approximation scheme, converting the hard (average) KL constraint to a soft one is indispensable, even though the optimal solution is the same without doing so. The reason is that, as we will see in section 3.3, we want to leave β as a hyperparameter of our choice, rather than a Lagrange multiplier out of our control.

The optimal policy π_{k+1} maximizes the approximate expected improvement $\mathbb{A}_k(\pi)$, but its true expected improvement (Eq.(4)) is unclear (even can be negative in some cases, then it's not "improvement" though). In the next subsection, we will show that π_{k+1} does achieve better performance than π_k for all k (i.e., monotonic policy improvements), thus making π_{k+1} the "target" policy we want to take.

However, a policy is generally represented by a function class, which doesn't always capture the target policy π_{k+1}. Therefore, we need to find a policy in the function class closest to π_{k+1} under a certain measure, which brings another approximation called projection. The specific projection measure we use here is the maximum KL divergence. First, let us define:

$$D_{\text{KL}}^{\text{max}}(\pi\|\pi') := \max_s D_{\text{KL}}\big(\pi(\cdot|s)\big\|\pi'(\cdot|s)\big). \tag{11}$$

Approximation II Suppose policy π is restricted in a policy class Π. The output policy of the current iteration $\widehat{\pi}_{k+1}$ can be obtained by projecting π_{k+1} onto Π:

$$\widehat{\pi}_{k+1} = \arg\min_{\pi\in\Pi} D_{\text{KL}}^{\text{max}}(\pi_{k+1}\|\pi), \tag{12}$$

in a sense that

$$D_{\text{KL}}^{\text{max}}(\pi_{k+1}\|\widehat{\pi}_{k+1}) \le \sup_{s,a}\left|\log\frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)}\right| := \delta_\Pi < +\infty. \tag{13}$$

Serving as a uniform upper bound on $D_{\text{KL}}\big(\pi_{k+1}(\cdot|s)\big\|\widehat{\pi}_{k+1}(\cdot|s)\big)$ over all states, δ_Π measures how well $\widehat{\pi}_{k+1}$ can approximate π_{k+1} in terms of the "worst-case scenario". Since this is a stand-alone projection problem, the richer the policy class Π, the smaller the upper bound δ_Π can be (i.e., monotonicity property).

Denote the uniform distribution over a set $\mathcal{X}$ by Unif$_\mathcal{X}$. The update Eq.(12) can be equivalently performed by solving the following *supervised regression* problem:

$$\widehat{\pi}_{k+1} = \arg\min_{\pi\in\Pi} \mathbb{E}_{s\sim\text{Unif}_\mathcal{S}}\left[D_{\text{KL}}\big(\pi_{k+1}(\cdot|s)\big\|\pi(\cdot|s)\big)\right] \tag{14}$$

$$= \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim\text{Unif}_\mathcal{S}}\mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\exp\left(\frac{1}{\beta}A^{\pi_k}(s,a)\right)\log\pi(a|s)\right]. \tag{15}$$

This is the update rule of our variant algorithm. A subtle yet important difference between this variant and the original AWR algorithm (Peng et al. [2019]) is that the original AWR uses $s\sim d_\mu^{\pi_k}$ as the state weighting scheme in Eq.(15) instead of $s\sim$ Unif$_\mathcal{S}$. However, a key observation is that, optimizing under $s\sim d_\mu^{\pi_k}$ is NOT sufficient[5] to guarantee the monotonicity property of δ_Π, which as we'll see later, is essential to the final global optimality guarantee.

A detailed derivation for this subsection is presented in Appendix A.1. In the following subsections, we will analyze each approximation scheme and provide global optimality guarantees for updates Eq.(10) and Eq.(15) individually. For clarity, we refer to π_{k+1} as the target policy and $\widehat{\pi}_{k+1}$ as the output policy of the current iteration in the rest of the paper.

3.2 Global Optimality for the Target Policy Update

In this subsection, we will dive into the approximation I and develop the global optimality guarantee for the target policy update Eq.(10). For any termination iterate π_K, we derive an upper bound on $V^\star(\rho) - V^{\pi_K}(\rho)$ as a function of K (i.e., convergence rate). Note that here we measure the performance using ρ while the target policy π_{k+1} is updated under μ (i.e., μ-restart setting).

Our proof strategy borrows ideas from Agarwal et al. [2021] and Even-Dar et al. [2009]. First, the following lemma is helpful, which shows that π_{k+1} achieves performance improvement over π_k for all k:

[5] Optimizing under $s\sim d_\mu^{\pi_k}$ won't guarantee a small KL divergence upper bound, because there may exist some states where $d_\mu^{\pi_k}(s)$ are (nearly) zero. A sufficient condition may require that each state can be reached reasonably often under π_k, which generally can't be guaranteed.

Lemma 1 (Improvement lower bound for target update)**.** *For the iterates π_k generated by the target policy update Eq.(10), and for all initial state distributions $\rho \in \Delta(\mathcal{S})$,*

$$V^{\pi_{k+1}}(\rho) - V^{\pi_k}(\rho) \geq \beta\, \mathbb{E}_{s\sim\rho}[\log Z_k(s)] \geq 0. \tag{16}$$

This lemma indicates that this update rule is bound to generate policies with monotonic performance improvements, even though the target policy π_{k+1} is originally designed to only maximize the approximate expected improvement. With this result, we now develop the global optimality of the termination policy π_K.

Theorem 1 (Global optimality for target update)**.** *Given initial policy $\pi_0(\cdot|s) = \mathrm{Unif}_{\mathcal{A}}, \forall s$, suppose the target policy update Eq.(10) generates a sequence of policies $\{\pi_0, \pi_1, \ldots, \pi_K\}$. For all initial state distribution $\rho \in \Delta(\mathcal{S})$ and all terminal iteration $K > 0$,*

$$V^{\star}(\rho) - V^{\pi_K}(\rho) \leq \frac{\beta \log|\mathcal{A}| + V_{\max}}{K(1-\gamma)}. \tag{17}$$

This theorem reveals that the target policy update rule can generate a near-optimal policy π_K with a convergence rate of $O(1/K)$. Interestingly, this sub-optimality upper bound has no dependence on the size[6] of the state space $|\mathcal{S}|$ and the distribution mismatch coefficient, even though the target policy is updated under a different measure μ.

The detailed proofs for the lemma and theorem are described in Appendix A.2. Next, we will establish analogous results for the variant AWR update rule.

3.3 Global Optimality for AWR

In this subsection, we will focus on the approximation II and develop the global optimality guarantee for the variant AWR update Eq.(15). The analysis in this part largely follows Kakade and Langford [2002]. As a reminder, in iteration k, AWR[7] takes in policy π_k as the input and outputs a (projected) policy $\widehat{\pi}_{k+1}$. We start from the following lemma which states that the output policy $\widehat{\pi}_{k+1}$ is close to the input policy π_k in terms of the KL divergence.[8]

Lemma 2. *For $\beta \geq V_{\max}$, we have:*

$$D_{\mathrm{KL}}^{\max}(\pi_k \| \widehat{\pi}_{k+1}) \leq 1 + \big(1 + \sqrt{2}\big)\delta_{\Pi}. \tag{18}$$

With policies close to each other, the inducing state distributions should be similar as well. The following lemma formalizes this idea, showing that if $\widehat{\pi}_{k+1}$ and π_k are close in terms of the KL divergence for all states, the total variation distance (also ℓ_1 distance) between the resulting state distributions from $\widehat{\pi}_{k+1}$ and π_k will be small up to an effective horizon amplification.

Lemma 3. *For $\beta \geq V_{\max}$, and for any k, we have:*

$$\big\| d_{\mu}^{\widehat{\pi}_{k+1}} - d_{\mu}^{\pi_k} \big\|_1 \leq \frac{\gamma}{1-\gamma}\sqrt{2\delta_{\Pi}^{+}}, \tag{19}$$

where $\delta_{\Pi}^{+} := 1 + \big(1+\sqrt{2}\big)\delta_{\Pi}$.

To avoid redundancy, we'll implicitly assume $\beta \geq V_{\max}$ and denote $\delta_{\Pi}^{+} := 1 + (1+\sqrt{2})\delta_{\Pi}$ for the rest of the analysis. In other words, unless stated otherwise, we will omit all the $\beta \geq V_{\max}$ conditions and δ_{Π}^{+} definitions in the following lemmas and theorems for conciseness. One should be aware that these conditions exist even though they are not explicitly stated. Note that δ_{Π}^{+} also has the monotonicity property that it will monotonically decrease as the size of Π grows.

The next lemma develops an analogous result to lemma 1, establishing an improvement lower bound for the output policy $\widehat{\pi}_{k+1}$.

Lemma 4 (Improvement lower bound for AWR)**.** *Denote $\mathbb{A}_k(\pi) := \mathbb{E}_{s\sim d_{\mu}^{\pi_k}} \mathbb{E}_{a\sim\pi(\cdot|s)}[A^{\pi_k}(s,a)]$. For any k,*

$$V^{\widehat{\pi}_{k+1}}(\mu) - V^{\pi_k}(\mu) \geq \frac{1}{1-\gamma}\left(\mathbb{A}_k(\widehat{\pi}_{k+1}) - \frac{\gamma V_{\max}}{1-\gamma}\sqrt{2\delta_{\Pi}^{+}}\right). \tag{20}$$

[6] This result can be readily extended to the continuous case where $|\mathcal{A}|$ represents the volume of $\mathcal{A}$, as shown in the proof.

[7] Unless stated otherwise, all the "AWR" in this part refers to the variant AWR for conciseness, which won't cause ambiguity since the meaning is clear from the context.

[8] The intuition is somewhat similar to the Conservative Policy Iteration algorithm.

Recall that $\mathbb{A}_k(\pi)$ is the objective function of the constrained policy search problem (Eq.(5)), and $\mathbb{A}_k(\widehat{\pi}_{k+1})$ measures the approximate expected improvement of $\widehat{\pi}_{k+1}$ over π_k. Although $\mathbb{A}_k(\widehat{\pi}_{k+1})$ seems to tell us little information about the policy improvement of AWR in each iteration since $\widehat{\pi}_{k+1}$ is only an approximate solution to the constrained policy search problem whose objective is yet another approximation to the true expected improvement, this lemma provides a measure of monotonic policy improvements of AWR using $\mathbb{A}_k(\widehat{\pi}_{k+1})$. That is, as long as we have:

$$\mathbb{A}_k(\widehat{\pi}_{k+1}) \geq \frac{\gamma V_{\max}}{1-\gamma}\sqrt{2\delta_{\Pi}^{+}}, \tag{21}$$

we can ensure that $\widehat{\pi}_{k+1}$ is making improvement over π_k under any measure μ. The quantity on the right-hand side can be viewed as a threshold of $\mathbb{A}_k(\widehat{\pi}_{k+1})$ for achieving improvement at iteration k. However, there is no guarantee that this condition will always be satisfied, thus, in contrast to lemma 1, AWR has neither a monotonic policy improvement guarantee nor a convergence rate due to the incorporation of approximation II.

One thing we can do is to enlarge the size of the policy class $|\Pi|$ to reduce δ_{Π}^{+} with the hope that it will be more likely for $\mathbb{A}_k(\widehat{\pi}_{k+1})$ to go beyond the threshold. This observation is consistent with our previous analysis because $\widehat{\pi}_{k+1}$ as a projection is meant to be close to π_{k+1} which is proved to be better than π_k (lemma 1), and increasing $|\Pi|$ will make $\widehat{\pi}_{k+1}$ closer to π_{k+1}. Based on this idea, it's natural and straightforward to come up with the following termination criteria for AWR, which allows a $\varepsilon \cdot V_{\max}$ margin of error for violating the condition Eq.(21).

Termination Criteria Return π_k, if $\widehat{\pi}_{k+1}$ satisfies that:

$$\mathbb{A}_k(\widehat{\pi}_{k+1}) \leq \frac{\gamma V_{\max}}{1-\gamma}\sqrt{2\delta_{\Pi}^{+}} - \varepsilon V_{\max}. \tag{22}$$

Intuitively, we want the algorithm to proceed even when there is "occasionally" no policy improvement, as long as the performance degeneration is not too severe (which is controlled by the parameter $\varepsilon \geq 0$). The "violation" error $\varepsilon \cdot V_{\max}$ represents the amount of performance degeneration in each iteration we can tolerate at most, beyond which we should consider terminating the algorithm. With the termination criteria, we now develop the global optimality guarantee for the variant AWR algorithm Eq.(15).

Theorem 2 (Global optimality for AWR)**.** *Upon termination, the update Eq.(15) returns a policy π_K such that:*

$$V^{\star}(\rho) - V^{\pi_K}(\rho) \leq \frac{V_{\max}}{(1-\gamma)^2}\left\|\frac{d_{\rho}^{\pi^{\star}}}{\mu}\right\|_{\infty}\left(\frac{\gamma}{1-\gamma}\sqrt{2\delta_{\Pi}^{+}} - \varepsilon + \varepsilon_{\Pi}\right), \tag{23}$$

where $\varepsilon_{\Pi} := \frac{1}{V_{\max}}\mathbb{E}_{s\sim d_{\mu}^{\pi_K}}[\max_{a\in\mathcal{A}} A^{\pi_K}(s,a) - A^{\pi_K}(s,\widehat{\pi}_{K+1})]$.

Recall in section 3.1 we have mentioned that the target policy π_{k+1} maximizes $\mathbb{A}_k(\pi)$, thus here we have:

$$\mathbb{A}_K(\pi_{K+1}) = \max_{\pi}\mathbb{A}_K(\pi) = \max_{\pi}\mathbb{E}_{s\sim d_{\mu}^{\pi_K}}[A^{\pi_K}(s,\pi)] = \max_{a}\mathbb{E}_{s\sim d_{\mu}^{\pi_K}}[A^{\pi_K}(s,a)]. \tag{24}$$

Therefore, ε_{Π} is essentially the difference between $\mathbb{A}_K(\pi_{K+1})$ and $\mathbb{A}_K(\widehat{\pi}_{K+1})$ (up to a normalization factor), which is specifically the projection error in iteration K measured by $\mathbb{A}_K(\cdot)$. More importantly, this reveals the nature of the regression Eq.(15): the output policy $\widehat{\pi}_{k+1}$ is trying to mimic the behavior of $\arg\max_{a\in\mathcal{A}}[A^{\pi_k}(s,a)]$ under $s \sim d_{\mu}^{\pi_k}$ in each iteration. As a comment, both δ_{Π} and ε_{Π} represent the projection error: δ_{Π} measures it by the "worst-case" KL divergence, while ε_{Π} measures it using $\mathbb{A}_K(\cdot)$. Again, enriching the policy class Π will reduce ε_{Π}.

This theorem indicates that the AWR algorithm guarantees to find a near-optimal policy if:

- The policy class Π is rich enough so that δ_{Π}^{+} and ε_{Π} are small.
- The restart distribution μ covers $d_{\rho}^{\pi^{\star}}$ in a sense that the distribution mismatch coefficient $\left\|\frac{d_{\rho}^{\pi^{\star}}}{\mu}\right\|_{\infty}$ is small.

The first one is the assumption about the function class. It's crucial that δ_{Π}^{+} and ε_{Π} can decrease when the size of the policy class $|\Pi|$ grows to achieve near-optimality. Note that this is an assumption with only a realizability concern for the policy class since the algorithm is merely a supervised regression. The second one is the assumption about the data distribution. Having a highly exploratory initial state distribution μ is vitally important for the algorithm to be effective. Note that there is no convergence guarantee for the AWR algorithm (as we can see that there is no explicit K in the upper bound) since in each iteration the output policy doesn't guarantee to achieve improvement. Theorem 2 only guarantees that the final output policy returned by AWR is near-optimal, but it doesn't claim how many iterations are needed to generate such a near-optimal policy.

Finally, as a remark, one can aggressively choose large ε to further reduce the sub-optimality bound, but at the cost of incurring more iterations (possibly the algorithm will never stop). If we set $\varepsilon = 0$, the algorithm may stop early but with a poorer performance. This reflects the trade-off between iteration complexity and near optimality since there is no convergence guarantee for AWR due to the incorporation of the approximation II.

The detailed proofs for the lemmas and theorem in this subsection are presented in Appendix A.3.

4 Conclusions and Future Work

We set out to develop a global optimality guarantee for the advantage-weighted regression algorithm in the on-policy setting for the tabular case. Specifically, we provide an upper bound on the sub-optimality of the policy returned by AWR upon termination, which has no dependence on the cardinality of the state and action space. We motivate the AWR algorithm as a projection (onto the policy class) of the optimal solution to a constrained policy search problem and analyze the global optimality of the policy update rule with and without the projection step. Our main finding is that without the projection step, the target policy update achieves global optimality with a convergence rate of $O(1/K)$. With the projection step, the (variant) AWR update rule can also achieve global optimality but with no convergence guarantee. Thus, a more refined analysis is needed to develop the convergence guarantees and iteration complexity for these regression-based methods. Moreover, the AWR algorithm has its off-policy generalization which can learn from fully off-policy datasets. Therefore, another interesting direction will be extending the results for the off-policy setting.

References

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. Approximate planning in large pomdps via reusable trajectories. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 1001–1007, Cambridge, MA, USA, 1999. MIT Press.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1889–1897. JMLR.org, 2015.

Sham Kakade. A natural policy gradient. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 1531–1538, Cambridge, MA, USA, 2001. MIT Press.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *CoRR*, abs/1910.00177, 2019. URL https://arxiv.org/abs/1910.00177.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

Eyal Even-Dar, Sham. M. Kakade, and Yishay Mansour. Online markov decision processes. *Math. Oper. Res.*, 34 (3):726–736, August 2009. ISSN 0364-765X. doi: 10.1287/moor.1090.0396. URL https://doi.org/10.1287/moor.1090.0396.

A Proofs

A.1 Proofs for Section 3.1: Derivation of the Variant AWR

According to the approximation I, we have the following optimization problem:

$$\max_{\pi} \; \mathbb{E}_{s\sim d_{\mu}^{\pi_k}}\mathbb{E}_{a\sim\pi(\cdot|s)}[A^{\pi_k}(s,a)] + \beta\Big(\delta - \mathbb{E}_{s\sim d_{\mu}^{\pi_k}}\big[D_{\mathrm{KL}}\big(\pi(\cdot|s)\big\|\pi_k(\cdot|s)\big)\big]\Big) \tag{25}$$

$$\text{s.t. } \sum_a \pi(a|s) = 1, \;\; \forall s. \tag{26}$$

It's not hard to verify that this is a *concave* optimization problem, since the objective Eq.(25) is concave with respect to $\pi(a|s)$ and the constraints are affine. It's also straightforward to show that this problem has *strong duality* according to Slater's constraint qualifications. Therefore, any solution satisfying the KKT conditions is the optimal solution. First, we form the Lagrangian:

$$\begin{aligned}\mathcal{L}(\pi,\lambda) = \sum_s d_{\mu}^{\pi_k}(s)\sum_a \pi(a|s)A^{\pi_k}(s,a) + \beta\left[\delta - \sum_s d_{\mu}^{\pi_k}(s)\sum_a \pi(a|s)\log\frac{\pi(a|s)}{\pi_k(a|s)}\right] \\ + \sum_s \lambda_s\Big(1 - \sum_a \pi(a|s)\Big),\end{aligned} \tag{27}$$

with $\lambda = \{\lambda_s \,|\, \forall s \in \mathcal{S}\}$ being the Lagrange multipliers. Differentiating $\mathcal{L}(\pi,\lambda)$ with respect to $\pi(a|s)$ results in:

$$\frac{\partial\mathcal{L}(\pi,\lambda)}{\partial\pi(a|s)} = d_{\mu}^{\pi_k}(s)A^{\pi_k}(s,a) - \beta d_{\mu}^{\pi_k}(s)\log\pi(a|s) + \beta d_{\mu}^{\pi_k}(s)\log\pi_k(a|s) - \beta d_{\mu}^{\pi_k}(s) - \lambda_s. \tag{28}$$

Next, let us inspect the KKT conditions:

$$\left.\frac{\partial\mathcal{L}(\pi,\lambda^\star)}{\partial\pi(a|s)}\right|_{\pi=\pi^\star} = 0, \tag{29}$$

$$\sum_a \pi^\star(a|s) = 1, \;\; \forall s. \tag{30}$$

Solving for Eq.(29) yields:

$$\pi^\star(a|s) = \pi_k(a|s)\exp\Big(\frac{1}{\beta}A^{\pi_k}(s,a)\Big)\exp\left(-1 - \frac{\lambda_s^\star}{\beta d_{\mu}^{\pi_k}(s)}\right). \tag{31}$$

Plugging it in Eq.(30), we know that the second exponential term serves as the partition function $Z_k(s)$ that normalizes the conditional action distribution,

$$Z_k(s) \coloneqq \exp\left(1 + \frac{\lambda_s^\star}{\beta d_{\mu}^{\pi_k}(s)}\right) = \sum_a \pi_k(a|s)\exp\Big(\frac{1}{\beta}A^{\pi_k}(s,a)\Big). \tag{32}$$

As a result, the optimal solution (denoted as π_{k+1}) is given as:

$$\pi_{k+1}(a|s) \coloneqq \pi^\star(a|s) = \frac{1}{Z_k(s)}\pi_k(a|s)\exp\Big(\frac{1}{\beta}A^{\pi_k}(s,a)\Big). \tag{33}$$

Before we move on, let us inspect $Z_k(s)$. First, according to Jensen's inequality,

$$Z_k(s) = \mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\exp\Big(\frac{1}{\beta}A^{\pi_k}(s,a)\Big)\right] \geq \exp\left(\mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\frac{1}{\beta}A^{\pi_k}(s,a)\right]\right) = 1, \tag{34}$$

where we have used that $\mathbb{E}_{a\sim\pi_k(\cdot|s)}[A^{\pi_k}(s,a)] = 0, \forall s$. Besides,

$$\begin{aligned}Z_k(s) = \mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\exp\Big(\frac{1}{\beta}A^{\pi_k}(s,a)\Big)\right] &\leq \sup_a \exp\Big(\frac{1}{\beta}A^{\pi_k}(s,a)\Big) \\ \leq \sup_{s,a}\exp\Big(\frac{1}{\beta}A^{\pi_k}(s,a)\Big) &\leq \exp\left(\frac{V_{\max}}{\beta}\right),\end{aligned} \tag{35}$$

where for the first inequality we use the fact that the convex average of each element is not greater than its maximum, and for the last inequality we use the observation that $\sup_{s,a}|A^{\pi}(s,a)| \leq V_{\max}, \forall\pi$.

Together we have a bound on $\log Z_k(s)$, which will be useful in the following proofs.

$$0 \le \log Z_k(s) \le \frac{V_{\max}}{\beta}, \ \ \forall s. \tag{36}$$

Next, we consider the projection part. If policy π is restricted in a policy class Π, the output policy of AWR of the current iteration $\widehat{\pi}_{k+1}$ can be obtained by projecting the optimal solution π_{k+1} onto the manifold of Π:

$$\widehat{\pi}_{k+1} = \arg\min_{\pi\in\Pi} D_{\mathrm{KL}}^{\max}(\pi_{k+1}\|\pi). \tag{37}$$

According to this update rule, $\widehat{\pi}_{k+1}(\cdot|s)$ and $\pi_{k+1}(\cdot|s)$ are meant to be similar for all states, i.e., for any $s \in \mathcal{S}$,

$$D_{\mathrm{KL}}\big(\pi_{k+1}(\cdot|s)\big\|\widehat{\pi}_{k+1}(\cdot|s)\big) \tag{38}$$

$$=\mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}\left[\log\frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)}\right] \tag{39}$$

$$\le\big\|\pi_{k+1}(\cdot|s)\big\|_1\cdot\sup_a\left|\log\frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)}\right| \tag{40}$$

$$\le\sup_{s,a}\left|\log\frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)}\right| := \delta_\Pi < +\infty, \tag{41}$$

where for the second step we use Hölder's inequality, and for the third step we use the fact that $\big\|\pi_{k+1}(\cdot|s)\big\|_1 = 1$ since π_{k+1} is a valid distribution. Note that δ_Π must exist because $\widehat{\pi}_{k+1}(a|s) \approx \pi_{k+1}(a|s), \forall s, a$. Serving as a uniform upper bound on $D_{\mathrm{KL}}\big(\pi_{k+1}(\cdot|s)\big\|\widehat{\pi}_{k+1}(\cdot|s)\big)$ over all states, δ_Π measures how well $\widehat{\pi}_{k+1}$ approximates π_{k+1} in terms of the "worst-case scenario". Since this is a stand-alone projection problem, the richer the policy class Π, the smaller the upper bound δ_Π can be (i.e., monotonicity property).

The update Eq.(37) can be equivalently performed by solving the following *supervised regression* problem:

$$\widehat{\pi}_{k+1} = \arg\min_{\pi\in\Pi}\mathbb{E}_{s\sim\mathrm{Unif}_{\mathcal{S}}}\left[D_{\mathrm{KL}}\big(\pi_{k+1}(\cdot|s)\big\|\pi(\cdot|s)\big)\right] \tag{42}$$

$$= \arg\min_{\pi\in\Pi}\mathbb{E}_{s\sim\mathrm{Unif}_{\mathcal{S}}}\left[D_{\mathrm{KL}}\left(\frac{1}{Z_k(s)}\pi_k(\cdot|s)\exp\left(\frac{1}{\beta}A^{\pi_k}(s,\cdot)\right)\middle\|\,\pi(\cdot|s)\right)\right] \tag{43}$$

$$= \arg\max_{\pi\in\Pi}\mathbb{E}_{s\sim\mathrm{Unif}_{\mathcal{S}}}\left[\frac{1}{Z_k(s)}\mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\exp\left(\frac{1}{\beta}A^{\pi_k}(s,a)\right)\log\pi(a|s)\right]\right] \tag{44}$$

$$= \arg\max_{\pi\in\Pi}\mathbb{E}_{s\sim\mathrm{Unif}_{\mathcal{S}}}\mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\exp\left(\frac{1}{\beta}A^{\pi_k}(s,a)\right)\log\pi(a|s)\right], \tag{45}$$

where the penultimate step omits all irrelevant terms, and the final step holds due to $Z_k(s) > 0, \forall s$.

Finally, we show that using the state weighting scheme $s \sim d_\mu^{\pi_k}$, as is presented in the original AWR algorithm, will break the monotonicity property.

$$\widehat{\pi}'_{k+1} = \arg\min_{\pi\in\Pi}\mathbb{E}_{s\sim d_\mu^{\pi_k}}\left[D_{\mathrm{KL}}\big(\pi_{k+1}(\cdot|s)\big\|\pi(\cdot|s)\big)\right] \tag{46}$$

$$= \arg\max_{\pi\in\Pi}\mathbb{E}_{s\sim d_\mu^{\pi_k}}\mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\exp\left(\frac{1}{\beta}A^{\pi_k}(s,a)\right)\log\pi(a|s)\right]. \tag{47}$$

Suppose $D_{\mathrm{KL}}^{\max}(\pi_{k+1}\|\widehat{\pi}'_{k+1}) \le \delta'_\Pi$. When optimizing under $\mathbb{E}_{s\sim d_\mu^{\pi_k}}[\cdot]$, it's possible that the KL divergence is large at one particular state, while the expected KL divergence is still small, for example, when the weighting probability for that state $d_\mu^{\pi_k}(s)$ is (nearly) zero. In this case, minimizing the expected KL divergence doesn't imply a small KL divergence upper bound δ'_Π. Therefore, increasing the size of the policy class Π can't guarantee a monotonic decrease of δ'_Π, even with infinite data. A sufficient condition ensuring the monotonicity property may require that each state can be reached reasonably often under π_k, which generally can't be guaranteed.

A.2 Proofs for Section 3.2: Global Optimality for the Target Policy Update

Lemma 1 (Improvement lower bound for target update)**.** *For the iterates π_k generated by the target policy update Eq.(10), and for all initial state distributions $\rho \in \Delta(\mathcal{S})$,*

$$V^{\pi_{k+1}}(\rho) - V^{\pi_k}(\rho) \ge \beta\,\mathbb{E}_{s\sim\rho}[\log Z_k(s)] \ge 0. \tag{48}$$

Proof. According to the performance-difference lemma,

$$\begin{aligned}
V^{\pi_{k+1}}(\rho) - V^{\pi_k}(\rho) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_{k+1}}}\mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)] && (49)\\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_{k+1}}}\mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}\left[\beta\log\frac{\pi_{k+1}(a|s)Z_k(s)}{\pi_k(a|s)}\right] && (50)\\
&= \frac{\beta}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_{k+1}}}\mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}\left[\log\frac{\pi_{k+1}(a|s)}{\pi_k(a|s)} + \log Z_k(s)\right] && (51)\\
&= \frac{\beta}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_{k+1}}}\left[D_{\mathrm{KL}}\big(\pi_{k+1}(\cdot|s)\big\|\pi_k(\cdot|s)\big) + \log Z_k(s)\right] && (52)\\
&\geq \frac{\beta}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_{k+1}}}[\log Z_k(s)] && (53)\\
&\geq \beta\,\mathbb{E}_{s\sim\rho}[\log Z_k(s)]\,, && (54)
\end{aligned}$$

where in the second step we back out $A^{\pi_k}(s,a)$ from the update rule Eq.(10), and in the last step we use the fact that $d_\rho^{\pi_{k+1}} \geq (1-\gamma)\rho$ by Eq.(2). The proof is completed by noticing that $\log Z_k(s) \geq 0, \forall s$, as mentioned in Eq.(36). □

Theorem 1 (Global optimality for target update)**.** *Given initial policy $\pi_0(\cdot|s) = \mathrm{Unif}_{\mathcal{A}}, \forall s$, suppose the target policy update Eq.(10) generates a sequence of policies $\{\pi_0, \pi_1, \ldots, \pi_K\}$. For all initial state distribution $\rho \in \Delta(\mathcal{S})$ and all terminal iteration $K > 0$,*

$$V^\star(\rho) - V^{\pi_K}(\rho) \leq \frac{\beta\log|\mathcal{A}| + V_{\max}}{K(1-\gamma)}. \tag{55}$$

Proof. Since ρ is fixed, for simplicity we denote $d^\star$ as shorthand for $d_\rho^{\pi^\star}$. By the performance difference lemma,

$$\begin{aligned}
V^\star(\rho) - V^{\pi_k}(\rho) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot|s)}[A^{\pi_k}(s,a)] && (56)\\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot|s)}\left[\beta\log\frac{\pi_{k+1}(a|s)Z_k(s)}{\pi_k(a|s)}\right] && (57)\\
&= \frac{\beta}{1-\gamma}\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot|s)}\left[\log\frac{\pi^\star(a|s)}{\pi_k(a|s)} - \log\frac{\pi^\star(a|s)}{\pi_{k+1}(a|s)} + \log Z_k(s)\right] && (58)\\
&= \frac{\beta}{1-\gamma}\mathbb{E}_{s\sim d^\star}\left[D_{\mathrm{KL}}\big(\pi^\star(\cdot|s)\big\|\pi_k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi^\star(\cdot|s)\big\|\pi_{k+1}(\cdot|s)\big) + \log Z_k(s)\right]. && (59)
\end{aligned}$$

According to lemma 1, with $d^\star$ as the initial state distribution, we have:

$$V^{\pi_{k+1}}(d^\star) - V^{\pi_k}(d^\star) \geq \beta\,\mathbb{E}_{s\sim d^\star}[\log Z_k(s)]\,. \tag{60}$$

Since for all k, $V^{\pi_{k+1}}(\rho) \geq V^{\pi_k}(\rho)$ (by lemma 1), we have:

$$\begin{aligned}
V^\star(\rho) - V^{\pi_{K-1}}(\rho) &\leq V^\star(\rho) - \frac{1}{K}\sum_{k=0}^{K-1}V^{\pi_k}(\rho) = \frac{1}{K}\sum_{k=0}^{K-1}\big(V^\star(\rho) - V^{\pi_k}(\rho)\big)\\
&= \frac{\beta}{K(1-\gamma)}\sum_{k=0}^{K-1}\mathbb{E}_{s\sim d^\star}\left[D_{\mathrm{KL}}\big(\pi^\star(\cdot|s)\big\|\pi_k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi^\star(\cdot|s)\big\|\pi_{k+1}(\cdot|s)\big) + \log Z_k(s)\right].
\end{aligned} \tag{61}$$

The above expression consists of two parts:

$$\begin{aligned}
\text{(I)} &:= \frac{\beta}{K(1-\gamma)}\sum_{k=0}^{K-1}\mathbb{E}_{s\sim d^\star}\left[D_{\mathrm{KL}}\big(\pi^\star(\cdot|s)\big\|\pi_k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi^\star(\cdot|s)\big\|\pi_{k+1}(\cdot|s)\big)\right] && (62)\\
&= \frac{\beta}{K(1-\gamma)}\mathbb{E}_{s\sim d^\star}\left[D_{\mathrm{KL}}\big(\pi^\star(\cdot|s)\big\|\pi_0(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi^\star(\cdot|s)\big\|\pi_K(\cdot|s)\big)\right] && (63)\\
&\leq \frac{\beta}{K(1-\gamma)}\mathbb{E}_{s\sim d^\star}\left[D_{\mathrm{KL}}\big(\pi^\star(\cdot|s)\big\|\pi_0(\cdot|s)\big)\right] && (64)\\
&\leq \frac{\beta\log|\mathcal{A}|}{K(1-\gamma)}, && (65)
\end{aligned}$$

where in the last step we use that $\pi_0(\cdot|s) = \text{Unif}_{\mathcal{A}}, \forall s$ as well as $\pi^\star(a|s) \leq 1, \forall s, a$.

$$\begin{aligned}
\text{(II)} &:= \frac{\beta}{K(1-\gamma)} \sum_{k=0}^{K-1} \mathbb{E}_{s\sim d^\star}[\log Z_k(s)] && (66)\\
&\leq \frac{1}{K(1-\gamma)} \sum_{k=0}^{K-1} \big(V^{\pi_{k+1}}(d^\star) - V^{\pi_k}(d^\star)\big) && (67)\\
&= \frac{1}{K(1-\gamma)} \big(V^{\pi_K}(d^\star) - V^{\pi_0}(d^\star)\big) && (68)\\
&\leq \frac{V_{\max}}{K(1-\gamma)}, && (69)
\end{aligned}$$

where the first inequality holds by Eq.(60), and the last inequality is based on that $0 \leq V^\pi(\mu) \leq V_{\max}, \forall \pi, \mu$. Together we have:

$$V^\star(\rho) - V^{\pi_{K-1}}(\rho) \leq \text{(I)} + \text{(II)} \leq \frac{\beta \log|\mathcal{A}| + V_{\max}}{K(1-\gamma)}. \tag{70}$$

The proof is completed by noticing that $V^\star(\rho) - V^{\pi_K}(\rho) \leq V^\star(\rho) - V^{\pi_{K-1}}(\rho)$. □

A.3 Proofs for Section 3.3: Global Optimality for AWR

Throughout this section, we denote $D_{\text{TV}}(P\|Q)$ as the total variation distance between distributions P and Q. The total variation distance is half of the ℓ_1 distance, $D_{\text{TV}}(P\|Q) = \frac{1}{2}\|P - Q\|_1$, and it relates to the KL divergence via $D_{\text{TV}}(P\|Q) \leq \sqrt{\frac{1}{2}D_{\text{KL}}(P\|Q)}$.

Lemma 2. *For $\beta \geq V_{\max}$, we have:*

$$D_{\text{KL}}^{\max}(\pi_k\|\widehat{\pi}_{k+1}) \leq 1 + \big(1 + \sqrt{2}\big)\delta_\Pi. \tag{71}$$

Proof. For all $s \in \mathcal{S}$,

$$\begin{aligned}
D_{\text{KL}}\big(\pi_k(\cdot|s)\big\|\pi_{k+1}(\cdot|s)\big) &= \mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\log \frac{Z_k(s)\pi_k(a|s)}{\pi_k(a|s)\exp\Big(\frac{1}{\beta}A^{\pi_k}(s,a)\Big)}\right]\\
&= \mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\log Z_k(s) - \frac{1}{\beta}A^{\pi_k}(s,a)\right] = \log Z_k(s) \leq \frac{V_{\max}}{\beta} \leq 1,
\end{aligned} \tag{72}$$

where the first inequality holds due to Eq.(36) and the last step results from $\beta \geq V_{\max}$.

For any s, we have:

$$\begin{aligned}
&D_{\text{KL}}\big(\pi_k(\cdot|s)\big\|\widehat{\pi}_{k+1}(\cdot|s)\big) - D_{\text{KL}}\big(\pi_k(\cdot|s)\big\|\pi_{k+1}(\cdot|s)\big) = \mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\log \frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)}\right] && (73)\\
=&\sum_a \pi_{k+1}(a|s)\log\frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)} + \sum_a \pi_k(a|s)\log\frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)} - \sum_a \pi_{k+1}(a|s)\log\frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)} && (74)\\
=&D_{\text{KL}}\big(\pi_{k+1}(\cdot|s)\big\|\widehat{\pi}_{k+1}(\cdot|s)\big) + \sum_a \big(\pi_k(a|s) - \pi_{k+1}(a|s)\big)\log\frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)} && (75)\\
\leq&D_{\text{KL}}\big(\pi_{k+1}(\cdot|s)\big\|\widehat{\pi}_{k+1}(\cdot|s)\big) + \big\|\pi_k(\cdot|s) - \pi_{k+1}(\cdot|s)\big\|_1 \cdot \sup_{s,a}\left|\log\frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)}\right|, && (76)
\end{aligned}$$

where the last step is based on Hölder's inequality.

By Eq.(72), for all $s \in \mathcal{S}$,

$$\begin{aligned}
\big\|\pi_k(\cdot|s) - \pi_{k+1}(\cdot|s)\big\|_1 &= 2D_{\text{TV}}\big(\pi_k(\cdot|s)\big\|\pi_{k+1}(\cdot|s)\big)\\
&\leq 2\sqrt{\frac{1}{2}D_{\text{KL}}\big(\pi_k(\cdot|s)\big\|\pi_{k+1}(\cdot|s)\big)} \leq \sqrt{2}.
\end{aligned} \tag{77}$$

Recall that in Eq.(13):

$$\delta_\Pi \coloneqq \sup_{s,a} \left| \log \frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)} \right|, \tag{78}$$

$$D_{\text{KL}}^{\max}(\pi_{k+1} \| \widehat{\pi}_{k+1}) \leq \delta_\Pi. \tag{79}$$

To put everything together, for all $s \in \mathcal{S}$, we have:

$$D_{\text{KL}}\big(\pi_k(\cdot|s) \big\| \widehat{\pi}_{k+1}(\cdot|s)\big) \leq D_{\text{KL}}\big(\pi_k(\cdot|s) \big\| \pi_{k+1}(\cdot|s)\big) + D_{\text{KL}}\big(\pi_{k+1}(\cdot|s) \big\| \widehat{\pi}_{k+1}(\cdot|s)\big) + \sqrt{2}\delta_\Pi \tag{80}$$

$$\leq 1 + \delta_\Pi + \sqrt{2}\delta_\Pi = 1 + \big(1+\sqrt{2}\big)\delta_\Pi. \tag{81}$$

Thus $D_{\text{KL}}^{\max}(\pi_k \| \widehat{\pi}_{k+1}) \leq 1 + \big(1+\sqrt{2}\big)\delta_\Pi$. □

Lemma 3. *For $\beta \geq V_{\max}$, and for any k, we have:*

$$\big\| d_\mu^{\widehat{\pi}_{k+1}} - d_\mu^{\pi_k} \big\|_1 \leq \frac{\gamma}{1-\gamma}\sqrt{2\delta_\Pi^+}, \tag{82}$$

where $\delta_\Pi^+ \coloneqq 1 + \big(1+\sqrt{2}\big)\delta_\Pi$.

Proof. First, by lemma 2, for all $s \in \mathcal{S}$, we have that:

$$\begin{aligned} \big\|\pi_k(\cdot|s) - \widehat{\pi}_{k+1}(\cdot|s)\big\|_1 &= 2D_{\text{TV}}\big(\pi_k(\cdot|s)\big\|\widehat{\pi}_{k+1}(\cdot|s)\big) \\ &\leq 2\sqrt{\frac{1}{2}D_{\text{KL}}\big(\pi_k(\cdot|s)\big\|\widehat{\pi}_{k+1}(\cdot|s)\big)} \leq \sqrt{2\delta_\Pi^+}. \end{aligned} \tag{83}$$

The last inequality holds because we have $\beta \geq V_{\max}$.

Denote $\mathbb{P}_t^\pi$ as a shorthand for $\Pr(s_t \mid \pi, \mu)$, which represents the state distribution at time step t induced by π with μ as the initial state distribution. We consider bounding $\big\|\mathbb{P}_t^{\widehat{\pi}_{k+1}} - \mathbb{P}_t^{\pi_k}\big\|_1$ for any t.

$$\mathbb{P}_t^{\widehat{\pi}_{k+1}}(s') - \mathbb{P}_t^{\pi_k}(s') \tag{84}$$

$$= \sum_{s,a}\Big(\mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s)\widehat{\pi}_{k+1}(a|s) - \mathbb{P}_{t-1}^{\pi_k}(s)\pi_k(a|s)\Big)P(s'|s,a) \tag{85}$$

$$= \sum_{s,a}\Big(\mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s)\widehat{\pi}_{k+1}(a|s) - \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s)\pi_k(a|s) + \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s)\pi_k(a|s) - \mathbb{P}_{t-1}^{\pi_k}(s)\pi_k(a|s)\Big)P(s'|s,a) \tag{86}$$

$$= \sum_{s,a}\mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s)\big(\widehat{\pi}_{k+1}(a|s) - \pi_k(a|s)\big)P(s'|s,a) + \sum_{s,a}\Big(\mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) - \mathbb{P}_{t-1}^{\pi_k}(s)\Big)\pi_k(a|s)P(s'|s,a). \tag{87}$$

Thus we have,

$$\begin{aligned} &\big\|\mathbb{P}_t^{\widehat{\pi}_{k+1}} - \mathbb{P}_t^{\pi_k}\big\|_1 = \sum_{s'}\Big|\mathbb{P}_t^{\widehat{\pi}_{k+1}}(s') - \mathbb{P}_t^{\pi_k}(s')\Big| \\ &\leq \sum_{s,a,s'}\mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s)\big|\widehat{\pi}_{k+1}(a|s) - \pi_k(a|s)\big|P(s'|s,a) + \sum_{s,a,s'}\Big|\mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) - \mathbb{P}_{t-1}^{\pi_k}(s)\Big|\pi_k(a|s)P(s'|s,a). \end{aligned} \tag{88}$$

The first term of the above expression:

$$\text{(I)} \coloneqq \sum_{s,a,s'}\mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s)\big|\widehat{\pi}_{k+1}(a|s) - \pi_k(a|s)\big|P(s'|s,a) \tag{89}$$

$$= \sum_s \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s)\sum_a\big|\widehat{\pi}_{k+1}(a|s) - \pi_k(a|s)\big|\sum_{s'}P(s'|s,a) \tag{90}$$

$$= \sum_s \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s)\big\|\widehat{\pi}_{k+1}(\cdot|s) - \pi_k(\cdot|s)\big\|_1 \tag{91}$$

$$\leq \sum_s \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s)\sqrt{2\delta_\Pi^+} \tag{92}$$

$$= \sqrt{2\delta_\Pi^+}, \tag{93}$$

where the penultimate step is due to Eq.(83), in which the upper bound holds for all $s \in \mathcal{S}$ (this is crucial).

And the second term:

$$\text{(II)} := \sum_{s,a,s'} \left| \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) - \mathbb{P}_{t-1}^{\pi_k}(s) \right| \pi_k(a|s) P(s'|s,a) \tag{94}$$

$$= \sum_{s} \left| \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) - \mathbb{P}_{t-1}^{\pi_k}(s) \right| \sum_a \pi_k(a|s) \sum_{s'} P(s'|s,a) \tag{95}$$

$$= \left\| \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}} - \mathbb{P}_{t-1}^{\pi_k} \right\|_1. \tag{96}$$

Putting everything together, we have:

$$\left\| \mathbb{P}_{t}^{\widehat{\pi}_{k+1}} - \mathbb{P}_{t}^{\pi_k} \right\|_1 \leq \text{(I)} + \text{(II)} \leq \sqrt{2\delta_\Pi^+} + \left\| \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}} - \mathbb{P}_{t-1}^{\pi_k} \right\|_1 \tag{97}$$

$$\leq 2\sqrt{2\delta_\Pi^+} + \left\| \mathbb{P}_{t-2}^{\widehat{\pi}_{k+1}} - \mathbb{P}_{t-2}^{\pi_k} \right\|_1 \leq \cdots \tag{98}$$

$$\leq t\sqrt{2\delta_\Pi^+} + \left\| \mathbb{P}_{0}^{\widehat{\pi}_{k+1}} - \mathbb{P}_{0}^{\pi_k} \right\|_1 \tag{99}$$

$$= t\sqrt{2\delta_\Pi^+}, \tag{100}$$

where $\mathbb{P}_0^{\widehat{\pi}_{k+1}} = \mathbb{P}_0^{\pi_k} = \mu$ by definition.

Now by the definition of d_μ^π (Eq.(2)), we have:

$$d_\mu^{\widehat{\pi}_{k+1}} - d_\mu^{\pi_k} = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \left(\mathbb{P}_t^{\widehat{\pi}_{k+1}} - \mathbb{P}_t^{\pi_k} \right). \tag{101}$$

Thus,

$$\left\| d_\mu^{\widehat{\pi}_{k+1}} - d_\mu^{\pi_k} \right\|_1 \leq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \left\| \mathbb{P}_t^{\widehat{\pi}_{k+1}} - \mathbb{P}_t^{\pi_k} \right\|_1 \tag{102}$$

$$\leq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t t \sqrt{2\delta_\Pi^+} \tag{103}$$

$$= \frac{\gamma}{1-\gamma} \sqrt{2\delta_\Pi^+}, \tag{104}$$

where in the last step we have used that $\sum_{t=0}^{\infty} \gamma^t t = \frac{\gamma}{(1-\gamma)^2}$. □

Lemma 4 (Improvement lower bound for AWR). *Denote* $\mathbb{A}_k(\pi) := \mathbb{E}_{s \sim d_\mu^{\pi_k}} \mathbb{E}_{a \sim \pi(\cdot|s)}[A^{\pi_k}(s,a)]$. *For any k,*

$$V^{\widehat{\pi}_{k+1}}(\mu) - V^{\pi_k}(\mu) \geq \frac{1}{1-\gamma} \left(\mathbb{A}_k(\widehat{\pi}_{k+1}) - \frac{\gamma V_{\max}}{1-\gamma} \sqrt{2\delta_\Pi^+} \right). \tag{105}$$

Proof. By performance difference lemma,

$$V^{\widehat{\pi}_{k+1}}(\mu) - V^{\pi_k}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\mu^{\widehat{\pi}_{k+1}}} \mathbb{E}_{a \sim \widehat{\pi}_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)]. \tag{106}$$

According to lemma 3, we have:

$$(1-\gamma)\big(V^{\widehat{\pi}_{k+1}}(\mu) - V^{\pi_k}(\mu)\big) = \mathbb{E}_{s \sim d_\mu^{\widehat{\pi}_{k+1}}} \mathbb{E}_{a \sim \widehat{\pi}_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)] \tag{107}$$

$$= \mathbb{E}_{s \sim d_\mu^{\pi_k}} \mathbb{E}_{a \sim \widehat{\pi}_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)] + \mathbb{E}_{s \sim d_\mu^{\widehat{\pi}_{k+1}}} \mathbb{E}_{a \sim \widehat{\pi}_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)] - \mathbb{E}_{s \sim d_\mu^{\pi_k}} \mathbb{E}_{a \sim \widehat{\pi}_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)] \tag{108}$$

$$\geq \mathbb{E}_{s \sim d_\mu^{\pi_k}} \mathbb{E}_{a \sim \widehat{\pi}_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)] - \left\| d_\mu^{\widehat{\pi}_{k+1}} - d_\mu^{\pi_k} \right\|_1 \cdot \sup_{s,a} |A^{\pi_k}(s,a)| \tag{109}$$

$$\geq \mathbb{E}_{s \sim d_\mu^{\pi_k}} \mathbb{E}_{a \sim \widehat{\pi}_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)] - \frac{\gamma V_{\max}}{1-\gamma} \sqrt{2\delta_\Pi^+} \tag{110}$$

$$= \mathbb{A}_k(\widehat{\pi}_{k+1}) - \frac{\gamma V_{\max}}{1-\gamma} \sqrt{2\delta_\Pi^+}, \tag{111}$$

where the first inequality holds due to: (Hölder's inequality)

$$\begin{aligned}&\Big|\mathbb{E}_{s\sim d_\mu^{\widehat{\pi}_{k+1}}}\mathbb{E}_{a\sim\widehat{\pi}_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)] - \mathbb{E}_{s\sim d_\mu^{\pi_k}}\mathbb{E}_{a\sim\widehat{\pi}_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)]\Big| \\ \le&\big\|d_\mu^{\widehat{\pi}_{k+1}} - d_\mu^{\pi_k}\big\|_1\cdot\sup_s\big|\mathbb{E}_{a\sim\widehat{\pi}_{k+1}(\cdot|s)}[A^{\pi_k}(s,a)]\big| \le \big\|d_\mu^{\widehat{\pi}_{k+1}} - d_\mu^{\pi_k}\big\|_1\cdot\sup_{s,a}|A^{\pi_k}(s,a)|,\end{aligned} \tag{112}$$

and the second inequality results from lemma 3 as well as the fact that $\sup_{s,a}|A^\pi(s,a)| \le V_{\max}, \forall\pi$. □

Theorem 2 (Global optimality for AWR). *Upon termination, the update Eq.(15) returns a policy π_K such that:*

$$V^\star(\rho) - V^{\pi_K}(\rho) \le \frac{V_{\max}}{(1-\gamma)^2}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty\left(\frac{\gamma}{1-\gamma}\sqrt{2\delta_\Pi^+} - \varepsilon + \varepsilon_\Pi\right), \tag{113}$$

where $\varepsilon_\Pi := \frac{1}{V_{\max}}\mathbb{E}_{s\sim d_\mu^{\pi_K}}[\max_{a\in\mathcal{A}}A^{\pi_K}(s,a) - A^{\pi_K}(s,\widehat{\pi}_{K+1})]$.

Proof. By the performance difference lemma,

$$\begin{aligned}V^\star(\rho) - V^{\pi_K}(\rho) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot|s)}[A^{\pi_K}(s,a)] && (114)\\ &\le \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\Big[\max_{a\in\mathcal{A}}A^{\pi_K}(s,a)\Big] && (115)\\ &\le \frac{1}{1-\gamma}\left\|\frac{d_\rho^{\pi^\star}}{d_\mu^{\pi_K}}\right\|_\infty\mathbb{E}_{s\sim d_\mu^{\pi_K}}\Big[\max_{a\in\mathcal{A}}A^{\pi_K}(s,a)\Big] && (116)\\ &\le \frac{1}{(1-\gamma)^2}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty\mathbb{E}_{s\sim d_\mu^{\pi_K}}\Big[\max_{a\in\mathcal{A}}A^{\pi_K}(s,a)\Big] && (117)\\ &= \frac{1}{(1-\gamma)^2}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty\left(\mathbb{A}_K(\widehat{\pi}_{K+1}) + \mathbb{E}_{s\sim d_\mu^{\pi_K}}\Big[\max_{a\in\mathcal{A}}A^{\pi_K}(s,a)\Big] - \mathbb{A}_K(\widehat{\pi}_{K+1})\right), && (118)\end{aligned}$$

where in the second inequality we use the fact that $\max_{a\in\mathcal{A}}A^{\pi_K}(s,a) \ge 0, \forall s$, and the change of measure for weighted ℓ_1 norm. The third inequality holds because $d_\mu^\pi \ge (1-\gamma)\mu, \forall\pi$ (by Eq.(2)).

According to the termination criteria Eq.(22), if the algorithm returns π_K, we have:

$$\mathbb{A}_K(\widehat{\pi}_{K+1}) \le \frac{\gamma V_{\max}}{1-\gamma}\sqrt{2\delta_\Pi^+} - \varepsilon V_{\max}. \tag{119}$$

Recall that $\mathbb{A}_K(\widehat{\pi}_{K+1}) := \mathbb{E}_{s\sim d_\mu^{\pi_K}}\mathbb{E}_{a\sim\widehat{\pi}_{K+1}(\cdot|s)}[A^{\pi_K}(s,a)] = \mathbb{E}_{s\sim d_\mu^{\pi_K}}[A^{\pi_K}(s,\widehat{\pi}_{K+1})]$.

Define ε_Π as:

$$\begin{aligned}\varepsilon_\Pi V_{\max} &:= \mathbb{E}_{s\sim d_\mu^{\pi_K}}\Big[\max_{a\in\mathcal{A}}A^{\pi_K}(s,a)\Big] - \mathbb{A}_K(\widehat{\pi}_{K+1}) \\ &= \mathbb{E}_{s\sim d_\mu^{\pi_K}}\Big[\max_{a\in\mathcal{A}}A^{\pi_K}(s,a) - A^{\pi_K}(s,\widehat{\pi}_{K+1})\Big].\end{aligned} \tag{120}$$

By the terminate condition Eq.(119) and the definition of ε_Π Eq.(120), we have:

$$\begin{aligned}V^\star(\rho) - V^{\pi_K}(\rho) &\le \frac{1}{(1-\gamma)^2}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty\left(\frac{\gamma V_{\max}}{1-\gamma}\sqrt{2\delta_\Pi^+} - \varepsilon V_{\max} + \varepsilon_\Pi V_{\max}\right) && (121)\\ &= \frac{V_{\max}}{(1-\gamma)^2}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty\left(\frac{\gamma}{1-\gamma}\sqrt{2\delta_\Pi^+} - \varepsilon + \varepsilon_\Pi\right). && (122)\end{aligned}$$

Thus finishes the proof. □