

---

# ON OPTIMALITY GUARANTEE FOR ADVANTAGE-WEIGHTED REGRESSION

---

Changzhi Yan\*

## ABSTRACT

In this work, we develop an 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 in which the policy is updated using standard regression. Focusing on the expected improvement of successive policy iterates, we formulate a constrained policy search problem and motivate AWR as its approximate optimization, where the approximation can be decomposed into two steps. In the first step, we establish a target policy update rule based on the optimal solution to the problem as an intermediate result and show that this target policy update not only ensures monotonic policy improvement but also enjoys a convergence rate of  $O(1/K)$  for near-optimal policies. In the second step, we develop the AWR policy update rule as the projection of the target policy onto the policy class and derive an upper bound on the sub-optimality of the policy output by AWR upon termination, which has no dependence on the size of the state and action space. We verify that this update rule has a crucial property for ensuring near-optimal termination policies. These results provide a theoretical understanding of the AWR algorithm and intuitions for obtaining effective policies.

## 1 Introduction

The concept of *reduction to supervised learning* plays an important role in reinforcement learning (RL) since it allows us to extend the well-established guarantees of supervised learning algorithms to RL methods. Arguably the most basic setting in supervised learning is agnostic learning, which aims to find the best classifier or hypothesis in a given class. In the context of RL, the goal of agnostic learning is 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 supervised learning? This question reveals the fundamental hardness of RL. Kearns et al. [1999] first introduced that, even though we can avoid dependence on the size of the state space, without further assumptions, agnostic learning is not possible in RL, unless we are willing to pay a sample complexity that is 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: first, we don't make assumptions about the function (policy) class, the realizability is the only concern; second, our final result is distribution-dependent; third, even though the sub-optimality upper bound of AWR's output policy has no dependence on the size of the state and action space, we fail to establish the convergence guarantee for this algorithm.

Another key idea 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]) 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. With a 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 improvement. TRPO forces the new policy to be close to the current policy by explicitly imposing a KL constraint in the optimization procedure and

---

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

directly maximizes the expected performance improvement of the new policy. Borrowing ideas from TRPO, AWR formulates a similar optimization problem with a subtle difference in the approximation scheme. Although AWR tries to maximize the (approximate) improvement of successive policy iterates, we'll show that AWR doesn't guarantee monotonic performance improvement. Note that TRPO gives an equivalent update procedure to the Natural Policy Gradient (NPG, Kakade [2001]), thus 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. 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 ideas 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 basic settings and notations and revisit the original formulation of advantage-weighted regression. In section 3, we first formulate a constrained policy search problem that aims to maximize the expected policy improvement. With a two-step approximation to its optimization, we derive a variant algorithm of AWR with an optimality guarantee. Then, we derive the sub-optimality upper bound for the results in each approximation step individually. Finally, section 4 gives our conclusions. Note that AWR is an off-policy method, and we restrict ourselves to the on-policy setting for the convenience of analysis.

## 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  $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ , where  $\Delta(\mathcal{S})$  is the space of probability distributions over  $\mathcal{S}$  (i.e., the probability simplex), reward function  $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, R_{\max}]$ , 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.<sup>2</sup> Given a (stationary and stochastic) policy  $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ , the interaction protocol of the agent with the environment can be 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  $\pi$  is defined as the expected discounted return:

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

where  $\mathbb{E}_\pi[\cdot]$  refers to the distribution of trajectories under policy  $\pi$ . The state-action value (or Q-value) function of policy  $\pi$  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  $\pi$ . 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  $\pi$  is defined as:

$$d_\mu^\pi(s) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s \mid \pi, \mu), \quad (2)$$

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

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

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

<sup>2</sup>The introduced results in this work can be easily adapted to the continuous case.

### 3 Optimality Guarantees

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

Suppose the AWR algorithm outputs a termination policy  $\pi_K$ . We aim to show that  $\pi_K$  is near-optimal by deriving an upper bound on the sub-optimality  $V^*(\rho) - V^{\pi_K}(\rho)$  with dependence on some MDP-specific quantities<sup>3</sup>, and under some conditions, the sub-optimality upper bound can be lowered. In the following subsections, we first derive a variant algorithm of AWR and then provide an analysis of its performance guarantee. This variant differs from the original one (Peng et al. [2019]) in a subtle but important way. As we will see, our variant algorithm has a crucial property that guarantees near-optimal policies which the original AWR doesn’t hold.

#### 3.1 Derivation of the AWR Variant

In this subsection, we motivate the AWR algorithm as an approximate optimization of a constrained policy search problem and develop a variant algorithm with an optimality guarantee by decomposing the approximation into two steps. In the beginning, by the idea of incremental policy updates, we want to find a policy  $\pi_{k+1}$  that maximizes the *expected improvement*  $V^{\pi_{k+1}}(\mu) - V^{\pi_k}(\mu)$  over a sampling policy  $\pi_k$  which is given as the output from the previous iteration. According to the performance difference lemma<sup>4</sup>, the expected improvement can be expressed as:

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

However, at the current iteration, we don’t know the state distribution of  $\pi_{k+1}$ , and all we have access to is  $d_{\mu}^{\pi_k}$ . Following Schulman et al. [2015], we force  $\pi_{k+1}$  to be close to  $\pi_k$  by explicitly imposing a KL constraint 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  $\pi_{k+1}$ . Thus, we can formulate the following *constrained policy search problem*:

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

$$\text{s.t. } \mathbb{E}_{s \sim d_{\mu}^{\pi_k}} [D_{\text{KL}}(\pi(\cdot|s) \parallel \pi_k(\cdot|s))] \leq \delta, \quad (6)$$

$$\sum_a \pi(a|s) = 1, \quad \forall s. \quad (7)$$

Denote the 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  $\pi$  over the current policy  $\pi_k$ . Note that we don’t necessarily restrict to a particular policy parameterization (such as the softmax), which is why we have the normalization constraint. To simplify the problem, we relax the hard KL constraint into a soft one, which results in the following optimization problem.

**Approximation I** Converting the hard KL constraint into a soft constraint with coefficient  $\beta$ :

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

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

It has a closed-form solution (denoted as  $\pi_{k+1}$ ):

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

where  $Z_k(s)$  is the partition function. Note by Eq.(8) that  $\beta$  is on the same scale as  $V_{\max}$ . In the above approximation scheme, converting the hard 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  $\beta$  as a parameter of our choice, rather than a Lagrange multiplier out of our control.

<sup>3</sup>Such as the effective horizon  $1/(1-\gamma)$ , the scale of the problem  $V_{\max}$ , and the initial state distribution  $\rho$  (or  $\mu$ ), preferably with no dependence on the size of state space  $|S|$  and action space  $|A|$ .

<sup>4</sup>Since this lemma is quite standard in the literature, I directly invoke it without providing proof in this work.

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

However, when a policy is represented by a policy class that doesn't capture the target policy  $\pi_{k+1}$ , we need to find its closest policy within the policy class under a certain measure, which brings another approximation called projection. The specific measure we use here is the maximum KL divergence:  $D_{\text{KL}}^{\max}(\pi \parallel \pi') := \max_s D_{\text{KL}}(\pi(\cdot|s) \parallel \pi'(\cdot|s))$ .

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

$$\hat{\pi}_{k+1} = \arg \min_{\pi \in \Pi} D_{\text{KL}}^{\max}(\pi_{k+1} \parallel \pi), \quad (11)$$

in a sense that

$$D_{\text{KL}}^{\max}(\pi_{k+1} \parallel \hat{\pi}_{k+1}) \leq \sup_{s,a} \left| \log \frac{\pi_{k+1}(a|s)}{\hat{\pi}_{k+1}(a|s)} \right| := \delta_{\Pi} < +\infty. \quad (12)$$

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

Denote the uniform distribution over a set  $\mathcal{X}$  by  $\text{Unif}_{\mathcal{X}}$ . The update Eq.(11) can be equivalently performed by solving the following *supervised regression* problem:

$$\hat{\pi}_{k+1} = \arg \min_{\pi \in \Pi} \mathbb{E}_{s \sim \text{Unif}_{\mathcal{S}}} [D_{\text{KL}}(\pi_{k+1}(\cdot|s) \parallel \pi(\cdot|s))] \quad (13)$$

$$= \arg \max_{\pi \in \Pi} \mathbb{E}_{s \sim \text{Unif}_{\mathcal{S}}} \mathbb{E}_{a \sim \pi_k(\cdot|s)} \left[ \exp \left( \frac{1}{\beta} A^{\pi_k}(s, a) \right) \log \pi(a|s) \right]. \quad (14)$$

This is the update rule of our variant algorithm. A subtle yet important difference between this variant and the original AWR update rule (Eq.(3)) is that the original AWR uses  $s \sim d_{\mu}^{\pi_k}$  as the state weighting scheme. However, optimizing under  $\mathbb{E}_{s \sim d_{\mu}^{\pi_k}}[\cdot]$  is NOT sufficient<sup>5</sup> to guarantee the monotonicity property of  $\delta_{\Pi}$ , which as we'll see later, is essential to the final optimality guarantee.

A detailed derivation for this subsection is presented in Appendix A.1. In the following subsections, we will derive the upper bound on the sub-optimality of the termination policy generated by updates Eq.(10) and Eq.(14) individually. For clarity, we refer to  $\pi_{k+1}$  as the target policy and  $\hat{\pi}_{k+1}$  as the output policy of the current iteration for the rest of the paper.

### 3.2 Optimality for Target Policy Update

In this subsection, we will dive into the approximation I and develop the optimality guarantee for the target policy update Eq.(10). For any termination iterate  $\pi_K$ , we derive an upper bound on  $V^*(\rho) - V^{\pi_K}(\rho)$  as a function of  $K$  (i.e., convergence rate). Note that here we measure the performance using  $\rho$  while the target policy  $\pi_{k+1}$  is updated under  $\mu$  (i.e.,  $\mu$ -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  $\pi_{k+1}$  achieves performance improvement over  $\pi_k$  for all  $k$ :

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

$$V^{\pi_{k+1}}(\rho) - V^{\pi_k}(\rho) \geq \beta \mathbb{E}_{s \sim \rho} [\log Z_k(s)] \geq 0. \quad (15)$$

This lemma indicates that this update rule is bound to generate policies with monotonic performance improvement, even though the target policy  $\pi_{k+1}$  is designed to only maximize the approximate expected improvement. With this result, we now develop the sub-optimality upper bound of the termination policy  $\pi_K$ .

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

$$V^*(\rho) - V^{\pi_K}(\rho) \leq \frac{\beta \log |\mathcal{A}| + V_{\max}}{K(1 - \gamma)}. \quad (16)$$

<sup>5</sup>When we enlarge the policy class  $\Pi$ , optimizing under  $\mathbb{E}_{s \sim d_{\mu}^{\pi_k}}[\cdot]$  won't guarantee a small  $\delta_{\Pi}$ , because there may exist some states where  $d_{\mu}^{\pi_k}(s)$  is zero. A sufficient condition may require that each state can be reached reasonably often under  $\pi_k$ , which generally can't be guaranteed.

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

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

### 3.3 Optimality for AWR

In this subsection, we will focus on the approximation  $\Pi$  and develop the optimality guarantee for the variant AWR update Eq.(14). The analysis in this part largely follows Kakade and Langford [2002]. As a reminder, at iteration  $k$ , AWR<sup>7</sup> takes in policy  $\pi_k$  as the input and outputs a (projected) policy  $\hat{\pi}_{k+1}$ . We start from the following lemma which states that the output policy  $\hat{\pi}_{k+1}$  is close to the input policy  $\pi_k$  in terms of the KL divergence.<sup>8</sup>

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

$$D_{\text{KL}}^{\max}(\pi_k \|\hat{\pi}_{k+1}) \leq 1 + (1 + \sqrt{2})\delta_{\Pi}. \quad (17)$$

With policies close to each other, the inducing state distributions should be similar as well. The following lemma formalizes this idea, showing that if  $\hat{\pi}_{k+1}$  and  $\pi_k$  are close in terms of the KL divergence for all states, the total variation distance (also  $\ell_1$  distance) between the resulting state distributions from  $\hat{\pi}_{k+1}$  and  $\pi_k$  will be small up to an effective horizon amplification.

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

$$\|d_{\mu}^{\hat{\pi}_{k+1}} - d_{\mu}^{\pi_k}\|_1 \leq \frac{\gamma}{1-\gamma} \sqrt{2\delta_{\Pi}^+}, \quad (18)$$

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

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  $\delta_{\Pi}^+$  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  $\delta_{\Pi}^+$  also has the monotonicity property that it will monotonically decrease as the size of  $\Pi$  grows. The next lemma develops an analogous result to lemma 1, establishing an improvement lower bound for the output policy  $\hat{\pi}_{k+1}$ .

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

$$V^{\hat{\pi}_{k+1}}(\mu) - V^{\pi_k}(\mu) \geq \frac{1}{1-\gamma} \left( \mathbb{A}_k(\hat{\pi}_{k+1}) - \frac{\gamma V_{\max}}{1-\gamma} \sqrt{2\delta_{\Pi}^+} \right). \quad (19)$$

Recall that  $\mathbb{A}_k(\pi)$  is the objective function of the constrained policy search problem (Eq.(5)), and  $\mathbb{A}_k(\hat{\pi}_{k+1})$  measures the approximate expected improvement of  $\hat{\pi}_{k+1}$  over  $\pi_k$ . Although  $\mathbb{A}_k(\hat{\pi}_{k+1})$  seems to tell us little information about the policy improvement of AWR in each iteration since  $\hat{\pi}_{k+1}$  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 criterion for monotonic policy improvement of AWR using  $\mathbb{A}_k(\hat{\pi}_{k+1})$ . That is, as long as we have:

$$\mathbb{A}_k(\hat{\pi}_{k+1}) \geq \frac{\gamma V_{\max}}{1-\gamma} \sqrt{2\delta_{\Pi}^+}, \quad (20)$$

we can ensure that  $\hat{\pi}_{k+1}$  is making improvement over  $\pi_k$  under any measure  $\mu$ . The quantity on the right-hand side can be viewed as a threshold of  $\mathbb{A}_k(\hat{\pi}_{k+1})$  for the output policy achieving improvement at iteration  $k$ . However, there is no guarantee that this criterion 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  $\Pi$ .

One thing we can do is to enlarge the size of the policy class  $|\Pi|$  to reduce  $\delta_{\Pi}^+$  with the hope that it will be more likely for  $\mathbb{A}_k(\hat{\pi}_{k+1})$  to go beyond the threshold. This observation is consistent with our previous analysis because  $\hat{\pi}_{k+1}$  as a projection is meant to be close to  $\pi_{k+1}$  which is proved to be better than  $\pi_k$  (lemma 1), and increasing  $|\Pi|$  will make  $\hat{\pi}_{k+1}$  closer to  $\pi_{k+1}$ . Based on this idea, it's natural and straightforward to develop the following termination criterion for the AWR algorithm, which allows a  $\varepsilon V_{\max}$  margin of error for violating the condition Eq.(20).

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

<sup>7</sup>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.

<sup>8</sup>The intuition is somewhat similar to the Conservative Policy Iteration algorithm.

**Termination Criterion** Return  $\pi_k$ , if  $\hat{\pi}_{k+1}$  satisfies that:

$$\mathbb{A}_k(\hat{\pi}_{k+1}) \leq \frac{\gamma V_{\max}}{1 - \gamma} \sqrt{2\delta_{\Pi}^+} - \varepsilon V_{\max}. \quad (21)$$

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  $\varepsilon V_{\max}$  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 criterion, we now develop the sub-optimality upper bound of the (variant) AWR’s termination policy.

**Theorem 2** (Optimality of AWR). *Upon termination, the update Eq.(14) returns a policy  $\pi_K$  such that:*

$$V^*(\rho) - V^{\pi_K}(\rho) \leq \frac{V_{\max}}{(1 - \gamma)^2} \left\| \frac{d_{\rho}^{\pi^*}}{\mu} \right\|_{\infty} \left( \frac{\gamma}{1 - \gamma} \sqrt{2\delta_{\Pi}^+} - \varepsilon + \varepsilon_{\Pi} \right), \quad (22)$$

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

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

$$\mathbb{A}_K(\pi_{K+1}) = \max_{\pi} \mathbb{A}_K(\pi) = \max_{\pi} \mathbb{E}_{s \sim d_{\mu}^{\pi_K}} [A^{\pi_K}(s, \pi)] = \max_a \mathbb{E}_{s \sim d_{\mu}^{\pi_K}} [A^{\pi_K}(s, a)]. \quad (23)$$

Therefore,  $\varepsilon_{\Pi}$  is essentially the difference between  $\mathbb{A}_K(\pi_{K+1})$  and  $\mathbb{A}_K(\hat{\pi}_{K+1})$  (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.(14): the output policy  $\hat{\pi}_{k+1}$  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  $\delta_{\Pi}$  and  $\varepsilon_{\Pi}$  represent the projection error:  $\delta_{\Pi}$  measures it by the “worst-case” KL divergence, while  $\varepsilon_{\Pi}$  measures it using  $\mathbb{A}_K(\cdot)$ . Again, enriching the policy class  $\Pi$  will reduce  $\varepsilon_{\Pi}$ .

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

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

The first one is about the function class. It’s crucial that  $\delta_{\Pi}^+$  and  $\varepsilon_{\Pi}$  can decrease when the size of the policy class  $|\Pi|$  grows to achieve near optimality. Note that this is a condition with only the realizability concern since the algorithm is merely a supervised regression. The second one is about the data distribution. Having a highly exploratory initial state distribution  $\mu$  to cover  $d_{\rho}^{\pi^*}$  is vitally important for the algorithm to be effective. Note that there is no convergence guarantee<sup>9</sup> 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.

Finally, as a remark, one can aggressively choose large  $\varepsilon$  to further lower the sub-optimality upper bound, at the cost of incurring more iterations (possibly the algorithm will never stop). If we set  $\varepsilon = 0$ , the algorithm may stop early with a poorer performance. This reflects the trade-off between iteration complexity and near optimality. As an extreme case, when the realizability is satisfied,  $\hat{\pi}_{k+1}$  becomes  $\pi_{k+1}$ , and the AWR’s update rule Eq.(14) degenerates into the target policy update rule Eq.(10). In this case,  $\delta_{\Pi}^+ = \varepsilon_{\Pi} = 0$  and  $\varepsilon$  must be set to 0, and the termination criterion will never be satisfied.

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

## 4 Conclusions and Future Work

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

<sup>9</sup>Theorem 2 only guarantees that the termination policy returned by AWR is near-optimal. It doesn’t claim how many iterations are needed to generate such a near-optimal policy.

## 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 AWR Variant

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

$$\max_{\pi} \mathbb{E}_{s \sim d_{\mu}^{\pi_k}} \mathbb{E}_{a \sim \pi(\cdot|s)} [A^{\pi_k}(s, a)] + \beta \left( \delta - \mathbb{E}_{s \sim d_{\mu}^{\pi_k}} [D_{\text{KL}}(\pi(\cdot|s) \parallel \pi_k(\cdot|s))] \right) \quad (24)$$

$$\text{s.t. } \sum_a \pi(a|s) = 1, \quad \forall s. \quad (25)$$

It's not hard to verify that this is a *concave* optimization problem, since the objective Eq.(24) 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:

$$\begin{aligned} \mathcal{L}(\pi, \lambda) = & \sum_s d_{\mu}^{\pi_k}(s) \sum_a \pi(a|s) A^{\pi_k}(s, a) + \beta \left[ \delta - \sum_s d_{\mu}^{\pi_k}(s) \sum_a \pi(a|s) \log \frac{\pi(a|s)}{\pi_k(a|s)} \right] \\ & + \sum_s \lambda_s \left( 1 - \sum_a \pi(a|s) \right), \end{aligned} \quad (26)$$

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

$$\frac{\partial \mathcal{L}(\pi, \lambda)}{\partial \pi(a|s)} = d_{\mu}^{\pi_k}(s) A^{\pi_k}(s, a) - \beta d_{\mu}^{\pi_k}(s) \log \pi(a|s) + \beta d_{\mu}^{\pi_k}(s) \log \pi_k(a|s) - \beta d_{\mu}^{\pi_k}(s) - \lambda_s. \quad (27)$$

Next, let us inspect the KKT conditions:

$$\left. \frac{\partial \mathcal{L}(\pi, \lambda^*)}{\partial \pi(a|s)} \right|_{\pi=\pi^*} = 0, \quad (28)$$

$$\sum_a \pi^*(a|s) = 1, \quad \forall s. \quad (29)$$

Solving for Eq.(28) yields:

$$\pi^*(a|s) = \pi_k(a|s) \exp\left(\frac{1}{\beta} A^{\pi_k}(s, a)\right) \exp\left(-1 - \frac{\lambda_s^*}{\beta d_{\mu}^{\pi_k}(s)}\right). \quad (30)$$

Plugging it in Eq.(29), we know that the second exponential term serves as the partition function  $Z_k(s)$  that normalizes the conditional action distribution,

$$Z_k(s) := \exp\left(1 + \frac{\lambda_s^*}{\beta d_{\mu}^{\pi_k}(s)}\right) = \sum_a \pi_k(a|s) \exp\left(\frac{1}{\beta} A^{\pi_k}(s, a)\right). \quad (31)$$

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

$$\pi_{k+1}(a|s) := \pi^*(a|s) = \frac{1}{Z_k(s)} \pi_k(a|s) \exp\left(\frac{1}{\beta} A^{\pi_k}(s, a)\right). \quad (32)$$

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

$$Z_k(s) = \mathbb{E}_{a \sim \pi_k(\cdot|s)} \left[ \exp\left(\frac{1}{\beta} A^{\pi_k}(s, a)\right) \right] \geq \exp\left(\mathbb{E}_{a \sim \pi_k(\cdot|s)} \left[ \frac{1}{\beta} A^{\pi_k}(s, a) \right]\right) = 1, \quad (33)$$

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

$$\begin{aligned} Z_k(s) &= \mathbb{E}_{a \sim \pi_k(\cdot|s)} \left[ \exp\left(\frac{1}{\beta} A^{\pi_k}(s, a)\right) \right] \leq \sup_a \exp\left(\frac{1}{\beta} A^{\pi_k}(s, a)\right) \\ &\leq \sup_{s,a} \exp\left(\frac{1}{\beta} A^{\pi_k}(s, a)\right) \leq \exp\left(\frac{V_{\max}}{\beta}\right), \end{aligned} \quad (34)$$

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.

$$0 \leq \log Z_k(s) \leq \frac{V_{\max}}{\beta}, \quad \forall s. \quad (35)$$

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

$$\hat{\pi}_{k+1} = \arg \min_{\pi \in \Pi} D_{\text{KL}}^{\max}(\pi_{k+1} \| \pi). \quad (36)$$

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

$$D_{\text{KL}}(\pi_{k+1}(\cdot|s) \| \hat{\pi}_{k+1}(\cdot|s)) \quad (37)$$

$$= \mathbb{E}_{a \sim \pi_{k+1}(\cdot|s)} \left[ \log \frac{\pi_{k+1}(a|s)}{\hat{\pi}_{k+1}(a|s)} \right] \quad (38)$$

$$\leq \|\pi_{k+1}(\cdot|s)\|_1 \cdot \sup_a \left| \log \frac{\pi_{k+1}(a|s)}{\hat{\pi}_{k+1}(a|s)} \right| \quad (39)$$

$$\leq \sup_{s,a} \left| \log \frac{\pi_{k+1}(a|s)}{\hat{\pi}_{k+1}(a|s)} \right| := \delta_{\Pi} < +\infty, \quad (40)$$

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

The update Eq.(36) can be equivalently performed by solving the following *supervised regression* problem:

$$\hat{\pi}_{k+1} = \arg \min_{\pi \in \Pi} \mathbb{E}_{s \sim \text{Unif}_{\mathcal{S}}} [D_{\text{KL}}(\pi_{k+1}(\cdot|s) \| \pi(\cdot|s))] \quad (41)$$

$$= \arg \min_{\pi \in \Pi} \mathbb{E}_{s \sim \text{Unif}_{\mathcal{S}}} \left[ D_{\text{KL}} \left( \frac{1}{Z_k(s)} \pi_k(\cdot|s) \exp \left( \frac{1}{\beta} A^{\pi_k}(s, \cdot) \right) \| \pi(\cdot|s) \right) \right] \quad (42)$$

$$= \arg \max_{\pi \in \Pi} \mathbb{E}_{s \sim \text{Unif}_{\mathcal{S}}} \left[ \frac{1}{Z_k(s)} \mathbb{E}_{a \sim \pi_k(\cdot|s)} \left[ \exp \left( \frac{1}{\beta} A^{\pi_k}(s, a) \right) \log \pi(a|s) \right] \right] \quad (43)$$

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

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.

$$\hat{\pi}'_{k+1} = \arg \min_{\pi \in \Pi} \mathbb{E}_{s \sim d_{\mu}^{\pi_k}} [D_{\text{KL}}(\pi_{k+1}(\cdot|s) \| \pi(\cdot|s))] \quad (45)$$

$$= \arg \max_{\pi \in \Pi} \mathbb{E}_{s \sim d_{\mu}^{\pi_k}} \mathbb{E}_{a \sim \pi_k(\cdot|s)} \left[ \exp \left( \frac{1}{\beta} A^{\pi_k}(s, a) \right) \log \pi(a|s) \right]. \quad (46)$$

Suppose  $D_{\text{KL}}^{\max}(\pi_{k+1} \| \hat{\pi}'_{k+1}) \leq \delta'_{\Pi}$ . 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 upper bound  $\delta'_{\Pi}$ . Therefore, increasing the size of the policy class  $\Pi$  can't guarantee a monotonic decrease of  $\delta'_{\Pi}$ . A sufficient condition ensuring the monotonicity property may require that each state can be reached reasonably often under  $\pi_k$ , which generally can't be guaranteed.

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

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

$$V^{\pi_{k+1}}(\rho) - V^{\pi_k}(\rho) \geq \beta \mathbb{E}_{s \sim \rho} [\log Z_k(s)] \geq 0. \quad (47)$$

*Proof.* According to the performance-difference lemma,

$$V^{\pi_{k+1}}(\rho) - V^{\pi_k}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi_{k+1}}} \mathbb{E}_{a \sim \pi_{k+1}(\cdot|s)} [A^{\pi_k}(s, a)] \quad (48)$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi_{k+1}}} \mathbb{E}_{a \sim \pi_{k+1}(\cdot|s)} \left[ \beta \log \frac{\pi_{k+1}(a|s) Z_k(s)}{\pi_k(a|s)} \right] \quad (49)$$

$$= \frac{\beta}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi_{k+1}}} \mathbb{E}_{a \sim \pi_{k+1}(\cdot|s)} \left[ \log \frac{\pi_{k+1}(a|s)}{\pi_k(a|s)} + \log Z_k(s) \right] \quad (50)$$

$$= \frac{\beta}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi_{k+1}}} [D_{\text{KL}}(\pi_{k+1}(\cdot|s) \parallel \pi_k(\cdot|s)) + \log Z_k(s)] \quad (51)$$

$$\geq \frac{\beta}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi_{k+1}}} [\log Z_k(s)] \quad (52)$$

$$\geq \beta \mathbb{E}_{s \sim \rho} [\log Z_k(s)], \quad (53)$$

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.(35).  $\square$

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

$$V^*(\rho) - V^{\pi_K}(\rho) \leq \frac{\beta \log |\mathcal{A}| + V_{\max}}{K(1-\gamma)}. \quad (54)$$

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

$$V^*(\rho) - V^{\pi_k}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^*} \mathbb{E}_{a \sim \pi^*(\cdot|s)} [A^{\pi_k}(s, a)] \quad (55)$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^*} \mathbb{E}_{a \sim \pi^*(\cdot|s)} \left[ \beta \log \frac{\pi_{k+1}(a|s) Z_k(s)}{\pi_k(a|s)} \right] \quad (56)$$

$$= \frac{\beta}{1-\gamma} \mathbb{E}_{s \sim d^*} \mathbb{E}_{a \sim \pi^*(\cdot|s)} \left[ \log \frac{\pi^*(a|s)}{\pi_k(a|s)} - \log \frac{\pi^*(a|s)}{\pi_{k+1}(a|s)} + \log Z_k(s) \right] \quad (57)$$

$$= \frac{\beta}{1-\gamma} \mathbb{E}_{s \sim d^*} [D_{\text{KL}}(\pi^*(\cdot|s) \parallel \pi_k(\cdot|s)) - D_{\text{KL}}(\pi^*(\cdot|s) \parallel \pi_{k+1}(\cdot|s)) + \log Z_k(s)]. \quad (58)$$

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

$$V^{\pi_{k+1}}(d^*) - V^{\pi_k}(d^*) \geq \beta \mathbb{E}_{s \sim d^*} [\log Z_k(s)]. \quad (59)$$

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

$$\begin{aligned} V^*(\rho) - V^{\pi_{K-1}}(\rho) &\leq V^*(\rho) - \frac{1}{K} \sum_{k=0}^{K-1} V^{\pi_k}(\rho) = \frac{1}{K} \sum_{k=0}^{K-1} (V^*(\rho) - V^{\pi_k}(\rho)) \\ &= \frac{\beta}{K(1-\gamma)} \sum_{k=0}^{K-1} \mathbb{E}_{s \sim d^*} [D_{\text{KL}}(\pi^*(\cdot|s) \parallel \pi_k(\cdot|s)) - D_{\text{KL}}(\pi^*(\cdot|s) \parallel \pi_{k+1}(\cdot|s)) + \log Z_k(s)]. \end{aligned} \quad (60)$$

The above expression consists of two parts:

$$\text{(I)} := \frac{\beta}{K(1-\gamma)} \sum_{k=0}^{K-1} \mathbb{E}_{s \sim d^*} [D_{\text{KL}}(\pi^*(\cdot|s) \parallel \pi_k(\cdot|s)) - D_{\text{KL}}(\pi^*(\cdot|s) \parallel \pi_{k+1}(\cdot|s))] \quad (61)$$

$$= \frac{\beta}{K(1-\gamma)} \mathbb{E}_{s \sim d^*} [D_{\text{KL}}(\pi^*(\cdot|s) \parallel \pi_0(\cdot|s)) - D_{\text{KL}}(\pi^*(\cdot|s) \parallel \pi_K(\cdot|s))] \quad (62)$$

$$\leq \frac{\beta}{K(1-\gamma)} \mathbb{E}_{s \sim d^*} [D_{\text{KL}}(\pi^*(\cdot|s) \parallel \pi_0(\cdot|s))] \quad (63)$$

$$\leq \frac{\beta \log |\mathcal{A}|}{K(1-\gamma)}, \quad (64)$$

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

$$\text{(II)} := \frac{\beta}{K(1-\gamma)} \sum_{k=0}^{K-1} \mathbb{E}_{s \sim d^*} [\log Z_k(s)] \quad (65)$$

$$\leq \frac{1}{K(1-\gamma)} \sum_{k=0}^{K-1} (V^{\pi_{k+1}}(d^*) - V^{\pi_k}(d^*)) \quad (66)$$

$$= \frac{1}{K(1-\gamma)} (V^{\pi_K}(d^*) - V^{\pi_0}(d^*)) \quad (67)$$

$$\leq \frac{V_{\max}}{K(1-\gamma)}, \quad (68)$$

where the first inequality holds by Eq.(59), and the last inequality is based on that  $0 \leq V^\pi(\mu) \leq V_{\max}, \forall \pi, \mu$ . Together we have:

$$V^*(\rho) - V^{\pi_K-1}(\rho) \leq \text{(I)} + \text{(II)} \leq \frac{\beta \log |\mathcal{A}| + V_{\max}}{K(1-\gamma)}. \quad (69)$$

The proof is completed by noticing that  $V^*(\rho) - V^{\pi_K}(\rho) \leq V^*(\rho) - V^{\pi_K-1}(\rho)$ .  $\square$

### A.3 Proofs for Section 3.3: 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  $\ell_1$  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:

$$D_{\text{KL}}^{\max}(\pi_k\|\hat{\pi}_{k+1}) \leq 1 + (1 + \sqrt{2})\delta_{\Pi}. \quad (70)$$

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

$$\begin{aligned} D_{\text{KL}}(\pi_k(\cdot|s)\|\pi_{k+1}(\cdot|s)) &= \mathbb{E}_{a \sim \pi_k(\cdot|s)} \left[ \log \frac{Z_k(s)\pi_k(a|s)}{\pi_k(a|s) \exp\left(\frac{1}{\beta}A^{\pi_k}(s, a)\right)} \right] \\ &= \mathbb{E}_{a \sim \pi_k(\cdot|s)} \left[ \log Z_k(s) - \frac{1}{\beta}A^{\pi_k}(s, a) \right] = \log Z_k(s) \leq \frac{V_{\max}}{\beta} \leq 1, \end{aligned} \quad (71)$$

where the first inequality holds due to Eq.(35) and the last step results from  $\beta \geq V_{\max}$ .

For any  $s$ , we have:

$$D_{\text{KL}}(\pi_k(\cdot|s)\|\hat{\pi}_{k+1}(\cdot|s)) - D_{\text{KL}}(\pi_k(\cdot|s)\|\pi_{k+1}(\cdot|s)) = \mathbb{E}_{a \sim \pi_k(\cdot|s)} \left[ \log \frac{\pi_{k+1}(a|s)}{\hat{\pi}_{k+1}(a|s)} \right] \quad (72)$$

$$= \sum_a \pi_{k+1}(a|s) \log \frac{\pi_{k+1}(a|s)}{\hat{\pi}_{k+1}(a|s)} + \sum_a \pi_k(a|s) \log \frac{\pi_{k+1}(a|s)}{\hat{\pi}_{k+1}(a|s)} - \sum_a \pi_{k+1}(a|s) \log \frac{\pi_{k+1}(a|s)}{\hat{\pi}_{k+1}(a|s)} \quad (73)$$

$$= D_{\text{KL}}(\pi_{k+1}(\cdot|s)\|\hat{\pi}_{k+1}(\cdot|s)) + \sum_a (\pi_k(a|s) - \pi_{k+1}(a|s)) \log \frac{\pi_{k+1}(a|s)}{\hat{\pi}_{k+1}(a|s)} \quad (74)$$

$$\leq D_{\text{KL}}(\pi_{k+1}(\cdot|s)\|\hat{\pi}_{k+1}(\cdot|s)) + \|\pi_k(\cdot|s) - \pi_{k+1}(\cdot|s)\|_1 \cdot \sup_{s,a} \left| \log \frac{\pi_{k+1}(a|s)}{\hat{\pi}_{k+1}(a|s)} \right|, \quad (75)$$

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

By Eq.(71), for all  $s \in \mathcal{S}$ ,

$$\begin{aligned} \|\pi_k(\cdot|s) - \pi_{k+1}(\cdot|s)\|_1 &= 2D_{\text{TV}}(\pi_k(\cdot|s)\|\pi_{k+1}(\cdot|s)) \\ &\leq 2\sqrt{\frac{1}{2}D_{\text{KL}}(\pi_k(\cdot|s)\|\pi_{k+1}(\cdot|s))} \leq \sqrt{2}. \end{aligned} \quad (76)$$

Recall that in Eq.(12):

$$\delta_{\Pi} := \sup_{s,a} \left| \log \frac{\pi_{k+1}(a|s)}{\widehat{\pi}_{k+1}(a|s)} \right|, \quad (77)$$

$$D_{\text{KL}}^{\max}(\pi_{k+1} \|\widehat{\pi}_{k+1}) \leq \delta_{\Pi}. \quad (78)$$

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

$$D_{\text{KL}}(\pi_k(\cdot|s) \|\widehat{\pi}_{k+1}(\cdot|s)) \leq D_{\text{KL}}(\pi_k(\cdot|s) \|\pi_{k+1}(\cdot|s)) + D_{\text{KL}}(\pi_{k+1}(\cdot|s) \|\widehat{\pi}_{k+1}(\cdot|s)) + \sqrt{2}\delta_{\Pi} \quad (79)$$

$$\leq 1 + \delta_{\Pi} + \sqrt{2}\delta_{\Pi} = 1 + (1 + \sqrt{2})\delta_{\Pi}. \quad (80)$$

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

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

$$\|d_{\mu}^{\widehat{\pi}_{k+1}} - d_{\mu}^{\pi_k}\|_1 \leq \frac{\gamma}{1-\gamma} \sqrt{2\delta_{\Pi}^+}, \quad (81)$$

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

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

$$\begin{aligned} \|\pi_k(\cdot|s) - \widehat{\pi}_{k+1}(\cdot|s)\|_1 &= 2D_{\text{TV}}(\pi_k(\cdot|s) \|\widehat{\pi}_{k+1}(\cdot|s)) \\ &\leq 2\sqrt{\frac{1}{2}D_{\text{KL}}(\pi_k(\cdot|s) \|\widehat{\pi}_{k+1}(\cdot|s))} \leq \sqrt{2\delta_{\Pi}^+}. \end{aligned} \quad (82)$$

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

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

$$\mathbb{P}_t^{\widehat{\pi}_{k+1}}(s') - \mathbb{P}_t^{\pi_k}(s') \quad (83)$$

$$= \sum_{s,a} \left( \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) \widehat{\pi}_{k+1}(a|s) - \mathbb{P}_{t-1}^{\pi_k}(s) \pi_k(a|s) \right) P(s'|s, a) \quad (84)$$

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

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

Thus we have,

$$\begin{aligned} \|\mathbb{P}_t^{\widehat{\pi}_{k+1}} - \mathbb{P}_t^{\pi_k}\|_1 &= \sum_{s'} \left| \mathbb{P}_t^{\widehat{\pi}_{k+1}}(s') - \mathbb{P}_t^{\pi_k}(s') \right| \\ &\leq \sum_{s,a,s'} \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) |\widehat{\pi}_{k+1}(a|s) - \pi_k(a|s)| P(s'|s, a) + \sum_{s,a,s'} \left| \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) - \mathbb{P}_{t-1}^{\pi_k}(s) \right| \pi_k(a|s) P(s'|s, a). \end{aligned} \quad (87)$$

The first term of the above expression:

$$(I) := \sum_{s,a,s'} \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) |\widehat{\pi}_{k+1}(a|s) - \pi_k(a|s)| P(s'|s, a) \quad (88)$$

$$= \sum_s \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) \sum_a |\widehat{\pi}_{k+1}(a|s) - \pi_k(a|s)| \sum_{s'} P(s'|s, a) \quad (89)$$

$$= \sum_s \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) \|\widehat{\pi}_{k+1}(\cdot|s) - \pi_k(\cdot|s)\|_1 \quad (90)$$

$$\leq \sum_s \mathbb{P}_{t-1}^{\widehat{\pi}_{k+1}}(s) \sqrt{2\delta_{\Pi}^+} \quad (91)$$

$$= \sqrt{2\delta_{\Pi}^+}, \quad (92)$$

where the penultimate step is due to Eq.(82), in which the upper bound holds for all  $s \in \mathcal{S}$  (this is crucial).

And the second term:

$$(\text{II}) := \sum_{s,a,s'} \left| \mathbb{P}_{t-1}^{\hat{\pi}^{k+1}}(s) - \mathbb{P}_{t-1}^{\pi^k}(s) \right| \pi_k(a|s) P(s'|s, a) \quad (93)$$

$$= \sum_s \left| \mathbb{P}_{t-1}^{\hat{\pi}^{k+1}}(s) - \mathbb{P}_{t-1}^{\pi^k}(s) \right| \sum_a \pi_k(a|s) \sum_{s'} P(s'|s, a) \quad (94)$$

$$= \left\| \mathbb{P}_{t-1}^{\hat{\pi}^{k+1}} - \mathbb{P}_{t-1}^{\pi^k} \right\|_1. \quad (95)$$

Putting everything together, we have:

$$\left\| \mathbb{P}_t^{\hat{\pi}^{k+1}} - \mathbb{P}_t^{\pi^k} \right\|_1 \leq (\text{I}) + (\text{II}) \leq \sqrt{2\delta_{\Pi}^+} + \left\| \mathbb{P}_{t-1}^{\hat{\pi}^{k+1}} - \mathbb{P}_{t-1}^{\pi^k} \right\|_1 \quad (96)$$

$$\leq 2\sqrt{2\delta_{\Pi}^+} + \left\| \mathbb{P}_{t-2}^{\hat{\pi}^{k+1}} - \mathbb{P}_{t-2}^{\pi^k} \right\|_1 \leq \dots \quad (97)$$

$$\leq t\sqrt{2\delta_{\Pi}^+} + \left\| \mathbb{P}_0^{\hat{\pi}^{k+1}} - \mathbb{P}_0^{\pi^k} \right\|_1 \quad (98)$$

$$= t\sqrt{2\delta_{\Pi}^+}, \quad (99)$$

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

Now by the definition of  $d_{\mu}^{\pi}$  (Eq.(2)), we have:

$$d_{\mu}^{\hat{\pi}^{k+1}} - d_{\mu}^{\pi^k} = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \left( \mathbb{P}_t^{\hat{\pi}^{k+1}} - \mathbb{P}_t^{\pi^k} \right). \quad (100)$$

Thus,

$$\left\| d_{\mu}^{\hat{\pi}^{k+1}} - d_{\mu}^{\pi^k} \right\|_1 \leq (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \left\| \mathbb{P}_t^{\hat{\pi}^{k+1}} - \mathbb{P}_t^{\pi^k} \right\|_1 \quad (101)$$

$$\leq (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t t \sqrt{2\delta_{\Pi}^+} \quad (102)$$

$$= \frac{\gamma}{1 - \gamma} \sqrt{2\delta_{\Pi}^+}, \quad (103)$$

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

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

$$V^{\hat{\pi}^{k+1}}(\mu) - V^{\pi^k}(\mu) \geq \frac{1}{1 - \gamma} \left( \mathbb{A}_k(\hat{\pi}_{k+1}) - \frac{\gamma V_{\max}}{1 - \gamma} \sqrt{2\delta_{\Pi}^+} \right). \quad (104)$$

*Proof.* By performance difference lemma,

$$V^{\hat{\pi}^{k+1}}(\mu) - V^{\pi^k}(\mu) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{\hat{\pi}^{k+1}}} \mathbb{E}_{a \sim \hat{\pi}_{k+1}(\cdot|s)} [A^{\pi^k}(s, a)]. \quad (105)$$

According to lemma 3, we have:

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

$$= \mathbb{E}_{s \sim d_{\mu}^{\pi^k}} \mathbb{E}_{a \sim \hat{\pi}_{k+1}(\cdot|s)} [A^{\pi^k}(s, a)] + \mathbb{E}_{s \sim d_{\mu}^{\hat{\pi}^{k+1}}} \mathbb{E}_{a \sim \hat{\pi}_{k+1}(\cdot|s)} [A^{\pi^k}(s, a)] - \mathbb{E}_{s \sim d_{\mu}^{\pi^k}} \mathbb{E}_{a \sim \hat{\pi}_{k+1}(\cdot|s)} [A^{\pi^k}(s, a)] \quad (107)$$

$$\geq \mathbb{E}_{s \sim d_{\mu}^{\pi^k}} \mathbb{E}_{a \sim \hat{\pi}_{k+1}(\cdot|s)} [A^{\pi^k}(s, a)] - \left\| d_{\mu}^{\hat{\pi}^{k+1}} - d_{\mu}^{\pi^k} \right\|_1 \cdot \sup_{s,a} |A^{\pi^k}(s, a)| \quad (108)$$

$$\geq \mathbb{E}_{s \sim d_{\mu}^{\pi^k}} \mathbb{E}_{a \sim \hat{\pi}_{k+1}(\cdot|s)} [A^{\pi^k}(s, a)] - \frac{\gamma V_{\max}}{1 - \gamma} \sqrt{2\delta_{\Pi}^+} \quad (109)$$

$$= \mathbb{A}_k(\hat{\pi}_{k+1}) - \frac{\gamma V_{\max}}{1 - \gamma} \sqrt{2\delta_{\Pi}^+}, \quad (110)$$

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

$$\begin{aligned} & \left| \mathbb{E}_{s \sim d_{\mu}^{\hat{\pi}_{k+1}}} \mathbb{E}_{a \sim \hat{\pi}_{k+1}(\cdot|s)} [A^{\pi_k}(s, a)] - \mathbb{E}_{s \sim d_{\mu}^{\pi_k}} \mathbb{E}_{a \sim \hat{\pi}_{k+1}(\cdot|s)} [A^{\pi_k}(s, a)] \right| \\ & \leq \|d_{\mu}^{\hat{\pi}_{k+1}} - d_{\mu}^{\pi_k}\|_1 \cdot \sup_s |\mathbb{E}_{a \sim \hat{\pi}_{k+1}(\cdot|s)} [A^{\pi_k}(s, a)]| \leq \|d_{\mu}^{\hat{\pi}_{k+1}} - d_{\mu}^{\pi_k}\|_1 \cdot \sup_{s,a} |A^{\pi_k}(s, a)|, \end{aligned} \quad (111)$$

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

**Theorem 2** (Optimality of AWR). *Upon termination, the update Eq.(14) returns a policy  $\pi_K$  such that:*

$$V^*(\rho) - V^{\pi_K}(\rho) \leq \frac{V_{\max}}{(1-\gamma)^2} \left\| \frac{d_{\rho}^{\pi^*}}{\mu} \right\|_{\infty} \left( \frac{\gamma}{1-\gamma} \sqrt{2\delta_{\Pi}^+} - \varepsilon + \varepsilon_{\Pi} \right), \quad (112)$$

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

*Proof.* By the performance difference lemma,

$$V^*(\rho) - V^{\pi_K}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\rho}^{\pi^*}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} [A^{\pi_K}(s, a)] \quad (113)$$

$$\leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\rho}^{\pi^*}} \left[ \max_{a \in \mathcal{A}} A^{\pi_K}(s, a) \right] \quad (114)$$

$$\leq \frac{1}{1-\gamma} \left\| \frac{d_{\rho}^{\pi^*}}{d_{\mu}^{\pi_K}} \right\|_{\infty} \mathbb{E}_{s \sim d_{\mu}^{\pi_K}} \left[ \max_{a \in \mathcal{A}} A^{\pi_K}(s, a) \right] \quad (115)$$

$$\leq \frac{1}{(1-\gamma)^2} \left\| \frac{d_{\rho}^{\pi^*}}{\mu} \right\|_{\infty} \mathbb{E}_{s \sim d_{\mu}^{\pi_K}} \left[ \max_{a \in \mathcal{A}} A^{\pi_K}(s, a) \right] \quad (116)$$

$$= \frac{1}{(1-\gamma)^2} \left\| \frac{d_{\rho}^{\pi^*}}{\mu} \right\|_{\infty} \left( \mathbb{A}_K(\hat{\pi}_{K+1}) + \mathbb{E}_{s \sim d_{\mu}^{\pi_K}} \left[ \max_{a \in \mathcal{A}} A^{\pi_K}(s, a) \right] - \mathbb{A}_K(\hat{\pi}_{K+1}) \right), \quad (117)$$

where in the second inequality we use the fact that  $\max_{a \in \mathcal{A}} A^{\pi_K}(s, a) \geq 0, \forall s$ , and the change of measure for weighted  $\ell_1$  norm. The third inequality holds because  $d_{\mu}^{\pi_K} \geq (1-\gamma)\mu, \forall \pi$  (by Eq.(2)).

According to the termination criterion Eq.(21), if the algorithm returns  $\pi_K$ , we have:

$$\mathbb{A}_K(\hat{\pi}_{K+1}) \leq \frac{\gamma V_{\max}}{1-\gamma} \sqrt{2\delta_{\Pi}^+} - \varepsilon V_{\max}. \quad (118)$$

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

$$\begin{aligned} \varepsilon_{\Pi} V_{\max} &:= \mathbb{E}_{s \sim d_{\mu}^{\pi_K}} \left[ \max_{a \in \mathcal{A}} A^{\pi_K}(s, a) \right] - \mathbb{A}_K(\hat{\pi}_{K+1}) \\ &= \mathbb{E}_{s \sim d_{\mu}^{\pi_K}} \left[ \max_{a \in \mathcal{A}} A^{\pi_K}(s, a) - A^{\pi_K}(s, \hat{\pi}_{K+1}) \right]. \end{aligned} \quad (119)$$

By the terminate condition Eq.(118) and the definition of  $\varepsilon_{\Pi}$  Eq.(119), we have:

$$V^*(\rho) - V^{\pi_K}(\rho) \leq \frac{1}{(1-\gamma)^2} \left\| \frac{d_{\rho}^{\pi^*}}{\mu} \right\|_{\infty} \left( \frac{\gamma V_{\max}}{1-\gamma} \sqrt{2\delta_{\Pi}^+} - \varepsilon V_{\max} + \varepsilon_{\Pi} V_{\max} \right) \quad (120)$$

$$= \frac{V_{\max}}{(1-\gamma)^2} \left\| \frac{d_{\rho}^{\pi^*}}{\mu} \right\|_{\infty} \left( \frac{\gamma}{1-\gamma} \sqrt{2\delta_{\Pi}^+} - \varepsilon + \varepsilon_{\Pi} \right). \quad (121)$$

Thus finishes the proof.  $\square$