GATED REGULARIZATION FOR OFFLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes a regularization method that carefully balances greedy policy optimization with the risks of out-of-distribution actions in offline reinforcement learning. The method introduces a simple and computationally efficient module, called the gate, which adaptively scales the strength of classical behavioral policy regularization based on the entropy of the behavioral policy. The behavioral policy's entropy serves as a proxy for how well the dynamics of a given state are represented in the available offline dataset: high entropy (i.e., more uniform behavior) permits aggressive off-policy updates, while low entropy enforces imitation-style conservatism. We show that gated regularization yields a value function that is provably sandwiched between those of classical regularized and optimal policies, while preserving the computational benefits of standard regularization methods. Furthermore, we prove that the resulting gated Q-learning algorithm converges almost surely, with non-asymptotic sample complexity no worse than that of classical Q-learning. Empirical experiments validate the theoretical findings and demonstrate gated regularization as a simple yet useful tool for achieving effective offline reinforcement learning.

1 INTRODUCTION

Offline Reinforcement Learning (ORL) (Levine et al., 2020; Prudencio et al., 2024), also referred to as Batch Reinforcement Learning (Lange et al., 2012), can be applied in settings where the Reinforcement Learning (RL) approach of learning through interaction and exploration is not possible, difficult, or even unethical (Laroche et al., 2019; Nachum et al., 2019; Jaques et al., 2019). For these settings, the toolbox of imitation learning provides tailored algorithms, allowing agents to learn by imitating behavior in the form of a dataset consisting of behavioral or expert demonstrations (Rashidinejad et al., 2021). Behavioral cloning (Ross & Bagnell, 2010; Ross et al., 2011) learns to copy a behavioral or expert policy. In contrast, ORL goes one step further by attempting to learn a policy that actually surpasses the performance of the policy which generated the training dataset. The concept of ORL is particularly promising in domains such as healthcare and robotics, where collecting logged data from practitioners is feasible and learning improved treatment or control policies is highly desirable. However, ORL introduces additional challenges beyond those encountered in standard RL settings. The most significant challenge in ORL is the problem of out-of-distribution (OOD) errors (Ross & Bagnell, 2010; Kumar et al., 2019; Fujimoto et al., 2019; Levine et al., 2020; Mao et al., 2024). OOD errors occur when agents choose actions that result in transitions to states that are not represented in the offline training data. A common strategy to mitigate the risk of OOD errors is to regularize the learned policy toward the behavioral policy (Wu et al., 2019; Levine et al., 2020). This regularization constrains the policy from choosing actions that could result in transitions to poorly covered or unknown regions of the state space. While effective at reducing distributional shift, this approach introduces a new limitation: by tethering the learned policy closely to the behavioral policy, it restricts the extent of possible improvement, even in areas where the offline dataset provides ample coverage.

Therefore, ORL faces a challenging trade-off: it has to greedily improve a policy while simultaneously safeguarding exploration against OOD errors. Striking the optimal balance clearly depends on the behavioral policy, and therefore the characteristics of the offline data. In this paper, we aim to provide a practical tool to help address this trade-off. More specifically, we seek to answer the following question:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

How can the learning process be optimized to most efficiently exploit the structure of offline data produced by a spatially inhomogeneous behavioral policy?

Our paper proposes a novel module (called "gate") that leverages the lower probability of OOD errors in thoroughly covered regions of the state-action space by adjusting the regularization based on the uniformity of the behavioral policy. Here, uniformity of the behavioral policy serves as a proxy for coverage of the environment's dynamics following a given state. This proxy is based on the assumption that, given a close-to-uniform behavioral policy and a sufficiently large offline dataset, most actions and subsequent states are likely to be covered in the data, allowing for greedy policy improvement and deviation from the behavioral policy. The less uniform the behavioral policy is, the more the regularization is required to bind the learned policy to the behavioral policy, since the offline data is also likely to contain only transitions based on the action chosen by the more deterministic behavioral policy. Thereby the gate transitions the algorithm from an off-policy learning regime to an imitation learning regime based on the structure of the data (Rashidinejad et al., 2021), with the behavioral policy serving as the proxy for the dataset's local structure. This is especially important in scenarios with spatially inhomogeneous behavioral policies, i.e., policies with different levels of optimality across different subsets of the state space (Li et al., 2023; Hong et al., 2023). In settings where the data is generated by a behavioral policy of consistent quality (or consistent lack of quality), it may be feasible to use a fixed regularization weight during training. However, when dealing with spatially inhomogeneous behavioral policy data, a fixed weight can become either overly optimistic or overly conservative in different regions of the state space. To that end, our proposed gate is a simple function that takes the behavioral policy as input and maps it to zero when the policy is uniform, and to one as the policy approaches determinism. The gate is a lightweight operation that can be integrated in front of explicit regularization schemes in any algorithm. It adapts learning to the effective distributional coverage "behind a state".

Related Literature. A general overview of ORL, its applications, methods, and problems is provided by Levine et al. (2020) and Prudencio et al. (2024). In addition, Rashidinejad et al. (2021) and Hong et al. (2023) provide insights into how the structure of the offline dataset connects off-policy learning with ORL.

Regularization. A straightforward way to reduce the probability of OOD actions occurring is to use regularization with respect to the behavioral policy. There are multiple ways in which regularization can be applied. Implicit regularization can be achieved by penalizing learned state-action values relative to the behavioral distribution (Kidambi et al., 2020; Lien et al., 2024). Alternatively, explicit regularization directly incorporates the behavioral policy (Wu et al., 2019; Li et al., 2023; Kostrikov et al., 2021; Ma et al., 2024), which can also be applied to diffusion-based policies (Gao et al., 2025; Guo et al., 2025). Occupation regularization aims to keep the difference between the state-action occupation induced by the learned policy and the behavioral policy small (Nachum et al., 2019; Fujimoto et al., 2019; Vlastelica et al., 2023). Our example method applies regularization by incorporating a value penalty in the Bellman update that results from deviating from the behavioral policy (Geist et al., 2019; Wu et al., 2019).

Relaxing Conservatism. Conservative learning is important in the ORL setting to avoid OOD actions caused by overly optimistic policy or value updates. Prior work suggests that learning a lower bounding Q-function can counteract such wrongly optimistic value estimates (Agarwal et al., 2020; Kumar et al., 2020). Compared to methods that directly target the overestimation bias using ensembles and robustness (Kumar et al., 2019; Lan et al., 2020; Chen et al., 2021; Rigter et al., 2022), regularization causes conservative updates by binding policy and value estimation to the behavioral policy. To improve the learned policy compared to the behavioral policy, it is required to relax the conservatism when possible (Levine et al., 2020; Prudencio et al., 2024; Cetin et al., 2024; Jing et al., 2026). Model-based approaches can penalize states not covered by the offline data by regularizing against actions that would lead to OOD states without binding the learned policy to the behavioral policy (Fujimoto et al., 2019; Kidambi et al., 2020). Modeling ORL as a two-player game, where one player finds a reward maximizing policy and the other player pessimistically evaluates that policy with respect to the behavioral policy or data, can be utilized to reduce absolute pessimism to relative pessimism when updating values and policies (Rajeswaran et al., 2020; Cheng et al., 2022; Bhardwaj et al., 2023; Wei et al., 2024). Another adversarial approach is proposed by Wang et al. (2024), who propose to adversarially split the offline data to maximize the distribution shift of the train-validation splits in order to learn generalization across distribution shifts. Finally, the level of conservatism can

be tuned for subsets of the state and action spaces based on one or more confidence metrics that guide the balance between cautious and greedy updates (Koppel et al., 2024; Liu et al., 2024; Kim & Oh, 2023; Yeom et al., 2024; Zhang et al., 2024a; Mao et al., 2025). Our gate falls under the umbrella of the latter line of relaxing conservatism and is conceptually related to Luo et al. (2025). However, the coefficients of Luo et al. (2025) depend on the learned policy and evolve with training, whereas our gate depends only on the behavioral policy and is therefore decoupled from the learning algorithm.

Constraining Policy Support. On top of regularizing values and policies, the learned policy can also be constrained to an approximation of the support of the behavioral data (Laroche et al., 2019; Kumar et al., 2019; Mao et al., 2023; 2025). When working only with actions sampled to match the offline data distribution, it is then possible to select the high-reward actions in a greedy manner (Wang et al., 2020; Hu et al., 2024). Similar to applying a penalty to states that are outside of the behavioral distribution of a learned model (Kidambi et al., 2020), one can also rank state-action pairs based on whether they have been seen in the offline data during training (Gulcehre et al., 2021). Our closed-form gated policy maximization found in Section 3 constrains the maximizing policy of the value update step to the support of the behavioral policy.

In practice, many of the described methods combine multiple elements such as models, ensembles, robustness, and regularization (Kumar et al., 2019; Kidambi et al., 2020; Yu et al., 2020; Rigter et al., 2022; Zhang et al., 2024b; Wang et al., 2026). Our gated example method is inspired by the comparative study of regularization strategies of Wu et al. (2019). They find that value penalties work slightly better than direct policy regularization and that the concrete choice of divergence does not affect the performance consistently. Finally, they observe that more advanced regularization methods do not necessarily outperform simple ones. Combined, these results motivate the choice of (i) basing the regularization as well as the gate on the Kullback–Leibler (KL) divergence, (ii) applying the regularization to the value estimate, and (iii) using a simple gate to make the regularization weight adaptive.

Contribution. We introduce an adaptive regularization scaling method in the form of a gate that can be used in the setting of ORL. The gate uses the uniformity of the behavioral policy as a proxy for how likely the environment's dynamics in a given state are to be contained in the offline dataset, thereby balancing greedy policy improvement with avoiding actions that are likely to cause out-of-distribution errors. In a nutshell, the main contributions of this paper can be summarized as follows.

- We propose a new gated regularization scheme that enables a smooth transition between off-policy learning and imitation learning.
- We prove that the gated regularization characterizes a value function that is consistently closer to the optimal value function than classical regularized approaches, while retaining their computational advantages (Theorem 2.1).
- We establish that, when applied to a single trajectory of data generated by a behavioral policy, the proposed method almost surely converges to the corresponding fixed point, with a non-asymptotic sample complexity no worse than that of classical Q-learning (Theorems 3.1 and 3.2).
- We show that the gated learning scheme remains computationally efficient, requiring only closed-form evaluations and introducing no additional computational overhead compared to classical Q-learning methods.

Notation. Let $\mathcal{S}$ and $\mathcal{A}$ denote finite state and action spaces and let $\Delta(\mathcal{A})$ be the corresponding probability simplex. For a stationary randomized policy $\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}$, we introduce the notation $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and $\pi_s \in \Delta(\mathcal{A})$, where $\pi_s(a) = \pi(a|s)$ for all $a \in \mathcal{A}$. Further, for a mapping $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ we introduce the vector notation $Q_s = Q(s, \cdot) \in \mathbb{R}^{|\mathcal{A}|}$ and the full vector of all Q-values as $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ with the corresponding value vector $V \in \mathbb{R}^{|\mathcal{S}|}$. The Euclidean inner product is denoted by $\langle \cdot, \cdot \rangle$. z denotes a generic scalar or vector and κ is a generic coefficient or small positive parameter that are locally defined where they appear and p, q are generic elements of $\Delta(\mathcal{A})$.

Structure. The work is structured as follows. Section 2 introduces the problem, our solution approach, and how the fixed points of our method compare to the optimal values as well as the behavioral values. Section 3 introduces the sample-based algorithm derived from our solution approach and its convergence to the established fixed points. Section 4 contains outlook experiments in the setting of continuous state-action spaces. Section 5 summarizes the findings, discusses

limitations, and provides an outlook. Appendix A lists selected existing results that have been used throughout this paper, adapted to fit our setting and notation. Appendix B contains all proofs and derivations of the results and contribution of this work. Appendix C provides the links to the code repositories and experiments in different settings.

2 Problem Setup and Fixed Points

Consider a Markov Decision Process (MDP) given by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, s_0)$ where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a kernel such that $S'_k = S_{k+1} \sim P(\cdot|s, a)$ governs the model's transition dynamics given $S_k = s$ and $A_k = a$, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a reward function, $\gamma \in (0, 1)$ is a discount factor, and $s_0 \in \mathcal{S}$ is an initial state. A randomized stationary policy is an element of $\Pi = \Delta(\mathcal{A})^{|\mathcal{S}|}$, the set of all possible policies, that assigns a probability distribution over all actions to each state with $A_k \sim \pi(\cdot|s) \in \Delta(\mathcal{A})$ given $S_k = s$.

A value function $V : \mathcal{S} \to \mathbb{R}$ maps each state to an associated value. In this setting, the infinite horizon discounted reward problem is then given by

$$V^\star(s_0) = \max_{\pi \in \Pi} \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r(S_k, A_k) \middle| S_0 = s_0 \right]. \tag{1}$$

Assuming the full model of the environment is known, the Bellman operator for a specific stationary policy $\pi \in \Pi$ is defined as $\mathcal{T}_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$, where for each $s \in \mathcal{S}$

$$\left(\mathcal{T}_\pi V\right)(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V(s') \right),$$

with $\mathcal{T}V = \max_\pi \mathcal{T}_\pi V$ being its greedy counterpart. For each value function $V(s)$, there exists an associated Q-function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for each $a \in \mathcal{A}$, defined as

$$Q(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V(s'),$$

such that we can rewrite the Bellman operator as $\left(\mathcal{T}_\pi V\right)(s) = \langle \pi_s, Q_s \rangle$. It is well known (Geist et al., 2019; Bertsekas, 2018; Bertsekas & Tsitsiklis, 2008) that the Bellman operators are γ-contractions with respect to the ℓ_∞-norm with unique fixed points $V^\pi(s) = \left(\mathcal{T}_\pi V^\pi\right)(s) = \langle \pi_s, Q^\pi_s \rangle$ and $V^\star(s) = \max_{\pi \in \Pi} \left(\mathcal{T}_\pi V^\star\right)(s) = \left(\mathcal{T}_{\pi^\star} V^\star\right)(s) = \langle \pi^\star_s, Q^\star_s \rangle$ for all $s \in \mathcal{S}$. Denote a behavioral policy as $\pi^b \in \Pi$, which in general differs from an optimal policy $\pi^\star$, such that $V^b(s) = V^{\pi^b}(s) \leq V^{\pi^\star}(s) = V^\star(s)$ for all $s \in \mathcal{S}$. The behavioral policy, applied to $\mathcal{M}$, induces a restricted MDP $\mathcal{M}_b = (\mathcal{S}_b, \mathcal{A}_b, P, r, \gamma, s_0)$, where $\mathcal{A}_b(s) = \{a \in \mathcal{A} | \pi^b(a|s) > 0\}$ and $\mathcal{S}_b = \{s \in \mathcal{S} | \exists k \geq 0 \text{ s.t. } \mathbb{P}_{\pi^b}[S_k = s | S_0 = s_0] > 0\}$ is all the states that are reachable from s_0 by following π^b. The set of allowed policies in $\mathcal{M}_b$ then becomes $\Pi_b \subseteq \Pi$ such that $\pi \in \Pi_b$ if and only if $\sum_{a \in \mathcal{A}_b(s)} \pi(a|s) = 1$ for all $s \in \mathcal{S}_b$. Therefore, $P(\cdot|s, a) \in \Delta(\mathcal{S}_b)$ for all $s \in \mathcal{S}_b$ and $a \in \mathcal{A}_b(s)$, which makes $\mathcal{M}_b$ closed under $\pi \in \Pi_b$. The maximum value function of the restricted MDP is obtained by maximizing the policy over the elements of Π_b, which yields the optimal value function $V^\star_b$, for which $V^b(s) \leq V^\star_b(s) \leq V^\star(s)$ for all $s \in \mathcal{S}_b$. The respective Bellman operator as well as the Q-function of the restricted MDP are obtained analogously by restricting the sums to the state and action spaces of $\mathcal{M}_b$.

2.1 Regularized Bellman Operator

A well-known technique in reinforcement learning for enhancing exploration, improving robustness, and thereby promoting better convergence behavior is entropic regularization; see Geist et al. (2019) for a detailed discussion.

Definition 2.1 (Regularization Function)**.** *A regularization function is a function* $\mathcal{R} : \Delta(\mathcal{A}) \times \Delta(\mathcal{A}) \to \mathbb{R} \cup \{+\infty\}$ *such that for any* $p, q \in \Delta(\mathcal{A})$*,* $\mathcal{R}(p, q) = 0$ *if and only if* $p = q$*.*

A concrete choice of regularization function can be given as the relative entropy.

Definition 2.2 (Regularized Bellman Operator)**.** *The regularized Bellman operator for a parameter $\beta > 0$, a specific stationary policy $\pi \in \Pi$, and a regularization function $\mathcal{R}$ is*

$$(\mathcal{T}_{\mathcal{R},\pi} V)(s) = (\mathcal{T}_\pi V)(s) - \beta \mathcal{R}(\pi_s, \pi_s^b), \tag{2}$$

with the corresponding greedy regularized Bellman operator being

$$(\mathcal{T}_{\mathcal{R}} V)(s) = \max_{\pi \in \Pi} (\mathcal{T}_{\mathcal{R},\pi} V)(s). \tag{3}$$

We can again use the associated Q-function to write the regularized Bellman operator as $(\mathcal{T}_{\mathcal{R},\pi} V)(s) = \langle \pi_s, Q_s \rangle - \beta \mathcal{R}(\pi_s, \pi_s^b)$. If we assume $\mathcal{R}$ to be strongly convex and Lower Semi-Continuous (LSC) in its first argument, we can follow the argument from Geist et al. (2019) and write the regularized greedy Bellman operator as the convex conjugate $\Omega_{\mathcal{R}}^* : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}$ of $\Omega_{\mathcal{R}} : \Delta(\mathcal{A}) \to \mathbb{R}$,

$$\Omega_{\mathcal{R}}^*(Q_s) = \max_{\pi \in \Pi} \{\langle \pi_s, Q_s \rangle - \Omega_{\mathcal{R}}(\pi_s)\},$$

where $\Omega_{\mathcal{R}}(\pi_s) = \beta \mathcal{R}(\pi_s, \pi_s^b)$, since π_s^b is fixed for each state $s \in \mathcal{S}$ and $\beta > 0$ is a constant. This allows us to rewrite the regularized Bellman operators (2) and (3) as $(\mathcal{T}_{\mathcal{R},\pi} V)(s) = (\mathcal{T}_\pi V)(s) - \Omega_{\mathcal{R}}(\pi_s)$ and $(\mathcal{T}_{\mathcal{R}} V)(s) = \Omega_{\mathcal{R}}^*(Q_s)$. Since Geist et al. (2019, Proposition 1) applies and Π is compact, there exists a unique greedy policy

$$\pi^{\mathcal{R}} = \arg\max_{\pi \in \Pi} (\mathcal{T}_{\mathcal{R},\pi} V)(s), \quad \forall s \in \mathcal{S}$$

associated with any value function V, and we denote $\left(\mathcal{T}_{\mathcal{R},\pi^{\mathcal{R}}} V\right)(s) = (\mathcal{T}_{\mathcal{R}} V)(s)$ for all $s \in \mathcal{S}$ as in Equation (3). Further, due to Geist et al. (2019, Proposition 2) the regularized Bellman operators are γ-contractions with respect to the ℓ_∞-norm and we denote by $V^{\mathcal{R},\pi}$ and $V^{\mathcal{R}}$ their unique fixed points, i.e., $V^{\mathcal{R},\pi}(s) = \left(\mathcal{T}_{\mathcal{R},\pi} V^{\mathcal{R},\pi}\right)(s)$ and $V^{\mathcal{R}}(s) = \left(\mathcal{T}_{\mathcal{R}} V^{\mathcal{R}}\right)(s)$ for all $s \in \mathcal{S}$.

2.2 Gated Bellman Operator

A limitation of the regularized Bellman operator discussed in Section 2.1 is that the magnitude β of the entropy regularization in (2) is fixed and does not adapt to the properties and local structure of the available offline data. Addressing this limitation is the key motivation behind introducing the gate: it can be interpreted as a behavioral-policy-dependent adjustment of the entropic regularization strength.

Definition 2.3 (Gate)**.** *A gate is a function $\Phi : \Delta(\mathcal{A}) \to [0, 1]$ such that for any $p \in \Delta(\mathcal{A})$*

(i) $\Phi(p) = 0$, if p is the uniform distribution on $\mathcal{A}$, and

(ii) $\Phi(p) = 1$, if there is an $i \in \{1, \ldots, |\mathcal{A}|\}$ such that $p_i = 1$.

The intuition for the gate Φ is that it maps a maximum entropy distribution to 0 whereas a minimum entropy distribution is mapped to 1. By denoting the Shannon entropy as $H(\cdot)$ and considering a policy $\pi \in \Pi$ and $|\mathcal{A}| \geq 2$, a concrete choice of a gate can be given as

$$\Phi(\pi_s) = 1 - \frac{H(\pi_s)}{\log(|\mathcal{A}|)}. \tag{4}$$

Definition 2.4 (Gated Bellman Operator)**.** *Let Φ be a gate. The gated regularized Bellman operator, which from now on will just be called the gated Bellman operator, for a parameter $\beta > 0$ and a specific stationary policy $\pi \in \Pi$ is*

$$(\mathcal{T}_{\Phi,\pi} V)(s) = (\mathcal{T}_\pi V)(s) - \beta \Phi(\pi_s^b) \mathcal{R}(\pi_s, \pi_s^b), \tag{5}$$

with the corresponding greedy gated Bellman operator being

$$(\mathcal{T}_{\Phi} V)(s) = \max_{\pi \in \Pi} (\mathcal{T}_{\Phi,\pi} V)(s). \tag{6}$$

Analogous to the Bellman operators above, we can rewrite the gated Bellman operator as $(\mathcal{T}_{\Phi,\pi} V)(s) = \langle \pi_s, Q_s \rangle - \Omega_\Phi(\pi_s)$ where $\Omega_\Phi(\pi_s) = \Phi(\pi_s^b)\Omega_{\mathcal{R}}(\pi_s) = \beta \Phi(\pi_s^b) \mathcal{R}(\pi_s, \pi_s^b)$, where we use the convention of $0 \cdot (+\infty) = 0$. In Appendix B.1, we show that the gated Bellman operator and its greedy counterpart satisfy Geist et al. (2019, Proposition 2). Therefore, both are γ-contractions with respect to the ℓ_∞-norm with the value functions $V^{\Phi,\pi}$ and V^{Φ} being their unique fixed points, respectively, and a corresponding policy exists due to $\mathcal{R}$ being LSC in its first argument and Π being compact such that

$$\pi^{\Phi} \in \arg\max_{\pi \in \Pi} \left(\mathcal{T}_{\Phi,\pi} V^{\Phi}\right)(s), \quad \forall s \in \mathcal{S}.$$

2.3 Bounds on Fixed Points

Knowing the theoretical fixed points of the regularized and gated Bellman operators, we are now interested in how they relate to each other and the optimal value function $V^\star$.

Lower bound on $V^{\mathcal{R}}$. Since the greedy regularized Bellman operator (Definition 2.2) includes a maximum, the value function after application of the operator can be bounded from below by any policy. We can use the behavioral policy π^b, which results in a vanishing regularization term $\mathcal{R}(\pi^b_s, \pi^b_s) = 0$, i.e., $\left(\mathcal{T}_{\mathcal{R}} V\right)(s) = \max_{\pi\in\Pi}\left(\mathcal{T}_{\mathcal{R},\pi} V\right)(s) \geq \left(\mathcal{T}_{\mathcal{R},\pi^b} V\right)(s) = \left(\mathcal{T}_{\pi^b} V\right)(s)$.

Proposition 2.1. *For any $V(s)$, $\beta > 0$, and $k \in \mathbb{N}_+$*

$$\left(\mathcal{T}^k_{\mathcal{R}} V\right)(s) \geq \left(\mathcal{T}^k_{\pi^b} V\right)(s), \quad \forall s \in \mathcal{S}.$$

Taking the limits on both sides results in a lower bound of the fixed point of the regularized Bellman operator $V^{\mathcal{R}}(s) = \lim_{k\to\infty}\left(\mathcal{T}^k_{\mathcal{R}} V\right)(s) \geq \lim_{k\to\infty}\left(\mathcal{T}^k_{\pi^b} V\right)(s) = V^b(s)\ \forall s \in \mathcal{S}$.

Lower bound on V^{Φ}. Following the same arguments as applied to the regularized Bellman operator, the gated Bellman operator also can be bounded from below by $\left(\mathcal{T}^k_{\Phi} V\right)(s) \geq \left(\mathcal{T}^k_{\pi^b} V\right)(s)\ \forall s \in \mathcal{S}$.

Proposition 2.2. *For any $V(s)$, $\beta > 0$, $k \in \mathbb{N}_+$, and any gate Φ*

$$\left(\mathcal{T}^k_{\Phi} V\right)(s) \geq \left(\mathcal{T}^k_{\mathcal{R}} V\right)(s), \quad \forall s \in \mathcal{S}.$$

Taking the limits on both sides yields $V^{\Phi}(s) = \lim_{k\to\infty}\left(\mathcal{T}^k_{\Phi} V\right)(s) \geq \lim_{k\to\infty}\left(\mathcal{T}^k_{\mathcal{R}} V\right)(s) = V^{\mathcal{R}}(s)$.

Theorem 2.1 (Bounds on the Gated Bellman Operator)**.** *For any gate Φ and any $\beta > 0$ the fixed points V^{Φ} of the greedy gated Bellman operator* (6)*, $V^{\mathcal{R}}$ of the greedy regularized Bellman operator* (3)*, and $V^\star$ of the classical Bellman operator satisfy*

$$V^{\mathcal{R}}(s) \leq V^{\Phi}(s) \leq V^\star(s) \quad \forall s \in \mathcal{S},$$

assuming the regularized Bellman operator as well as the gated Bellman operator use the same β.

Based on the same arguments, but applied to the restricted MDP $\mathcal{M}_b$ and using Π_b as the feasibility set for all maximization operations, we obtain the same ordering for the corresponding fixed points of the restricted setting. Namely,

$$V^{\mathcal{R}}_b(s) \leq V^{\Phi}_b(s) \leq V^\star_b(s) \quad \forall s \in \mathcal{S}_b. \tag{7}$$

3 Gated Q-learning

In this section, we relax the assumption of full knowledge of the transition kernel P and instead assume access to a dataset of state-action trajectories generated by a behavioral policy $\mathcal{D}_{\pi^b} = \{(s_k, a_k, s'_k)\}_{k=0}^{K-1}$, where $s'_k \sim P(\cdot|s_k, a_k)$ and $a_k \sim \pi^b(\cdot|s_k)$. Unlike the setting with full knowledge of the model, where Bellman updates can be applied simultaneously to all state-action pairs, updates have to be performed based on the individual samples. This is known as the asynchronous setting (Sutton & Barto, 2018; Meyn, 2022). Throughout this section, we assume that the behavioral policy $\pi^b \in \Pi$ is known and that the Markov state-action chain induced by π^b is irreducible and aperiodic.

Definition 3.1 (Gated Bellman Operator for Q-values)**.** *The gated regularized Bellman operator for Q-values*[1] *for a parameter $\beta > 0$, a specific stationary policy $\pi \in \Pi$, all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ is*

$$\left(\Gamma_{\Phi,\pi} Q\right)(s,a) = r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a)\left(\langle \pi_{s'}, Q_{s'}\rangle - \Omega_{\Phi}(\pi_{s'})\right), \tag{8}$$

where $\Omega_{\Phi}(\pi_{s'}) = \beta\Phi(\pi^b_{s'})\mathcal{R}(\pi_{s'}, \pi^b_{s'})$ and the corresponding greedy gated Bellman operator for Q-values is

$$\left(\Gamma_{\Phi} Q\right)(s,a) = \max_{\pi\in\Pi}\left(\Gamma_{\Phi,\pi} Q\right)(s,a). \tag{9}$$

[1] From now on we refer to this as the gated Bellman operator for Q-values.

Without access to the transition kernel P, a natural approach is to introduce an empirical gated Bellman operator to be used with the offline dataset $\mathcal{D}_{\pi^b}$.

Definition 3.2 (Gated Bellman Operator Single Sample Approximation)**.** *For a sample $(s, a, s') \in \mathcal{D}_{\pi^b}$, the single sample approximation for the greedy gated Bellman operator* (9) *is*

$$\left(\hat{\Gamma}_\Phi Q\right)(s, a, s') = r(s, a) + \gamma \max_{\pi \in \Pi} \left\{\langle \pi_{s'}, Q_{s'} \rangle - \Omega_\Phi(\pi_{s'})\right\}, \tag{10}$$

where $\Omega_\Phi(\pi_{s'}) = \beta\Phi(\pi^b_{s'})\mathcal{R}(\pi_{s'}, \pi^b_{s'})$.

Definition 3.3 (Gated Q-learning)**.** *Given an offline dataset $\mathcal{D}_{\pi^b} = \{(s_k, a_k, s'_k)\}_{k=0}^{K-1}$ and a learning rate sequence $(\alpha_k(s, a))_{k\geq 0}$, the gated Q-learning update rule is*

$$\hat{Q}_{k+1}(s_k, a_k) = (1 - \alpha_k(s_k, a_k))\hat{Q}_k(s_k, a_k) + \alpha_k(s_k, a_k)\left(\hat{\Gamma}_\Phi \hat{Q}_k\right)(s_k, a_k, s'_k), \tag{11}$$

where $\hat{Q}_0$ is initialized to zero.

Theorem 3.1 (Convergence of Gated Q-learning)**.** *If the offline set of transitions $\{(s_k, a_k, s'_k)\}_{k\geq 0}$ is sampled from an ergodic Markov state-action chain and the learning rates $(\alpha_k(s, a))_{k\geq 0}$ satisfy the standard Robbins-Monro conditions (Robbins & Monro, 1951), then the sequence $(\hat{Q}_k)_{k\geq 0}$ generated by the gated Q-learning update rule* (11) *converges to Q^Φ almost surely as $k \to \infty$, i.e., $\lim_{k\to\infty} \|\hat{Q}_k - Q^\Phi\|_\infty = 0 \quad \mathbb{P}^{\pi^b}$-a.s., where Q^Φ is such that $\langle \pi^\Phi_s, Q^\Phi_s \rangle - \Omega_\Phi(\pi^\Phi_s) = V^\Phi(s)$ for all $s \in \mathcal{S}$.*

For the analysis of finite sample complexity, which is adapted from Li et al. (2021, Theorem 3), let $\vartheta_k(s, a)$ be the visitation count of state-action pair (s, a) up to timestep k, and use a global learning rate sequence $(\alpha_k)_{k\geq 0}$ given by

$$\alpha_k = \min\left\{1, c_\alpha \exp\left(\left\lfloor \log\left(\frac{\log(k)}{\hat{\eta}_{\min,k}(1-\gamma)\gamma^2 k}\right)\right\rfloor\right)\right\} \tag{12}$$

with $\alpha_0 = \alpha_1 = 1$, $c_\alpha > 0$ and

$$\hat{\eta}_{\min,k} = \begin{cases} \frac{1}{|\mathcal{S}||\mathcal{A}|} & \text{if} \quad \min_{(s,a)} \vartheta_k(s, a) = 0, \\ \hat{\eta}_{\min,k-1} & \text{if} \quad \frac{1}{3} < \frac{\frac{1}{k}\min_{(s,a)} \vartheta_k(s,a)}{\hat{\eta}_{\min,k-1}} < 3, \quad \text{and} \\ \frac{1}{k}\min_{(s,a)} \vartheta_k(s, a) & \text{otherwise.} \end{cases}$$

Theorem 3.2 (Non-Asymptotic Sample Complexity)**.** *Let the Markov state-action chain induced by π^b be ergodic and its stationary distribution be denoted by η_{π^b}. Further, let the offline set of transitions $\{(s_k, a_k, s'_k)\}_{k=0}^{K-1}$ be sampled from this chain and the learning rate schedule be given by* (12)*. And finally, let*

$$\tilde{Q}_k = \begin{cases} \hat{Q}_k & \text{if} \quad \alpha_{k+1} \neq \alpha_k \\ \tilde{Q}_{k-1} & \text{else.} \end{cases}$$

Then, for the gated Q-learning update rule of (11)*, there exists a universal constant $C > 0$, such that for any $\delta \in (0, 1)$ and $\varepsilon \in (0, \frac{1}{1-\gamma}]$,* (12) *satisfies the Robbins-Monro conditions (Robbins & Monro, 1951) and*

$$\|\tilde{Q}_K - Q^\Phi\|_\infty \leq \varepsilon,$$

where Q^Φ is such that $\langle \pi^\Phi_s, Q^\Phi_s \rangle - \Omega_\Phi(\pi^\Phi_s) = V^\Phi(s)$ for all $s \in \mathcal{S}$, with probability at least $1 - \delta$, given that the update count K satisfies $K \geq 24$ and

$$K \geq \frac{C}{\eta_{min}} \max\left\{\frac{1}{(1-\gamma)^5\varepsilon^2}, \frac{k_{mix}}{(1-\gamma)\gamma^2}\right\} \log\left(\frac{|\mathcal{S}||\mathcal{A}|K}{\delta}\right) \log(K)$$

where $\eta_{min} = \min_{s,a} \eta_{\pi^b}(s, a)$ and k_{mix} is the mixing time of the chain based on Paulin (2015, Definition 1.3).

The same results of Theorem 3.1 and Theorem 3.2 hold, with Q^Φ_b instead of Q^Φ and using the respective state and action spaces for the bounds, under equivalent assumptions, when the samples are obtained from an ergodic Markov chain induced by π^b on $\mathcal{M}_b$ and by using Π_b as the feasibility set. Although the restricted MDP realizes different fixed points, the relative order of the regularized and gated fixed points remains unchanged from the unrestricted setting.

Closed-Form Gate. Using the KL divergence as the regularization function and gating with any gate Φ following Definition 2.3, we express $\bar{\pi}(\cdot|s) \in \arg\max_{\pi \in \Pi} \{\langle \pi_s, Q_s \rangle - \Omega_\Phi(\pi_s)\}$, when $\Phi(\pi^b_s) > 0$, in closed form. We follow Sutter et al. (2019, Lemma 2) and Nesterov (2005, Section 5.2) and express $\bar{\pi}$ as

$$\bar{\pi}(a|s) = \frac{e^{z(s,a)}\pi^b(a|s)}{\sum_{a' \in \mathcal{A}} e^{z(s,a')}\pi^b(a'|s)} \quad \text{with} \quad z(s,a) = \frac{Q(s,a) - \max_{a' \in \mathcal{A}} Q(s,a')}{\beta \Phi(\pi^b_s)}. \tag{13}$$

The non-zero gate requirement is non-restrictive in practice since it is unlikely for π^b to be exactly uniform. For the edge case, it would either be possible to apply an ϵ-perturbation to π^b, or treat the case separately and use a greedy policy with $\bar{\pi}(a|s) = 1$ for $a = \arg\max_{a' \in \mathcal{A}} Q(s,a')$.

Continuous Setting. To extend the concept for application in the case of continuous action spaces, we have to update Definition 2.3 to $\Phi^c : \Delta(\mathcal{A}) \to [0,1]$, where $\Delta(\mathcal{A})$ now denotes the set of probability distributions over the continuous action space $\mathcal{A}$. As in the discrete case, we base the mapping of the gate on properties of the distribution over actions. In order to achieve the desired functionality of the gate, it has to measure the degree of uniformity of the distribution over action values. Similar to the discrete setting, this can be done in many different ways. Assume a policy that draws its actions from a parametrized Gaussian, as is common practice in continuous-action based algorithms, such that $a \sim \pi_c(\cdot|s) = \mathcal{N}(\mu(s), \sigma^2(s))$. In this case, a simple choice of a gate can be given by

$$\Phi^c(\pi_c(\cdot|s)) = e^{-\frac{\sigma(s)}{\rho}}, \tag{14}$$

where $\rho > 0$ is a tunable parameter that controls how fast we want the gate to shut down the regularization for increasing values of $\sigma(s)$. This simple gate is computationally cheap and matches the desired shape of a gating function. That is, the larger the standard deviation of the behavioral policy, the wider the range of actions we expect to be contained in the behavioral data, analogous to more uniform behavioral policies in the discrete setting. Consequently, the state-dependent standard deviation of the behavioral policy can be used as a proxy for the coverage of state dynamics in the offline data.

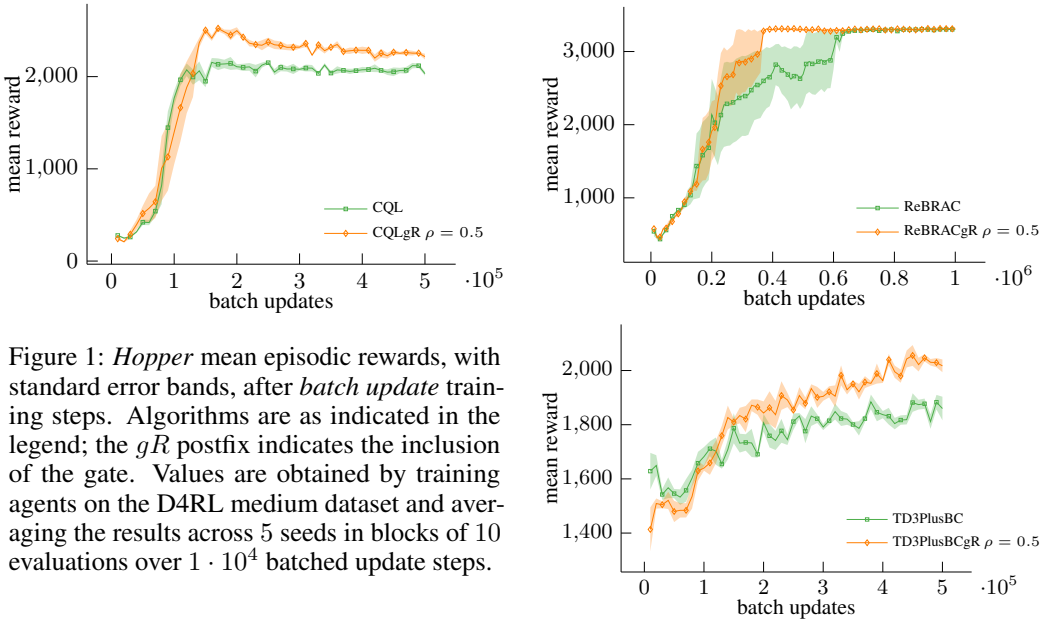


Figure 1: *Hopper* mean episodic rewards, with standard error bands, after *batch update* training steps. Algorithms are as indicated in the legend; the gR postfix indicates the inclusion of the gate. Values are obtained by training agents on the D4RL medium dataset and averaging the results across 5 seeds in blocks of 10 evaluations over $1 \cdot 10^4$ batched update steps.

4 Experiments

Appendix C contains experiments intended to demonstrate different aspects of the behavioral-policy-based gating concept. These include tabular experiments that verify the theory and experiments

with continuous states and actions to demonstrate viability in more practical applications. It also provides the hardware specifications as well as the links to the code repositories for both the discrete and continuous experiments. In this section, to provide an outlook in the direction of more applied algorithms and the incorporation of the gate into established learning methods, we run experiments in Gymnasium MuJoCo environments (Towers et al., 2024; Todorov et al., 2012) using offline datasets from D4RL (Fu et al., 2021).

The offline datasets are annotated with their corresponding behavioral policies based on models reconstructed from the policy model parameters in the metadata. As baseline algorithms we chose a set of explicitly regularized learning methods, namely CQL (Kumar et al., 2020), TD3+BC (Fujimoto & Gu, 2021), and ReBRAC (Tarasov et al., 2023), and their respective implementations from the d3rlpy library (Seno & Imai, 2022). When running the experiments, we use the exact hyperparameters provided in their d3rlpy implementation for each environment and for both variants of each algorithm, the original as well as the gated version. The gate (14) with $\rho = 0.5$ is used across all methods and plugged into each algorithm's respective regularization step. No additional algorithmic or code-level optimizations are included in the methods used throughout this experiment (Engstrom et al., 2020). Figure 1 shows the comparison between the algorithms and their gated variants trained on D4RL *Hopper* medium offline data. It is an example of what the gate can deliver when the conditions of the behavioral data are inhomogeneous. Appendix C.2 contains examples where the behavioral data is more homogeneous. It is clearly visible that the effect of including the gate varies strongly across algorithms and environments. The effect can include significantly improved learning performance in the early stages of training, a policy with higher reward, or no significant change relative to the baseline algorithm's performance. This aligns with the expectations implied by the theory underlying the gate.

5 Limitations & Conclusion

Limitations. A major limitation of gated methods is their reliance on sufficient data for the "behavioral policy as a proxy for dynamics knowledge" assumption to hold. In sparse data regimes, one could encounter a state with stochastic transition dynamics that is visited exactly $|\mathcal{A}|$ times with each action chosen once due to π^b being uniform. In such cases, the gate permits unjustified optimistic greediness. Additionally, it is not always the case that the behavioral policy is recorded and therefore readily available. Instead, it might have to be estimated from the offline data first. However, one might argue that in the settings where ORL is intended to be deployed in the long term, such as the medical sector, autonomous vehicles, or robotics, the behavioral policy is very likely to be available. This also highlights an interesting perspective on whether or not to use the estimated behavioral policy, even if the true behavioral policy is available. In states where π^b supports more actions than have been recorded in the data, regularization with π^b draws the learned policy closer to potential OOD actions than an estimated behavioral policy that, even though less accurate, may be a safer policy to use as a regularization target.

Conclusion. In this work, we introduce gated regularization, a simple and computationally cheap module that allows regularized Q-learning to exploit more effectively the inherent structure of offline data. It is based on using the entropy of the behavioral policy as a proxy for how much of a state's actions and corresponding dynamics are covered by the offline dataset. We prove that, in the sampled tabular setting, the proposed method achieves the same non-asymptotic sample complexity as classical Q-learning and yields a value function that is never further from the optimal than its non-gated counterpart. Crucially, the KL gate allows for a closed-form solution for the gated policy-maximization step, making each gated update as cheap as a standard Q-learning update while strictly constraining the learned policy to the support of the behavioral policy, thereby safeguarding against OOD actions in the value updates. Finally, we validate our theory through numerical simulations, providing evidence that the gate could be incorporated into explicitly regularized methods to improve performance in settings with inhomogeneous behavioral policies. Therefore, the gate is a lightweight and useful concept to be added to the ORL toolbelt.

Outlook. Based on the observations in Section 4, tuning the hyperparameters and deploying code-level optimizations with the gate included may further improve learning speed and final policy performance. Further, the theoretical properties of using the gate in regularized policy-iteration methods (Wu et al., 2019; Sutton & Barto, 2018) could be analyzed.

AI USE STATEMENT

In this work, we used generative AI tools to assist in checking our proofs, both to find mistakes and to obtain a "third-party" point of view, which helped us in making assumptions explicit that we had left implicit due to the context. We did not use generative AI tools for the development of our ideas, methods, claims, formulation of hypotheses, or the analysis and interpretation of results. No translation or formatting of datasets was required. Additionally, we used generative AI tools to make documentation of the code more thorough and detailed and to help find typos, grammatical mistakes, and notation inconsistencies, and to check that all figures load the correct data. Each issue flagged by the generative AI tools was evaluated for necessity and, where applicable, verified mathematically; no edits to this manuscript were directly made by any of the used tools. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is theoretical and not intended for direct real-world application. While the work does have a clear real-world motivation, there is a large gap between the synthetic setting assumed throughout and any practical deployment. Since the work involves neither human subjects nor collection or release of data, and instead uses synthetic data and open-source environments, algorithms, and data, concerns regarding privacy, security, fairness, and legal compliance do not arise.

REPRODUCIBILITY STATEMENT

In order to ensure reproducibility, all assumptions and requirements are properly stated, and proofs for all statements are provided in Appendix B. Further, Appendix A restates the cited results used throughout this paper in our notation. For the experiments, all source code is available at github.com/gatedregularization/code for the tabular experiments and github.com/gatedregularization/ccode for the continuous experiments. A description of hardware and hyperparameters used can be found in Appendix C and in the documentation of the respective repositories. The documentation further contains detailed system information and hash-locked package dependencies as well as Dockerfiles and instructions on how to run the experiments.

ACKNOWLEDGMENTS

Anonymous acknowledgments during double-blind review.

REFERENCES

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An Optimistic Perspective on Offline Reinforcement Learning. *International Conference on Machine Learning*, 2020.

Dimitri P. Bertsekas. *Approximate Dynamic Programming*. Athena Scientific, fourth edition, 2018.

Dimitri P. Bertsekas and John N. Tsitsiklis. Neuro-Dynamic Programming. In *Encyclopedia of Optimization*. Springer US, 2008.

Mohak Bhardwaj, Tengyang Xie, Byron Boots, Nan Jiang, and Ching-An Cheng. Adversarial Model for Offline Reinforcement Learning. *Conference on Neural Information Processing Systems*, 2023.

Edoardo Cetin, Andrea Tirinzoni, Matteo Pirotta, Alessandro Lazaric, Yann Ollivier, and Ahmed Touati. Simple Ingredients for Offline Reinforcement Learning. *International Conference on Machine Learning*, 2024.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized Ensembled Double Q-learning: Learning Fast Without a Model. *International Conference on Learning Representations*, 2021.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially Trained Actor Critic for Offline Reinforcement Learning. *International Conference on Machine Learning*, 2022.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation Matters in Deep RL: A case study on PPO and TRPO. *International Conference on Learning Representations*, 2020.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. *arXiv:2004.07219*, 2021.

Scott Fujimoto and Shixiang Shane Gu. A Minimalist Approach to Offline Reinforcement Learning. *Conference on Neural Information Processing Systems*, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-Policy Deep Reinforcement Learning without Exploration. *International Conference on Machine Learning*, 2019.

Chen-Xiao Gao, Chenyang Wu, Mingjun Cao, Chenjun Xiao, Yang Yu, and Zongzhang Zhang. Behavior-Regularized Diffusion Policy Optimization for Offline Reinforcement Learning. *International Conference on Machine Learning*, 2025.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A Theory of Regularized Markov Decision Processes. *International Conference on Machine Learning*, 2019.

Caglar Gulcehre, Sergio Gómez Colmenarejo, Ziyu Wang, Jakub Sygnowski, Thomas Paine, Konrad Zolna, Yutian Chen, Matthew Hoffman, Razvan Pascanu, and Nando de Freitas. Regularized Behavior Value Estimation. *arXiv:2103.09575*, 2021.

Junyu Guo, Zhi Zheng, Donghao Ying, Ming Jin, Shangding Gu, Costas Spanos, and Javad Lavaei. Don't trade off safety: Diffusion regularization for constrained offline RL. *Conference on Neural Information Processing Systems*, 2025.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Berlin Heidelberg, 2001.

Zhang-Wei Hong, Aviral Kumar, Sathwik Karnik, Abhishek Bhandwaldar, Akash Srivastava, Joni Pajarinen, Romain Laroche, Abhishek Gupta, and Pulkit Agrawal. Beyond Uniform Sampling: Offline Reinforcement Learning with Imbalanced Datasets. *Conference on Neural Information Processing Systems*, 2023.

Shengchao Hu, Ziqing Fan, Chaoqin Huang, Li Shen, Ya Zhang, Yanfeng Wang, and Dacheng Tao. Q-value Regularized Transformer for Offline Reinforcement Learning. *International Conference on Machine Learning*, 2024.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog. *arXiv:1907.00456*, 2019.

Tan Jing, Xiaorui Li, Chao Yao, Xiaojuan Ban, Yuetong Fang, Renjing Xu, and Zhaolin Yuan. Adaptive scaling of policy constraints for offline reinforcement learning. *International Conference on Learning Representations*, 2026.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-Based Offline Reinforcement Learning. *Conference on Neural Information Processing Systems*, 2020.

Byeongchan Kim and Min-hwan Oh. Model-based Offline Reinforcement Learning with Count-based Conservatism. *International Conference on Machine Learning*, 2023.

Alec Koppel, Sujay Bhatt, Jiacheng Guo, Joe Eappen, Mengdi Wang, and Sumitra Ganesh. Information-Directed Pessimism for Offline Reinforcement Learning. *International Conference on Machine Learning*, 2024.

Ilya Kostrikov, Jonathan Tompson, Rob Fergus, and Ofir Nachum. Offline Reinforcement Learning with Fisher Divergence Critic Regularization. *International Conference on Machine Learning*, 2021.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing Off-Policy Q-learning via Bootstrapping Error Reduction. *Conference on Neural Information Processing Systems*, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for Offline Reinforcement Learning. *Conference on Neural Information Processing Systems*, 2020.

Qingfeng Lan, Yangchen Pan, Alona Fyshe, and Martha White. Maxmin Q-learning: Controlling the Estimation Bias of Q-learning. *International Conference on Learning Representations*, 2020.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement Learning: State-of-the-Art*, volume 12. Springer Berlin Heidelberg, 2012.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe Policy Improvement with Baseline Bootstrapping. *International Conference on Machine Learning*, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv:2005.01643*, 2020.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction. *arXiv:2006.03041*, 2021.

Jiachen Li, Edwin Zhang, Ming Yin, Qinxun Bai, Yu-Xiang Wang, and William Yang Wang. Offline Reinforcement Learning with Closed-Form Policy Improvement Operators. *International Conference on Machine Learning*, 2023.

Yun-Hsuan Lien, Ping-Chun Hsieh, Tzu-Mao Li, and Yu-Shuen Wang. Enhancing Value Function Estimation through First-Order State-Action Dynamics in Offline Reinforcement Learning. *International Conference on Machine Learning*, 2024.

Tenglong Liu, Yang Li, Yixing Lan, Hao Gao, Wei Pan, and Xin Xu. Adaptive Advantage-Guided Policy Regularization for Offline Reinforcement Learning. *International Conference on Machine Learning*, 2024.

Qin-Wen Luo, Ming-Kun Xie, Ye-Wen Wang, and Sheng-Jun Huang. Learning to Trust Bellman Updates: Selective State-Adaptive Regularization for Offline RL. *International Conference on Machine Learning*, 2025.

Yi Ma, Jianye Hao, Xiaohan Hu, Yan Zheng, and Chenjun Xiao. Iteratively refined behavior regularization for offline reinforcement learning. *Conference on Neural Information Processing Systems*, 2024.

Yixiu Mao, Hongchang Zhang, Chen Chen, Yi Xu, and Xiangyang Ji. Supported Trust Region Optimization for Offline Reinforcement Learning. *International Conference on Machine Learning*, 2023.

Yixiu Mao, Qi Wang, Chen Chen, Yun Qu, and Xiangyang Ji. Offline Reinforcement Learning with OOD State Correction and OOD Action Suppression. *Conference on Neural Information Processing Systems*, 2024.

Yixiu Mao, Yun Qu, Qi Wang, and Xiangyang Ji. Adaptive Neighborhood-Constrained Q-learning for Offline Reinforcement Learning. *Conference on Neural Information Processing Systems*, 2025.

Arthur Mensch and Mathieu Blondel. Differentiable Dynamic Programming for Structured Prediction and Attention. *International Conference on Machine Learning*, 2018.

Sean Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, first edition, 2022.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. AlgaeDICE: Policy Gradient from Arbitrary Experience. *Optimization Foundations for Reinforcement Learning Workshop at NeurIPS*, 2019.

Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1), 2005.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *arXiv:1212.2015*, 2015.

Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems. *Transactions on Neural Networks and Learning Systems*, 2024.

Aravind Rajeswaran, Igor Mordatch, and Vikash Kumar. A Game Theoretic Framework for Model Based Reinforcement Learning. *International Conference on Machine Learning*, 2020.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism. *Conference on Neural Information Processing Systems*, 2021.

Marc Rigter, Bruno Lacerda, and Nick Hawes. RAMBO-RL: Robust Adversarial Model-Based Offline Reinforcement Learning. *Conference on Neural Information Processing Systems*, 2022.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3), 1951.

Stephane Ross and J Andrew Bagnell. Efficient Reductions for Imitation Learning. *International Conference on Artificial Intelligence and Statistics*, 2010.

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. *International Conference on Artificial Intelligence and Statistics*, 2011.

Takuma Seno and Michita Imai. d3rlpy: An offline deep reinforcement learning library. *Journal of Machine Learning Research*, 2022.

Tobias Sutter, David Sutter, Peyman Mohajerin Esfahani, and John Lygeros. Generalized maximum entropy estimation. *Journal of Machine Learning Research*, 20, 2019.

Richard S. Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the Minimalist Approach to Offline Reinforcement Learning. *Conference on Neural Information Processing Systems*, 2023.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. *International Conference on Intelligent Robots and Systems*, 2012.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A Standard Interface for Reinforcement Learning Environments. *Conference on Neural Information Processing Systems*, 2024.

John N. Tsitsiklis. Asynchronous Stochastic Approximation and Q-learning. *Machine Learning*, 16, 1994.

Marin Vlastelica, Jin Cheng, Georg Martius, and Pavel Kolev. Offline Diversity Maximization Under Imitation Constraints. *Reinforcement Learning Journal*, 3, 2023.

Da Wang, Lin Li, Wei Wei, Qixian Yu, Jianye Hao, and Jiye Liang. Improving Generalization in Offline Reinforcement Learning via Adversarial Data Splitting. *International Conference on Machine Learning*, 2024.

Qingjun Wang, Hongtu Zhou, Hang Yu, Junqiao Zhao, Yanping Zhao, Chen Ye, Ziqiao Wang, and Guang Chen. Beyond Penalization: Diffusion-based out-of-distribution detection and selective regularization in offline reinforcement learning. *International Conference on Learning Representations*, 2026.

Ziyu Wang, Alexander Novikov, Konrad Zolna, Jost Tobias Springenberg, Scott Reed, Bobak Shahriari, Noah Siegel, Josh Merel, Caglar Gulcehre, Nicolas Heess, and Nando de Freitas. Critic Regularized Regression. *Conference on Neural Information Processing Systems*, 2020.

Honghao Wei, Xiyue Peng, Arnob Ghosh, and Xin Liu. Adversarially trained weighted actor-critic for safe offline reinforcement learning. *Conference on Neural Information Processing Systems*, 2024.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior Regularized Offline Reinforcement Learning. *arXiv:1911.11361*, 2019.

Junghyuk Yeom, Yonghyeon Jo, Jungmo Kim, Sanghyeon Lee, and Seungyul Han. Exclusively penalized Q-learning for offline reinforcement learning. *Conference on Neural Information Processing Systems*, 2024.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based Offline Policy Optimization. *Conference on Neural Information Processing Systems*, 2020.

Jing Zhang, Linjiajie Fang, Kexin Shi, Wenjia Wang, and Bing-Yi Jing. Q-distribution guided Q-learning for offline reinforcement learning: Uncertainty penalized Q-value via consistency model. *Conference on Neural Information Processing Systems*, 2024a.

Ruoqi Zhang, Ziwei Luo, Jens Sjölund, Thomas B. Schön, and Per Mattsson. Entropy-regularized diffusion policy with Q-ensembles for offline reinforcement learning. *Conference on Neural Information Processing Systems*, 2024b.

A Auxiliary Results

The following subsections contain referenced results, adjusted to match our notation, that are used or required in the respective sections of the main paper. Throughout this section, let $\mathcal{X} = \mathcal{S} \times \mathcal{A}$, $d = |\mathcal{X}|$, and $\mathbb{1}_d$ be a vector of length d that is filled with ones, and let $\mathbb{I}_d$ be a $d \times d$ identity matrix. ϵ is used as a local accuracy of intermediate results and ε is the final accuracy target of the main theorems. Further, let the absolute value $|\cdot|$ as well as inequalities operate entrywise.

A.1 Problem Setup and Fixed Points

Legendre-Fenchel Transform. Let $\Omega : \Delta(\mathcal{A}) \to \mathbb{R}$ be a strongly convex function. The Legendre-Fenchel transform (or convex conjugate) of Ω is $\Omega^* : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}$ defined as

$$\Omega^*(Q_s) = \max_{\pi \in \Pi} \{\langle \pi_s, Q_s \rangle - \Omega(\pi_s)\}, \quad \forall s \in \mathcal{S}.$$

Based on this we have the following proposition (Geist et al., 2019; Hiriart-Urruty & Lemaréchal, 2001; Mensch & Blondel, 2018).

Proposition A.1 (Geist et al., 2019, Proposition 1). *Let Ω be strongly convex. Then we have the following properties.*

1. *Unique maximizing argument: The maximizer given by*

$$\pi_s^\diamond = \arg\max_{\pi \in \Pi} \{\langle \pi_s, Q_s \rangle - \Omega(\pi_s)\}$$

   is unique and the mapping $Q_s \mapsto \pi_s^\diamond$ is Lipschitz continuous.

2. *Boundedness: If there are constants L_Ω and U_Ω such that for all $\pi \in \Pi$, we have $L_\Omega \leq \Omega(\pi_s) \leq U_\Omega$, then*

$$\max_{a \in \mathcal{A}} Q(s,a) - U_\Omega \leq \Omega^*(Q_s) \leq \max_{a \in \mathcal{A}} Q(s,a) - L_\Omega.$$

3. *Distributivity: For any $z \in \mathbb{R}$ and $\mathbb{1}_{|\mathcal{A}|}$ the vector of ones, we have*

$$\Omega^*\left(Q_s + z\mathbb{1}_{|\mathcal{A}|}\right) = \Omega^*(Q_s) + z.$$

4. *Monotonicity: If $Q_s^{(1)} \leq Q_s^{(2)}$ then $\Omega^*\left(Q_s^{(1)}\right) \leq \Omega^*\left(Q_s^{(2)}\right)$.*

Assume the vector $V \in \mathbb{R}^{|\mathcal{S}|}$ contains the $V(s)$ $\forall s \in \mathcal{S}$ as its elements.

Proposition A.2 (Geist et al., 2019, Proposition 2). *The operator $\mathcal{T}_{\mathcal{R},\pi}$ is affine and has the following properties.*

1. *Monotonicity: Let $V^{(1)}, V^{(2)} \in \mathbb{R}^{|\mathcal{S}|}$ such that $V^{(1)} \geq V^{(2)}$. Then,*

$$\mathcal{T}_{\mathcal{R},\pi} V^{(1)} \geq \mathcal{T}_{\mathcal{R},\pi} V^{(2)} \quad \textit{and} \quad \mathcal{T}_{\mathcal{R}} V^{(1)} \geq \mathcal{T}_{\mathcal{R}} V^{(2)}.$$

2. *Distributivity: For any $z \in \mathbb{R}$ and $\mathbb{1}_{|\mathcal{S}|}$ the vector of ones*

$$\mathcal{T}_{\mathcal{R},\pi}\left(V + z\mathbb{1}_{|\mathcal{S}|}\right) = \mathcal{T}_{\mathcal{R},\pi} V + \gamma z \mathbb{1}_{|\mathcal{S}|} \quad \textit{and} \quad \mathcal{T}_{\mathcal{R}}\left(V + z\mathbb{1}_{|\mathcal{S}|}\right) = \mathcal{T}_{\mathcal{R}} V + \gamma z \mathbb{1}_{|\mathcal{S}|}.$$

3. *Contraction: Both operators are γ-contractions in supremum norm. For any $V^{(1)}, V^{(2)} \in \mathbb{R}^{|\mathcal{S}|}$,*

$$\|\mathcal{T}_{\mathcal{R},\pi} V^{(1)} - \mathcal{T}_{\mathcal{R},\pi} V^{(2)}\|_\infty \leq \gamma \|V^{(1)} - V^{(2)}\|_\infty \quad \textit{and}$$
$$\|\mathcal{T}_{\mathcal{R}} V^{(1)} - \mathcal{T}_{\mathcal{R}} V^{(2)}\|_\infty \leq \gamma \|V^{(1)} - V^{(2)}\|_\infty.$$

A.2 GATED LEARNING

With

$$\theta_k = \begin{bmatrix} \theta_k(1,1) \\ \theta_k(1,2) \\ \vdots \\ \theta_k(s,a) \\ \vdots \\ \theta_k(|\mathcal{S}|,|\mathcal{A}|) \end{bmatrix} \in \mathbb{R}^d \quad \text{and} \quad \|\theta\|_v = \max_{(s,a)} \frac{|\theta(s,a)|}{v(s,a)} \text{ for } v \in \mathbb{R}^d,$$

we can follow Tsitsiklis (1994) and construct the stochastic process $(\theta_k)_{k\geq 0}$ defined as

$$\theta_{k+1}(s,a) = \theta_k(s,a) + \alpha_k(s,a)\left(F(\theta_k)(s,a) - \theta_k(s,a) + \omega_k(s,a)\right), \tag{15}$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$, and $\omega_k(s,a)$ is a noise term based on θ_k and s'_k. In this setting, the following holds.

Assumption A.1 (Tsitsiklis, 1994, Assumption 1). *For each $(s,a) \in \mathcal{X}$ the visitation count $\vartheta_k(s,a)$ at timestep k goes to infinity with probability* 1.

Assumption A.2 (Tsitsiklis, 1994, Assumption 2). *Let $(\mathcal{F}_k)_{k\geq 0}$ be a sequence of increasing σ-algebras containing the entire history up to k, such that θ_k is $\mathcal{F}_k$-measurable.*

a) θ_0 is $\mathcal{F}_0$-measurable.

b) For every (s,a) and k, $\omega_k(s,a)$ is $\mathcal{F}_{k+1}$-measurable.

c) For every (s,a) and k, $\alpha_k(s,a)$ and $\vartheta_k(s,a)$ are $\mathcal{F}_k$-measurable.

d) For every (s,a) and k, $\mathbb{E}[\omega_k(s,a)|\mathcal{F}_k] = 0$.

e) There exist (deterministic) constants c_A and c_B such that

$$\mathbb{E}[\omega_k^2(s,a)|\mathcal{F}_k] \leq c_A + c_B \max_{s,a} \max_{i\leq k} |\theta_i(s,a)|^2,$$
$$\forall \theta_k \in \mathbb{R}^d, \forall (s,a) \in \mathcal{X}, s'_k \in \mathcal{S}, k \in \mathbb{N}.$$

Assumption A.3 (Tsitsiklis, 1994, Assumption 3).

a) $\sum_{k=1}^{\infty} \alpha_k(s,a) = \infty \quad \forall (s,a) \in \mathcal{X}$ with probability 1.

b) $\sum_{k=1}^{\infty} \alpha_k^2(s,a) < \infty \quad \forall (s,a) \in \mathcal{X}$ with probability 1.

Assumption A.4 (Tsitsiklis, 1994, Assumption 5). *The mapping F has a unique fixed point $\theta^\star \in \mathbb{R}^d$ for which there exists a positive vector v and a scalar $\kappa \in [0,1)$, such that*

$$\|F(\theta) - \theta^\star\|_v \leq \kappa \|\theta - \theta^\star\|_v, \quad \forall \theta \in \mathbb{R}^d.$$

Theorem A.1 (Tsitsiklis, 1994, Theorem 3). *Let Assumptions A.1, A.2, A.3, and A.4 hold. Then, θ_k converges to $\theta^\star$ with probability* 1.

The following paragraph establishes a set of definitions and conditions, and also additional notation required for the auxiliary results from Li et al. (2021) adapted to our notation.

Let k_{mix} denote the mixing time of an ergodic Markov state-action chain according to Paulin (2015, Definition 1.3), $\eta_{\min} = \min_{(s,a)\in\mathcal{X}} \eta_{\pi^b}(s,a)$ the smallest occupancy probability of the chain's stationary distribution η_{π^b}, and $K \in \mathbb{N}_+$ the overall number of updates. Define

$$k_{\text{frame}} = \frac{443 k_{\text{mix}}}{\eta_{\min}} \log\left(\frac{4dK}{\delta}\right),$$

$$k_{\text{threshold}} = \max\left\{\frac{2\log\left(\frac{1}{(1-\gamma)^2\epsilon}\right)}{\alpha\eta_{\min}}, k_{\text{frame}}\right\},$$

$$\varrho = (1-\gamma)(1-(1-\alpha)^{\frac{1}{2}\eta_{\min} k_{\text{frame}}}), \quad \text{and}$$
$$l(k) = \max\left\{0, \left\lfloor \frac{k - k_{\text{threshold}}}{k_{\text{frame}}} \right\rfloor\right\}$$

using a constant learning rate $\alpha \in (0,1)$, probability of failure $\delta \in (0, \frac{1}{2})$, and accuracy $\epsilon \in (0, \frac{1}{1-\gamma}]$. With $x_k = (s_k, a_k) \in \mathcal{X}$, let Λ_k be the diagonal $d \times d$ matrix that has zeros everywhere except for the diagonal entry corresponding to (x_k, x_k), which is set to the learning rate, such that $[\Lambda_k]_{x_k,x_k}$ is α_k or the constant learning rate α depending on the setting. Further, let T be the full $d \times |\mathcal{S}|$ transition matrix and T_k be a $d \times |\mathcal{S}|$ matrix that is zero everywhere except for the entry that corresponds to $((s_k, a_k), s'_k)$, which is set to one indicating the transition of sample k, such that $[\mathrm{T}_k]_{x_k,s'_k} = 1$. Finally, let Q be the vector of length d that contains all the Q-values, such that $\Delta_k = \hat{Q}_k - Q^\Phi$, and V the corresponding vector of length $|\mathcal{S}|$ representing the value function.

Lemma A.1 (Li et al., 2021, Lemma 1). *Consider any fixed vector $V^\Phi \in \mathbb{R}^{|\mathcal{S}|}$. Then there exists a universal constant $c > 0$ such that for any $\delta, \alpha \in (0,1)$ such that $\alpha \log \frac{dK}{\delta} \in (0,1)$, one has*

$$\left| \gamma \sum_{i=0}^{k} \left[\prod_{j=i+1}^{k} (\mathbb{I}_d - \Lambda_j) \right] \Lambda_i \left(\mathrm{T}_i - \mathrm{T}\right) V^\Phi \right| \leq c\gamma \sqrt{\alpha \log \frac{dK}{\delta}} \mathbb{1}_d \|V^\Phi\|_\infty$$

for all $k+1 \in [1, K]$, with probability at least $1-\delta$.

Lemma A.2 (Li et al., 2021, Lemma 2). *For any $\delta, \alpha \in (0,1)$, suppose that $K > k_{frame}$. Then with probability at least $1-\delta$ one has*

$$\left| \left[\prod_{i=0}^{k} (\mathbb{I}_d - \Lambda_i) \right] \Delta_0 \right| \leq (1-\alpha)^{\frac{1}{2}(k+1)\eta_{min}} |\Delta_0| \leq (1-\alpha)^{\frac{1}{2}(k+1)\eta_{min}} \mathbb{1}_d \|\Delta_0\|_\infty$$

uniformly over all $k+1 \in [k_{frame}, K]$ and all vectors $\Delta_0 \in \mathbb{R}^d$.

Moreover, in the case where $k+1 < k_{frame}$, the bound becomes

$$\left| \left[\prod_{i=0}^{k} (\mathbb{I}_d - \Lambda_i) \right] \Delta_0 \right| \leq \mathbb{1}_d \|\Delta_0\|_\infty,$$

since $\mathbb{I}_d - \Lambda_i$ is a diagonal non-negative matrix whose entries lie in $[1-\alpha, 1]$.

Lemma A.3 (Li et al., 2021, Lemma 3). *Let the process $(u_k)_{k=0}^K$ be defined as*

$$u_{k+1} = \begin{cases} \gamma \sum_{i=0}^{k} \prod_{j=i+1}^{k} (\mathbb{I}_d - \Lambda_j)\Lambda_i \mathbb{1}_d \|u_i\|_\infty + \mathbb{1}_d \|\Delta_0\|_\infty & \text{for } k+1 \in [1, k_{threshold}] \\ \gamma \sum_{i=0}^{k} \prod_{j=i+1}^{k} (\mathbb{I}_d - \Lambda_j)\Lambda_i \mathbb{1}_d \|u_i\|_\infty & \text{for } k+1 > k_{threshold} \end{cases}$$

with $u_0 = \frac{\Delta_0}{1-\gamma}$. Consider any fixed vector $V^\Phi \in \mathbb{R}^{|\mathcal{S}|}$, some universal constant $c > 0$, and $\delta, \alpha \in (0,1)$ such that $\alpha \log\left(\frac{dK}{\delta}\right) \in (0,1)$. Then with probability at least $1-2\delta$ the elements of $(u_k)_{k=0}^K$ are such that

$$\|\Delta_k\|_\infty \leq c\gamma \sqrt{\alpha \log \frac{dK}{\delta}} \frac{\|V^\Phi\|_\infty}{1-\gamma} + \|u_k\|_\infty + \epsilon$$

is satisfied for all $k \in [0, K]$.

Lemma A.4 (Li et al., 2021, Lemma 4). *Let the sequence $(w_l)_{l \geq 0}$ be defined as*

$$w_l = (1-\varrho)^l \frac{\|\Delta_0\|_\infty}{1-\gamma} = (1-\varrho)^l \frac{\|\hat{Q}_0 - Q^\Phi\|_\infty}{1-\gamma}.$$

For any $\alpha \in (0,1)$, $\delta \in \left(0, \frac{1}{2}\right)$ such that $\alpha \log\left(\frac{dK}{\delta}\right) \in (0,1)$, with probability at least $1-2\delta$ it holds that

$$\|u_k\|_\infty \leq w_{l(k)}$$

for all $k \in [0, K]$.

Theorem A.2 (Li et al., 2021, Theorem 5). *Let the Markov state-action chain induced by π^b be ergodic and its stationary distribution be denoted by η_{π^b}. Further, let the offline set of transitions $\{(s_k, a_k, s'_k)\}_{k=0}^{K-1}$ be sampled from this chain. Consider the Gated Q-learning update rule of Equation* (11) *with constant learning rate α. For any $\alpha \in (0,1)$, $\delta \in (0, \frac{1}{2})$ such that $\alpha \log\left(\frac{dK}{\delta}\right) \in (0,1)$ and any $\epsilon \in (0, \frac{1}{1-\gamma}]$, there exists a universal constant $c > 0$ such that with probability at least $1 - 6\delta$,*

$$\|\hat{Q}_k - Q^\Phi\|_\infty \leq (1-\varrho)^{l(k)} \frac{\|\hat{Q}_0 - Q^\Phi\|_\infty}{1-\gamma} + \frac{c\gamma}{1-\gamma}\|V^\Phi\|_\infty \sqrt{\alpha \log\left(\frac{dK}{\delta}\right)} + \epsilon$$

holds for all $k \leq K$.

The following theorem resembles Theorem 1 of Li et al. (2021) up to the factor 3 in the $\log$ and the requirement of $K \geq 24$.

Theorem A.3. *Let the Markov state-action chain induced by π^b be ergodic and its stationary distribution be denoted by η_{π^b}. Further, let the offline set of transitions $\{(s_k, a_k, s'_k)\}_{k=0}^{K-1}$ be sampled from this chain. Then, for the gated Q-learning update rule of* (11)*, there exist universal constants $c_0, c_1 > 0$, such that for any $\delta \in (0,1)$ and $\varepsilon \in (0, \frac{1}{1-\gamma}]$*

$$\|\hat{Q}_K - Q^\Phi\|_\infty \leq \varepsilon$$

with probability at least $1-\delta$, given that the update count K and the constant learning rate α satisfy $K \geq 24$,

$$K \geq \frac{c_0}{\eta_{min}} \left(\frac{1}{(1-\gamma)^5 \varepsilon^2} + \frac{k_{mix}}{1-\gamma}\right) \log\left(\frac{dK}{\delta}\right) \log\left(\frac{3}{(1-\gamma)^2 \varepsilon}\right), \quad \textit{and}$$

$$\alpha = \frac{c_1}{\log\left(\frac{dK}{\delta}\right)} \min\left\{\frac{(1-\gamma)^4 \varepsilon^2}{\gamma^2}, \frac{1}{k_{mix}}\right\},$$

where $\eta_{min} = \min_{s,a} \eta_{\pi^b}(s,a)$ and k_{mix} is the mixing time of the chain based on Paulin (2015, Definition 1.3).

A.3 Derivation of Equation (13)

The derivation starts with Equation (6), that is

$$(\mathcal{T}_\Phi V)(s) = \max_{\pi \in \Pi} \left\{ \left[\sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V(s') \right) \right] - \beta \Phi(\pi_s^b) \mathcal{R}(\pi_s, \pi_s^b) \right\}. \tag{16}$$

We use the KL divergence as the regularization function

$$\mathcal{R}(\pi_s, \pi_s^b) = D_{KL}(\pi_s \| \pi_s^b) = \sum_{a \in \mathcal{A}} \pi(a|s) \log\left(\frac{\pi(a|s)}{\pi^b(a|s)}\right)$$

as well as a gate of the form

$$\Phi(\pi_s) = 1 - \frac{H(\pi_s)}{\log(|\mathcal{A}|)},$$

where $H(\cdot)$ denotes the Shannon entropy. Since $\Phi(\pi_s^b)$ just scales the regularizing function, and $D_{KL}(\pi_s \| \pi_s^b)$ is strongly convex in the first argument for fixed π_s^b, the entire gated regularization block is strongly convex when $\Phi(\pi_s^b) > 0$. We can now invoke standard results from entropy maximization theory to show that the maximization problem (16) admits a closed-form solution.

Lemma A.5 ((Sutter et al., 2019, Lemma 2))**.** *For any $p \in \Delta(\mathcal{A})$ and $z \in \mathbb{R}^{|\mathcal{A}|}$ the unique optimizer to a problem of the form*

$$\min_{q \in \Delta(\mathcal{A})} \{D_{KL}(q\|p) - \langle q, z \rangle\}$$

is given by the Gibbs distribution

$$q_i^\star = \frac{p_i e^{z_i}}{\sum_{j=1}^{|\mathcal{A}|} p_j e^{z_j}} \quad \textit{for } i = 1, \ldots, |\mathcal{A}|,$$

which leads to the optimal value of $-\log \sum_{i=1}^{|\mathcal{A}|} p_i e^{z_i}$.

Equipped with Lemma A.5, we can formulate the closed-form solution for π that maximizes Equation (6). Rewriting the sum in (6) as

$$\sum_{a\in\mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a)V(s') \right) = \langle \pi_s, Q_s \rangle$$

where $Q_s = r(s,\cdot) + \gamma P_s V$, with $P_s \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{S}|}$ such that $(P_s)_{a,s'} = P(s'|s,a)$, yields

$$\bar{\pi}(\cdot|s) \in \arg\max_{\pi\in\Pi} \left\{ \langle \pi_s, Q_s \rangle - \beta\Phi(\pi_s^b) D_{KL}(\pi_s \| \pi_s^b) \right\} \quad \forall s \in \mathcal{S}.$$

To bring this expression into the form required to apply the Lemma, we flip the signs to transform it into a minimization problem and divide by the scaling factor $\beta\Phi(\pi_s^b)$

$$\bar{\pi}(\cdot|s) \in \arg\min_{\pi\in\Pi} \left\{ D_{KL}(\pi_s \| \pi_s^b) - \left\langle \pi_s, \frac{Q_s}{\beta\Phi(\pi_s^b)} \right\rangle \right\} \quad \forall s \in \mathcal{S}.$$

This form matches the problem from Lemma A.5, and therefore the maximizer of $\bar{\pi}_s$ is given by

$$\bar{\pi}(\cdot|s) = \frac{e^{\frac{Q(s,\cdot)}{\beta\Phi(\pi_s^b)}} \pi^b(\cdot|s)}{\sum_{a'\in\mathcal{A}} e^{\frac{Q(s,a')}{\beta\Phi(\pi_s^b)}} \pi^b(a'|s)}.$$

The same can be applied for the regularized update step without the gate, where we simply replace $\beta\Phi(\pi_s^b)$ with a constant β.

Implementation Remark. To avoid numerical instabilities when the exponential function causes values to become very large, apply the stabilizing trick from Nesterov (2005, Section 5.2)

$$\bar{\pi}(a|s) = \frac{e^{z(s,a)}\pi^b(a|s)}{\sum_{a'\in\mathcal{A}} e^{z(s,a')}\pi^b(a'|s)} \quad \text{with} \quad z(s,a) = \frac{Q(s,a) - \max_{a'\in\mathcal{A}} Q(s,a')}{\beta\Phi(\pi_s^b)}.$$

Finally, replace $\beta\Phi(\pi_s^b)$ with $\beta|\Phi(\pi_s^b)| + \kappa$. For values close to zero, numerical precision may flip the sign of the value returned by the gate which would blow up the exponential function and cause further instabilities, which is mitigated with an absolute value. The tiny constant $\kappa > 0$ avoids division by zero without reducing much of the weight given to the Q-values, in practice not hindering an almost regularization-free greedy policy. An alternative option would be to add a separate case that selects $\bar{\pi}(a|s) = 1$ for $a = \arg\max_{a'\in\mathcal{A}} Q(s,a')$ when $\Phi(\pi_s^b) = 0$.

B PROOFS

Throughout this section, let $\mathcal{X} = \mathcal{S} \times \mathcal{A}$, $d = |\mathcal{X}|$, $\mathbb{1}_d$ be a vector of length d that is filled with ones, and $\mathbb{I}_d$ a $d \times d$ identity matrix. ϵ is used as a local accuracy of intermediate results and ε is the final accuracy target of the main theorems. Further, let the absolute value $|\cdot|$ as well as inequalities operate entrywise.

B.1 PROOFS OF SECTION 2

Proof that $\mathcal{M}_b$ is closed under $\pi \in \Pi_b$. Let $s \in \mathcal{S}_b$, then $\exists k \geq 0$ such that $\mathbb{P}^{\pi^b}[S_k = s | S_0 = s_0] > 0$. Suppose $s' \in \mathcal{S}$ such that $\mathbb{P}^{\pi}[S_1 = s' | S_0 = s] > 0$ for some $\pi \in \Pi_b$, this implies $\exists a \in \mathcal{A}$ for which $\pi(a|s)P(s'|s,a) > 0$, hence $a \in \mathcal{A}_b(s)$ and consequently $\pi^b(a|s) > 0$. Therefore, $\mathbb{P}^{\pi^b}[S_{k+1} = s' | S_0 = s_0] \geq \mathbb{P}^{\pi^b}[S_k = s | S_0 = s_0]\, \pi^b(a|s)P(s'|s,a) > 0$ which implies that $s' \in \mathcal{S}_b$. □

Proof of Geist et al. (2019, Proposition 2) holding for the gated Bellman operator. We follow the steps used in Geist et al. (2019). Let π be any policy with $\pi \in \Pi$. First, rewrite Equation (5)

$$\begin{aligned}
(\mathcal{T}_{\Phi,\pi} V)(s) &= (\mathcal{T}_\pi V)(s) - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) \\
&= \left[\sum_{a\in\mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a)V(s') \right) \right] - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) \\
&= \left[\sum_{a\in\mathcal{A}} \pi(a|s) r(s,a) \right] - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) + \gamma \left[\sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} P(s'|s,a)V(s') \right]
\end{aligned}$$

for any $s \in \mathcal{S}$, which shows that the gated Bellman operator is affine in $V(\cdot)$.

Next, assume $V^{(1)}(\cdot), V^{(2)}(\cdot)$ such that $V^{(1)}(s) \geq V^{(2)}(s)$ for all $s \in \mathcal{S}$, then we have

$$\begin{aligned} V^{(1)}(s) \geq V^{(2)}(s) &\implies \left(\mathcal{T}_\pi V^{(1)}\right)(s) \geq \left(\mathcal{T}_\pi V^{(2)}\right)(s) \\ &\iff \left(\mathcal{T}_\pi V^{(1)}\right)(s) - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) \geq \left(\mathcal{T}_\pi V^{(2)}\right)(s) - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) \\ &\iff \left(\mathcal{T}_{\Phi,\pi} V^{(1)}\right)(s) \geq \left(\mathcal{T}_{\Phi,\pi} V^{(2)}\right)(s), \end{aligned}$$

for all $s \in \mathcal{S}$, which shows monotonicity for the gated Bellman operator. The same applies analogously to the greedy gated Bellman operator since the maximum over policies applied to the monotone $\mathcal{T}_{\Phi,\pi}$ preserves the pointwise inequality, as in

$$\left(\mathcal{T}_\Phi V^{(1)}\right)(s) = \max_{\pi \in \Pi} \left(\mathcal{T}_{\Phi,\pi} V^{(1)}\right)(s) \geq \max_{\pi \in \Pi} \left(\mathcal{T}_{\Phi,\pi} V^{(2)}\right)(s) = \left(\mathcal{T}_\Phi V^{(2)}\right)(s)$$

for all $s \in \mathcal{S}$.

Then, let $z \in \mathbb{R}$, then distributivity for the gated Bellman operator follows from

$$\begin{aligned} \left(\mathcal{T}_{\Phi,\pi}(V + z\mathbb{1}_{|\mathcal{S}|})\right)(s) &= \left(\mathcal{T}_\pi(V + z\mathbb{1}_{|\mathcal{S}|})\right)(s) - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) \\ &= \left[\sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)(V(s') + z)\right)\right] \\ &\quad - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) \\ &= \left[\sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)V(s')\right)\right] \\ &\quad + \gamma z - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) \\ &= \left(\mathcal{T}_\pi V\right)(s) - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) + \gamma z \\ &= \left(\mathcal{T}_{\Phi,\pi} V\right)(s) + \gamma z \end{aligned}$$

for all $s \in \mathcal{S}$. The same applies to the greedy gated Bellman operator since

$$\begin{aligned} \left(\mathcal{T}_\Phi(V + z\mathbb{1}_{|\mathcal{S}|})\right)(s) &= \max_{\pi \in \Pi} \left(\mathcal{T}_{\Phi,\pi}(V + z\mathbb{1}_{|\mathcal{S}|})\right)(s) \\ &= \max_{\pi \in \Pi} \left\{\left(\mathcal{T}_{\Phi,\pi} V\right)(s) + \gamma z\right\} \\ &= \max_{\pi \in \Pi} \left(\mathcal{T}_{\Phi,\pi} V\right)(s) + \gamma z \\ &= \left(\mathcal{T}_\Phi V\right)(s) + \gamma z \end{aligned}$$

for all $s \in \mathcal{S}$.

Finally, assume any $V^{(1)}(\cdot), V^{(2)}(\cdot)$, then

$$\begin{aligned} \left(\mathcal{T}_{\Phi,\pi} V^{(1)}\right)(s) - \left(\mathcal{T}_{\Phi,\pi} V^{(2)}\right)(s) &= \left(\mathcal{T}_\pi V^{(1)}\right)(s) - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) \\ &\quad - \left(\mathcal{T}_\pi V^{(2)}\right)(s) + \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s, \pi_s^b) \\ &= \left(\mathcal{T}_\pi V^{(1)}\right)(s) - \left(\mathcal{T}_\pi V^{(2)}\right)(s) \end{aligned}$$

shows that the gated Bellman operator contracts in the same way as the unregularized Bellman operator and it therefore holds true that

$$\|\mathcal{T}_{\Phi,\pi} V^{(1)} - \mathcal{T}_{\Phi,\pi} V^{(2)}\|_\infty \leq \gamma \|V^{(1)} - V^{(2)}\|_\infty.$$

For the greedy gated Bellman operator, let

$$\|\mathcal{T}_\Phi V^{(1)} - \mathcal{T}_\Phi V^{(2)}\|_\infty = \max_{s\in\mathcal{S}} \left| \left(\mathcal{T}_\Phi V^{(1)}\right)(s) - \left(\mathcal{T}_\Phi V^{(2)}\right)(s) \right|.$$

For any $s \in \mathcal{S}$ where, without loss of generality, it is assumed that $\left(\mathcal{T}_\Phi V^{(1)}\right)(s) \geq \left(\mathcal{T}_\Phi V^{(2)}\right)(s)$, further let

$$\pi^{(1)} \in \arg\max_{\pi\in\Pi} \left(\mathcal{T}_{\Phi,\pi} V^{(1)}\right)(s) \qquad \text{and} \qquad \pi^{(2)} \in \arg\max_{\pi\in\Pi} \left(\mathcal{T}_{\Phi,\pi} V^{(2)}\right)(s),$$

then

$$\begin{aligned}
\left| \left(\mathcal{T}_\Phi V^{(1)}\right)(s) - \left(\mathcal{T}_\Phi V^{(2)}\right)(s) \right| &= \left(\mathcal{T}_\Phi V^{(1)}\right)(s) - \left(\mathcal{T}_\Phi V^{(2)}\right)(s) \\
&= \left(\mathcal{T}_{\Phi,\pi^{(1)}} V^{(1)}\right)(s) - \left(\mathcal{T}_{\Phi,\pi^{(2)}} V^{(2)}\right)(s) \\
&\leq \left(\mathcal{T}_{\Phi,\pi^{(1)}} V^{(1)}\right)(s) - \left(\mathcal{T}_{\Phi,\pi^{(1)}} V^{(2)}\right)(s) \\
&\leq \gamma \|V^{(1)} - V^{(2)}\|_\infty,
\end{aligned}$$

where the first inequality follows from $\pi^{(2)}$ being the maximizer of $\left(\mathcal{T}_{\Phi,\pi} V^{(2)}\right)(s)$ and the second inequality holds due to

$$\begin{aligned}
\left(\mathcal{T}_{\Phi,\pi^{(1)}} V^{(1)}\right)(s) - \left(\mathcal{T}_{\Phi,\pi^{(1)}} V^{(2)}\right)(s) &= \left(\mathcal{T}_{\pi^{(1)}} V^{(1)}\right)(s) - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s^{(1)}, \pi_s^b) \\
&\quad - \left(\mathcal{T}_{\pi^{(1)}} V^{(2)}\right)(s) + \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s^{(1)}, \pi_s^b) \\
&= \left[\sum_{a\in\mathcal{A}} \pi^{(1)}(a|s) \left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) V^{(1)}(s')\right)\right] \\
&\quad - \left[\sum_{a\in\mathcal{A}} \pi^{(1)}(a|s) \left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) V^{(2)}(s')\right)\right] \\
&= \gamma \sum_{a\in\mathcal{A}} \pi^{(1)}(a|s) \sum_{s'\in\mathcal{S}} P(s'|s,a) \left(V^{(1)}(s') - V^{(2)}(s')\right) \\
&\leq \gamma \max_{s'\in\mathcal{S}} |V^{(1)}(s') - V^{(2)}(s')|
\end{aligned}$$

for all $s \in \mathcal{S}$, shows that the greedy gated Bellman operator is also a contraction with respect to the ℓ_∞-norm which concludes the proof. □

Proof of Proposition 2.1. The induction base case is given by

$$\left(\mathcal{T}_\mathcal{R} V\right)(s) = \max_{\pi\in\Pi} \left\{ \left(\mathcal{T}_\pi V\right)(s) - \beta\mathcal{R}(\pi_s, \pi_s^b) \right\} \geq \left(\mathcal{T}_{\pi^b} V\right)(s) - \beta\mathcal{R}(\pi_s^b, \pi_s^b) = \left(\mathcal{T}_{\pi^b} V\right)(s)$$

for all $s \in \mathcal{S}$ and any V, where the inequality holds due to the $\max$ operation in combination with its lower bound attained at π^b due to $\mathcal{R}(\pi_s^b, \pi_s^b) = 0$ following its definition.

The induction step is then taken over k

$$\begin{aligned}
\left(\mathcal{T}_\mathcal{R}^{k+1} V\right)(s) &= \max_{\pi\in\Pi} \left\{ \left(\mathcal{T}_\pi \mathcal{T}_\mathcal{R}^k V\right)(s) - \beta\mathcal{R}(\pi_s, \pi_s^b) \right\} \\
&\geq \max_{\pi\in\Pi} \left\{ \left(\mathcal{T}_\pi \mathcal{T}_{\pi^b}^k V\right)(s) - \beta\mathcal{R}(\pi_s, \pi_s^b) \right\} \\
&\geq \left(\mathcal{T}_{\pi^b} \mathcal{T}_{\pi^b}^k V\right)(s) - \beta\mathcal{R}(\pi_s^b, \pi_s^b) \\
&= \left(\mathcal{T}_{\pi^b}^{k+1} V\right)(s)
\end{aligned}$$

for all $s \in \mathcal{S}$, where the first inequality follows from the induction hypothesis and monotonicity of $\mathcal{T}_\pi$ and the second one from the lower bound on the $\max$ operation as in the base case. □

Proof of Proposition 2.2. The induction base case is given by

$$\begin{aligned}(\mathcal{T}_\Phi V)(s) &= \max_{\pi\in\Pi}\left\{(\mathcal{T}_\pi V)(s) - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s,\pi_s^b)\right\}\\ &\geq \max_{\pi\in\Pi}\left\{(\mathcal{T}_\pi V)(s) - \beta\mathcal{R}(\pi_s,\pi_s^b)\right\}\\ &= (\mathcal{T}_\mathcal{R} V)(s)\end{aligned}$$

for all $s\in\mathcal{S}$ and any V, where the inequality holds due to $\Phi(\pi_s)\in[0,1]$ for any $\pi_s\in\Delta(\mathcal{A})$.

The induction step is then taken over k

$$\begin{aligned}\left(\mathcal{T}_\Phi^{k+1}V\right)(s) &= \max_{\pi\in\Pi}\left\{\left(\mathcal{T}_\pi\mathcal{T}_\Phi^k V\right)(s) - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s,\pi_s^b)\right\}\\ &\geq \max_{\pi\in\Pi}\left\{\left(\mathcal{T}_\pi\mathcal{T}_\mathcal{R}^k V\right)(s) - \beta\Phi(\pi_s^b)\mathcal{R}(\pi_s,\pi_s^b)\right\}\\ &\geq \max_{\pi\in\Pi}\left\{\left(\mathcal{T}_\pi\mathcal{T}_\mathcal{R}^k V\right)(s) - \beta\mathcal{R}(\pi_s,\pi_s^b)\right\}\\ &= \left(\mathcal{T}_\mathcal{R}^{k+1}V\right)(s)\end{aligned}$$

for all $s\in\mathcal{S}$, where the first inequality follows from the induction hypothesis and monotonicity of $\mathcal{T}$ and the second one from $\Phi(\pi_s)\in[0,1]$ for any $\pi_s\in\Delta(\mathcal{A})$ as in the base case. □

Proof of Theorem 2.1. Inequality

$$V^\mathcal{R}(s)\leq V^\Phi(s)\quad\forall s\in\mathcal{S}$$

holds due to Proposition 2.2 and using the contraction properties of both operators when taking the limits $V^\mathcal{R}(s)=\lim_{k\to\infty}\left(\mathcal{T}_\mathcal{R}^k V\right)(s)\leq\lim_{k\to\infty}\left(\mathcal{T}_\Phi^k V\right)(s)=V^\Phi(s)$. Since $\Omega_\Phi(\pi_s)\geq 0$ for any $\pi_s\in\Delta(\mathcal{A})$, we get the base case

$$(\mathcal{T}_\Phi V)(s)=\max_{\pi\in\Pi}\left\{(\mathcal{T}_\pi V)(s)-\Omega_\Phi(\pi_s)\right\}\leq\max_{\pi\in\Pi}(\mathcal{T}_\pi V)(s)=(\mathcal{T}V)(s)$$

for all $s\in\mathcal{S}$ and any V. The induction step is then taken over k

$$\left(\mathcal{T}_\Phi^{k+1}V\right)(s)\leq\left(\mathcal{T}\mathcal{T}_\Phi^k V\right)(s)\leq\left(\mathcal{T}\mathcal{T}^k V\right)(s)=\left(\mathcal{T}^{k+1}V\right)(s)$$

for all $s\in\mathcal{S}$, where the first inequality is due to the base case and the second inequality follows from the induction hypothesis and monotonicity of $\mathcal{T}$. Then, using the contraction properties of both operators and taking the limits yields

$$V^\Phi(s)=\lim_{k\to\infty}\left(\mathcal{T}_\Phi^k V\right)(s)\leq\lim_{k\to\infty}\left(\mathcal{T}^k V\right)(s)=V^\star(s),$$

which concludes the proof. □

B.2 Proofs of Section 3

This section requires the Markov state-action chain induced by the MDP and behavioral policy π^b to be ergodic. We can assume full action support of π^b across the MDP since we can use the behavioral policy restricted MDP $\mathcal{M}_b$ or assume the MDP to be a behavioral policy restriction of another MDP such that full action support by π^b is given. Since the state-action space of the Markov state-action chain is finite, we only need to assume it to consist of a single communicating class in order to satisfy irreducibility and positive recurrence. Aperiodicity is then ensured by a single state-action pair with a self-loop of positive probability.

Remark B.1. *If π^b instead induces a unichain, we can further refine our analysis to just consider the single communicating class obtained from that unichain and omit the transient state-action pairs. The ORL agent is unable to learn meaningful representations in the transient states, so trying to improve upon the behavioral policy can only be done for states inside the positive recurrent class.*

Remark B.2. *If the underlying environment is episodic, we can construct a continuing environment by assuming the terminal states transition into the initial state on every action. This allows us to record state-action trajectories as if following along a continuing Markov state-action chain. It is important to ignore the resetting transitions during data collection, or not use them for updates, to ensure no values pass through this imaginary transition.*

Proof of Theorem 3.1. To show that the update scheme (11) converges to the Φ-optimal Q-values, we adapt the method used in Tsitsiklis (1994). The update rule (11) can be written as a stochastic process of the form of Equation (15) by setting $\theta_k(s,a) = \hat{Q}_k(s,a)$, each element of F to

$$F(\theta)(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{\pi \in \Pi} \left\{ \langle \pi_{s'}, \theta(s', \cdot) \rangle - \Omega_\Phi(\pi_{s'}) \right\},$$

and $\omega_k(s,a) = U_k - Z_k(s,a)$, where

$$U_k = \max_{\pi \in \Pi} \left\{ \gamma \left\langle \pi_{s'_k}, \theta_k(s'_k, \cdot) \right\rangle - \gamma \Omega_\Phi(\pi_{s'_k}) \right\} \text{ and}$$

$$Z_k(s,a) = \max_{\pi \in \Pi} \left\{ \sum_{s' \in \mathcal{S}} P(s'|s,a) \left(\gamma \langle \pi_{s'}, \theta_k(s', \cdot) \rangle - \gamma \Omega_\Phi(\pi_{s'}) \right) \right\},$$

with $\theta_k(s,\cdot) = \begin{bmatrix} \theta_k(s,1) & \cdots & \theta_k(s,a) & \cdots & \theta_k(s,|\mathcal{A}|) \end{bmatrix}^\top$. To apply the convergence Theorem A.1, Assumptions A.1, A.2, A.3, and A.4 must hold. Assuming an ergodic Markov state-action chain and that we can choose the sequence $(\alpha_k(s,a))_{k \geq 0}$, Assumptions A.1 and A.3 trivially hold. If we let $(\mathcal{F}_k)_{k \geq 0}$ be a sequence of increasing σ-algebras containing the entire history up to k, such that $\alpha_k(s,a)$, θ_k are $\mathcal{F}_k$-measurable and $\omega_k(s,a)$ is $\mathcal{F}_{k+1}$-measurable, since U_k contains $s'_k = s_{k+1}$, then Assumptions A.2 a)-c) also hold. To show that Assumption A.2 d) holds, we use the fact that π_s can be chosen independently from $\pi_{s'}$ for all $s, s' \in \mathcal{S}, s \neq s'$. Therefore, it does not matter if we choose the maximizing $\pi \in \Pi$ before taking the expectation or if we update the π_s individually for each state

$$\begin{aligned} Z_k(s,a) &= \max_{\pi \in \Pi} \left\{ \sum_{s' \in \mathcal{S}} P(s'|s,a) \left[\gamma \langle \pi_{s'}, \theta_k(s', \cdot) \rangle - \gamma \Omega_\Phi(\pi_{s'}) \right] \right\} \\ &= \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{\pi \in \Pi} \left\{ \gamma \langle \pi_{s'}, \theta_k(s', \cdot) \rangle - \gamma \Omega_\Phi(\pi_{s'}) \right\}. \end{aligned}$$

Since the transition $s' \sim P(\cdot|s,a)$ only depends on (s,a) it does not matter if we have the entire history, captured by $\mathcal{F}_k$, available or just the last state-action pair (s,a). Therefore, $\mathbb{P}[S'_k = s' | S_k, A_k] = \mathbb{P}[S'_k = s' | \mathcal{F}_k]$ and

$$\begin{aligned} Z_k(s,a) &= \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{\pi \in \Pi} \left\{ \gamma \langle \pi_{s'}, \theta_k(s', \cdot) \rangle - \gamma \Omega_\Phi(\pi_{s'}) \right\} \\ &= \mathbb{E}\left[U_k | S_k = s, A_k = a \right] \\ &= \mathbb{E}\left[U_k | \mathcal{F}_k \right], \end{aligned}$$

which ensures that $\mathbb{E}\left[\omega_k(s,a) | \mathcal{F}_k\right] = 0$. Since $\theta_0 = \hat{Q}_0 \in \mathbb{R}^d$ and

$$\gamma \left\langle \pi^b_{s'}, \theta_k(s', \cdot) \right\rangle \leq \max_{\pi \in \Pi} \left\{ \gamma \langle \pi_{s'}, \theta_k(s', \cdot) \rangle - \gamma \Omega_\Phi(\pi_{s'}) \right\} \leq \gamma \max_{a \in \mathcal{A}} \theta_k(s', a), \tag{17}$$

for any $s' \in \mathcal{S}$, where the lower bound follows from $\Omega_\Phi(\pi^b_{s'}) = 0$ for any $s' \in \mathcal{S}$ and the upper bound follows from $\Omega_\Phi(\pi_{s'}) \geq 0$ for any $s' \in \mathcal{S}$ and $\pi \in \Pi$, as well as $r \in [0,1]$ it is guaranteed that $\left(\hat{\Gamma}_\Phi \hat{Q}_k\right)(s_k, a_k, s'_k) \in \mathbb{R}$ and consequently $\theta_{k+1} \in \mathbb{R}^d$ whenever $\theta_k \in \mathbb{R}^d$. This yields $\theta_k \in \mathbb{R}^d$ for all $k \geq 0$ by induction. Further we have that the absolute values of the upper and lower bound in (17) are themselves bounded by $\gamma \|\theta_k\|_\infty$, which implies $|U_k| \leq \gamma \|\theta_k\|_\infty$. The bound on $|U_k|$ then implies $|Z_k(s,a)| \leq \gamma \|\theta_k\|_\infty$, since $Z_k(s,a)$ is the expected value of U_k.

Therefore,

$$\omega_k^2(s,a) = (U_k - Z_k(s,a))^2 \leq 4\gamma^2 \|\theta_k\|_\infty^2 \leq 4\gamma^2 \max_{(s',a')} \max_{i \leq k} \theta_i^2(s',a') \quad \forall (s,a) \in \mathcal{X}.$$

Since the right-hand side is $\mathcal{F}_k$-measurable, taking the conditional expectation given $\mathcal{F}_k$ leaves it unchanged, and Assumption A.2 e) is satisfied with $c_A = 0$ and $c_B = 4\gamma^2$. Finally, by using the results from Section 2.2, we know that $F(\theta)$, which is the greedy gated Bellman operator for Q-values, is a γ-contraction and Assumption A.4 holds. Convergence of the single sample approximation update rule (11) can therefore be inferred based on Theorem A.1. $\square$

Proof that the Learning Rate of Theorem 3.2 satisfies the Robbins-Monro conditions. Let $\vartheta_k(x)$ be the visitation count of state-action pair x up to timestep k. Then the learning rate

$$\alpha_k = \min\left\{1, c_\alpha \exp\left(\left\lfloor \log\left(\frac{\log(k)}{\hat{\eta}_{\min,k}(1-\gamma)\gamma^2 k}\right)\right\rfloor\right)\right\}$$

with $\alpha_0 = \alpha_1 = 1$ and

$$\hat{\eta}_{\min,k} = \begin{cases} \frac{1}{d} & \text{if} \quad \min_{x\in\mathcal{X}} \vartheta_k(x) = 0, \\ \hat{\eta}_{\min,k-1} & \text{if} \quad \frac{1}{3} < \frac{\frac{1}{k}\min_{x\in\mathcal{X}} \vartheta_k(x)}{\hat{\eta}_{\min,k-1}} < 3, \quad \text{and} \\ \frac{1}{k}\min_{x\in\mathcal{X}} \vartheta_k(x) & \text{otherwise} \end{cases}$$

satisfies the Robbins-Monro conditions (Robbins & Monro, 1951) with probability at least $1-\delta$.

Following Li et al. (2021, Equation (67)), there exists a $c_{\hat{\eta}} \in [\frac{1}{6}, \frac{9}{2}]$ such that

$$\hat{\eta}_{\min,k} = c_{\hat{\eta}}\eta_{\min} \qquad \text{for} \quad K \geq k \geq \frac{443 k_{\text{mix}} \log\left(\frac{4dk}{\delta}\right)}{\eta_{\min}} \tag{18}$$

with probability at least $1-\delta$. Since the right side of that condition on k grows in the logarithm of k, there exists a threshold value $\bar{k}$, that satisfies $\bar{k} = 443 k_{\text{mix}} \log\left(\frac{4d\bar{k}}{\delta}\right)\eta_{\min}^{-1}$, such that all $k \geq \bar{k}$ satisfy the requirement of (18) and we get

$$\frac{1}{6}\eta_{\min} \leq \hat{\eta}_{\min,k} \leq \frac{9}{2}\eta_{\min} \quad \forall k \geq \bar{k}$$

with probability at least $1-\delta$. The next steps will use these two bounds on $\hat{\eta}_{\min,k}$ to evaluate the sum of a lower bound on α_k and the sum of a squared upper bound on α_k, to check the two Robbins-Monro conditions.

Since $\lfloor z \rfloor > z - 1$ for all $z \in \mathbb{R}$, we get that for all $k \geq \bar{k}$

$$\begin{aligned} c_\alpha \exp\left(\left\lfloor \log\left(\frac{\log(k)}{\hat{\eta}_{\min,k}(1-\gamma)\gamma^2 k}\right)\right\rfloor\right) &\geq c_\alpha \exp\left(\left\lfloor \log\left(\frac{\log(k)}{\frac{9}{2}\eta_{\min}(1-\gamma)\gamma^2 k}\right)\right\rfloor\right) \\ &> c_\alpha \exp\left(\log\left(\frac{\log(k)}{\frac{9}{2}\eta_{\min}(1-\gamma)\gamma^2 k}\right) - 1\right) \\ &= \frac{2c_\alpha}{9e\eta_{\min}(1-\gamma)\gamma^2}\frac{\log(k)}{k}. \end{aligned}$$

Ignoring the scaling constant for brevity, we observe that for all $k > \exp(1)$, which is satisfied for $\bar{k}$ since $k_{\text{mix}} \geq 1$, $\eta_{\min} \leq 1$, and $\log\left(\frac{4d\bar{k}}{\delta}\right) > 1$, given when $d > 1$ and $k > 1$, the sequence $\left(\frac{\log(k)}{k}\right)_{k\geq\bar{k}}$ is elementwise lower bounded by the harmonic sequence $\left(\frac{1}{k}\right)_{k\geq\bar{k}}$, which has a diverging sum, and therefore

$$\sum_{k=\bar{k}}^{\infty}\frac{\log(k)}{k} \geq \sum_{k=\bar{k}}^{\infty}\frac{1}{k} = \infty.$$

As a result, we have that

$$\sum_{k=1}^{\infty}\frac{2c_\alpha}{9e\eta_{\min}(1-\gamma)\gamma^2}\frac{\log(k)}{k} = \infty$$

since the tail of the sum diverges to infinity and the first $\bar{k}-1$ elements of the sum are finite. This sum is a lower bound on the sum over α_k and therefore implies that

$$\sum_{k=1}^{\infty} c_\alpha \exp\left(\left\lfloor \log\left(\frac{\log(k)}{\hat{\eta}_{\min,k}(1-\gamma)\gamma^2 k}\right)\right\rfloor\right) = \infty \quad \implies \sum_{k=1}^{\infty}\alpha_k = \infty$$

with probability at least $1-\delta$.

Since $\lfloor z \rfloor \leq z$ for all $z \in \mathbb{R}$, we get that for all $k \geq \bar{k}$

$$\begin{aligned} c_\alpha \exp\left(\left\lfloor \log\left(\frac{\log(k)}{\hat{\eta}_{\min,k}(1-\gamma)\gamma^2 k}\right)\right\rfloor\right) &\leq c_\alpha \exp\left(\left\lfloor \log\left(\frac{\log(k)}{\frac{1}{6}\eta_{\min}(1-\gamma)\gamma^2 k}\right)\right\rfloor\right) \\ &\leq c_\alpha \exp\left(\log\left(\frac{\log(k)}{\frac{1}{6}\eta_{\min}(1-\gamma)\gamma^2 k}\right)\right) \\ &= \frac{6c_\alpha}{\eta_{\min}(1-\gamma)\gamma^2}\frac{\log(k)}{k}. \end{aligned}$$

Ignoring the scaling constant for brevity, we use

$$\log(z) = \log(z) - \log(1) = \int_1^z \frac{1}{t}\mathrm{d}t \leq \int_1^z \frac{1}{t^{1-\kappa}}\mathrm{d}t = \int_1^z \frac{t^\kappa}{t}\mathrm{d}t = \frac{z^\kappa - 1}{\kappa} < \frac{z^\kappa}{\kappa}$$

that holds for any $\kappa > 0$ and $z \in \mathbb{R}_+$, to obtain

$$\frac{\log(k)^2}{k^2} < \frac{\left(\frac{k^\kappa}{\kappa}\right)^2}{k^2} = \frac{k^{2\kappa}}{\kappa^2 k^2} = \frac{1}{\kappa^2 k^{2(1-\kappa)}}.$$

For any $\kappa \in (0, \frac{1}{2})$

$$\sum_{k=1}^{\infty} \frac{1}{\kappa^2 k^{2(1-\kappa)}}$$

is a p-series that converges with $\mathrm{p} = 2(1-\kappa) > 1$. From

$$\sum_{k=1}^{\infty} \frac{\log(k)^2}{k^2} < \sum_{k=1}^{\infty} \frac{1}{\kappa^2 k^{2(1-\kappa)}} < \infty$$

we get

$$\sum_{k=1}^{\infty} \left(\frac{6c_\alpha}{\eta_{\min}(1-\gamma)\gamma^2}\frac{\log(k)}{k}\right)^2 < \infty$$

since the tail of the sum converges to a finite value and the first $\bar{k} - 1$ elements of the sum are finite. This sum is an upper bound on the sum over α_k^2 and therefore implies that

$$\sum_{k=1}^{\infty} \left(c_\alpha \exp\left(\left\lfloor \log\left(\frac{\log(k)}{\hat{\eta}_{\min,k}(1-\gamma)\gamma^2 k}\right)\right\rfloor\right)\right)^2 < \infty \quad \implies \sum_{k=1}^{\infty} (\alpha_k)^2 < \infty,$$

which proves that the learning rate satisfies the Robbins-Monro conditions (Robbins & Monro, 1951) as long as the lower bound on $\hat{\eta}_{\min,k}$ holds, which it does with probability at least $1 - \delta$.

□

Proof of Theorem 3.2. The proof adapts the proof of Li et al. (2021, Theorem 3). To that end we first show Theorem A.2, based on which we can show Theorem A.3, which then allows us to prove Theorem 3.2. As such, we begin with assuming a constant learning rate α. In order to apply the analysis, the Markov state-action chain induced by the behavioral policy π^b on the MDP must be time-homogeneous and uniformly ergodic. Time-homogeneity is given by the behavioral policy π^b as well as the underlying MDP being constant, which ensures that the induced Markov state-action chain is time-homogeneous. The Markov state-action chain is assumed to be ergodic based on the arguments given in the first paragraph of Section B.2. As a result of the Markov state-action chain being finite and ergodic, it is uniformly ergodic based on Paulin (2015).

Under these conditions, there exists a stationary distribution induced by the behavioral policy π^b, denoted by η_{π^b}, with the smallest occupancy probability being $\eta_{\min} = \min_{(s,a)} \eta_{\pi^b}(s,a)$. Let the k-step transition probability from state $y \in \mathcal{X}$ to state $x \in \mathcal{X}$ be $P^k_{\pi^b}(x|y) = \mathbb{P}_{P,\pi^b}[X_k = x | X_0 = y]$. We can then define the mixing time following Paulin (2015, Definition 1.3), such that

$$k_{\text{mix}} = \min\left\{k \middle| \max_{x \in \mathcal{X}} \mathrm{d}_{\text{TV}}\left(P^k_{\pi^b}(\cdot|x), \eta_{\pi^b}\right) \leq \frac{1}{4}\right\},$$

where $\mathrm{d_{TV}}$ is the total variation distance.

Note that,

$$0 \leq V^{\Phi}(s) \leq V^{\star}(s) \leq \frac{1}{1-\gamma} \qquad \forall s \in \mathcal{S} \tag{19}$$

where the first and the second inequality hold due to $r \in [0,1]$, Theorem 2.1, and $V^b \leq V^{\mathcal{R}}$ and the last inequality holds due to $r \in [0,1]$.

Further, let Q be the vector of length d that contains all the Q-values

$$Q = \begin{bmatrix} Q(1,1) & Q(1,2) & \cdots & Q(1,|\mathcal{A}|) & Q(2,1) & Q(2,2) & \cdots & Q(|\mathcal{S}|,|\mathcal{A}|) \end{bmatrix}^{\top},$$

and V be the corresponding value vector of length $|\mathcal{S}|$

$$V = \begin{bmatrix} V(1) & V(2) & \cdots & V(|\mathcal{S}|) \end{bmatrix}^{\top}.$$

The gated Q-learning update rule, given by equations (10) and (11) and used with a constant learning rate α as required by Theorem A.3 and Theorem A.2

$$\hat{\Gamma}_{\Phi} Q(s,a,s') = r(s,a) + \gamma \max_{\pi \in \Pi} \{\langle \pi_{s'}, Q_{s'} \rangle - \Omega_{\Phi}(\pi_{s'})\}$$

$$\hat{Q}_{k+1}(s_k,a_k) = (1-\alpha)\hat{Q}_k(s_k,a_k) + \alpha \hat{\Gamma}_{\Phi} \hat{Q}_k(s_k,a_k,s'_k)$$

where $\Omega_{\Phi}(\pi_{s'}) = \beta \Phi(\pi^b_{s'}) \mathcal{R}(\pi_{s'}, \pi^b_{s'})$, can then be written in matrix form

$$\hat{Q}_{k+1} = (\mathbb{I}_d - \Lambda_k)\hat{Q}_k + \Lambda_k \left(r + \gamma \mathrm{T}_k \max_{\pi \in \Pi} \left\{ \mathrm{D}(\pi) \hat{Q}_k - \Omega_{\Phi}(\pi) \right\} \right)$$

by using the fact that π_s can be chosen independently from $\pi_{s'}$ for all $s, s' \in \mathcal{S}, s \neq s'$ and the following objects. Let $x_k = (s_k, a_k) \in \mathcal{X}$, Λ_k is a sparse diagonal $d \times d$ matrix that has zeros everywhere except for the diagonal entry corresponding to (x_k, x_k), which is set to the learning rate, such that $[\Lambda_k]_{x_k,x_k}$ is α_k or the constant learning rate α depending on the setting. T_k is a $d \times |\mathcal{S}|$ matrix that is zero everywhere except for the entry that corresponds to $((s_k,a_k),s'_k)$, which is set to one indicating the transition of sample k, such that $[\mathrm{T}_k]_{x_k,s'_k} = 1$. $\mathrm{D}(\pi)$ is an $|\mathcal{S}| \times d$ block matrix that has the policy vectors for each state as its diagonal elements

$$\mathrm{D}(\pi) = \begin{bmatrix} \pi_1^{\top} & 0 & \cdots & 0 \\ 0 & \pi_2^{\top} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_{|\mathcal{S}|}^{\top} \end{bmatrix}.$$

And finally, $\Omega_{\Phi}(\pi) = \begin{bmatrix} \Omega_{\Phi}(\pi_1) & \cdots & \Omega_{\Phi}(\pi_{|\mathcal{S}|}) \end{bmatrix}^{\top}$ is a vector of length $|\mathcal{S}|$.

As the first step, we show that Theorem A.2 applies for gated Q-learning.

Proof of Theorem A.2. The above allows us to write the element-wise error at update step $k+1$ as

$$\Delta_{k+1} = \hat{Q}_{k+1} - Q^{\Phi} = (\mathbb{I}_d - \Lambda_k)\hat{Q}_k + \Lambda_k \left(r + \gamma \mathrm{T}_k \max_{\pi \in \Pi} \left\{ \mathrm{D}(\pi) \hat{Q}_k - \Omega_{\Phi}(\pi) \right\} \right) - Q^{\Phi}$$

$$= (\mathbb{I}_d - \Lambda_k)(\hat{Q}_k - Q^{\Phi}) + \Lambda_k \left(r + \gamma \mathrm{T}_k \max_{\pi \in \Pi} \left\{ \mathrm{D}(\pi) \hat{Q}_k - \Omega_{\Phi}(\pi) \right\} - Q^{\Phi} \right).$$

Plugging in

$$Q^{\Phi} = r + \gamma \mathrm{T} \max_{\pi \in \Pi} \left\{ \mathrm{D}(\pi) Q^{\Phi} - \Omega_{\Phi}(\pi) \right\} = r + \gamma \mathrm{T} V^{\Phi}, \tag{20}$$

where T is the complete $d \times |\mathcal{S}|$ transition matrix for all state-action pairs, and using

$$\max_{\pi \in \Pi} \left\{ \mathrm{D}(\pi) \hat{Q}_k - \Omega_{\Phi}(\pi) \right\} = \hat{V}_k$$

yields

$$\begin{aligned}
\Delta_{k+1} = \hat{Q}_{k+1} - Q^{\Phi} &= (\mathbb{I}_d - \Lambda_k)(\hat{Q}_k - Q^{\Phi}) + \Lambda_k \left(r + \gamma \mathrm{T}_k \max_{\pi \in \Pi} \left\{ \mathrm{D}(\pi) \hat{Q}_k - \Omega_{\Phi}(\pi) \right\} - Q^{\Phi} \right) \\
&= (\mathbb{I}_d - \Lambda_k)(\hat{Q}_k - Q^{\Phi}) + \Lambda_k \left(r + \gamma \mathrm{T}_k \hat{V}_k - r - \gamma \mathrm{T} V^{\Phi} \right) \\
&= (\mathbb{I}_d - \Lambda_k)\Delta_k + \gamma \Lambda_k \left(\mathrm{T}_k \hat{V}_k - \mathrm{T} V^{\Phi} \right) \\
&= (\mathbb{I}_d - \Lambda_k)\Delta_k + \gamma \Lambda_k \left(\mathrm{T}_k \hat{V}_k - \mathrm{T} V^{\Phi} + \mathrm{T}_k V^{\Phi} - \mathrm{T}_k V^{\Phi} \right) \\
&= (\mathbb{I}_d - \Lambda_k)\Delta_k + \gamma \Lambda_k \left(\mathrm{T}_k - \mathrm{T} \right) V^{\Phi} + \gamma \Lambda_k \mathrm{T}_k \left(\hat{V}_k - V^{\Phi} \right).
\end{aligned}$$

Unrolling the recursion by plugging in the definition of Δ_k, and distributing the terms gives

$$\begin{aligned}
\Delta_{k+1} = (\mathbb{I}_d - \Lambda_k) &\left((\mathbb{I}_d - \Lambda_{k-1})\Delta_{k-1} + \gamma \Lambda_{k-1} \left(\mathrm{T}_{k-1} - \mathrm{T} \right) V^{\Phi} + \gamma \Lambda_{k-1} \mathrm{T}_{k-1} \left(\hat{V}_{k-1} - V^{\Phi} \right) \right) \\
&+ \gamma \Lambda_k \left(\mathrm{T}_k - \mathrm{T} \right) V^{\Phi} \\
&+ \gamma \Lambda_k \mathrm{T}_k \left(\hat{V}_k - V^{\Phi} \right) \\
= (\mathbb{I}_d - \Lambda_k)&(\mathbb{I}_d - \Lambda_{k-1})\Delta_{k-1} \\
&+ \gamma \Lambda_k \left(\mathrm{T}_k - \mathrm{T} \right) V^{\Phi} + (\mathbb{I}_d - \Lambda_k)\gamma \Lambda_{k-1} \left(\mathrm{T}_{k-1} - \mathrm{T} \right) V^{\Phi} \\
&+ \gamma \Lambda_k \mathrm{T}_k \left(\hat{V}_k - V^{\Phi} \right) + (\mathbb{I}_d - \Lambda_k)\gamma \Lambda_{k-1} \mathrm{T}_{k-1} \left(\hat{V}_{k-1} - V^{\Phi} \right) \\
&\vdots \\
= &\left[\prod_{i=0}^{k} (\mathbb{I}_d - \Lambda_i) \right] \Delta_0 \\
&+ \gamma \left[\sum_{i=0}^{k} \left[\prod_{j=i+1}^{k} (\mathbb{I}_d - \Lambda_j) \right] \Lambda_i \left(\mathrm{T}_i - \mathrm{T} \right) V^{\Phi} \right] \\
&+ \gamma \left[\sum_{i=0}^{k} \left[\prod_{j=i+1}^{k} (\mathbb{I}_d - \Lambda_j) \right] \Lambda_i \mathrm{T}_i \left(\hat{V}_i - V^{\Phi} \right) \right].
\end{aligned}$$

Let us denote the three elements of this sum as

$$\begin{aligned}
\Psi_{k+1}^{(1)} &= \left[\prod_{i=0}^{k} (\mathbb{I}_d - \Lambda_i) \right] \Delta_0 \\
\Psi_{k+1}^{(2)} &= \gamma \sum_{i=0}^{k} \left[\prod_{j=i+1}^{k} (\mathbb{I}_d - \Lambda_j) \right] \Lambda_i \left(\mathrm{T}_i - \mathrm{T} \right) V^{\Phi} \\
\Psi_{k+1}^{(3)} &= \gamma \sum_{i=0}^{k} \left[\prod_{j=i+1}^{k} (\mathbb{I}_d - \Lambda_j) \right] \Lambda_i \mathrm{T}_i \left(\hat{V}_i - V^{\Phi} \right),
\end{aligned}$$

such that

$$\Delta_{k+1} = \Psi_{k+1}^{(1)} + \Psi_{k+1}^{(2)} + \Psi_{k+1}^{(3)}. \tag{21}$$

The first element $\Psi_{k+1}^{(1)}$ is the decaying vector of differences resulting from the Q-value initialization. As such, the dynamics of this term only depend on the sampling process of the Markov state-action chain as well as the learning rate α. Using the same assumptions on the Markov state-action chain as well as the learning rate is sufficient for Lemma A.2 to hold.

The second element $\Psi_{k+1}^{(2)}$ is the stochastic noise that captures the differences in realized transitions from the true transition probabilities T. Consequently, this term also depends only on the conditions of the Markov state-action chain as well as the learning rate sequence. Using the same assumptions on the Markov state-action chain, the learning rate, and $\|V^{\Phi}\|_{\infty} \leq \frac{1}{1-\gamma}$, which is given by (19), is sufficient for Lemma A.1 to hold.

This also enables us to use the same values for k_{frame} and $k_{\text{threshold}}$ as Li et al. (2021, Equation (37a) and Equation (37b)),

$$k_{\text{frame}} = \frac{443 k_{\text{mix}}}{\eta_{\min}} \log\left(\frac{4dK}{\delta}\right)$$

$$k_{\text{threshold}} = \max\left\{\frac{2\log\left(\frac{1}{(1-\gamma)^2\epsilon}\right)}{\alpha\eta_{\min}}, k_{\text{frame}}\right\},$$

with $\delta \in (0,1)$ and K being the total number of update steps.

For the last element $\Psi_{k+1}^{(3)}$, which is the propagation of the value function iterates through the recursion, we have that

$$\|\mathrm{T}_i\left(\hat{V}_i - V^{\Phi}\right)\|_{\infty} \leq \|\hat{V}_i - V^{\Phi}\|_{\infty} \leq \|\hat{Q}_i - Q^{\Phi}\|_{\infty} = \|\Delta_i\|_{\infty}$$

where the first inequality is due to $\mathrm{T}_i z$ selecting a single element from the vector $z \in \mathbb{R}^{|\mathcal{S}|}$ and therefore $\|\mathrm{T}_i z\|_{\infty} \leq \|z\|_{\infty}$ and the second inequality is due to

$$\begin{aligned}
\|\hat{V}_i - V^{\Phi}\|_{\infty} &= \left\|\max_{\pi\in\Pi}\left\{\mathrm{D}(\pi)\hat{Q}_i - \Omega_{\Phi}(\pi)\right\} - \max_{\pi\in\Pi}\left\{\mathrm{D}(\pi)Q^{\Phi} - \Omega_{\Phi}(\pi)\right\}\right\|_{\infty} \\
&\leq \left\|\max_{\pi\in\Pi}\left\{\left|\mathrm{D}(\pi)\hat{Q}_i - \Omega_{\Phi}(\pi) - \mathrm{D}(\pi)Q^{\Phi} + \Omega_{\Phi}(\pi)\right|\right\}\right\|_{\infty} \\
&= \left\|\max_{\pi\in\Pi}\left\{\left|\mathrm{D}(\pi)\left(\hat{Q}_i - Q^{\Phi}\right)\right|\right\}\right\|_{\infty} \\
&\leq \max_{(s,a)\in\mathcal{X}}\left\{\left|\hat{Q}_i(s,a) - Q^{\Phi}(s,a)\right|\right\} \\
&= \left\|\hat{Q}_i - Q^{\Phi}\right\|_{\infty},
\end{aligned}$$

that holds because of $\|\mathrm{D}(\pi)z\|_{\infty} \leq \|z\|_{\infty}$ for every $\pi \in \Pi$ and $z \in \mathbb{R}^d$.

Since Λ_i as well as $\mathbb{I}_d - \Lambda_j$ are both non-negative diagonal matrices, the above result leads to the bound

$$\begin{aligned}
|\Psi_{k+1}^{(3)}| &\leq \gamma \sum_{i=0}^{k}\left[\prod_{j=i+1}^{k}(\mathbb{I}_d - \Lambda_j)\right]\Lambda_i\left|\mathrm{T}_i\left(\hat{V}_i - V^{\Phi}\right)\right| \\
&\leq \gamma \sum_{i=0}^{k}\left[\prod_{j=i+1}^{k}(\mathbb{I}_d - \Lambda_j)\right]\Lambda_i \mathbb{1}_d \|\Delta_i\|_{\infty}
\end{aligned}$$

just like in Li et al. (2021). Substituting this bound as well as Li et al. (2021, Lemma 2) and Li et al. (2021, Lemma 1) into

$$|\Delta_{k+1}| \leq |\Psi_{k+1}^{(1)}| + |\Psi_{k+1}^{(2)}| + |\Psi_{k+1}^{(3)}|,$$

which is obtained by applying the triangle inequality to Equation (21), results in the same bounds as Li et al. (2021, Equation (47)). For $k+1 < k_{\text{frame}}$, the bound is

$$|\Delta_{k+1}| \leq \mathbb{1}_d\|\Delta_0\|_{\infty} + \mathbb{1}_d c\gamma\sqrt{\alpha\log\frac{dK}{\delta}}\|V^{\Phi}\|_{\infty} + \gamma\sum_{i=0}^{k}\left[\prod_{j=i+1}^{k}(\mathbb{I}_d - \Lambda_j)\right]\Lambda_i\mathbb{1}_d\|\Delta_i\|_{\infty} \quad (22)$$

and for $k+1 \in [k_{\text{frame}}, K]$, the bound is

$$\begin{aligned} |\Delta_{k+1}| \leq\ & \mathbb{1}_d (1-\alpha)^{\frac{1}{2}(k+1)\eta_{\min}} \|\Delta_0\|_\infty \\ & + \mathbb{1}_d c\gamma \sqrt{\alpha \log \frac{dK}{\delta}} \|V^\Phi\|_\infty \\ & + \gamma \sum_{i=0}^{k} \left[\prod_{j=i+1}^{k} (\mathbb{I}_d - \Lambda_j) \right] \Lambda_i \mathbb{1}_d \left\|\Delta_i\right\|_\infty , \end{aligned} \tag{23}$$

with probability of at least $1-2\delta$.

The upper bound

$$\|\Delta_k\|_\infty \leq \frac{c\gamma \sqrt{\alpha \log \frac{dK}{\delta}} \|V^\Phi\|_\infty + \|\Delta_0\|_\infty}{1-\gamma} \tag{24}$$

is shown in Li et al. (2021, Section 6.2.2) using the bounds of (22) and (23), the structure of Λ_k, and the aggregation of state-action visitation counts, all of which are the same in their and our setting. Therefore, the bound (24) holds across all $k \in [1, K]$.

Using this and the fact that $\|V^\Phi\|_\infty \leq \frac{1}{1-\gamma}$ and $\|Q^\Phi\|_\infty \leq \frac{1}{1-\gamma}$, where the bound on Q^Φ holds due to $Q^\Phi = r + \gamma \mathrm{T} V^\Phi$ with $\|V^\Phi\|_\infty \leq \frac{1}{1-\gamma}$, we satisfy all conditions required by Lemma A.3 and Lemma A.4, which combined yield Theorem A.2 from which we obtain

$$\|\hat{Q}_k - Q^\Phi\|_\infty \leq (1-\varrho)^{l(k)} \frac{\|\hat{Q}_0 - Q^\Phi\|_\infty}{1-\gamma} + \frac{c\gamma}{1-\gamma} \|V^\Phi\|_\infty \sqrt{\alpha \log\left(\frac{dK}{\delta}\right)} + \epsilon, \tag{25}$$

that holds with probability of at least $1-6\delta$ as obtained via a union bound over the required Lemmas A.1, A.2, A.3, and A.4. While $1-6\delta$ is conservative, due to Lemma A.3 holding based on Lemmas A.1 and A.2 rather than on its own event, it does not affect the final result since the value gets absorbed into the universal constants. The only artifact resulting from this choice is the $K \geq 24$ requirement that is non-restrictive, given the bounds on the horizon K. □

As the second step, we show that Theorem A.3 holds for gated Q-learning based on the fact that Theorem A.2 applies.

Proof of Theorem A.3. Proving Theorem A.3 requires showing that the following two bounds on the terms of (25) hold, given that we invoke Theorem A.2 with the learning rate set to be

$$\alpha \leq \min\left\{ \frac{(1-\gamma)^4 \epsilon^2}{c^2 \gamma^2 \log\left(\frac{dK}{\delta}\right)}, \frac{1}{\frac{1}{2}\eta_{\min} k_{\text{frame}}} \right\}, \tag{26}$$

since

$$\begin{aligned} \alpha \leq \frac{1}{\frac{1}{2}\eta_{\min} k_{\text{frame}}} &= \frac{1}{\frac{1}{2}\eta_{\min} \frac{443 k_{\text{mix}}}{\eta_{\min}} \log\left(\frac{4dK}{\delta}\right)} = \frac{2}{443 k_{\text{mix}} \log\left(\frac{4dK}{\delta}\right)} \\ \implies \quad \alpha \log\left(\frac{dK}{\delta}\right) &\leq \frac{2}{443 k_{\text{mix}}} \frac{\log\left(\frac{dK}{\delta}\right)}{\log\left(\frac{4dK}{\delta}\right)} \leq \frac{2}{443 k_{\text{mix}}} < 1 \end{aligned} \tag{27}$$

satisfies the requirements for Lemmas A.1, A.3, and A.4 as well as for Theorem A.2.

The first term of (25) satisfies

$$(1-\varrho)^{l(k)} \frac{\|\hat{Q}_0 - Q^\Phi\|_\infty}{1-\gamma} = (1-\varrho)^{l(k)} \frac{\|\Delta_0\|_\infty}{1-\gamma} \leq \epsilon \tag{28}$$

whenever

$$k \geq k_{\text{threshold}} + k_{\text{frame}} + \frac{4}{(1-\gamma)\alpha\eta_{\min}} \log\left(\frac{\|\Delta_0\|_\infty}{(1-\gamma)\epsilon}\right) \tag{29}$$

and that the second term of (25) satisfies

$$\frac{c\gamma}{1-\gamma}\|V^{\Phi}\|_{\infty}\sqrt{\alpha\log\left(\frac{dK}{\delta}\right)} \leq \epsilon. \tag{30}$$

Given the conditions (26) and (29), $k_{\text{threshold}}$, and k_{frame} combined with $\|V^{\Phi}\|_{\infty} \leq \frac{1}{1-\gamma}$ allows us to follow the same steps as given in Li et al. (2021, Section 6.3) to prove the two bounds (28) and (30). Combining both with the ϵ of (25) results in

$$\|\hat{Q}_k - Q^{\Phi}\|_{\infty} \leq \epsilon + \epsilon + \epsilon = 3\epsilon.$$

Therefore, the proof of Theorem A.3 is concluded when ϵ is replaced by $\frac{\varepsilon}{3}$ and the failure probability of Theorem A.2 is set to $\frac{\delta}{6}$.

The following derivation is done using the ε and δ substitution. Note that since $\frac{\delta}{6} < \frac{1}{2}$ for any $\delta \in (0,1)$, as required by Theorem A.3, $\frac{\delta}{6} \in (0, \frac{1}{2})$ and (27) remain satisfied as it is required by Theorem A.2.

To obtain the learning rate as given in Theorem A.3 we start with the currently given learning rate of

$$\min\left\{\frac{(1-\gamma)^4\frac{\varepsilon^2}{9}}{c^2\gamma^2\log\left(\frac{6dK}{\delta}\right)}, \frac{1}{\frac{1}{2}\eta_{\text{min}}k_{\text{frame}}}\right\}$$

which results from applying the substitutions to (26). Expanding the second element of the minimum by plugging in k_{frame} and expanding results in

$$\alpha \leq \frac{1}{\frac{1}{2}\eta_{\text{min}}k_{\text{frame}}} = \frac{1}{\frac{1}{2}\eta_{\text{min}}\frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right)} = \frac{2}{443k_{\text{mix}}\log\left(\frac{24dK}{\delta}\right)}.$$

Since $K \geq 24$, $d \geq 1$, and $\delta < 1$, we have $\frac{dK}{\delta} > 24$ and hence

$$\frac{dK}{\delta} > 24 \iff \log\left(\frac{dK}{\delta}\right) > \log(24) \iff 2\log\left(\frac{dK}{\delta}\right) > \log\left(\frac{24dK}{\delta}\right). \tag{31}$$

With $c > 0$ being a universal constant, c_1 is required to satisfy both

$$c_1 \leq \frac{1}{9c^2}\frac{\log\left(\frac{dK}{\delta}\right)}{\log\left(\frac{6dK}{\delta}\right)} \quad \text{and} \quad c_1 \leq \frac{2}{443}\frac{\log\left(\frac{dK}{\delta}\right)}{\log\left(\frac{24dK}{\delta}\right)}$$

which, by Equation (31), is implied by

$$c_1 \leq \frac{1}{9c^2}\frac{\log\left(\frac{dK}{\delta}\right)}{2\log\left(\frac{dK}{\delta}\right)} \quad \text{and} \quad c_1 \leq \frac{2}{443}\frac{\log\left(\frac{dK}{\delta}\right)}{2\log\left(\frac{dK}{\delta}\right)}$$

such that the universal constant c_1 can be given by

$$c_1 \leq \min\left\{\frac{1}{18c^2}, \frac{1}{443}\right\}.$$

This c_1 allows a choice of learning rate, that satisfies (26), to be

$$\frac{c_1}{\log\left(\frac{dK}{\delta}\right)}\min\left\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2}, \frac{1}{k_{\text{mix}}}\right\} \leq \min\left\{\frac{(1-\gamma)^4\frac{\varepsilon^2}{9}}{c^2\gamma^2\log\left(\frac{6dK}{\delta}\right)}, \frac{2}{443k_{\text{mix}}\log\left(\frac{24dK}{\delta}\right)}\right\},$$

and we therefore can set the learning rate α to

$$\alpha = \frac{c_1}{\log\left(\frac{|\mathcal{S}||\mathcal{A}|K}{\delta}\right)}\min\left\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2}, \frac{1}{k_{\text{mix}}}\right\},$$

as stated in the theorem.

To obtain the bound on K as given in Theorem A.3 we again plug in the values of $k_{\text{threshold}}$ and k_{frame} into (29) to obtain a sufficient lower bound on k

$$k \geq k_{\text{threshold}} + k_{\text{frame}} + \frac{4}{(1-\gamma)\alpha\eta_{\text{min}}} \log\left(\frac{3\|\Delta_0\|_\infty}{(1-\gamma)\varepsilon}\right) \tag{32}$$

$$= \max\left\{\frac{2\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right)}{\alpha\eta_{\text{min}}}, \frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right)\right\} + \frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) + \frac{4}{(1-\gamma)\alpha\eta_{\text{min}}}\log\left(\frac{3\|\Delta_0\|_\infty}{(1-\gamma)\varepsilon}\right).$$

The next steps are a series of substitutions that all increase the lower bound on k, and therefore imply (32) at every step, in order to derive the sample condition as stated in Theorem A.3. Using

$$\log\left(\frac{3\|\Delta_0\|_\infty}{(1-\gamma)\varepsilon}\right) \leq \log\left(\frac{3\frac{1}{1-\gamma}}{(1-\gamma)\varepsilon}\right) = \log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right), \tag{33}$$

which holds due to $\|\Delta_0\|_\infty \leq \frac{1}{1-\gamma}$ because of $\hat{Q}_0 = 0$ and $\|Q^\Phi\|_\infty \leq \frac{1}{1-\gamma}$, we get

$$k \geq \max\left\{\frac{2\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right)}{\alpha\eta_{\text{min}}}, \frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right)\right\} + \frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) + \frac{4}{(1-\gamma)\alpha\eta_{\text{min}}}\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right).$$

Using $\max\{z_1, z_2\} \leq z_1 + z_2$ for $z_1, z_2 \in \mathbb{R}_+$ yields

$$\begin{aligned} k &\geq \frac{2\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right)}{\alpha\eta_{\text{min}}} + \frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) + \frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) + \frac{4}{(1-\gamma)\alpha\eta_{\text{min}}}\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) \\ &= \frac{2}{\alpha\eta_{\text{min}}}\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) + 2\frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) + \frac{4}{(1-\gamma)\alpha\eta_{\text{min}}}\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right). \end{aligned}$$

Since $(1-\gamma) \in (0,1)$ we have

$$\frac{2}{\alpha\eta_{\text{min}}} \leq \frac{2}{(1-\gamma)\alpha\eta_{\text{min}}}$$

and therefore

$$\left(\frac{2}{\alpha\eta_{\text{min}}} + \frac{4}{(1-\gamma)\alpha\eta_{\text{min}}}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) \leq \frac{6}{(1-\gamma)\alpha\eta_{\text{min}}}\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right)$$

which results in

$$k \geq 2\frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) + \frac{1}{\alpha}\frac{6}{(1-\gamma)\eta_{\text{min}}}\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right).$$

Plugging in

$$\frac{1}{\alpha} = \frac{\log\left(\frac{dK}{\delta}\right)}{c_1} \max\left\{\frac{\gamma^2}{(1-\gamma)^4\varepsilon^2}, k_{\text{mix}}\right\}$$

gives

$$\begin{aligned}
k &\geq 2\frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) \\
&\quad + \frac{\log\left(\frac{dK}{\delta}\right)}{c_1}\max\left\{\frac{\gamma^2}{(1-\gamma)^4\varepsilon^2}, k_{\text{mix}}\right\}\frac{6}{(1-\gamma)\eta_{\text{min}}}\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) \\
&= 2\frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) \\
&\quad + \max\left\{\frac{\gamma^2}{(1-\gamma)^4\varepsilon^2}, k_{\text{mix}}\right\}\frac{6}{c_1(1-\gamma)\eta_{\text{min}}}\log\left(\frac{dK}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right),
\end{aligned}$$

which allows us to use $\max\{z_1, z_2\} \leq z_1 + z_2$ for $z_1, z_2 \in \mathbb{R}_+$ again

$$\begin{aligned}
k &\geq 2\frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) \\
&\quad + \frac{\gamma^2}{(1-\gamma)^4\varepsilon^2}\frac{6}{c_1(1-\gamma)\eta_{\text{min}}}\log\left(\frac{dK}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) \\
&\quad + \frac{6k_{\text{mix}}}{c_1(1-\gamma)\eta_{\text{min}}}\log\left(\frac{dK}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) \\
&= 2\frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) \\
&\quad + \frac{6\gamma^2}{c_1(1-\gamma)^5\eta_{\text{min}}\varepsilon^2}\log\left(\frac{dK}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) \\
&\quad + \frac{6k_{\text{mix}}}{c_1(1-\gamma)\eta_{\text{min}}}\log\left(\frac{dK}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) \\
&= 2\frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{24dK}{\delta}\right) \\
&\quad + \frac{6}{c_1\eta_{\text{min}}}\left(\frac{\gamma^2}{(1-\gamma)^5\varepsilon^2} + \frac{k_{\text{mix}}}{1-\gamma}\right)\log\left(\frac{dK}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right).
\end{aligned}$$

Due to (31)

$$k \geq 4\frac{443k_{\text{mix}}}{\eta_{\text{min}}}\log\left(\frac{dK}{\delta}\right) + \frac{6}{c_1\eta_{\text{min}}}\left(\frac{\gamma^2}{(1-\gamma)^5\varepsilon^2} + \frac{k_{\text{mix}}}{1-\gamma}\right)\log\left(\frac{dK}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) \quad (34)$$

and due to $\varepsilon \leq \frac{1}{1-\gamma}$

$$\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) \geq \log(3) > 1$$

we can pull the first term of (34) into the parentheses that contain the other summands

$$\begin{aligned}
k &\geq \frac{6}{c_1\eta_{\text{min}}}\left(\frac{c_1\eta_{\text{min}}}{6}4\frac{443k_{\text{mix}}}{\eta_{\text{min}}} + \frac{\gamma^2}{(1-\gamma)^5\varepsilon^2} + \frac{k_{\text{mix}}}{1-\gamma}\right)\log\left(\frac{dK}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right) \\
&= \frac{6}{c_1\eta_{\text{min}}}\left(\frac{4c_1443k_{\text{mix}}}{6} + \frac{\gamma^2}{(1-\gamma)^5\varepsilon^2} + \frac{k_{\text{mix}}}{1-\gamma}\right)\log\left(\frac{dK}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon}\right).
\end{aligned}$$

Since $c_1 \leq \frac{1}{443}$, we can use

$$\frac{4c_1443k_{\text{mix}}}{6} \leq \frac{4}{6}k_{\text{mix}} \leq \frac{k_{\text{mix}}}{1-\gamma}$$

to obtain

$$k \geq \frac{6}{c_1 \eta_{\text{min}}} \left(\frac{k_{\text{mix}}}{1-\gamma} + \frac{\gamma^2}{(1-\gamma)^5 \varepsilon^2} + \frac{k_{\text{mix}}}{1-\gamma} \right) \log\left(\frac{dK}{\delta}\right) \log\left(\frac{3}{(1-\gamma)^2 \varepsilon}\right)$$

and hence

$$k \geq \frac{12}{c_1 \eta_{\text{min}}} \left(\frac{\gamma^2}{(1-\gamma)^5 \varepsilon^2} + \frac{k_{\text{mix}}}{1-\gamma} \right) \log\left(\frac{dK}{\delta}\right) \log\left(\frac{3}{(1-\gamma)^2 \varepsilon}\right).$$

Choosing $c_0 = \frac{12}{c_1}$ and using $1 > \gamma^2$ gives

$$k \geq \frac{c_0}{\eta_{\text{min}}} \left(\frac{1}{(1-\gamma)^5 \varepsilon^2} + \frac{k_{\text{mix}}}{1-\gamma} \right) \log\left(\frac{dK}{\delta}\right) \log\left(\frac{3}{(1-\gamma)^2 \varepsilon}\right),$$

which implies (32) and is the condition imposed at $k = K$ as stated in the theorem

$$K \geq \frac{c_0}{\eta_{\text{min}}} \left(\frac{1}{(1-\gamma)^5 \varepsilon^2} + \frac{k_{\text{mix}}}{1-\gamma} \right) \log\left(\frac{|\mathcal{S}||\mathcal{A}|K}{\delta}\right) \log\left(\frac{3}{(1-\gamma)^2 \varepsilon}\right).$$

□

As both auxiliary results, Theorem A.2 and Theorem A.3, apply to gated Q-learning, we can continue to prove Theorem 3.2. To that end, we switch from the constant learning rate α as used up to this point to a learning rate sequence $(\alpha_k)_{k \geq 0}$ as given by (12).

Using the stabilized estimate (Li et al., 2021, Equation (67))

$$\hat{\eta}_{\text{min},k} = c_{\hat{\eta}} \eta_{\text{min}} \quad \text{with} \quad c_{\hat{\eta}} \in \left[\frac{1}{6}, \frac{9}{2}\right] \quad \text{for} \quad K \geq k \geq \frac{443 k_{\text{mix}} \log\left(\frac{4dk}{\delta}\right)}{\eta_{\text{min}}} \tag{35}$$

that holds true with probability at least $1 - \delta$, the learning rate (12) can be simplified to

$$\alpha_k = c_\alpha \exp\left(\left\lfloor \log\left(\frac{\log(k)}{c_{\hat{\eta}} \eta_{\text{min}} (1-\gamma) \gamma^2 k} \right) \right\rfloor \right) \qquad \text{for any} \quad k \geq \frac{6 c_\alpha k_{\text{mix}} \log\left(\frac{4dk}{\delta}\right)}{\eta_{\text{min}}(1-\gamma)\gamma^2}, \tag{36}$$

for $c_\alpha \geq 11$ which corresponds to Li et al. (2021, Equation (68)). The requirement of $c_\alpha \geq 11$ results from setting the second branch of the learning rate to be ≤ 1 and simultaneously having k satisfy the condition of (35).

The learning rate (36) is made up of constant segments that decrease stepwise. Let $(k_n)_{n \geq 1}$ be the sequence of endpoints of the constant segments with

$$k_1 \leq \frac{6 e c_\alpha k_{\text{mix}} \log\left(\frac{4dk_1}{\delta}\right)}{\eta_{\text{min}}(1-\gamma)\gamma^2} \qquad \text{and} \quad 2k_n < k_{n+1} < 3k_n. \tag{37}$$

Then, the piecewise constant learning rate can be expressed as in Li et al. (2021, Equation (70)), that is

$$\alpha_{(n)} = \alpha_k = \frac{c_{(n)} \log(k_n)}{\eta_{\text{min}}(1-\gamma)\gamma^2 k_n} \quad \text{with} \quad c_{(n)} \in \left[\frac{2c_\alpha}{9e}, 6c_\alpha\right], \tag{38}$$

for all $k \in (k_{n-1}, k_n]$. With

$$\tilde{Q}_k = \begin{cases} \hat{Q}_k & \text{if} \quad \alpha_{k+1} \neq \alpha_k \\ \tilde{Q}_{k-1} & \text{else,} \end{cases}$$

we obtain a sequence of Q-values $(\tilde{Q}_k)_{k \geq 0}$ that is piecewise constant with

$$\tilde{Q}_k = \hat{Q}_{k_n} \qquad \forall k \in [k_n, k_{n+1})$$

in the same way as the learning rate (38). As such it tracks the current Q-value iterate $\hat{Q}_k$ by reading its value based on the sequence $(k_n)_{n \geq 1}$. Since Theorem A.3 is a statement of gated Q-learning based on an initial estimate, a number of update steps, and a learning rate, this construction of piecewise constant learning rate segments enables the application of the theorem on segments. In particular, assuming the last segment satisfies the theorem's requirements on initial estimate, number of update steps, and learning rate allows us to apply the theorem to the last segment to obtain our desired result.

As such, the next steps show how the construction of the piecewise constant learning rate segments combined with the reward structure ensures that the last segment does indeed satisfy the required properties to use Theorem A.3. As its first element, we use the learning rate from (38).

For the rewards $r \in [0,1]$ our gated update rule, given by equations (10) and (11), implies

$$0 \leq \hat{Q}_k(s,a) \leq \frac{1}{1-\gamma} \qquad \forall k \geq 0, \forall (s,a) \in \mathcal{X}.$$

This implication can be shown by induction on k. The base case holds because $\hat{Q}_0 = 0$ is initialized to a zero vector. At any update step, only the entry (s_k, a_k) is updated and set to be a convex combination (Equation (11)) of $\hat{Q}_k(s_k,a_k)$ and $\hat{\Gamma}_\Phi \hat{Q}_k(s_k,a_k,s'_k)$ (Equation (10)), since $\alpha_k \in (0,1]$. For the updated entry, we can use the fact that $\Omega_\Phi(\pi^b_s) = 0$ for all $s \in \mathcal{S}$ is a lower bound on the greedy gated update to obtain for all $s \in \mathcal{S}$

$$0 \leq \left\langle \pi^b(\cdot|s), \hat{Q}_k(s,\cdot) \right\rangle \leq \max_{\pi \in \Pi} \left\{ \left\langle \pi(\cdot|s), \hat{Q}_k(s,\cdot) \right\rangle - \Omega_\Phi(\pi(\cdot|s)) \right\} \leq \max_{a \in \mathcal{A}} \hat{Q}_k(s,a) \leq \frac{1}{1-\gamma},$$

where the first and the last inequality are due to the induction hypothesis. Therefore,

$$0 \leq r(s_k,a_k) \leq \hat{\Gamma}_\Phi \hat{Q}_k(s_k,a_k,s'_k) \leq 1 + \frac{\gamma}{1-\gamma} = \frac{1}{1-\gamma}.$$

Since $r \in [0,1]$ and the update rule of Equation (11) is a convex combination, the possible range of values of all $\hat{Q}_k$ is preserved to be in $\left[0, \frac{1}{1-\gamma}\right]$.

As a result we get the bound

$$\|\tilde{Q}_k - Q^\Phi\|_\infty \leq \frac{1}{1-\gamma} \qquad \forall k \geq 0.$$

Setting our piecewise constant learning rate (38) equal to the learning rate required by Theorem A.3

$$\frac{c_{(n)} \log(k_n)}{\eta_{\min}(1-\gamma)\gamma^2 k_n} = \frac{c_1}{\log\left(\frac{dk_n}{\delta}\right)} \min\left\{ \frac{(1-\gamma)^4 \varepsilon^2}{\gamma^2}, \frac{1}{k_{\text{mix}}} \right\}$$

and solving for ε while assuming the first branch of the minimum is active

$$\frac{c_{(n)} \log(k_n)}{\eta_{\min}(1-\gamma)\gamma^2 k_n} = \frac{c_1}{\log\left(\frac{dk_n}{\delta}\right)} \frac{(1-\gamma)^4 \varepsilon_n^2}{\gamma^2}$$

$$\iff \frac{c_{(n)} \log\left(\frac{dk_n}{\delta}\right) \log(k_n)}{c_1 \eta_{\min}(1-\gamma)\gamma^2 k_n} = \frac{(1-\gamma)^4 \varepsilon_n^2}{\gamma^2}$$

$$\iff \sqrt{\frac{c_{(n)} \log\left(\frac{dk_n}{\delta}\right) \log(k_n)}{c_1 \eta_{\min}(1-\gamma)^5 k_n}} = \varepsilon_n$$

yields

$$\varepsilon_n = \sqrt{\frac{c_{0,(n)} \log\left(\frac{dk_n}{\delta}\right) \log(k_n)}{\eta_{\min}(1-\gamma)^5 k_n}},$$

where $c_{0,(n)} = \frac{c_{(n)}}{c_1} > 0$. This holds true for

$$\frac{(1-\gamma)^4 \varepsilon_n^2}{\gamma^2} \leq \frac{1}{k_{\text{mix}}} \tag{39}$$

$$\iff \frac{c_{0,(n)} \log\left(\frac{dk_n}{\delta}\right) \log(k_n)}{\eta_{\min}(1-\gamma)\gamma^2 k_n} \leq \frac{1}{k_{\text{mix}}}$$

$$\iff k_n \geq \frac{c_{0,(n)} k_{\text{mix}} \log\left(\frac{dk_n}{\delta}\right) \log(k_n)}{\eta_{\min}(1-\gamma)\gamma^2}. \tag{40}$$

Since Theorem A.3 requires $\varepsilon_n \leq \frac{1}{1-\gamma}$, we get the additional condition on k_n, which is

$$\begin{aligned} & \varepsilon_n^2 \leq \frac{1}{(1-\gamma)^2} \\ \iff & \frac{c_{0,(n)} \log\left(\frac{dk_n}{\delta}\right) \log(k_n)}{\eta_{\min}(1-\gamma)^5 k_n} \leq \frac{1}{(1-\gamma)^2} \\ \iff & k_n \geq \frac{c_{0,(n)} \log\left(\frac{dk_n}{\delta}\right) \log(k_n)}{\eta_{\min}(1-\gamma)^3}. \end{aligned} \tag{41}$$

However, since the theorem will only be applied to the final segment, this condition is also only required to hold for the final segment, which we will show later in the proof when we arrive at the final segment and have access to all its properties.

Since $2k_n < k_{n+1} < 3k_n$ (Li et al., 2021, Equation (69)), we have

$$k_n - k_{n-1} > k_{n-1} > \frac{1}{3}k_n = \frac{c_{0,(n)} \log\left(\frac{dk_n}{\delta}\right) \log(k_n)}{3\eta_{\min}(1-\gamma)^5 \varepsilon_n^2},$$

where the equality results from plugging in

$$\varepsilon_n = \sqrt{\frac{c_{0,(n)} \log\left(\frac{dk_n}{\delta}\right) \log(k_n)}{\eta_{\min}(1-\gamma)^5 k_n}} \iff k_n = \frac{c_{0,(n)} \log\left(\frac{dk_n}{\delta}\right) \log(k_n)}{\eta_{\min}(1-\gamma)^5 \varepsilon_n^2}.$$

Since each segment has its Q-values initialized with $\hat{Q}_{k_{n-1}}$ which, as shown above, does satisfy

$$\|\hat{Q}_{k_{n-1}} - Q^{\Phi}\|_\infty \leq \frac{1}{1-\gamma},$$

the final segment will have an initial Q-value that satisfies the requirements as needed for the application of Theorem A.3. Theorem A.3 states that the Q-values are initialized with zero which ensures that $\|\Delta_0\|_\infty \leq \frac{1}{1-\gamma}$. Since the above states that $\|\Delta_{k_{n-1}}\|_\infty \leq \frac{1}{1-\gamma}$, the requirement implied by the zero initialization transfers to an initial Q-value of $\hat{Q}_{k_{n-1}}$. Therefore we replace the Δ_0 of Bound (33) with this $\Delta_{k_{n-1}}$, which results in

$$\log\left(\frac{3\|\Delta_{k_{n-1}}\|_\infty}{(1-\gamma)\varepsilon_n}\right) \leq \log\left(\frac{3\frac{1}{1-\gamma}}{(1-\gamma)\varepsilon_n}\right) = \log\left(\frac{3}{(1-\gamma)^2\varepsilon_n}\right).$$

From the above we get that, when $\varepsilon_n \leq \frac{1}{1-\gamma}$,

$$\|\tilde{Q}_{k_n} - Q^{\Phi}\|_\infty \leq \varepsilon_n,$$

with probability of at least $1-2\delta$, where one $1-\delta$ is from Theorem A.3 applied to that segment and one is from the stabilization of $\hat{\eta}_{\min,k}$ in (35), given that

$$\frac{1}{3}k_n \geq \frac{c_0}{\eta_{\min}}\left(\frac{1}{(1-\gamma)^5\varepsilon_n^2} + \frac{k_{\text{mix}}}{1-\gamma}\right)\log\left(\frac{dk_n}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon_n}\right), \tag{42}$$

where c_0 is the universal constant of Theorem A.3, and $\frac{1}{3}k_n \geq 24$. This provides the number of samples a segment requires for the application of Theorem A.3 as well as the probability of the accuracy holding for a segment. Under this update sample count, the last segment's number of updates does satisfy the update count requirement needed to apply Theorem A.3.

Next, we will set the constants of the segments into relation to the universal constants of Theorem A.3. Because of requirement (39), we have that

$$\frac{(1-\gamma)^4\varepsilon_n^2}{\gamma^2} \leq \frac{1}{k_{\text{mix}}} \quad \iff \quad k_{\text{mix}} \leq \frac{\gamma^2}{(1-\gamma)^4\varepsilon_n^2}$$

which implies

$$\frac{2}{(1-\gamma)^5\varepsilon_n^2} \geq \left(\frac{1}{(1-\gamma)^5\varepsilon_n^2} + \frac{k_{\text{mix}}}{1-\gamma}\right)$$

and therefore

$$\frac{c_0 2\log\left(\frac{dk_n}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon_n}\right)}{\eta_{\min}(1-\gamma)^5\varepsilon_n^2} \geq \frac{c_0}{\eta_{\min}}\left(\frac{1}{(1-\gamma)^5\varepsilon_n^2}+\frac{k_{\text{mix}}}{1-\gamma}\right)\log\left(\frac{dk_n}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon_n}\right).$$

By plugging in the larger lower bound and solving for $c_{0,(n)}$ we obtain

$$\begin{aligned}\frac{1}{3}k_n = \frac{c_{0,(n)}\log\left(\frac{dk_n}{\delta}\right)\log(k_n)}{3\eta_{\min}(1-\gamma)^5\varepsilon_n^2} &\geq \frac{c_0 2\log\left(\frac{dk_n}{\delta}\right)\log\left(\frac{3}{(1-\gamma)^2\varepsilon_n}\right)}{\eta_{\min}(1-\gamma)^5\varepsilon_n^2}\\ \iff \frac{c_{0,(n)}\log(k_n)}{6} &\geq c_0\log\left(\frac{3}{(1-\gamma)^2\varepsilon_n}\right)\\ \iff c_{0,(n)} &\geq 6c_0\frac{1}{\log(k_n)}\log\left(\frac{3}{(1-\gamma)^2\varepsilon_n}\right).\end{aligned}$$

Then we use the fact that

$$\begin{aligned}\varepsilon_n &= \sqrt{\frac{c_{0,(n)}\log\left(\frac{dk_n}{\delta}\right)\log(k_n)}{\eta_{\min}(1-\gamma)^5 k_n}}\\ \iff \frac{1}{(1-\gamma)^2\varepsilon_n} &= \sqrt{\frac{\eta_{\min}(1-\gamma)k_n}{c_{0,(n)}\log\left(\frac{dk_n}{\delta}\right)\log(k_n)}} \leq \sqrt{k_n},\end{aligned}$$

where the inequality follows from $\eta_{\min} \leq 1$, $\gamma \in (0,1)$ and $c_{0,(n)}\log\left(\frac{dk_n}{\delta}\right)\log(k_n) \geq 1$ because $k_n \geq 24$ from the requirements of Theorem A.3, $d \geq 1$, $\delta \in (0,1)$ and $c_{0,(n)} = \frac{c_{(n)}}{c_1} \geq \frac{2c_\alpha 443}{9e} \geq \frac{2\cdot 11\cdot 443}{9\cdot 3} \geq \frac{9746}{27} > 1$, to obtain the lower bound

$$c_{0,(n)} \geq 6c_0\frac{1}{\log(k_n)}\log\left(3\sqrt{k_n}\right) = 6c_0\frac{\log(3)+\frac{1}{2}\log(k_n)}{\log(k_n)} = 6c_0\left(\frac{\log(3)}{\log(k_n)}+\frac{1}{2}\right).$$

$k_n \geq 24$ allows us to use the upper bound $\frac{\log(3)}{\log(9)} \geq \frac{\log(3)}{\log(k_n)}$

$$c_{0,(n)} \geq 6c_0\left(\frac{\log(3)}{\log(9)}+\frac{1}{2}\right) \geq 6c_0\frac{\log(3)+\frac{1}{2}\log(k_n)}{\log(k_n)} \tag{43}$$

combined with the bound on c_α in (38), to get

$$c_{0,(n)} = \frac{c_{(n)}}{c_1} \geq \frac{2c_\alpha}{9ec_1} \geq 6c_0.$$

This provides the condition that makes c_α dependent on the two constants c_0, c_1 of Theorem A.3, that is

$$c_\alpha \geq 27ec_0c_1$$

which ensures that (42) holds under the sample size condition (40).

Letting $N = \max\{n | k_n \leq K\}$ and therefore k_N be the largest interval endpoint that is smaller than K, the total number of timesteps, we know that

$$\frac{1}{3}K \leq k_N \leq K$$

because of the bounds on k_n in (37). Therefore we get

$$\begin{aligned}\|\tilde{Q}_K - Q^\Phi\|_\infty = \|\hat{Q}_{k_N} - Q^\Phi\|_\infty \leq \varepsilon_N &= \sqrt{\frac{c_{0,(N)}\log\left(\frac{dk_N}{\delta}\right)\log(k_N)}{\eta_{\min}(1-\gamma)^5 k_N}}\\ &\leq \sqrt{\frac{3c_{0,(N)}\log\left(\frac{dK}{\delta}\right)\log(K)}{\eta_{\min}(1-\gamma)^5 K}}.\end{aligned}$$

Finally, substituting δ with $\frac{\delta}{2}$ to account for the $1-2\delta$ probability from (42), we get

$$\varepsilon_N^2 \leq \frac{3c_{0,(N)} \log\left(\frac{2dK}{\delta}\right) \log(K)}{\eta_{\min}(1-\gamma)^5 K} \leq \varepsilon^2$$

$$\implies K \geq \frac{3c_{0,(N)} \log\left(\frac{2dK}{\delta}\right) \log(K)}{\eta_{\min}(1-\gamma)^5 \varepsilon^2}$$

$$= \frac{3c_{0,(N)}}{\eta_{\min}} \frac{1}{(1-\gamma)^5 \varepsilon^2} \log\left(\frac{2dK}{\delta}\right) \log(K),$$

using (31), we get

$$K \geq \frac{6c_{0,(N)}}{\eta_{\min}} \frac{1}{(1-\gamma)^5 \varepsilon^2} \log\left(\frac{dK}{\delta}\right) \log(K) \tag{44}$$

and, required by condition (39), that

$$K \geq \frac{3c_{0,(N)} k_{\text{mix}} \log\left(\frac{2dK}{\delta}\right) \log(K)}{\eta_{\min}(1-\gamma)\gamma^2} = \frac{3c_{0,(N)}}{\eta_{\min}} \frac{k_{\text{mix}}}{(1-\gamma)\gamma^2} \log\left(\frac{2dK}{\delta}\right) \log(K),$$

which by (31) yields

$$K \geq \frac{6c_{0,(N)}}{\eta_{\min}} \frac{k_{\text{mix}}}{(1-\gamma)\gamma^2} \log\left(\frac{dK}{\delta}\right) \log(K). \tag{45}$$

Combining both bounds, (44) and (45), results in

$$K \geq \frac{6c_{0,(N)}}{\eta_{\min}} \max\left\{\frac{1}{(1-\gamma)^5 \varepsilon^2}, \frac{k_{\text{mix}}}{(1-\gamma)\gamma^2}\right\} \log\left(\frac{dK}{\delta}\right) \log(K). \tag{46}$$

Since we have $\frac{1}{3}K \leq k_N$ and (44), which is the first branch of the maximum operation in (46) such that the value will never be smaller, we get

$$k_N \geq \frac{2c_{0,(N)} \log\left(\frac{dK}{\delta}\right) \log(K)}{\eta_{\min}(1-\gamma)^5 \varepsilon^2} \geq \frac{c_{0,(N)} \log\left(\frac{dk_N}{\delta}\right) \log(k_N)}{\eta_{\min}(1-\gamma)^3}$$

where the third term is the right-hand side of (41). This implies that $\varepsilon_N \leq \frac{1}{1-\gamma}$, as required.

Applying the upper bound (38) to the definition of $c_{0,(N)}$ and setting $C = 6 \cdot \frac{6c_\alpha}{c_1}$, we obtain the sample bound as stated in Theorem 3.2

$$K \geq \frac{C}{\eta_{\min}} \max\left\{\frac{1}{(1-\gamma)^5 \varepsilon^2}, \frac{k_{\text{mix}}}{(1-\gamma)\gamma^2}\right\} \log\left(\frac{|\mathcal{S}||\mathcal{A}|K}{\delta}\right) \log(K)$$

which concludes the proof. □

C EXPERIMENTS

The experiments were run on a workstation with an AMD Ryzen 9 5950X CPU (16 cores), 128 GB of RAM, and an RTX 3090 Ti, and the code for all experiments is available at github.com/gatedregularization/code for the tabular experiments and github.com/gatedregularization/ccode for the continuous experiments. Tabular experiments use the entropy-based gate, Equation (4), with the KL divergence as the regularization function, while continuous experiments use the standard-deviation-based gate, Equation (14), applied to the regularization of the respective algorithms.

C.1 TABULAR ENVIRONMENTS

The tabular *CliffWalking* and *Taxi* environments, from the Gymnasium library (Towers et al., 2024), provide numerical evidence for Section 3, since the tabular setting allows for an implementation that matches the theory exactly. For all tabular experiments, we train offline agents on $K = 2.5 \times 10^6$ samples collected under a behavioral policy π^b. This number of samples is well beyond the number required to obtain a sufficiently converged training result, where additional data no longer noticeably alters the learned policy.

The behavioral policies π^b reflect suboptimal, spatially inhomogeneous behavior obtained by perturbing the optimal policy $\pi^\star$, which is computed via synchronous value iteration (Sutton & Barto, 2018). Perturbation is done by letting π^b be, in every perturbed state, the convex combination of $\pi^\star$ and a noise vector over the actions with convex coefficient $\kappa \in [0, 1]$. We use two types of noise.

Random-Dirichlet Perturbed Policies. Behavioral policies underlying the results of Table 1(a) and Table 1(b) are obtained by random-Dirichlet perturbation of the optimal policy in every state of the environment with convex coefficient κ for both environments. In order to generate the noise, one concentration parameter α^D is uniformly sampled at random from $\{10^i | i \in \{-3, -2, -1, 0, 1, 2, 3\}\}$ per perturbed state and used for all actions when parameterizing the Dirichlet distribution over the action probabilities. A large α^D (e.g., $\alpha^D = 10^3$) yields a noise vector that is close to uniform, whereas a small α^D (e.g., $\alpha^D = 10^{-3}$) produces a probability vector that is close to deterministic. This yields a suboptimal and inhomogeneous policy.

In the *CliffWalking* environment, Table 1(a), stepping off the cliff makes OOD actions potentially dangerous and regularization relevant when learning from offline data. In the tabular setting, actions absent from the dataset may appear deceptively valuable, causing greedy Q-learners to become stuck in a loop or deterministically walk off the cliff, which makes regularization a necessity. Within the group of regularized methods, it can be seen that the performance of the behavioral policy lower bounds the performance of regularized and gated policies. Further, the gated method achieves higher expected rewards than the regularized method in most cases, especially in terms of worst-case performance. This performance ordering closely connects with Theorem 2.1, which is verified empirically by the check-marks that indicate whenever $V^\Phi(s) \geq V^\mathcal{R}(s)\ \forall s \in \mathcal{S}$. There are a few instances at which $V^\Phi(s) \geq V^\mathcal{R}(s)$ is violated at a small number of states, this indicates that the value tables have not yet fully converged since they can be gradually reduced by increasing the size of the offline datasets.

Table 1(b) shows the *Taxi* experiment with offline data generated by the random-Dirichlet behavioral policy described above. A major difference from the *CliffWalking* experiment is that Watkins' Q-learning works perfectly in this environment, an observation consistent across all *Taxi* experiments, independent of the behavioral policy. A plausible explanation of this can be given by the environment's lack of actions with catastrophic consequences. Even if the agent performs one or more OOD actions, the only costs are the extra steps taken. Taken together with the convex perturbation used to generate suboptimal behavioral policies and the resulting structure of the offline datasets, the learning process can be reasonably categorized within the neighborhood of off-policy learning. However, within the subset of regularized learning methods, the gated method matches or outperforms the regularized method and is more robust to the choice of β. Moreover, this experiment provides additional empirical validation of Theorem 2.1, as indicated by the check-marks.

Uniformly Perturbed Policies. The behavioral policies in Table 2(a) and Table 2(b) are generated by applying a convex combination of the optimal policy with the uniform distribution in every state

	π	$\kappa = 0.2$			$\kappa = 0.5$			$\kappa = 0.8$		
	b	-97.83	$[-235, -13]$		-340.19	$[-793, -23]$		-672.59	$[-1090, -257]$	
	Q	-100.00	$[-100, -100]$		-100.00	$[-100, -100]$		-100.00	$[-100, -100]$	
$\beta = 1$	$\mathcal{R}$	-17.62	$[-15, -13]$		-14.07	$[-17, -13]$		-32.32	$[-83, -13]$	
	Φ	-92.45	$[-100, -100]$	0.0021	-13.05	$[-13, -13]$	✓	-100.00	$[-100, -100]$	✓
$\beta = 2$	$\mathcal{R}$	-14.08	$[-17, -13]$		-16.06	$[-19, -13]$		-19.05	$[-25, -15]$	
	Φ	-14.96	$[-15, -13]$	✓	-13.38	$[-15, -13]$	✓	-13.93	$[-17, -13]$	✓
$\beta = 4$	$\mathcal{R}$	-14.76	$[-17, -13]$		-18.76	$[-25, -15]$		-24.84	$[-34, -17]$	
	Φ	-14.25	$[-17, -13]$	✓	-14.60	$[-17, -13]$	✓	-17.27	$[-19, -17]$	✓
$\beta = 8$	$\mathcal{R}$	-15.33	$[-19, -13]$		-21.44	$[-29, -15]$		-31.76	$[-45, -21]$	
	Φ	-14.94	$[-17, -13]$	✓	-16.86	$[-21, -13]$	✓	-17.96	$[-21, -17]$	✓
$\beta = 16$	$\mathcal{R}$	-15.76	$[-19, -13]$		-23.62	$[-33, -16]$		-38.70	$[-57, -23]$	
	Φ	-15.45	$[-19, -13]$	✓	-19.57	$[-26, -15]$	✓	-18.96	$[-22, -17]$	✓
$\beta = 32$	$\mathcal{R}$	-17.68	$[-20, -13]$		-28.20	$[-37, -17]$		-48.93	$[-73, -26]$	
	Φ	-15.93	$[-19, -13]$	✓	-22.09	$[-30, -15]$	✓	-21.21	$[-26, -17]$	✓

(a) Gymnasium *CliffWalking*

	π	$\kappa = 0.2$			$\kappa = 0.5$			$\kappa = 0.8$		
	b	-4.75	$[-22, 9]$		-50.40	$[-104, -6]$		-280.65	$[-496, -82]$	
	Q	7.93	$[5, 11]$		7.93	$[5, 11]$		7.93	$[5, 11]$	
$\beta = 1$	$\mathcal{R}$	7.61	$[4, 11]$		7.15	$[3, 11]$		-4.12	$[0, 11]$	
	Φ	7.79	$[4, 11]$	✓	7.82	$[4, 11]$	✓	7.71	$[4, 11]$	0.0066
$\beta = 2$	$\mathcal{R}$	7.10	$[3, 11]$		5.83	$[1, 10]$		-17.37	$[-50, 10]$	
	Φ	7.46	$[4, 11]$	✓	7.45	$[4, 11]$	0.0002	6.90	$[3, 11]$	0.012
$\beta = 4$	$\mathcal{R}$	5.96	$[1, 11]$		2.53	$[-6, 10]$		-23.60	$[-93, 8]$	
	Φ	6.81	$[3, 11]$	✓	6.53	$[2, 11]$	✓	3.49	$[0, 11]$	0.0108
$\beta = 8$	$\mathcal{R}$	3.68	$[-6, 10]$		-4.98	$[-21, 8]$		-31.36	$[-76, 4]$	
	Φ	5.29	$[-1, 11]$	✓	4.44	$[-2, 10]$	✓	-3.69	$[-10, 10]$	0.0036
$\beta = 16$	$\mathcal{R}$	0.93	$[-11, 10]$		-16.01	$[-41, 5]$		-52.15	$[-107, -7]$	
	Φ	2.72	$[-7, 10]$	✓	-0.39	$[-13, 9]$	✓	-14.91	$[-100, 9]$	0.001
$\beta = 32$	$\mathcal{R}$	-1.37	$[-16, 10]$		-27.39	$[-62, 1]$		-91.90	$[-177, -23]$	
	Φ	0.05	$[-13, 10]$	✓	-9.29	$[-30, 7]$	✓	-28.12	$[-100, 7]$	✓

(b) Gymnasium *Taxi*

Table 1: Mean episodic rewards of the converged policies under random-Dirichlet perturbations with convex coefficient κ. The labels b, Q, $\mathcal{R}$, and Φ correspond to the behavioral, Watkins', regularized, and gated agents, respectively. All means are obtained by evaluating 10 individual agents on $1 \cdot 10^5$ episodes with a maximum length of 100 steps. Brackets denote the $[10^{\text{th}}, 90^{\text{th}}]$ percentiles. Check-marks indicate where $V^{\Phi}(s) \geq V^{\mathcal{R}}(s)\ \forall s \in \mathcal{S}$ for all 10 agents, and the numbers indicate the fraction of states for which $V^{\Phi}(s) < V^{\mathcal{R}}(s)$. The mean rewards of the optimal policies for *CliffWalking* and *Taxi* are $-13[-13, -13]$ and $7.94[5, 11]$, respectively.

of the environment with convex coefficient κ. If all state-action pairs are reachable via random play, random-uniform perturbation with a large κ eventually results in a covering offline dataset viable for off-policy learning. As expected, it can be seen that high κ values allow the behavioral policy to cover most of the state-action space and make off-policy learning viable, as evident from the strong Watkins' Q-learner performance with the exception of the anomaly at $\kappa = 0.2$. In this setting, the gate nearly deactivates regularization with increasing κ, enabling the gated agent to remain near-optimal across a broad range of β values. Additionally, both experiments provide empirical validation of Theorem 2.1, as indicated by the check-marks.

	π	$\kappa = 0.2$			$\kappa = 0.5$			$\kappa = 0.8$		
	b	-83.58	$[-229, -13]$		-256.60	$[-648, -23]$		-714.32	$[-1288, -154]$	
	Q	-100.00	$[-100, -100]$		-13.00	$[-13, -13]$		-13.00	$[-13, -13]$	
$\beta = 1$	$\mathcal{R}$	-13.48	$[-15, -13]$		-14.33	$[-17, -13]$		-15.31	$[-18, -13]$	
	Φ	-13.14	$[-13, -13]$	✓	-13.01	$[-13, -13]$	✓	-13.00	$[-13, -13]$	✓
$\beta = 2$	$\mathcal{R}$	-14.23	$[-17, -13]$		-16.65	$[-21, -13]$		-19.70	$[-25, -15]$	
	Φ	-13.61	$[-15, -13]$	✓	-13.18	$[-14, -13]$	✓	-13.00	$[-13, -13]$	✓
$\beta = 4$	$\mathcal{R}$	-15.03	$[-18, -13]$		-19.60	$[-25, -15]$		-25.91	$[-35, -18]$	
	Φ	-14.39	$[-17, -13]$	✓	-14.08	$[-16, -13]$	✓	-13.00	$[-13, -13]$	✓
$\beta = 8$	$\mathcal{R}$	-15.66	$[-19, -13]$		-22.22	$[-29, -16]$		-32.83	$[-46, -22]$	
	Φ	-15.18	$[-18, -13]$	✓	-16.24	$[-20, -13]$	✓	-13.08	$[-13, -13]$	✓
$\beta = 16$	$\mathcal{R}$	-16.11	$[-20, -13]$		-24.21	$[-32, -17]$		-39.49	$[-57, -25]$	
	Φ	-15.75	$[-19, -13]$	✓	-19.16	$[-25, -15]$	✓	-13.79	$[-15, -13]$	✓
$\beta = 32$	$\mathcal{R}$	-17.86	$[-21, -13]$		-28.24	$[-36, -17]$		-48.86	$[-71, -27]$	
	Φ	-16.24	$[-20, -13]$	✓	-21.87	$[-29, -15]$	✓	-16.45	$[-20, -13]$	✓

(a) Gymnasium *CliffWalking*

	π	$\kappa = 0.2$			$\kappa = 0.5$			$\kappa = 0.8$		
	b	-4.67	$[-21, 9]$		-46.67	$[-92, -8]$		-228.50	$[-352, -90]$	
	Q	7.93	$[5, 11]$		7.93	$[5, 11]$		7.93	$[5, 11]$	
$\beta = 1$	$\mathcal{R}$	7.61	$[4, 11]$		7.17	$[3, 11]$		6.82	$[3, 11]$	
	Φ	7.82	$[4, 11]$	✓	7.92	$[5, 11]$	✓	7.93	$[5, 11]$	✓
$\beta = 2$	$\mathcal{R}$	7.08	$[3, 11]$		5.86	$[1, 10]$		4.88	$[0, 10]$	
	Φ	7.50	$[4, 11]$	✓	7.77	$[4, 11]$	✓	7.93	$[5, 11]$	✓
$\beta = 4$	$\mathcal{R}$	5.90	$[1, 11]$		2.47	$[-7, 10]$		-0.95	$[-13, 8]$	
	Φ	6.86	$[3, 11]$	✓	7.19	$[3, 11]$	✓	7.92	$[5, 11]$	✓
$\beta = 8$	$\mathcal{R}$	3.58	$[-6, 10]$		-5.32	$[-21, 8]$		-17.00	$[-41, 4]$	
	Φ	5.40	$[-1, 11]$	✓	5.90	$[1, 10]$	✓	7.82	$[4, 11]$	✓
$\beta = 16$	$\mathcal{R}$	0.81	$[-11, 10]$		-16.41	$[-41, 5]$		-46.61	$[-91, -8]$	
	Φ	2.86	$[-7, 10]$	✓	2.58	$[-6, 10]$	✓	7.30	$[4, 11]$	✓
$\beta = 32$	$\mathcal{R}$	-1.46	$[-16, 9]$		-27.22	$[-59, 0]$		-88.75	$[-163, -26]$	
	Φ	0.15	$[-13, 10]$	✓	-5.13	$[-21, 8]$	✓	5.94	$[1, 10]$	✓

(b) Gymnasium *Taxi*

Table 2: Mean episodic rewards of the converged policies under random-uniform perturbations with convex coefficient κ. The labels b, Q, $\mathcal{R}$, and Φ correspond to the behavioral, Watkins', regularized, and gated agents, respectively. All means are obtained by evaluating 10 individual agents on $1 \cdot 10^5$ episodes with a maximum length of 100 steps. Brackets denote the $[10^{\text{th}}, 90^{\text{th}}]$ percentiles. Checkmarks indicate cases where $V^{\Phi}(s) \geq V^{\mathcal{R}}(s)\ \forall s \in \mathcal{S}$ for all 10 agents. The mean rewards of the optimal policies for *CliffWalking* and *Taxi* are $-13[-13, -13]$ and $7.94[5, 11]$, respectively.

C.2 Environments with Continuous States and Actions

This section expands the experiment described in Section 4 with more environments from the Gymnasium MuJoCo library (Towers et al., 2024; Todorov et al., 2012), namely *Hopper* (Figure 2), *Ant* (Figure 3), and *HalfCheetah* (Figure 4). All models are trained on the policy-annotated D4RL (Fu et al., 2021) *medium* offline datasets, where *medium* refers to the quality of the behavioral policy. To annotate the datasets with their behavioral policies for each step, the underlying policy model is reconstructed from the model parameters provided in each dataset's metadata and applied to all data. A set of explicitly regularized learning methods, CQL (Kumar et al., 2020), TD3+BC (Fujimoto & Gu, 2021), and ReBRAC (Tarasov et al., 2023), is used to allow direct integration of the gate into their respective regularization steps. For all of these algorithms, their implementations provided by the d3rlpy library (Seno & Imai, 2022) are used. The gated variants are built based on these implementations by modifying the regularization of each algorithm with the addition of the continuous gate given by Equation (14) with scaling parameter $\rho = 0.5$. For comparison, we additionally run the experiments with $\rho = 2$. No additional algorithmic or code-level optimizations are included in the methods used throughout this experiment. Apart from the scaling parameter ρ, all hyperparameters are set to their d3rlpy default values, which adapt to each environment type, for all algorithms and their gated variants.

Across all plots, it can be seen that the effect of including the gate varies substantially depending on the algorithm, the environment, and the stage of training. However, one consistent observation is that the inclusion of the gate does not significantly degrade performance. Instead, it either significantly improves learning in the early stages of training, yields a policy with higher reward, or results in no significant change relative to the baseline algorithm's performance. A possible explanation for the differences in the observed effect is given by Table 3, which shows statistics of the Gaussian-policy standard deviations over states of the respective environments. It can be seen that the ordering of the spread of the policies' standard deviations across states matches the order in which the gated

Dataset	*Hopper*	*Ant*	*HalfCheetah*
std mean	0.366	0.138	0.414
std std	0.072	0.036	0.062
p10	0.288	0.106	0.347
p90	0.462	0.178	0.488

Table 3: Statistics of the Gaussian-policy standard deviations over the states in each environment's dataset, averaged over action dimensions.

variants positively differ from their baselines. That is, the more inhomogeneous the behavioral policy, the more improvement is obtained from the inclusion of the gate. Based on this observation, tuning the hyperparameters and deploying code-level optimizations with the gate included may yield faster learning and higher-performing policies than those obtained in the current experimental setup.

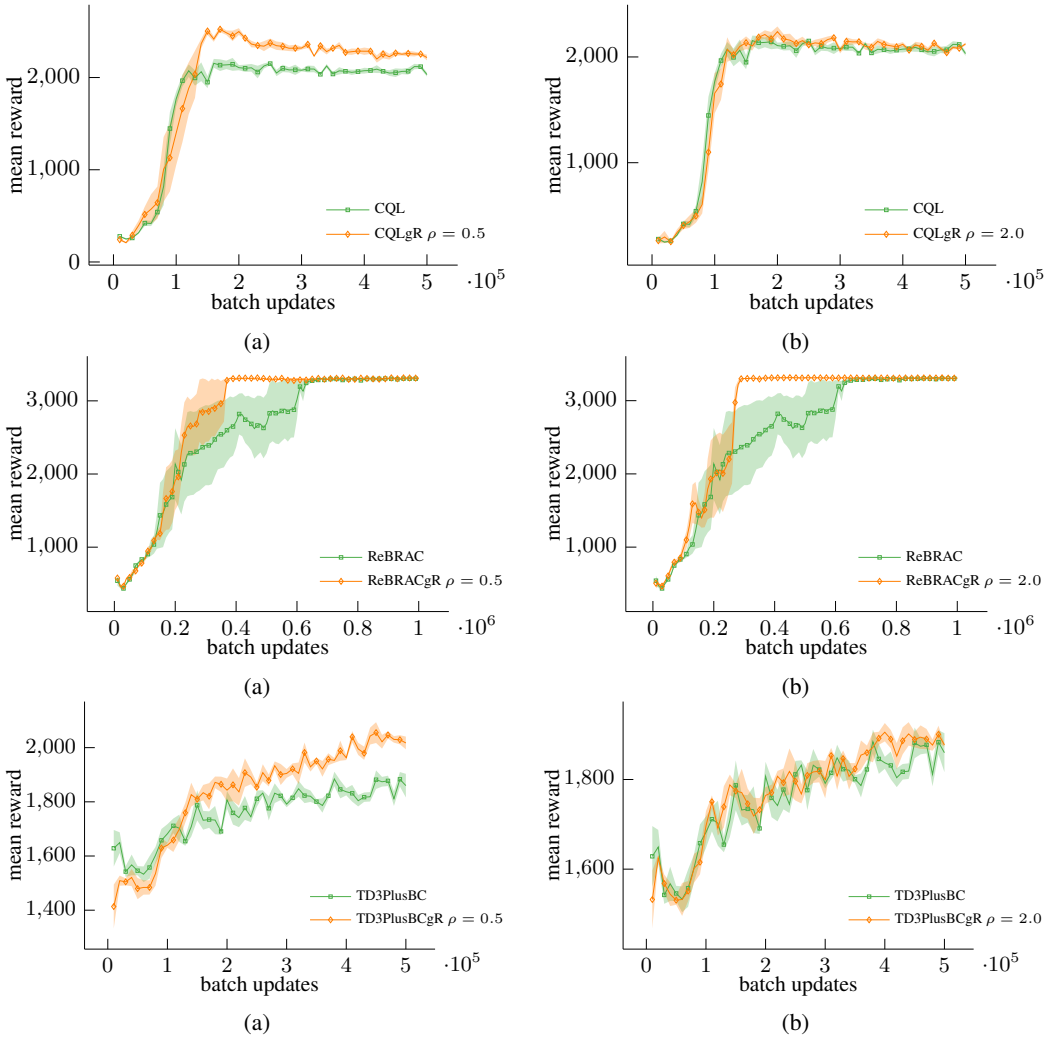


Figure 2: MuJoCo *Hopper* mean episodic rewards, with standard error bands, after *batch update* training steps with $\rho = 0.5$ (a) and $\rho = 2$ (b). Algorithms are as indicated in the legend, the gR postfix indicates the inclusion of the gate. Values are obtained by training agents on the D4RL medium dataset and averaging the results across 5 seeds in blocks of 10 evaluations over $1 \cdot 10^4$ batched update steps.

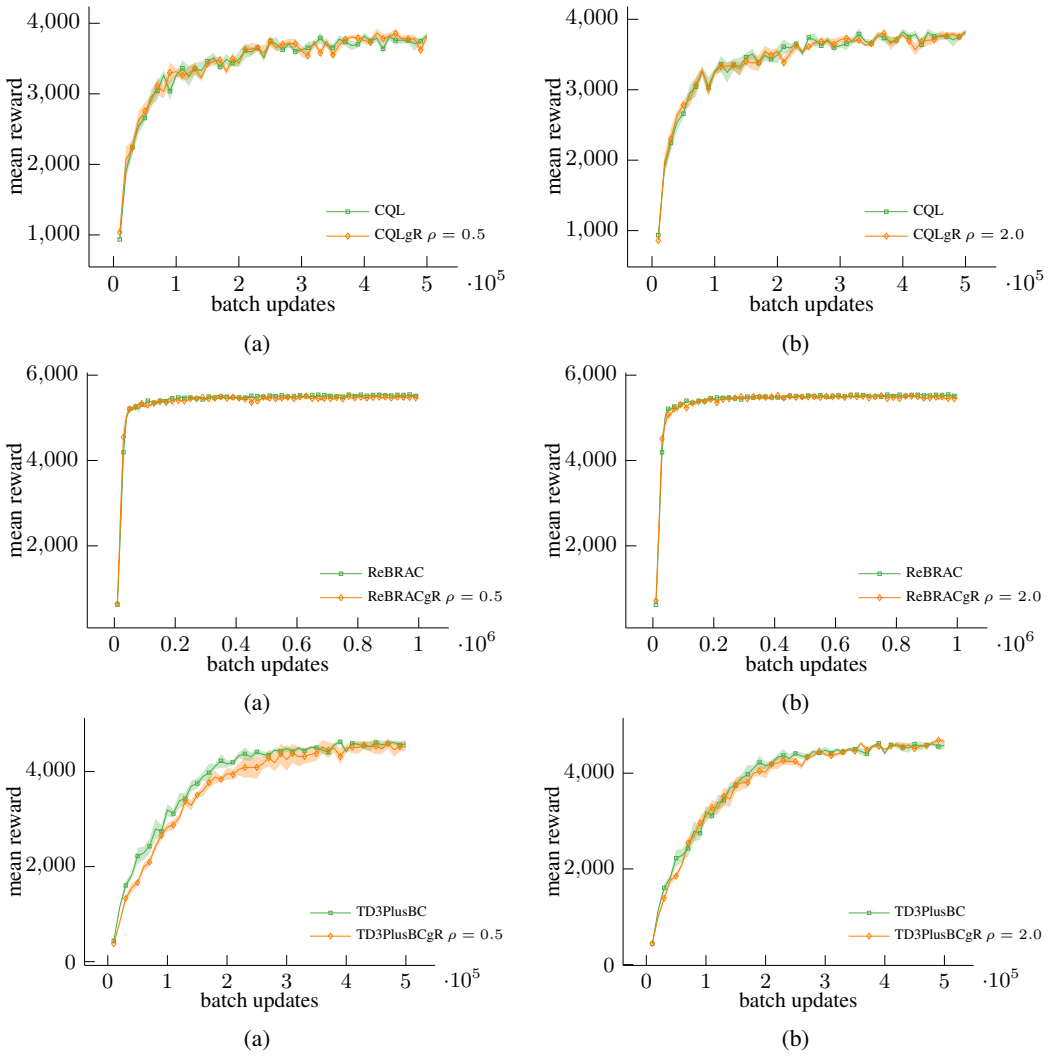


Figure 3: MuJoCo *Ant* mean episodic rewards, with standard error bands, after *batch update* training steps with $\rho = 0.5$ (a) and $\rho = 2$ (b). Algorithms are as indicated in the legend, the gR postfix indicates the inclusion of the gate. Values are obtained by training agents on the D4RL medium dataset and averaging the results across 5 seeds in blocks of 10 evaluations over $1 \cdot 10^4$ batched update steps.

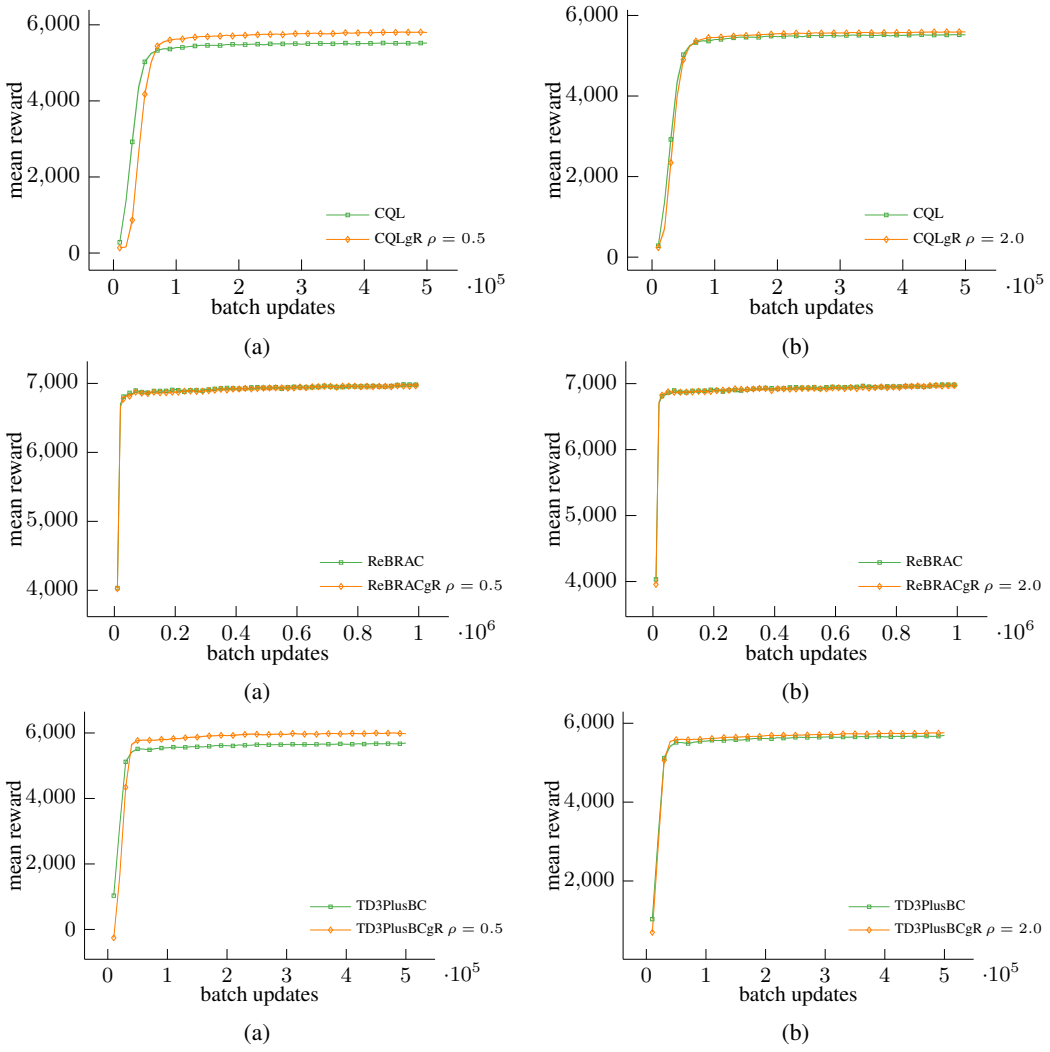


Figure 4: MuJoCo *HalfCheetah* mean episodic rewards, with standard error bands, after *batch update* training steps with $\rho = 0.5$ (a) and $\rho = 2$ (b). Algorithms are as indicated in the legend, the gR postfix indicates the inclusion of the gate. Values are obtained by training agents on the D4RL medium dataset and averaging the results across 5 seeds in blocks of 10 evaluations over $1 \cdot 10^4$ batched update steps.