MIXING REGULARIZATION FOR AVERAGE-REWARD Q-LEARNING AND Q-PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The average-reward setting for reinforcement learning is challenging because its Bellman operator is known to contract in span only when the underlying Markov chain is mixing. Whereas existing work typically *assumes* a contraction property, e.g. through environment mixing, we consider the setting where we are given a reward-agnostic m-step mixing policy, and we *control* the contraction and exploration properties of the learned policy by regularizing toward the mixing policy. We provide finite-time analysis of Q-learning algorithms that can be instantiated with multiple choices of regularization, with guarantees measured with respect to the regularized fixed point. Our framework exposes novel regularizer-dependent objects that control contraction, exploration, and operator bias. In the one-step setting, our rates match the best known rates for Q-iteration and synchronous Q-learning, and for m-step mixing we derive new rates. We also study Q-planning variants, and show that planning acts as time-localized learning, where the planning horizon H trades off the bias from the boundary value against the sample complexity. Finally, we present numerical studies validating the analysis: (i) knowledge of a strong mixing policy improves convergence and exploration, and its advantage increases with bottleneck severity and (ii) our theory predicts the convergence and bias rankings of regularizers from the α-divergence family.

1 INTRODUCTION

The average-reward setting of reinforcement learning is challenging because, without a discount factor, the contraction of the Bellman operator is solely determined by the closed-loop dynamics, formed by the transitions and the policy. The classical requirement for learning algorithms is the existence of a policy that mixes the induced Markov chain, which holds when the Markov decision process (MDP) is communicating and its transition graph is aperiodic. In that case, every full-support policy mixes, albeit at potentially very different rates.

However, the design of algorithms that exploit knowledge of a strongly mixing policy to improve the constants of their own convergence and exploration remains relatively understudied. This setting is natural in domains like robotics and fine-tuning, where new tasks are learned starting from an existing policy. Although regularization toward the mixing policy is a natural approach, its effects on average-reward learning are not understood, raising the question:

How does regularization toward a mixing policy control contraction, exploration, bias and, ultimately, sample complexity of average-reward Q-learning? How does this analysis inform regularization design?

We give the first finite-time analysis of average-reward Q-learning in which regularization toward a mixing policy provides both the contraction of the Bellman operator and the exploration needed to learn. Specifically, our main contributions are:

- **A framework for mixing regularization.** Given a reward-agnostic policy μ that mixes in m steps with parameter β_m, we identify regularizer-dependent quantities that drive contraction, exploration, and bias: the *domination profile* γ_Ω controls contraction, at rate $\tilde{\beta} = \beta_m \bar{\gamma}^{m-1}$, and exploration, guaranteeing coverage with probability $\beta_m \bar{\gamma}^m \eta_{\min}$ every m steps where $\eta_{\min}$ is the

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.

smallest mass of the minorizing measure. The *regularization modulus* controls the operator bias with a term scaling with $m/\tilde{\beta}$, thereby exposing a trade-off between faster contraction and larger bias. Unlike existing methods, which assume contraction conditions like uniform mixing or reachability, our assumption exploits knowledge from a single user-chosen policy.

- **Sample complexity of regularized Q-learning.** We analyze algorithms that isolate, in turn, contraction, statistical error, exploration and localization. Regularized Q-iteration converges linearly, with $O(|\mathcal{S}|^2|\mathcal{A}|\, m\, \tilde{\beta}^{-1} \log(1/\varepsilon))$ model queries; synchronous regularized Q-learning reaches ε-accuracy with high probability after $\tilde{O}(|\mathcal{S}||\mathcal{A}|\, m^4 \bar{w}^2 \tilde{\beta}^{-3} \varepsilon^{-2})$ samples where $\bar{w}$ bounds the noise; epoch regularized Q-learning, which learns from a single trajectory of its own greedy policies, needs $\tilde{O}(m^4 \bar{w}^2 \tilde{\beta}^{-4} (\bar{\gamma}\eta_{\min})^{-1} \varepsilon^{-2})$ transitions; regularized Q-planning, which restricts the learning process to the state-action pairs within H steps of the reset distribution, needs $\tilde{O}\big((h_1 + m/\tilde{\beta})^4 \bar{w}^2 q^{-1} \varepsilon^{-2}\big)$ transitions where h_1 is the depth after which the states reachable from the reset distribution no longer change and q is the minimum visitation probability across depth. In the last three cases, we match the optimal ε^{-2} dependence. Although direct comparison to state-of-the-art analysis is challenging because there is not a standard setting, we achieve comparable rates, summarized in Table 2.
- **Empirical validation.** In our first experiment, we empirically demonstrate that knowledge of a strong mixing policy can improve convergence and exploration properties, and its advantage increases with bottleneck severity. In our second experiment, we empirically demonstrate that our theory predicts the convergence and bias rankings of regularizers from the α-divergence family.

2 Background on Regularized Average-Reward MDPs

2.1 Average-Reward MDPs

We consider an infinite-horizon average-reward MDP $\mathcal{M} := (\mathcal{S}, \mathcal{A}, p, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ the transition probability and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the deterministic reward function. We assume that both $\mathcal{S}$ and $\mathcal{A}$ are finite. An action-value function is any $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

A policy π maps each state s to an action distribution: $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, inducing the state-action kernel $K_\pi(s', a'|s, a) := p(s'|s, a)\pi(a'|s')$ and the m-step state-action kernel $K_\pi^m(s', a'|s, a) := \sum_{\tilde{s},\tilde{a}} p(\tilde{s}|s, a)\, \pi(\tilde{a}|\tilde{s})\, K_\pi^{m-1}(s', a'|\tilde{s}, \tilde{a})$ with $K_\pi^0 := I_{\mathcal{S}\times\mathcal{A}}$.

The Bellman (optimality) operator is $\mathcal{T}(Q)(s, a) := r(s, a) + \mathbb{E}_{s'\sim p(\cdot|s,a)}[\max_{a'} Q(s', a')]$. The gain of a policy is $\rho^\pi(s) := \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}_\pi\big[\sum_{n=0}^{N-1} r(s_n, a_n) \,\big|\, s_0 = s\big]$, and the optimal gain is $\rho^*(s) := \sup_\pi \rho^\pi(s)$. Without discounting, $\mathcal{T}$ is not a contraction, and the relevant notion of fixed point is the optimality equation $\mathcal{T}(Q^*) = Q^* + \rho^*\mathbf{1}$: for communicating MDPs, ρ^* is constant and this equation has solutions (Schweitzer & Federgruen, 1978).

Because the optimality equation is translation invariant, it is standard to use the span seminorm:

Definition 1 (Span seminorms)**.** For a finite set $\mathcal{X}$ and $U \in \mathbb{R}^{\mathcal{X}}$, the span seminorm is $\|U\|_{\mathcal{X},\mathrm{sp}} := \max_{x\in\mathcal{X}} U(x) - \min_{x\in\mathcal{X}} U(x)$. This is used on $\mathbb{R}^{\mathcal{S}}$ and $\mathbb{R}^{\mathcal{A}}$, and for $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ we write $\|Q\|_{\mathrm{sp}} := \|Q\|_{\mathcal{S}\times\mathcal{A},\mathrm{sp}}$. We also define the *action-span seminorm* $\|Q\|_{\mathrm{asp}} := \max_{s\in\mathcal{S}} \|Q(s,\cdot)\|_{\mathcal{A},\mathrm{sp}}$. Both vanish on constants, $\|\cdot\|_{\mathrm{asp}}$ vanishes on every Q that depends only on the state, and $\|\cdot\|_{\mathrm{asp}} \le \|\cdot\|_{\mathrm{sp}}$.

2.2 Regularized MDPs

We use the regularized MDP framework of Geist et al. (2019). Let $\tau > 0$ be a temperature and, for each $s \in \mathcal{S}$, let $\Omega_s : \Delta(\mathcal{A}) \to \mathbb{R} \cup \{+\infty\}$ be a strongly convex function. With the shorthands $Q_s(a) := Q(s, a)$ and $\pi_s(a) := \pi(a|s)$, the regularized value function $\Omega^*(Q) \in \mathbb{R}^{\mathcal{S}}$ and the greedy policy $\pi_{\Omega,Q} : \mathcal{S} \to \Delta(\mathcal{A})$ are

$$\Omega^*(Q)(s) := \max_{\pi_s\in\Delta(\mathcal{A})} \{\langle \pi_s, Q_s\rangle - \tau\, \Omega_s(\pi_s)\}, \quad \pi_{\Omega,Q}(\cdot|s) := \arg\max_{\pi_s\in\Delta(\mathcal{A})} \{\langle \pi_s, Q_s\rangle - \tau\, \Omega_s(\pi_s)\},$$

so that $\Omega^*(\cdot)(s)$ is the convex conjugate of $\tau\Omega_s$, and the regularized Bellman operator is $\mathcal{T}_\Omega(Q)(s, a) := r(s, a) + \mathbb{E}_{s'\sim p(\cdot|s,a)}[\Omega^*(Q)(s')]$. Under Assumption 2 below, the maximizing policy is unique and equals $\nabla\Omega^*(Q)(s)$.

Examples include the α-divergences to a policy μ, such as the forward and reverse KL.

3 MIXING-REGULARIZED MDPS

We state our assumptions, define the objects that govern mixing regularization, and derive their effect on contraction, exploration and bias (Lemma 1); proofs are deferred to Appendix G.1.

3.1 ASSUMPTIONS

Assumption 1 (The reference policy μ is m-step Doeblin mixing)**.** There exists a stationary full-support policy $\mu : \mathcal{S} \to \Delta(\mathcal{A})$, an integer $m \geq 1$, a parameter $\beta_m \in (0, 1]$, and a full-support distribution $\nu \in \Delta(\mathcal{S})$ such that, $\forall (s, a, s')$, $\sum_{a'} K_\mu^m(s', a'|s, a) \geq \beta_m \nu(s')$. Furthermore, let $\eta(s, a) := \nu(s)\mu(a|s)$ and $\eta_{\min} := \min_{(s,a)} \eta(s, a)$. The m-step Doeblin condition implies that the MDP is communicating so ρ^* is constant and $\mathcal{T}(Q^*) = Q^* + \rho^* \mathbf{1}$ has solutions.

Assumption 2 (Convex regularity of Ω)**.** The temperature τ is positive and, for every $s \in \mathcal{S}$, Ω_s is strongly convex, closed and proper.[1]

Assumption 3 (Admissibility of the mixing policy)**.** For every $s \in \mathcal{S}$ we have $\mu_s \in \mathrm{dom}(\Omega_s)$, the effective domain of Ω_s.

3.2 MIXING PROPERTIES

Definition 2 (Mixing-regularization objects)**.** Let Ω satisfy Assumption 2, let μ be the mixing policy of Assumption 1, and let $R \geq 0$.

(i) *Radius*: for $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $R[Q, Q'] := \max\{\|Q\|_{\mathrm{asp}}, \|Q'\|_{\mathrm{asp}}\}$ is the largest action span on the segment $[Q', Q]$ (by convexity of $\|\cdot\|_{\mathrm{asp}}$).

(ii) *Domination profile*: $\gamma_\Omega(R) := \inf\left\{\pi_{\Omega,Q}(a|s)/\mu(a|s) \mid (s, a) \in \mathcal{S} \times \mathcal{A},\ \|Q\|_{\mathrm{asp}} \leq R\right\}$ is the largest factor such that $\pi_{\Omega,Q} \geq \gamma_\Omega(R)\,\mu$ for every Q of action span at most R. It lies in $[0, 1]$ and is non-increasing in R.

(iii) *Regularization modulus*: $\Lambda_\Omega(R) := \sup\left\{\|g_Q\|_{\mathcal{S},\mathrm{sp}} \mid \|Q\|_{\mathrm{asp}} \leq R\right\}$, where $g_Q(s) := \Omega^*(Q)(s) - \max_a Q(s, a)$ is the regularization gap. It is finite and non-decreasing in R.

Our finite-time results also require that greedy policies are full support. This is equivalent to the following property:

$$\gamma_\Omega(R) > 0, \quad \text{for every finite } R \geq 0. \tag{P1}$$

Many regularizers satisfy (P1) by default, including the α-divergences to μ with $\alpha \leq 1$. For those whose greedy policies can be sparse, such as α-divergences with $\alpha > 1$, an ϵ-floor that mixes the greedy policy with μ restores it[2].

Under Assumptions 1 to 3, the regularized optimality equation $\mathcal{T}_\Omega(Q^*_\Omega) = Q^*_\Omega + \rho^*_\Omega \mathbf{1}$ has a solution, ρ^*_Ω is unique, and under (P1) Q^*_Ω is unique up to an additive constant (Theorem G.1)[3].

Lemma 1 (Mixing properties of Ω-regularization)**.** Under Assumptions 1 and 2 and for any Q, Q':

(i) *Contraction*: with $\mathcal{T}^k_\Omega$ the k-th composition of the regularized Bellman operator,

$$\|\mathcal{T}^m_\Omega(Q) - \mathcal{T}^m_\Omega(Q')\|_{\mathrm{sp}} \leq \left(1 - \beta_m \prod_{k=1}^{m-1} \gamma_\Omega(R[\mathcal{T}^k_\Omega(Q), \mathcal{T}^k_\Omega(Q')])\right) \|Q - Q'\|_{\mathrm{sp}}.$$

(ii) *Coverage*: let $(s_n, a_n)_{n\geq 0}$ be a trajectory with $a_n \sim \pi_{\Omega,Q_n}(\cdot|s_n)$ and $s_{n+1} \sim p(\cdot|s_n, a_n)$, where Q_n is $\mathcal{F}_{n-1}$-measurable for $\mathcal{F}_n := \sigma(Q_0, (s_k, a_k)_{k\leq n})$. If $\gamma_1, \ldots, \gamma_m \geq 0$ are $\mathcal{F}_t$-measurable and $\pi_{\Omega,Q_{t+k}} \geq \gamma_k \mu$ almost surely, then almost surely

$$\mathbb{P}\big((s_{t+m}, a_{t+m}) = (s, a) \mid \mathcal{F}_t\big) \geq \beta_m \left(\prod_{k=1}^{m} \gamma_k\right) \eta(s, a), \quad \text{for every } (s, a).$$

[1] Strict convexity suffices for all our results; strong convexity only provides the Lipschitz constant of the greedy map in Lemma F.1.(i) (Remark F.1). [2] Appendix G.5 [3] Existence holds more generally, Appendix G.3

(iii) *Bias*: the one step operator gap is: $\|\mathcal{T}_\Omega(Q) - \mathcal{T}(Q)\|_{\mathrm{sp}} \le \Lambda_\Omega(\|Q\|_{\mathrm{asp}})$. Moreover, if Assumption 3 and (P1) hold, then for every solution (Q^*, ρ^*) of $\mathcal{T}(Q^*) = Q^* + \rho^* \mathbf{1}$,

$$\|Q^*_\Omega - Q^*\|_{\mathrm{sp}} \le \frac{m\,\Lambda_\Omega(R^*)}{\beta_m \gamma_\Omega(R^*)^{m-1}}, \qquad R^* := \max\Big\{\|Q^*_\Omega\|_{\mathrm{asp}}, \max_{0\le k<m} \|\mathcal{T}^k_\Omega(Q^*)\|_{\mathrm{asp}}\Big\},$$

and, if $|\mathcal{S}| \ge 2$ and all the Ω_s have the same minimum $\underline{\Omega}$, as divergences to μ do with $\underline{\Omega} = 0$,

$$-\Lambda_\Omega(\|Q^*\|_{\mathrm{asp}}) \le \rho^*_\Omega - \rho^* + \tau\underline{\Omega} \le 0.$$

The key observation behind the proof is that the difference of the operators applied to any two Q-functions can be written exactly as a single kernel applied to their difference: $\mathcal{T}_\Omega(Q) - \mathcal{T}_\Omega(Q') = K_{\bar\pi}(Q - Q')$, where $\bar\pi$ is the average of the greedy policies along the segment $[Q', Q]$.

Although existing work factorizes differences of hard maxima with a single policy kernel (Chen et al., 2026, Lemma B.6), ours makes this policy explicit and uses its properties: $\bar\pi$ inherits the domination of μ by the greedy policies, so composing m such kernels transfers the Doeblin condition of Assumption 1, which gives (i), and the same domination along a trajectory gives (ii). The profile γ_Ω thus controls contraction and exploration at once.

The last item exposes the trade-off between bias and convergence. The greedy policy only depends on Q/τ, so $\gamma_\Omega(R)$ only depends on R/τ, and a larger temperature raises it, hence the contraction and coverage rates; but it weakens the control of the modulus: for regularizers bounded on the simplex, $\Lambda_\Omega \le \tau\Delta_\Omega$ uniformly in the radius, with Δ_Ω the range of Ω (Remark G.1), a bound that grows linearly in τ. The gain gap, in contrast, involves neither the effective horizon $H_{\mathrm{eff}} := m/(\beta_m \gamma_\Omega^{m-1})$ nor γ_Ω, only the modulus at the τ-independent radius $\|Q^*\|_{\mathrm{asp}}$: as $\tau \to 0$, unlike the value bound, it vanishes linearly for bounded regularizers.

4 Algorithms and Finite-Time Analysis

We analyze increasingly realistic algorithms, each isolating one mechanism of Lemma 1: Q-iteration the contraction, synchronous Q-learning the statistical error, epoch Q-learning the cost of exploration, and Q-planning the effect of localization. Missing pseudocodes are given in Appendix B and proofs in Appendices H and J.

Throughout this section, let $\delta \in (0,1)$ be the failure probability, T be the budget of the algorithm and $e_0 := \|Q_0 - Q^*_\Omega\|_{\mathrm{sp}} > 0$ be the initial error. The iterates stay, with high probability, within span distance $2e_0$ of Q^*_Ω. On that ball, $\tilde\beta := \beta_m \bar\gamma^{\,m-1}$ with $\bar\gamma := \gamma_\Omega(\|Q^*_\Omega\|_{\mathrm{asp}} + 2e_0)$ is the contraction rate of Lemma 1.(i), $\bar w := \|Q^*_\Omega\|_{\mathrm{sp}} + 2e_0 + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}}$ bounds the noise, and $\iota := \log(6|\mathcal{S}||\mathcal{A}|(T+1)/\delta)$ is the price of a union bound over pairs and time.

4.1 Synchronous Ω-Regularized Q-Iteration

Synchronous Ω-regularized Q-iteration applies the update:

$$Q_{t+1}(s,a) = \Pi\left[\mathcal{T}_\Omega(Q_t)\right](s,a), \qquad \text{for all } (s,a),$$

where Π is the mean-centering projection. This algorithm isolates the core convergence mechanism of regularization under m-step mixing, on which the model-free analyses below build.

An immediate corollary of Lemma 1.(i) is the following.

Corollary 1 (Synchronous Ω-regularized Q-iteration)**.** Let Assumptions 1 to 3 and (P1) hold. Then $(\|Q_t - Q^*_\Omega\|_{\mathrm{sp}})_{t\ge0}$ is non-increasing and, for every $t \ge 0$,

$$\|Q_{t+m} - Q^*_\Omega\|_{\mathrm{sp}} \le \Big(1 - \beta_m\,\gamma_\Omega\big(\|Q^*_\Omega\|_{\mathrm{asp}} + \|Q_t - Q^*_\Omega\|_{\mathrm{sp}}\big)^{m-1}\Big)\|Q_t - Q^*_\Omega\|_{\mathrm{sp}}. \tag{1}$$

Consequently, for $\varepsilon \in (0, e_0)$, $\|Q_T - Q^*_\Omega\|_{\mathrm{sp}} \le \varepsilon$ for every $T \ge m\lceil\tilde\beta^{-1}\log(e_0/\varepsilon)\rceil$, i.e. after $O\big(|\mathcal{S}|^2|\mathcal{A}|\,m\,\tilde\beta^{-1}\log(e_0/\varepsilon)\big)$ model queries, one query being an evaluation of $p(s'|s,a)$.

For $m = 1$, the domination profile is not needed: $\tilde\beta = \beta_1$ and the rate is $1 - \beta_1$ per sweep, i.e. $O(|\mathcal{S}|^2|\mathcal{A}|\,\beta_1^{-1}\log(e_0/\varepsilon))$ model queries, which recovers the rate of relative value iteration under a one-step minorization of the kernel (Puterman, 1994).

Our result in (1) reveals a contraction rate that increases as the error shrinks: from $\tilde{\beta}$ to $\beta_m \gamma_\Omega(\|Q_\Omega^*\|_{\text{asp}})^{m-1}$. Although all of our analysis shares this feature, we present sample complexity results assuming the slowest convergence rate for simplicity.

4.2 Synchronous Ω-Regularized Q-Learning

Synchronous Ω-regularized Q-learning applies the update:

$$Q_{t+1}(s,a) = (1-\alpha_t)Q_t(s,a) + \alpha_t(r(s,a) + \Omega^*(Q_t)(s_t'(s,a))), \qquad \text{for all } (s,a),$$

where $s_t(s,a)$ is sampled from $p(.|s,a)$ for each (s,a) at each iteration. This algorithm isolates the statistical cost of learning from data.

Theorem 1 (Synchronous Ω-regularized Q-learning)**.** Let Assumptions 1 to 3 and (P1) hold, and set $\alpha(u) := (\tilde{\beta}^2 u^2)(9m(2+\tilde{\beta})^2(1+9m)^2\bar{w}^2\iota)^{-1}$ for $u > 0$. Run Algorithm 3 for T iterations with a constant step size $\alpha \le \alpha(e_0)$. Then, with probability at least $1-\delta$, $\sup_{t \le T} \|Q_t - Q_\Omega^*\|_{\text{sp}} \le 2e_0$ and

$$\|Q_T - Q_\Omega^*\|_{\text{sp}} \le \frac{(1-\tilde{\beta}/2)^{\frac{2\alpha T}{9m}}}{1-\tilde{\beta}/2}\|Q_0 - Q_\Omega^*\|_{\text{sp}} + \frac{3(2+\tilde{\beta})(1+9m)}{\tilde{\beta}}\bar{w}\sqrt{m\alpha\iota}. \tag{2}$$

In particular, for $\varepsilon \in (0, e_0)$, the choices $\alpha = \alpha(\varepsilon/2)$ and $T \ge \frac{9m}{\alpha\tilde{\beta}}\log\left(\frac{2e_0}{(1-\tilde{\beta}/2)\varepsilon}\right)$ give $\|Q_T - Q_\Omega^*\|_{\text{sp}} \le \varepsilon$, i.e. $\tilde{O}\big(|\mathcal{S}||\mathcal{A}|\, m^4\bar{w}^2/(\tilde{\beta}^3\varepsilon^2)\big)$ sampled transitions.

Because this algorithm updates every state-action pair at every iteration, the effect of γ_Ω is limited to contraction, not exploration. The sample complexity's dependence on contraction is $\tilde{\beta}_m^{-3}$ or, equivalently, $(\beta\bar{\gamma}^{m-1})^{-3}$. In the comparable setting where $m = 1$, the dependence on γ disappears, and, for unit rewards, we match[4] the rate of Zhang et al. (2021) i.e., $\tilde{O}(|\mathcal{S}||\mathcal{A}|\beta_1^{-5}\varepsilon^{-2})$.

4.3 Epoch Ω-Regularized Q-Learning

The pseudocode of epoch Ω-regularized Q-learning is given in Algorithm 1. It adds the effect of exploration because it learns from a single trajectory, rather than querying a generative model at arbitrary (s,a) pairs.

Algorithm 1 Epoch Ω-Regularized Q-Learning

Input: initial Q_0, regularizer Ω, step sizes $\{\alpha_j\}$, batch sizes $\{B_j\}$, initial state s_0.
Set $n \leftarrow 0$ and $j \leftarrow 0$.
while true do
  $N_j \leftarrow 0$ and $\widehat{\mathcal{T}_\Omega Q_j} \leftarrow 0$ on $\mathcal{S} \times \mathcal{A}$.
  while $\min_{(s,a)} N_j(s,a) < B_j$ **do**
    Sample $a_n \sim \pi_{\Omega,Q_j}(\cdot|s_n)$ and $s_{n+1} \sim p(\cdot|s_n, a_n)$.
    if $N_j(s_n,a_n) < B_j$ **then**
      $\widehat{\mathcal{T}_\Omega Q_j}(s_n,a_n) \mathrel{+}= r(s_n,a_n) + \Omega^*(Q_j)(s_{n+1})$ and $N_j(s_n,a_n) \mathrel{+}= 1$.
    end if
    $n \leftarrow n+1$.
  end while
  $Q_{j+1} \leftarrow \Pi\big[(1-\alpha_j)Q_j + \alpha_j\widehat{\mathcal{T}_\Omega Q_j}/B_j\big]$ and $j \leftarrow j+1$.
end while

Theorem 2 (Epoch Ω-regularized Q-learning with $\alpha_j \equiv 1$)**.** Let Assumptions 1 to 3 and (P1) hold, and set $B(u) := \frac{8m^2\bar{w}^2\iota}{(\tilde{\beta}u)^2}$ for $u > 0$. Run Algorithm 1 with $\alpha_j \equiv 1$ and a constant batch size $B \ge B(e_0)$, and let $J(T)$ be the number of epochs completed within T transitions. Then, with probability at least $1-\delta$, $\sup_{j \le J(T)} \|Q_j - Q_\Omega^*\|_{\text{sp}} \le 2e_0$, each epoch completed within the budget lasts at most $L := m\lceil(\sqrt{B}+\sqrt{\iota})^2/(\bar{\gamma}\tilde{\beta}\eta_{\min})\rceil + 1$ transitions, so that $J(T) \ge \lfloor T/L \rfloor$, and

$$\|Q_{J(T)} - Q_\Omega^*\|_{\text{sp}} \le (1-\tilde{\beta})^{\lfloor J(T)/m \rfloor}\|Q_0 - Q_\Omega^*\|_{\text{sp}} + \frac{2\sqrt{2}\,m\,\bar{w}}{\tilde{\beta}}\sqrt{\frac{\iota}{B}}. \tag{3}$$

[4] Corollary H.3

In particular, for $\varepsilon \in (0, e_0)$, the choices $B = \lceil B(\varepsilon/2) \rceil$ and $T \geq mL\lceil \tilde{\beta}^{-1} \log(2e_0/\varepsilon) \rceil$ give $\|Q_{J(T)} - Q^*_\Omega\|_{\mathrm{sp}} \leq \varepsilon$, i.e. $\tilde{O}\big(m^4 \bar{w}^2/(\tilde{\beta}^4 \bar{\gamma} \eta_{\min} \varepsilon^2)\big)$ sampled transitions.

Compared to synchronous Q-learning, the sample complexity worsens by a factor of $(\bar{\gamma}\tilde{\beta}\eta_{\min})^{-1}/|\mathcal{S}||\mathcal{A}|$. This is a natural quantity to pay in the case of exploration because $\bar{\gamma}\tilde{\beta}\eta_{\min}$ controls the per-pair visitation probability under greedy policy rollout, whereas a generative model can query arbitrary state-action pairs. The exploration price can be reduced from $\bar{\gamma}\tilde{\beta}\eta_{\min}$ to $\beta_m \eta_{\min}$ by rolling out μ, but the mixing policy μ will not seek to optimize the reward[5].

4.4 Ω-Regularized Q-Planning

We present theorems and proof sketches from Appendices I and J. From a reset distribution $\xi \in \Delta(\mathcal{S})$, planning rolls out trajectories of length H and backpropagates a value boundary condition Q_B, similar to Monte Carlo Tree Search (Grill et al., 2020). Planning isolates the role of localization: instead of targeting Q^*_Ω, it targets the planning solution defined by the backwards recursion: $Q^*_{\Omega,h} := \mathcal{T}_{\Omega,h}(Q^*_{\Omega,h+1})$ and $Q^*_{\Omega,H} := Q_B|_{\mathcal{S}_H \times \mathcal{A}}$, where Q_h and $\mathcal{T}_{\Omega,h}$ are the depth-dependent value function and Bellman operator, and $\mathcal{S}_h$ is the set of h-step reachable states. In other words, the planning solution is the best update possible with the information available in the first H steps.

Given this setting, we relax the mixing requirement from Assumption 1 to only consider the reachable states from ξ.

Assumption 4 (Localized m-step Doeblin mixing)**.** There exist integers $m \geq 1$ and $h_1 \geq 0$, a full-support policy μ, $\beta_m \in (0, 1]$ and $\nu \in \Delta(\mathcal{S})$ with support $\mathcal{S}_\infty$ such that (i) $\mathcal{S}_h = \mathcal{S}_\infty$ for every $h \geq h_1$, and (ii) $\sum_{a'} K^m_\mu(s', a'|s, a) \geq \beta_m \nu(s')$ for all $(s, a) \in \mathcal{S}_\infty \times \mathcal{A}$ and $s' \in \mathcal{S}_\infty$.

The first planning algorithm is the single backward sweep: $Q_h = \mathcal{T}_{\Omega,h}(Q_{h+1})$. This case terminates and returns the planning solution exactly after $O(H(\max_h |\mathcal{S}_h|)^2|\mathcal{A}|)$ model queries. Its root level is ε-accurate for Q^*_Ω once $H \geq h_1 + m + H_{\mathrm{eff}} \log_+(\Delta_B/\varepsilon)$[6]. If H is too small, there is insufficient backup information, compared to the boundary mismatch Δ_B.

The last algorithm we consider estimates each depth operator from rollouts. Whereas the learning algorithms iterate the stationary update $Q_{j+1} = \mathcal{T}_\Omega(Q_j)$, the planning algorithm applies, up to sampling noise, the depth-shifted update $Q_{j+1,h} = \mathcal{T}_{\Omega,h}(Q_{j,h+1})$: level h at epoch $j+1$ is computed from level $h+1$ at epoch j, which is frozen during epoch j. In particular, choosing $H \geq h_1 + m + H_{\mathrm{eff}} \log_+(2\Delta_B/\varepsilon)$, gives $\|Q_{J,0} - Q^*_\Omega\|_{\mathrm{sp},\mathcal{S}_0 \times \mathcal{A}} \leq \varepsilon$ with high probability, after $\tilde{O}\big((h_1 + H_{\mathrm{eff}})^4 \bar{w}^2/(q\,\varepsilon^2)\big)$ sampled transitions[7], where q is the minimum visitation probability across depth.

The root error is a sum of level residuals damped at rate $\tilde{\beta}$ every m levels, so noise does not accumulate over the horizon and only the $h_1 + H_{\mathrm{eff}}$ levels closest to the root contribute. Merging the H tables into a single table learned from resets removes the extra factor $1/\tilde{\beta}$.

5 Instances and Experiments

5.1 Instances of Base Regularizers

We instantiate our regularizers with the α-divergences from Amari (Cichocki & Amari, 2010):

$$D_\alpha(p\|q) := \left(\sum_a p_a^\alpha q_a^{1-\alpha} - 1\right) / \big(\alpha(\alpha - 1)\big).$$

These include $D_1 = \mathrm{KL}(p\|q)$ and $D_0 = \mathrm{KL}(q\|p)$ by continuity, which have been used for KL-regularized reinforcement learning (Vieillard et al., 2020), the χ^2-divergence underlying advantage-weighted updates (Belousov & Peters, 2019), and, for a uniform μ, the Tsallis entropies up to a rescaling of τ (Lee et al., 2018).

We define $\Omega_{\alpha,s} := D_\alpha(\cdot\|\mu_s)$ and we compute the relevant quantities in Table 1[8]. In particular, on the window $0 < \alpha \leq 1$, the divergences exhibit two favorable properties: the modulus is bounded uniformly in R and the greedy policy is full support.

[5] Remark H.7 [6] Proposition J.1 [7] Theorem I.1 [8] Notation and proofs in Appendix K.

Table 1: The α-divergence regularizers

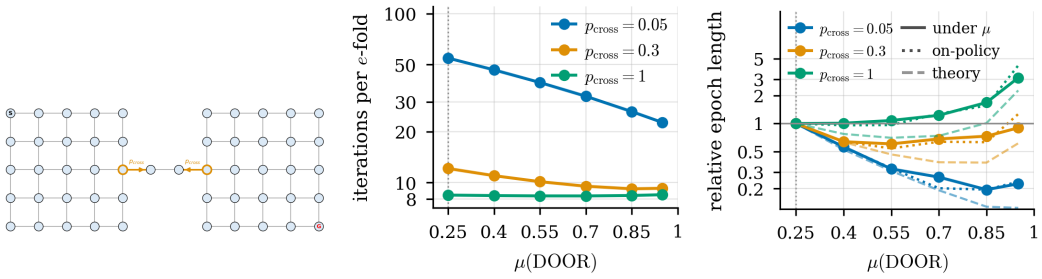
	general α	reverse KL $\alpha = 1$	forward KL $\alpha = 0$	Hellinger $\alpha = \frac{1}{2}$	χ^2 $\alpha = 2$
$\Omega_{\alpha,s}(\pi)$	$D_\alpha(\pi\|\mu_s)$	$\mathrm{KL}(\pi\|\mu_s)$	$\mathrm{KL}(\mu_s\|\pi)$	$2\|\sqrt{\pi} - \sqrt{\mu_s}\|_2^2$	$\frac{1}{2}\chi^2(\pi\|\mu_s)$
strong convexity	$\mu_{\min}^{(1-\alpha)_+}$ if $\alpha \le 2$	1	$\mu_{\min}$	$\sqrt{\mu_{\min}}$	1
$\pi_{\Omega,Q}(a\|s)/\mu_a$	$\left[1 + \frac{(\alpha-1)(Q_a-\zeta_s)}{\tau}\right]_+^{\frac{1}{\alpha-1}}$	$\frac{e^{Q_a/\tau}}{\sum_b \mu_b e^{Q_b/\tau}}$	$\left(1 + \frac{\zeta_s - Q_a}{\tau}\right)^{-1}$	$\left(1 + \frac{\zeta_s - Q_a}{2\tau}\right)^{-2}$	$\left[1 + \frac{Q_a - \zeta_s}{\tau}\right]_+$
$\gamma_\Omega(R) \ge$	$\exp_{2-\alpha}(-R/\tau)$	$e^{-R/\tau}$	$\frac{\tau}{\tau+R}$	$\left(1 + \frac{R}{2\tau}\right)^{-2}$	$\left[1 - \frac{R}{\tau}\right]_+$
$\Lambda_\Omega(R) \le$	$\tau \frac{\mu_{\min}^{1-\alpha}-1}{\alpha(\alpha-1)}$ if $\alpha > 0$	$\tau \log \frac{1}{\mu_{\min}}$	$\tau\left(1 + \log_+ \frac{R}{\tau}\right)$	$4\tau(1 - \sqrt{\mu_{\min}})$	$\frac{\tau}{2}\left(\frac{1}{\mu_{\min}} - 1\right)$
(P1) holds	iff $\alpha \le 1$	yes	yes	yes	no
Λ_Ω bounded in R	iff $\alpha > 0$	yes	no	yes	yes

5.2 Experiments

Does knowledge of μ improve convergence and exploration? We consider the *Two Rooms* MDP (Figure 1a): two 5×5 gridworld rooms joined by a two-cell doorway, with the four cardinal moves as actions. Moves are deterministic, except entering the doorway, which succeeds with probability p_{cross} and otherwise leaves the agent in place. The agent starts in the top-left corner and receives a reward of 1 in the bottom-right corner of the right room. The reference policy μ is uniform except at the two doorway entrances, where the entering action has probability $\mu(\text{DOOR})$ (0.25 is uniform): the doorway is a bottleneck, and μ can increase the flow through it. We use the ϵ-floored reverse KL ($\tau = 2$, $\epsilon = 0.05$) and sweep $\mu(\text{DOOR}) \in \{0.25, 0.4, 0.55, 0.7, 0.85, 0.95\}$ and $p_{\text{cross}} \in \{0.05, 0.3, 1\}$; details and the bias are in Appendix C.1.

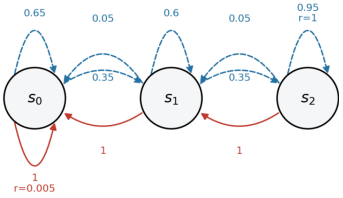
Fig. 1b measures convergence with synchronous regularized Q-iteration (Algorithm 2), which isolates contraction from sampling noise, through the number of iterations per e-fold, $1/\rho$, where $\|Q_k - Q^*_\Omega\|_{\text{sp}} \propto e^{-\rho k}$ once the rate has settled. It grows with the bottleneck severity and decreases with $\mu(\text{DOOR})$ when the doorway is a bottleneck ($\times 2.4$ at $p_{\text{cross}} = 0.05$), but is flat without one, consistent with the direction predicted by the contraction mechanism.

Fig. 1c measures exploration with epoch regularized Q-learning (Algorithm 1), through the epoch length relative to the uniform μ: solid, data collected under μ; dotted, on-policy; dashed, the coverage constant $\min_m m/(\beta_m \eta_{\min})$. Biasing μ towards the doorway shortens epochs by up to $\times 5$ at $p_{\text{cross}} = 0.05$ but triples them at $p_{\text{cross}} = 1$: knowledge of μ helps only when it targets an actual bottleneck. The coverage constant reproduces the shape of this dependence, though not its scale.

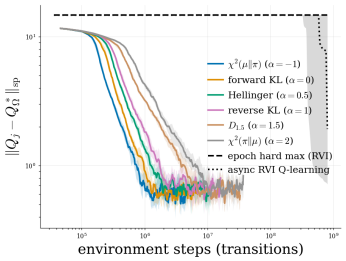


(1a) The Two-Rooms MDP. (1b) μ's effect on convergence. (1c) μ's effect on exploration.

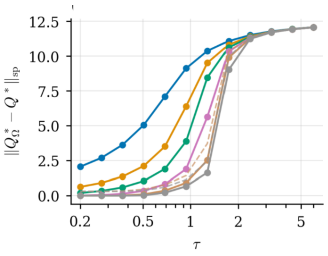
Does theory predict regularizer ranking? We consider the RiverSwim MDP (Strehl & Littman, 2008), with the parameterization of Osband et al. (2013) (Fig. 2a), where $\mathcal{S} = \{0, \ldots, n-1\}$ and $\mathcal{A} = \{\text{LEFT}, \text{RIGHT}\}$. LEFT deterministically moves the agent to $s - 1$, with a small reward $r(0, \text{LEFT}) = 0.005$ at the leftmost state. RIGHT swims against the current: it moves to $s + 1$ with probability 0.35, stays with probability 0.6 and moves to $s - 1$ with probability 0.05, with a large reward $r(n-1, \text{RIGHT}) = 1$ at the rightmost state. The reference policy μ is uniform and $n = 6$. We compare regularizers from the α-divergence family $D_\alpha(\pi\|\mu)$: $\chi^2(\mu\|\pi)$ ($\alpha = -1$), forward KL ($\alpha = 0$), Hellinger ($\alpha = 0.5$), reverse KL ($\alpha = 1$), $D_{1.5}$ and $\chi^2(\pi\|\mu)$ ($\alpha = 2$).



(2a) River Swim MDP ($n = 3$).



(2b) Ω's effect on convergence.



(2c) Ω's effect on bias.

Fig. 2c plots the bias $\|Q^*_\Omega - Q^*\|_{\mathrm{sp}}$ against τ (solid: without floor; dashed: the $\alpha > 1$ regularizers wrapped by the ϵ-floor, $\epsilon = 0.05$). At a common temperature, the bias decreases as α increases, consistent with the regularization modulus Λ_Ω decreasing in α (Table 1), and all curves merge at large τ, where every regularized policy approaches μ. Comparing convergence at a common τ would therefore confound speed with bias.

We thus choose, for each regularizer, the temperature giving the same bias $\|Q^*_\Omega - Q^*\|_{\mathrm{sp}} = 2$, and run epoch regularized Q-learning ($B = 64$, 20 seeds). Since increasing α lowers the domination profile γ_Ω, our theory predicts that at equal bias convergence slows as α increases. Fig. 2b confirms this ranking: it plots the span error of each method to its own target (Q^*_Ω for the regularized methods, Q^* for the baselines) against environment steps, and convergence slows as α increases. At equal bias, the choice of regularizer is thus not a trade-off but a ranking, in favour of small α. As baselines, the unregularized epoch algorithm never completes an epoch since the greedy policy does not cover the state-action space, and asynchronous RVI Q-learning only escapes the local reward trap after $\sim 10^8$ steps (for aysnchronous RVI Q-learning, we plot the span distance between the iterates and the hard fixed point).

6 Related Work

Where the contraction comes from. The average-reward Bellman operator is not a contraction, and finite-time analyses of Q-learning differ in how they deal with this. Zhang et al. (2021) and Chen (2025) assume that the Bellman operator itself is a span contraction. Chen et al. (2026) derive a contraction from a reachability assumption by running Q-learning on lazy dynamics, in which the system stays in its current state with a fixed probability. Jiao et al. (2026) analyze Q-learning in weakly communicating MDPs without any contraction, at the price of an ε^{-3} dependence, and other works rely on discounted approximations (Jin et al., 2024) or Halpern iterations with a generative model (Lee et al., 2025). With a generative model, model-based methods are minimax optimal and set the target rates (Wang et al., 2024; Zurek & Chen, 2024). We take a different route: the user supplies a policy that mixes, and regularizing toward it yields a contraction whose rate is regularizer dependent and leads to an ε^{-2} sample complexity, at the price of a bias that we bound (Table 2). The same policy also provides the coverage that asynchronous analyses usually pay through a covering or mixing time of the behavior policy (Even-Dar & Mansour, 2003; Li et al., 2020).

Regularization toward a reference policy. Regularizing toward a default policy is widespread: KL regularization averages errors in approximate value iteration (Vieillard et al., 2020), behavior priors speed up exploration and transfer (Teh et al., 2017; Tirumala et al., 2022; Ouhamma & Kamgarpour, 2026), and KL penalties to a reference policy underpin offline RL and RLHF (Wu et al., 2019; Xiong et al., 2024). Closest to our instances, Labbi et al. (2026) regularize policy gradient toward a reference policy with the same α-divergences, in the discounted setting. Their convergence constants are driven by the inverse link $(f')^{-1}$ at large negative arguments, which decays exponentially for the KL and polynomially for $\alpha < 1$; our domination profile is bounded below by this same quantity (Table 1). In the average-reward setting, Neu et al. (2017) and Adamczyk et al. (2025) study entropy-regularized objectives from the viewpoints of policy optimization and deep RL. These works use the reference policy for stability, prior knowledge, conservatism or conditioning; none uses it to make the dynamics mix, which is the mechanism we exploit.

Table 2: Sample complexity of average-reward methods, in sampled transitions and up to logarithmic factors; for Q-iteration we count evaluations of $p(s'|s,a)$. Let t_{mix} be the mixing time, K the worst expected hitting time of a reference state, and $\theta < 1$ a span-contraction factor. Whereas other methods target Q^*, we target the regularized fixed point Q^*_Ω.

Paper	Assumptions	Sample complexity
Q-iteration (known transitions)		
Puterman (1994)	one-step minorization $p(\cdot\|s,a) \geq \beta\nu$	$\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert\,\beta^{-1}\log(1/\varepsilon)$
Ours, Corollary 1	m-step Doeblin policy μ; regularizer satisfying (P1)	$\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert\,m\,\tilde{\beta}^{-1}\log(1/\varepsilon)$
Generative model, model-based		
Wang et al. (2024)	all stationary policies uniformly mixing	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\,t_{\mathrm{mix}}\,\varepsilon^{-2}$
Zurek & Chen (2024)	weakly communicating	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\,\lVert\max_a Q^*\rVert_{\mathrm{sp}}\,\varepsilon^{-2}$
Generative model, model-free		
Lee et al. (2025)	weakly communicating	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\,\lVert\max_a Q^*\rVert_{\mathrm{sp}}^2\,\varepsilon^{-2}$
Zhang et al. (2021)	unichain; J-step span contraction at rate θ	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\,J^3(1-\theta)^{-5}\,\varepsilon^{-2}$
Chen et al. (2026)	reachability, hitting time K	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\,2^{6K}\,\varepsilon^{-2}$
Ours, Theorem 1	same as for Q-iteration	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\,m^4\,\bar{w}^2\,\tilde{\beta}^{-3}\,\varepsilon^{-2}$
Single trajectory, model-free		
Chen (2025) (asynchronous)	one-step span contraction with factor θ; fixed full-support behavior policy; irreducible aperiodic behavior chain	$\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert^3\,\eta_{\min}^{-2}(1-\theta)^{-5}\,\varepsilon^{-2}$
Chen et al. (2026) (asynchronous)	reachability, hitting time K; fixed full-support behavior policy; aperiodic chain	$\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert^3\,2^{6K}\,\eta_{\min}^{-2}\,\varepsilon^{-2}$
Ours, Theorem 2 (epoch)	same as for Q-iteration	$m^4\bar{w}^2\tilde{\beta}^{-3}(\bar{\gamma}\tilde{\beta}\eta_{\min})^{-1}\varepsilon^{-2}$
Planning (resets to ξ), model-free		
Ours, Theorem I.1	m-step Doeblin policy μ on the states reachable from ξ (Assumption 4); regularizer satisfying (P1)	$(h_1 + H_{\mathrm{eff}})^4\,\bar{w}^2\,q^{-1}\,\varepsilon^{-2}$

Planning depth. In discounted planning, the depth of the lookahead tree is set by the discount factor (Kearns et al., 2002). Regularized tree search enjoys fast convergence at the root (Xiao et al., 2019) and polynomial sample complexity (Grill et al., 2019), but it targets the regularized optimum, whose greedy actions can be suboptimal for the original objective (Painter et al., 2023). In average-reward problems, finite-horizon values minus their linear growth converge as the horizon grows (Schweitzer & Federgruen, 1977). We make the depth needed for a given accuracy explicit in the mixing and the regularizer, with $H_{\mathrm{eff}} = m/\tilde{\beta}$ in the role of $1/(1-\gamma)$, and control the regularization bias separately.

7 Discussion and Limitations

The Doeblin pair (m, β_m) and the profile γ_Ω are uniform over states, so a bottleneck anywhere dictates the rate everywhere; state-dependent constants, weighted along the paths of μ, could yield much faster rates. Our criterion, uniform accuracy on all pairs, also favors exploring with μ and does not credit acting with the learned policy, which instance-dependent approaches may be able to capture. Future directions include relaxing Assumption 1 to allow transient states for single-trajectory learning, matching the $(1-\gamma)^{-4}\varepsilon^{-2}$ rate of discounted Q-learning (Li et al., 2024) with H_{eff} in place of $1/(1-\gamma)$, an asynchronous analysis (Chen, 2025), function approximation as well as temperature schedules to remove the bias while keeping early contraction.

AI use statement We used generative AI tools to identify relevant literature, polish writing, improve readability, and help implement the experiments. We have not used generative AI tools to

develop the theoretical framework, to propose hypotheses, or for proving mathematical claims. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jacob Adamczyk, Volodymyr Makarenko, Stas Tiomkin, and Rahul V. Kulkarni. Average-reward soft actor-critic. *Reinforcement Learning Journal*, 6:412–430, 2025.

Aristotle Arapostathis, Vivek S. Borkar, Emmanuel Fernández-Gaucherand, Mrinal K. Ghosh, and Steven I. Marcus. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM Journal on Control and Optimization*, 31(2):282–344, 1993.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Conference on Uncertainty in Artificial Intelligence*, 2009.

Boris Belousov and Jan Peters. Entropic regularization of Markov decision processes. *Entropy*, 21 (7):674, 2019.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 4th edition, 2012.

Zaiwei Chen. Non-asymptotic guarantees for average-reward Q-learning with adaptive stepsizes. In *Advances in Neural Information Processing Systems*, 2025.

Zaiwei Chen and Phalguni Nanda. From set convergence to pointwise convergence: Finite-time guarantees for average-reward Q-learning with adaptive stepsizes. *arXiv preprint arXiv:2504.18743v2*, 2026.

Zijun Chen, Zaiwei Chen, Nian Si, and Shengbo Wang. Achieving ε^{-2} dependence for average-reward Q-learning with a new contraction principle. *arXiv preprint arXiv:2601.21301*, 2026.

Andrzej Cichocki and Shun-ichi Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

Michael G. Crandall and Luc Tartar. Some relations between nonexpansive and order preserving mappings. *Proceedings of the American Mathematical Society*, 78(3):385–390, 1980.

Giulia De Pasquale, Kevin D. Smith, Francesco Bullo, and Maria Elena Valcher. Dual seminorms, ergodic coefficients and semicontraction theory. *IEEE Transactions on Automatic Control*, 69(5): 3040–3053, 2024.

Jean Dieudonné. *Foundations of Modern Analysis*, volume 10 of *Pure and Applied Mathematics*. Academic Press, 1960.

Roland L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability and Its Applications*, 1(1):65–80, 1956.

Eyal Even-Dar and Yishay Mansour. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5:1–25, 2003.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Stéphane Gaubert and Zheng Qu. Dobrushin's ergodicity coefficient for Markov operators on cones. *Integral Equations and Operator Theory*, 81(1):127–150, 2015.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, 2019.

Jean-Bastien Grill, Omar Darwiche Domingues, Pierre Ménard, Rémi Munos, and Michal Valko. Planning in entropy-regularized Markov decision processes and games. In *Advances in Neural Information Processing Systems*, 2019.

Jean-Bastien Grill, Florent Altché, Yunhao Tang, Thomas Hubert, Michal Valko, Ioannis Antonoglou, and Rémi Munos. Monte-Carlo tree search as regularized policy optimization. In *International Conference on Machine Learning*, 2020.

Onésimo Hernández-Lerma and Jean B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, volume 30 of *Applications of Mathematics*. Springer, 1996.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer, 2001.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Yuchen Jiao, Jiin Woo, Gen Li, Gauri Joshi, and Yuejie Chi. Sample complexity of average-reward Q-learning: From single-agent to federated reinforcement learning. *arXiv preprint arXiv:2601.13642*, 2026.

Ying Jin, Ramki Gummadi, Zhengyuan Zhou, and Jose Blanchet. Feasible Q-learning for average reward reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13(59):1865–1890, 2012.

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, 49(2–3):193–208, 2002.

William Kuszmaul and Qi Qi. The multiplicative version of Azuma's inequality, with an application to contention analysis. *arXiv preprint arXiv:2102.05077*, 2021.

Safwan Labbi, Daniil Tiapkin, Paul Mangold, and Éric Moulines. Beyond softmax and entropy: Convergence rates of policy gradients with f-SoftArgmax parameterization & coupled regularization. In *International Conference on Learning Representations*, 2026.

Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-optimal sample complexity for MDPs via anchoring. In *International Conference on Machine Learning*, 2025.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Sparse Markov decision processes with causal sparse Tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018.

Bas Lemmens and Roger Nussbaum. *Nonlinear Perron–Frobenius Theory*, volume 189 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2012.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. In *Advances in Neural Information Processing Systems*, 2020.

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 2024.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

Phalguni Nanda and Zaiwei Chen. A minimal-assumption analysis of Q-learning with time-varying policies. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 10(2), 2026.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, 2013.

Reda Ouhamma and Maryam Kamgarpour. Learning in zero-sum Markov games: Relaxing strong reachability and mixing time assumptions. In *AAAI Conference on Artificial Intelligence*, 2026.

Michael Painter, Mohamed Baioumy, Nick Hawes, and Bruno Lacerda. Monte Carlo tree search with Boltzmann exploration. In *Advances in Neural Information Processing Systems*, 2023.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

Sébastien Roch. *Modern Discrete Probability: An Essential Toolkit*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2024.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Vincent Roulet, Tianlin Liu, Nino Vieillard, Michael Eli Sander, and Mathieu Blondel. Loss functions and operators generated by f-divergences. In *International Conference on Machine Learning*, 2025.

Paul J. Schweitzer. On the existence of relative values for undiscounted Markovian decision processes with a scalar gain rate. *Journal of Mathematical Analysis and Applications*, 104(1):67–78, 1984.

Paul J. Schweitzer. On undiscounted Markovian decision processes with compact action spaces. *RAIRO – Operations Research*, 19(1):71–86, 1985.

Paul J. Schweitzer and Awi Federgruen. The asymptotic behavior of undiscounted value iteration in Markov decision problems. *Mathematics of Operations Research*, 2(4):360–381, 1977.

Paul J. Schweitzer and Awi Federgruen. The functional equations of undiscounted Markov renewal programming. *Mathematics of Operations Research*, 3(4):308–321, 1978.

Eugene Seneta. *Non-negative Matrices and Markov Chains*. Springer Series in Statistics. Springer, 2006.

Alexander L. Strehl and Michael L. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 2008.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Yee Whye Teh, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.

Dhruva Tirumala, Alexandre Galashov, Hyeonwoo Noh, Leonard Hasenclever, Razvan Pascanu, Jonathan Schwarz, Guillaume Desjardins, Wojciech Marian Czarnecki, Arun Ahuja, Yee Whye Teh, and Nicolas Heess. Behavior priors for efficient reinforcement learning. *Journal of Machine Learning Research*, 23(221):1–68, 2022.

Emanuel Todorov. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems*, 2007.

John N. Tsitsiklis. NP-hardness of checking the unichain condition in average cost MDPs. *Operations Research Letters*, 35(3):319–323, 2007.

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist. Leverage the average: An analysis of KL regularization in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Yi Wan, Huizhen Yu, and Richard S. Sutton. On convergence of average-reward Q-learning in weakly communicating Markov decision processes. *arXiv preprint arXiv:2408.16262*, 2024.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward Markov decision processes. In *International Conference on Learning Representations*, 2024.

Helmut Wielandt. Unzerlegbare, nicht negative Matrizen. *Mathematische Zeitschrift*, 52:642–648, 1950.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Chenjun Xiao, Ruitong Huang, Jincheng Mei, Dale Schuurmans, and Martin Müller. Maximum entropy Monte-Carlo planning. In *Advances in Neural Information Processing Systems*, 2019.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *International Conference on Machine Learning*, 2024.

Sheng Zhang, Zhe Zhang, and Siva Theja Maguluri. Finite sample analysis of average-reward TD learning and Q-learning. In *Advances in Neural Information Processing Systems*, 2021.

Lingwei Zhu, Zheng Chen, Matthew Schlegel, and Martha White. General Munchausen reinforcement learning with Tsallis Kullback–Leibler divergence. In *Advances in Neural Information Processing Systems*, 2023.

Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for weakly communicating and general average reward MDPs. In *Advances in Neural Information Processing Systems*, 2024.

APPENDIX CONTENTS

A Notation Summary **16**

B Algorithm Pseudocodes **17**

B.1 Learning 17

B.2 Planning 17

C Additional Experiments **18**

C.1 Two Rooms: experimental details and bias 18

C.2 Choice of the regularizer 19

D Discussion on Assumption 1 **19**

E Mathematical Background **21**

E.1 Concentration inequalities 21

E.2 On span seminorms and the Dobrushin coefficient 22

E.3 A mean value inequality for the span seminorm 23

F Proofs for Section 2 **24**

G Proofs for Section 3 **25**

G.1 Proofs for Subsection 3.2 25

G.2 Strict domination 29

G.3 Existence and uniqueness of the fixed point of $\mathcal{T}_\Omega$ 30

G.4 Control of the bias 35

G.5 ϵ-Floor-Wrapped Regularization 37

H Proofs for the learning algorithms of Section 4 **40**

H.1 Proof of Corollary 1 40

H.2 Proofs for Subsection 4.2 (Synchronous Ω-regularized Q-learning) 41

H.3 Proofs for Subsection 4.3 (Epoch Ω-regularized Q-learning) 49

I Planning: Setting and Main Results **57**

J Analysis of the Planning Algorithms **58**

J.1 Reachability and the extended target 59

J.2 Depth-dependent objects 59

J.3 The master lemma 60

J.4 Coverage 61

J.5 Model-based planning 62

J.6 Epoch planning 62

J.7 Merging the tables: epoch Q-learning with resets 65

K The α-divergence regularizers **66**

A NOTATION SUMMARY

Table 3: Notation used across sections.

MDP and kernels	
$\mathcal{S}$, $\mathcal{A}$, p, r	finite state and action sets, transition kernel, deterministic reward
$\mathbb{P}_\pi(\cdot \mid s_0 = s)$, $\mathbb{E}_\pi[\cdot \mid s_0 = s]$	law and expectation of the chain started at s, with $a_t \sim \pi(\cdot\|s_t)$ and $s_{t+1} \sim p(\cdot\|s_t, a_t)$
P	rectangular kernel $(s,a) \mapsto p(\cdot\|s,a)$, from $\mathcal{S} \times \mathcal{A}$ to $\mathcal{S}$
P_π	state kernel $P_\pi(s'\|s) = \sum_a \pi(a\|s)p(s'\|s,a)$, acting on $\mathbb{R}^{\mathcal{S}}$
K_π	state-action kernel $K_\pi(s',a'\|s,a) = p(s'\|s,a)\pi(a'\|s')$, acting on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$
$K_{\pi_1:\pi_m}$, K_μ^m	m-step state-action kernels, (G.2); $K_\mu^m(s'\|s,a)$ is the state marginal
Π	centering projection $\Pi[Q] = Q - \big(\mathbf{1}^\top Q/(\|\mathcal{S}\|\|\mathcal{A}\|)\big)\mathbf{1}$
Norms and seminorms	
$\|\cdot\|_{\mathrm{sp}}$, $\|Q\|_{\mathrm{asp}}$	span seminorm, action-span seminorm, Definition 1
$\|\cdot\|_{\mathrm{sp,op}}$	operator span seminorm, Definition E.3
$\|\cdot\|_{\mathrm{TV}}$, $\alpha(P)$	total variation distance; Dobrushin coefficient, Definition E.2
Regularization	
τ, Ω_s, $\mathrm{dom}(\Omega_s)$	temperature; regularizer at s; its effective domain
Ω^*, $\pi_{\Omega,Q}$	conjugate of $\tau\Omega$ and greedy policy, Subsection 2.2
$\mathcal{T}$, $\mathcal{T}_\Omega$	hard and regularized Bellman operators
Q^*, ρ^*; Q^*_Ω, ρ^*_Ω	optimal and regularized relative value and gain
$\tilde{r}_\pi$	regularized reward $\langle \pi_s, r_s\rangle - \tau\Omega_s(\pi_s)$
$\underline{\Omega}_s$, $\underline{\Omega}$, $\overline{\Omega}$, Δ_Ω	$\min \Omega_s$; $\min_s \underline{\Omega}_s$; $\max_{s,a} \Omega_s(\delta_a)$; $\overline{\Omega} - \underline{\Omega}$
$\overline{\pi}_{\Omega,[Q,Q']}$, $R[Q,Q']$	average-path policy and radius, Definition G.1.(i)–(ii)
$\gamma_\Omega(R)$	domination profile, Definition 2.(ii)
g_Q, $\Lambda_\Omega(R)$, $\Lambda^{\mathrm{pt}}_\Omega(R)$	regularization gap and moduli, Definition 2.(iii), Definition G.2
Mixing and rates	
μ, m, β_m, ν	mixing policy and its m-step Doeblin constants, Assumption 1
η, $\eta_{\min}$	$\eta(s,a) = \nu(s)\mu(a\|s)$ and its minimum
$\mu_{\min}$, $\nu_{\min}$	$\min_{s,a} \mu(a\|s)$ and $\min_s \nu(s)$
$\tilde{\beta}_m(A)$	regularized m-step contraction rate $\beta_m \gamma_\Omega(A)^{m-1}$
Learning analysis	
δ, T, e_0, ι	failure probability, budget, initial error $\|Q_0 - Q^*_\Omega\|_{\mathrm{sp}}$, $\log(6\|\mathcal{S}\|\|\mathcal{A}\|(T+1)/\delta)$, Section 4
$\bar{\gamma}$, $\bar{\beta}$, $\bar{w}$	domination constant $\gamma_\Omega(\|Q^*_\Omega\|_{\mathrm{asp}} + 2e_0)$, contraction rate $\beta_m \bar{\gamma}^{\,m-1}$ and noise scale on the ball of radius $2e_0$, Section 4
A_R, C_R, $\bar{\beta}(R)$, $\bar{w}(R)$	localized radii and constants, Definition H.2
p_m, κ_0	probability of m effective updates per block and sub-block budget, Lemma H.6
σ_R, e_t	exit time from the ball of radius R and error at time t, Definition H.2
$\mathcal{B}_i$, b_i, $\mathcal{B}_{\mathrm{res}}$, $I(T)$	blocks, their first indices, last incomplete block, number of completed blocks, Definition H.1 and Appendix H.2.2
$J(T)$, ℓ_j, $L_j(\delta')$, q_R	completed epochs, epoch length and its bound, coverage constant, Lemma H.15
Planning analysis	
ξ, H, Q_B	reset distribution, planning horizon, boundary value, Subsection 4.4
$\mathcal{S}_h$, $\mathcal{X}_h$, $\mathcal{S}'$, $\bar{S}_H$	reachable layers, $\mathcal{S}_h \times \mathcal{A}$, reachable set $\bigcup_h \mathcal{S}_h$, $\max_{h \le H} \|\mathcal{S}_h\|$, Appendix I
h_1, $\mathcal{S}_\infty$	depth after which the layers no longer change, and their common value, Assumption 4
$\mathcal{T}_{\Omega,h}$, $Q^*_{\Omega,h}$	depth-dependent operator from $\mathbb{R}^{\mathcal{X}_{h+1}}$ to $\mathbb{R}^{\mathcal{X}_h}$, planning solution, Subsection 4.4
Δ_B, D_0	boundary mismatch $\|Q_B - Q^*_\Omega\|_{\mathrm{sp},\mathcal{X}_H}$, initialization error $\max_{h<H} \|Q_{0,h} - Q^*_{\Omega,h}\|_{\mathrm{sp}}$; in planning, $e_0 := D_0 + \Delta_B$
$H_{\mathrm{eff}}(A)$, $M(h,A)$	effective horizon $m/\tilde{\beta}_m(A)$ and $(h_1 - h)_+ + H_{\mathrm{eff}}(A)$, (J.2); H_{eff} and M denote their values at $A = A_{2e_0}$ and $h = 0$, Appendix I
q_μ, $q_\pi(A)$, q	coverage constants under μ and under a behavior dominating $\gamma_\Omega(A)\mu$, (I.1); q is the constant of Theorem I.1
ι_H	$\log(4JH\bar{S}_H\|\mathcal{A}\|/\delta)$, Theorem I.1

B Algorithm Pseudocodes

B.1 Learning

All the learning algorithms apply the centering projection $\Pi[Q] := Q - \frac{1}{|\mathcal{S}||\mathcal{A}|}\mathbf{1}^\top Q\mathbf{1}$, which only shifts the iterates by a constant and does not affect the analysis in span seminorm.

Algorithm 2 Synchronous Ω-Regularized Q-Iteration

Input: initial Q_0, regularizer Ω, transition kernel p.
for $t = 0, 1, \dots$ **do**
  $\mathcal{T}_\Omega(Q_t)(s,a) \leftarrow r(s,a) + \sum_{s'} p(s'|s,a)\,\Omega^*(Q_t)(s')$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$.
  $Q_{t+1} \leftarrow \Pi[\mathcal{T}_\Omega](Q_t)$.
end for

Algorithm 3 Synchronous Ω-Regularized Q-Learning

Input: initial Q_0, regularizer Ω, step sizes $\{\alpha_t\}_{t\geq 0}$, generative model of p.
for $t = 0, 1, \dots$ **do**
  Sample $s'_t(s,a) \sim p(\cdot|s,a)$ independently for every $(s,a) \in \mathcal{S} \times \mathcal{A}$.
  $\widehat{\mathcal{T}_\Omega Q_t}(s,a) \leftarrow r(s,a) + \Omega^*(Q_t)\big(s'_t(s,a)\big)$ for every (s,a).
  $Q_{t+1} \leftarrow \Pi\big[(1-\alpha_t)Q_t + \alpha_t \widehat{\mathcal{T}_\Omega Q_t}\big]$.
end for

B.2 Planning

The planning algorithms take the reachable layers $\mathcal{S}_0, \dots, \mathcal{S}_H$ of Appendix I as input; they depend only on ξ and on the support of p, and the stopping rules of the model-free algorithms require them.

Algorithm 4 Backward Synchronous Ω-Regularized Q-Planning

Input: boundary value Q_B, regularizer Ω, transition kernel p, layers $\mathcal{S}_0, \dots, \mathcal{S}_H$.
Set $Q_H \leftarrow Q_B|_{\mathcal{X}_H}$.
for $h = H-1, \dots, 0$ **do**
  $Q_h(s,a) \leftarrow r(s,a) + \sum_{s' \in \mathcal{S}_{h+1}} p(s'|s,a)\,\Omega^*(Q_{h+1})(s')$ for every $(s,a) \in \mathcal{X}_h$.
end for
Return $(Q_h)_{h=0}^H$.

Algorithm 5 Epoch Ω-Regularized Q-Planning

Input: initial Q_0, boundary value Q_B, regularizer Ω, batch sizes $\{B_j\}$, reset distribution ξ, layers $\mathcal{S}_0, \dots, \mathcal{S}_H$.
Set $Q_{0,h} \leftarrow Q_0|_{\mathcal{X}_h}$ for $h < H$, $Q_{j,H} := Q_B|_{\mathcal{X}_H}$ for every j, and $j \leftarrow 0$.
while true do
  $N_{j,h} \leftarrow 0$ and $\widehat{\mathcal{T}_{\Omega,h} Q_{j,h+1}} \leftarrow 0$ on $\mathcal{X}_h$, for every $h < H$.
  while $\min_{h<H} \min_{(s,a)\in\mathcal{X}_h} N_{j,h}(s,a) < B_j$ **do**
    Sample $s_0 \sim \xi$.
    for $h = 0, \dots, H-1$ **do**
      Sample $a_h \sim \pi_{\Omega, Q_{j,h}}(\cdot|s_h)$ and $s_{h+1} \sim p(\cdot|s_h, a_h)$.
      if $N_{j,h}(s_h, a_h) < B_j$ **then**
        $\widehat{\mathcal{T}_{\Omega,h} Q_{j,h+1}}(s_h, a_h)$ += $r(s_h, a_h) + \Omega^*(Q_{j,h+1})(s_{h+1})$ and $N_{j,h}(s_h, a_h)$ += 1.
      end if
    end for
  end while
  $Q_{j+1,h} \leftarrow \widehat{\mathcal{T}_{\Omega,h} Q_{j,h+1}}/B_j$ for every $h < H$, and $j \leftarrow j+1$.
end while

Algorithm 6 Epoch Ω-Regularized Q-Learning with Resets

Input: initial $Q_0 \in \mathbb{R}^{\mathcal{S}_\infty \times \mathcal{A}}$, regularizer Ω, step sizes $\{\alpha_j\}$, batch sizes $\{B_j\}$, reset distribution ξ, burn-in depth h_1 and core $\mathcal{S}_\infty$, rollout length $H > h_1$.
Set $j \leftarrow 0$.
while true do
  $N_j \leftarrow 0$ and $\widehat{\mathcal{T}_\Omega Q_j} \leftarrow 0$ on $\mathcal{S}_\infty \times \mathcal{A}$.
  while $\min_{(s,a)\in\mathcal{S}_\infty\times\mathcal{A}} N_j(s,a) < B_j$ **do**
    Sample $s_0 \sim \xi$.
    for $h = 0, \dots, H-1$ **do**
      Sample $a_h \sim \mu(\cdot|s_h)$ if $h < h_1$ and $a_h \sim \pi_{\Omega,Q_j}(\cdot|s_h)$ otherwise, and $s_{h+1} \sim p(\cdot|s_h, a_h)$.
      if $h \geq h_1$ **and** $N_j(s_h, a_h) < B_j$ **then**
        $\widehat{\mathcal{T}_\Omega Q_j}(s_h, a_h) \mathrel{+}= r(s_h, a_h) + \Omega^*(Q_j)(s_{h+1})$ and $N_j(s_h, a_h) \mathrel{+}= 1$.
      end if
    end for
  end while
  $Q_{j+1} \leftarrow \Pi\big[(1-\alpha_j)Q_j + \alpha_j \widehat{\mathcal{T}_\Omega Q_j}/B_j\big]$ and $j \leftarrow j+1$.
end while

C Additional Experiments

C.1 Two Rooms: experimental details and bias

Setting. At the two doorway entrances, μ puts mass $\mu(\text{DOOR})$ on the action entering the doorway and $(1-\mu(\text{DOOR}))/3$ on each of the three others; everywhere else it is uniform. The regularizer is the ϵ-floored reverse KL with $\tau = 2$ and $\epsilon = 0.05$ in all panels.

Convergence. Synchronous regularized Q-iteration is exact, so no seeds are involved. The rate is fitted on the part of the error curve with $10^{-9} < \|Q_k - Q^*_\Omega\|_{\text{sp}}/\|Q_0 - Q^*_\Omega\|_{\text{sp}} < 10^{-3}$. The convergence accelerates before settling: early iterates have a larger action-span, hence a sharper policy, a smaller domination profile γ_Ω and a slower contraction. On the slowest configuration ($p_{\text{cross}} = 0.05$, uniform μ) the first e-fold takes about 170 iterations against 54.5 once settled, so a rate fitted on a fixed initial window depends on the window; we report the settled rate, which does not.

Exploration. Epoch lengths are for $B_j = 1$ accepted sample per state-action pair, with a cap of $4 \cdot 10^5$ transitions per epoch. Under μ: median over 12 seeds. On-policy: 4 epochs, median over 5 seeds of the mean of the last 2 epochs, the first epochs being dominated by the initialization $Q_0 = 0$. All curves are normalized by their value at the uniform μ. In absolute terms, one epoch under the uniform μ costs 12 304, 3 096 and 1 901 transitions for $p_{\text{cross}} = 0.05, 0.3, 1$, and the coverage constant sits 69, 52 and 39 times above these values, which is why only its shape is compared.

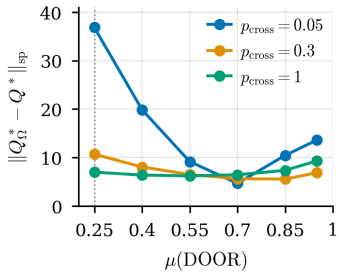


Figure 3: Bias $\|Q^*_\Omega - Q^*\|_{\text{sp}}$ on Two Rooms against $\mu(\text{DOOR})$, exact. Dotted line: uniform μ.

Bias. Figure 3 reports the bias $\|Q^*_\Omega - Q^*\|_{\text{sp}}$, computed exactly. When the doorway is a bottleneck, biasing μ towards it reduces the bias from 36.8 to 4.7, but past $\mu(\text{DOOR}) \approx 0.7$ the bias rises again,

likely because the same knob also pushes the agent from the goal room back through the doorway. The convergence and exploration panels reward a large $\mu(\text{DOOR})$; the bias is what prevents pushing it indefinitely.

C.2 Choice of the regularizer

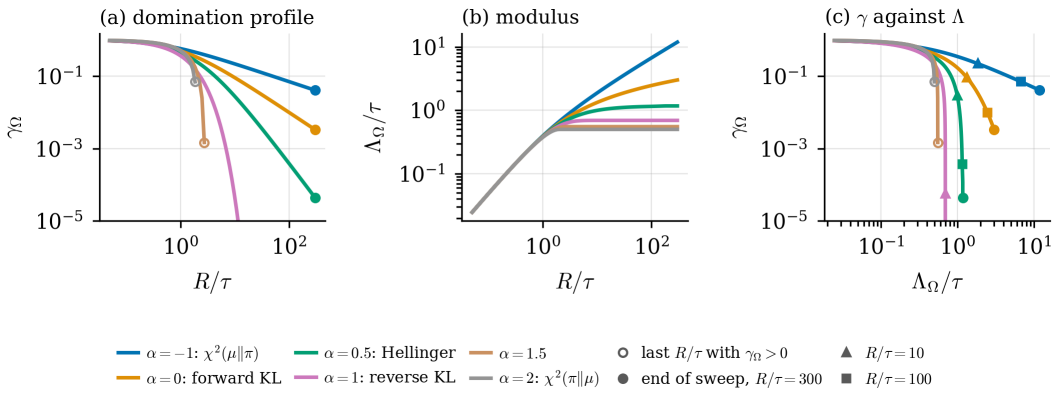


Figure 4: The objects of Definition 2 for the α-divergences to a uniform μ over two actions, and the resulting bias. (a) Domination profile γ_Ω and (b) regularization modulus Λ_Ω/τ, both functions of R/τ only. (c) The pair $(\Lambda_\Omega/\tau, \gamma_\Omega)$ as R/τ grows.

Figure 4 computes the two objects that govern our rates and bias for the α-divergences to μ. Since the greedy policy only depends on Q/τ, $\gamma_\Omega(R)$ and $\Lambda_\Omega(R)/\tau$ only depend on R/τ, so panels (a) and (b) hold for every temperature.

Panel (a) illustrates Theorem K.1.(iv): for $\alpha > 1$, the profile vanishes at a finite radius (open circles), at $R/\tau = \mu_{\max}^{1-\alpha}/(\alpha-1)$, that is 2 for $\alpha = 2$ and $2\sqrt{2}$ for $\alpha = 1.5$, so that (P1) fails; for $\alpha \le 1$, it stays positive, with an exponential decay for the reverse KL and a polynomial one for $\alpha < 1$, slower for smaller α. Panel (b) illustrates Theorem K.1.(v): the modulus saturates for $\alpha > 0$, at $\log 2$ for the reverse KL and $4(1-2^{-1/2})$ for the Hellinger regularizer, and grows without bound for $\alpha \le 0$, as $\log(R/\tau)$ for the forward KL and $(R/\tau)^{1/2}$ for $\alpha = -1$.

Panel (c) traces both objects together as the radius grows: a regularizer is preferable when its curve lies higher, for a faster contraction and coverage, and further left, for a smaller bias. No value of α dominates. Divergences with $\alpha \le 0$ keep a large profile at the price of an unbounded modulus, and those with $\alpha > 1$ have a small modulus but lose (P1); only the window $0 < \alpha \le 1$ keeps the modulus bounded and the profile positive. Within it, the Hellinger regularizer keeps a larger profile at large R/τ, whereas the reverse KL has the smaller modulus for this balanced μ, as discussed in Remark K.1.

The value bias on RiverSwim (Figure 2c) follows the ordering of the moduli in panel (b): at a given temperature, it decreases with α, and as $\tau \to 0$ it vanishes fastest for the bounded regularizers $\alpha > 0$, consistently with Lemma 1.(iii) and Theorem K.1.(v). For large τ, all greedy policies approach μ and the biases coincide. The ϵ-floor restores (P1) for $\alpha > 1$ at the price of a slightly larger bias at small temperature.

D Discussion on Assumption 1

We characterize the MDPs that satisfy Assumption 1 and its constants. Throughout, μ has full support and $K_\mu^n(s'|s,a) = \sum_{s_1} p(s_1|s,a) P_\mu^{n-1}(s'|s_1)$ (Appendix E).

Canonical constants. Let $\underline{V}_\mu^n(s') := \min_{(s,a)} K_\mu^n(s'|s,a)$. Since the minorization of Assumption 1 is uniform in the starting pair, a pair $(\beta, \tilde{\nu})$ satisfies it at horizon m if and only if $\beta\tilde{\nu} \le \underline{V}_\mu^m$ pointwise. The assumption thus holds at horizon m if and only if $\underline{V}_\mu^m > 0$, and the canonical pair

$\beta_m := \|\underline{V}_\mu^m\|_1$, $\nu := \underline{V}_\mu^m/\beta_m$, which we always use, dominates every other one, both in the contraction constant β and in the coverage profile $\beta\tilde{\nu} \otimes \mu$. Averaging $K_\mu^n(s'|s_1, a_1) \geq \underline{V}_\mu^n(s')$ over $a_1 \sim \mu(\cdot|s_1)$ and decomposing K_μ^{n+1} on its first step gives $\underline{V}_\mu^{n+1} \geq \underline{V}_\mu^n$: the assumption then holds at every horizon $n \geq m$, with canonical constants at least as good.

In fact, the contraction and the exploration rely on two different properties of K_μ^m. The contraction only requires a bound on its Dobrushin coefficient, which the minorization provides as $\alpha(K_\mu^m) \leq 1 - \beta_m$, possibly strictly (Lemma E.3), whereas the exploration requires the minorization itself, which bounds below the probability of reaching each pair. We state both through the Doeblin pair (β_m, ν), so that the domination profile transfers them at once: the greedy policies contract at rate $\beta_m\gamma_\Omega^{m-1}$ and visit every pair with probability at least $\beta_m\gamma_\Omega^m\eta_{\min}$ every m steps (Lemma 1).

Definition D.1 (Union graph)**.** The *union graph* G is the directed graph on $\mathcal{S}$ with an edge $s \to s'$ whenever $p(s'|s, a) > 0$ for some $a \in \mathcal{A}$. For $X \subseteq \mathcal{S}$ and $k \geq 0$, XG^k is the set of states reached from X by a walk of length exactly k. G is *primitive* if it is strongly connected and aperiodic, and its *exponent* $\mathrm{ex}(G)$ is then the smallest $k \geq 1$ such that $\{s\}G^k = \mathcal{S}$ for every s.

Since μ has full support, $P_\mu^n(s'|s) > 0$ if and only if $s' \in \{s\}G^n$: G is the support graph of P_μ for every full-support μ.

Proposition D.1 (Which MDPs satisfy Assumption 1)**.** For every full-support μ, Assumption 1 holds for some m if and only if P_μ is irreducible and aperiodic, if and only if G is primitive, if and only if the MDP is communicating and G is aperiodic. In particular, Assumption 1 holds with some full-support policy if and only if it holds with every full-support policy: whether it holds only depends on the support of p, and μ only affects its constants β_m and ν.

Proof. The second equivalence holds because G is the support graph of P_μ, and the third because G is strongly connected if and only if the MDP is communicating. If $\underline{V}_\mu^m > 0$, then $P_\mu^m > 0$ and $P_\mu^{m+1} > 0$ entrywise, so P_μ is irreducible, with closed walks of lengths m and $m+1$ through every state, hence aperiodic. Conversely, if G is primitive, then $\{s_1\}G^{\mathrm{ex}(G)} = \mathcal{S}$ for every s_1, so that, for $m := \mathrm{ex}(G) + 1$ and any s_1 with $p(s_1|s, a) > 0$, $K_\mu^m(\cdot|s, a) \geq p(s_1|s, a)P_\mu^{m-1}(\cdot|s_1) > 0$, i.e. $\underline{V}_\mu^m > 0$. □

Proposition D.2 (The minimal horizon)**.** Under Assumption 1, let $e(X) := \min\{k \geq 0 \mid XG^k = \mathcal{S}\}$ for nonempty $X \subseteq \mathcal{S}$. The smallest horizon at which Assumption 1 holds is

$$m^* = 1 + \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} e\big(\operatorname{supp} p(\cdot|s, a)\big),$$

and $\mathrm{ex}(G) \leq m^* \leq \mathrm{ex}(G) + 1 \leq (|\mathcal{S}| - 1)^2 + 2$.

Proof. By the first-step decomposition and the full support of μ, $K_\mu^m(\cdot|s, a) > 0$ if and only if $X_{s,a}G^{m-1} = \mathcal{S}$, with $X_{s,a} := \operatorname{supp} p(\cdot|s, a)$. As G is strongly connected, every state has a predecessor, so $XG^k = \mathcal{S}$ implies $XG^{k+1} = \mathcal{S}$, and $\{k \mid XG^k = \mathcal{S}\} = \{k \geq e(X)\}$. Hence $\underline{V}_\mu^m > 0$ if and only if $m - 1 \geq e(X_{s,a})$ for every (s, a), which gives the formula. Since $e(X) \leq \mathrm{ex}(G)$ for every nonempty X, $m^* \leq \mathrm{ex}(G) + 1$; since $P_\mu^{m^*}(s'|s) = \sum_a \mu(a|s)K_\mu^{m^*}(s'|s, a) > 0$ for every (s, s'), $m^* \geq \mathrm{ex}(G)$; and $\mathrm{ex}(G) \leq (|\mathcal{S}| - 1)^2 + 1$ by Wielandt (1950). □

The threshold m^* is thus a property of the transition graph alone, and it is small when the dynamics can idle: if every state stays in place with positive probability under some action, then $\mathrm{ex}(G)$ is the diameter of G, since any path can be lengthened with self-loops, and m^* is at most the diameter plus one. The horizon m of our rates, in contrast, is not a property of the MDP: any $m \geq m^*$ is valid and, since the algorithms do not use it, our guarantees hold at every such m. Larger horizons can only improve β_m and $\beta_m\eta_{\min}$, but they weaken the transfer factors γ_Ω^{m-1} for the contraction and γ_Ω^m for the exploration, so the best horizon to read our guarantees balances the mixing of μ against the domination of the greedy policies.

Remark D.1 (Comparison with other assumptions)**.** Under Assumption 1, the MDP is communicating (Corollary G.1) and every full-support policy induces an irreducible aperiodic chain (Proposition D.1). The assumption is thus equivalent to the standard one on the behavior policy of asynchronous Q-learning, a full-support policy inducing an irreducible aperiodic chain (Chen, 2025,

Assumption 3.1). Chen (2025) additionally assumes that the Bellman operator is a span contraction (his Assumption 2.1), whereas we obtain the contraction from regularization toward μ, and the constants of μ enter the rates. It is incomparable with the unichain condition: two states swapped by a single action form a unichain MDP whose G is periodic, and adding a self-loop action makes G primitive but the MDP not unichain. Unlike the unichain condition, whose verification is NP-hard (Tsitsiklis, 2007), it is checked in polynomial time on G, and it is preserved by adding actions or transitions. Periodicity can be removed by laziness, i.e. replacing $p(\cdot|s,a)$ by $\frac{1}{2}(\delta_s + p(\cdot|s,a))$, which preserves the gain of every stationary policy. Finally, if P_μ is unichain and aperiodic but has transient states, then for m large enough β_m is still positive but ν vanishes exactly on these states, which a trajectory following full-support policies visits finitely often: full support of ν is what makes single-trajectory learning possible on every pair.

E Mathematical Background

The following convention will be useful in the proofs.

Convention E.1. Throughout, $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ denote finite sets, typically $\mathcal{S}$ or $\mathcal{S}\times\mathcal{A}$, and P is a row-stochastic matrix from $\mathcal{X}$ to $\mathcal{Y}$. It maps functions $f\in\mathbb{R}^{\mathcal{Y}}$ to $Pf\in\mathbb{R}^{\mathcal{X}}$ and distributions $\phi\in\Delta(\mathcal{X})$ to $\phi P\in\Delta(\mathcal{Y})$:

$$(Pf)(x)=\sum_{y\in\mathcal{Y}}P(y|x)f(y),\qquad(\phi P)(y)=\sum_{x\in\mathcal{X}}\phi(x)P(y|x).$$

The product of P_1 from $\mathcal{X}$ to $\mathcal{Y}$ and P_2 from $\mathcal{Y}$ to $\mathcal{Z}$ is $(P_1P_2)(z|x)=\sum_{y\in\mathcal{Y}}P_1(y|x)P_2(z|y)$, so that $(P_1P_2)f=P_1(P_2f)$ and $\phi(P_1P_2)=(\phi P_1)P_2$. In a product $P_1P_2\cdots P_n$, the Markov chain therefore takes its first step with P_1 and its last with P_n, whereas a function is transformed first by P_n and last by P_1.

For a policy π, we also write $P_\pi(s'|s):=\sum_a\pi(a|s)\,p(s'|s,a)$ for its state kernel, and $K^m_\pi(s'|s,a):=\sum_{a'}K^m_\pi(s',a'|s,a)$ for the law of the state reached after m steps from (s,a); in this notation, Assumption 1 reads $K^m_\mu(\cdot|s,a)\geq\beta_m\nu$.

E.1 Concentration inequalities

The analysis of the model-free algorithms controls two different random quantities.

The number of visits to a state-action pair is a sum of indicators whose conditional probabilities are only known to be bounded below, on an event, by a small constant such as $q_R=\beta_m\gamma_\Omega(A_R)^m\eta_{\min}$; Lemma E.1 bounds its lower tail multiplicatively, which keeps the cover times of Lemma H.15 linear in $1/q_R$.

The noise in the Bellman targets is a martingale with bounded increments, whose running maximum is controlled by Lemma E.2.

Lemma E.1 (Self-bounding lower tail)**.** Let $(Z_i)_{i=1}^B$ be adapted to a filtration $(\mathcal{G}_i)$ with $Z_i\in[0,b]$ almost surely. Let $S:=\sum_{i=1}^BZ_i$ and $\bar S:=\sum_{i=1}^B\mathbb{E}[Z_i\mid\mathcal{G}_{i-1}]$. For any deterministic $a>0$ and $c\in(0,a)$,

$$\mathbb{P}\big(S<c,\ \bar S\geq a\big)\leq\exp\Big(-\frac{1}{b}\Big(a-c-c\log\frac{a}{c}\Big)\Big)\tag{E.1}$$

Proof. This is the lower tail of the multiplicative Azuma inequality (Kuszmaul & Qi, 2021, App. A), stated on the event $\{\bar S\geq a\}$ in the manner of Freedman (1975) rather than under an almost-sure lower bound on $\bar S$; we recall the argument. Let $\theta>0$ and $p_i:=\mathbb{E}[Z_i/b\mid\mathcal{G}_{i-1}]$. Convexity of $y\mapsto e^{-\theta y}$ on $[0,1]$ and $1-x\leq e^{-x}$ give $\mathbb{E}[e^{-\theta Z_i/b}\mid\mathcal{G}_{i-1}]\leq 1-p_i(1-e^{-\theta})\leq e^{-p_i(1-e^{-\theta})}$, so that $M_k:=\exp\big(-\frac{\theta}{b}\sum_{i\leq k}Z_i+(1-e^{-\theta})\sum_{i\leq k}p_i\big)$ is a non-negative supermartingale with $M_0=1$. On the event $\{S<c,\ \bar S\geq a\}$ we have $M_B\geq\exp\big(-\frac{\theta c}{b}+(1-e^{-\theta})\frac{a}{b}\big)$, so Markov's inequality bounds its probability by $\exp\big(\frac{1}{b}(\theta c-(1-e^{-\theta})a)\big)$, which is minimized at $\theta=\log(a/c)$; this is (E.1). □

Lemma E.2 (Maximal Azuma–Hoeffding inequality with predictable ranges)**.** Let $(M_k)_{0\le k\le n}$ be a martingale with $M_0 = 0$, adapted to $(\mathcal{F}_k)$, such that for each k the increment $M_k - M_{k-1}$ lies a.s. in an interval $[\mathcal{L}_k, \mathcal{L}_k + c_k]$ with $\mathcal{L}_k$ $\mathcal{F}_{k-1}$-measurable and $c_k > 0$ deterministic. Then for every $u > 0$,

$$\mathbb{P}\left(\max_{0\le k\le n} |M_k| \ge u\right) \le 2\exp\left(-\frac{2u^2}{\sum_{k=1}^n c_k^2}\right).$$

Proof. Apply Roch (2024, Theorem 3.2.1) to (M_k) with $A_k = \mathcal{L}_k$ and $B_k = \mathcal{L}_k + c_k$, and to $(-M_k)$ with $A_k = -\mathcal{L}_k - c_k$ and $B_k = -\mathcal{L}_k$, then take a union bound. □

E.2 On span seminorms and the Dobrushin coefficient

This section provides additional definitions and lemmas that are useful when working with the span seminorm.

Definition E.1 (Total-variation distance)**.** The total-variation distance between two probability distributions $\phi, \phi' \in \Delta(\mathcal{X})$ is defined as:

$$\|\phi - \phi'\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\mathcal{X}} |\phi(x) - \phi'(x)|. \tag{E.2}$$

Definition E.2 (Dobrushin coefficient)**.** The Dobrushin coefficient of a row-stochastic matrix $P \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ is defined as:

$$\alpha(P) = \sup_{\phi\neq\phi'\in\Delta(\mathcal{X})} \frac{\|\phi P - \phi' P\|_{\mathrm{TV}}}{\|\phi - \phi'\|_{\mathrm{TV}}} = \sup_{x,x'\in\mathcal{X}} \|P(\cdot|x) - P(\cdot|x')\|_{\mathrm{TV}}, \tag{E.3}$$

the equality of the two expressions being due to Dobrushin (1956).

The following lemma turns a Doeblin minorization, such as Assumption 1, into a bound on the Dobrushin coefficient, hence into a contraction in span seminorm by Lemma E.6.

Lemma E.3 (Doeblin and Dobrushin coefficients)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and let P be a row-stochastic matrix from $\mathcal{X}$ to $\mathcal{Y}$. If there exist $\beta \in (0, 1]$ and $\nu \in \Delta(\mathcal{Y})$ with $P(y|x) \ge \beta\nu(y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, then $\alpha(P) \le 1 - \beta$.

Proof. This is classical (Seneta, 2006, §4.3); we recall the argument. Since $|a - b| = a + b - 2\min\{a, b\}$ and both rows sum to one, $\|P(\cdot|x) - P(\cdot|x')\|_{\mathrm{TV}} = 1 - \sum_y \min\{P(y|x), P(y|x')\}$ for all $x, x' \in \mathcal{X}$. Each minimum is at least $\beta\nu(y)$ by the minorization, so the sum is at least β and every row distance is at most $1 - \beta$. □

Lemma E.4 (The terminal action does not affect the contraction)**.** Let $n \ge 1$, let $\pi_1, \dots, \pi_n$ be policies and $K := K_{\pi_1} \cdots K_{\pi_n}$, the chain taking its first step with K_{π_1} (Convention E.1). If $\pi_k \ge \gamma_k\mu$ for $k \in [\![1, n-1]\!]$ and $K_\mu^n(\cdot|s, a) \ge \beta\nu$ for every (s, a), then $\alpha(K) \le 1 - \beta\prod_{k=1}^{n-1}\gamma_k$.

Proof. The rows of K factor as $K(s', a'|x) = A(s'|x)\,\pi_n(a'|s')$, where $A(\cdot|x)$ is the law of the n-th state. The terminal factor being common to all rows, $1 - \alpha(K) = \min_{x,y}\sum_{s'}\min\{A(s'|x), A(s'|y)\} = 1 - \alpha(A)$. Only the $n - 1$ intermediate actions enter A, so the transport inequality $P_{\pi_k} \ge \gamma_k P_\mu$ gives $A(\cdot|x) \ge \big(\prod_{k=1}^{n-1}\gamma_k\big)K_\mu^n(\cdot|x) \ge \beta\big(\prod_{k=1}^{n-1}\gamma_k\big)\nu$, and Lemma E.3 concludes. □

Definition E.3 (Operator span seminorm)**.** The operator span seminorm of a linear operator $A : \mathbb{R}^p \to \mathbb{R}^q$ is defined as:

$$\|A\|_{\mathrm{sp,op}} = \sup_{\|x\|_{\mathrm{sp}}\le 1} \|Ax\|_{\mathrm{sp}} \in \mathbb{R}^+ \cup \{+\infty\}. \tag{E.4}$$

Lemma E.5 (Operator span seminorm)**.** Let $A : \mathbb{R}^p \to \mathbb{R}^q$ be linear. The following two conditions are equivalent:

(i) $\|A\|_{\mathrm{sp,op}} < +\infty$;
(ii) $A \cdot \mathrm{Ker}\|\cdot\|_{\mathrm{sp}} \subseteq \mathrm{Ker}\|\cdot\|_{\mathrm{sp}}$, i.e. $A\mathbf{1} \in \mathbb{R}\mathbf{1}$, i.e. A has constant row sums.

When they hold, for every $f \in \mathbb{R}^p$ and every linear map $B : \mathbb{R}^{p'} \to \mathbb{R}^p$ satisfying (ii),

$$\|Af\|_{\mathrm{sp}} \le \|A\|_{\mathrm{sp,op}} \|f\|_{\mathrm{sp}}, \qquad \|AB\|_{\mathrm{sp,op}} \le \|A\|_{\mathrm{sp,op}} \|B\|_{\mathrm{sp,op}}. \tag{E.5}$$

Proof. Assume $A\mathbf{1} \notin \mathbb{R}\mathbf{1}$, i.e. $\|A\mathbf{1}\|_{\mathrm{sp}} > 0$. For every $c > 0$ the vector $c\mathbf{1}$ satisfies $\|c\mathbf{1}\|_{\mathrm{sp}} = 0 \le 1$, while $\|A(c\mathbf{1})\|_{\mathrm{sp}} = c\|A\mathbf{1}\|_{\mathrm{sp}} \to +\infty$ as $c \to +\infty$. Hence $\|A\|_{\mathrm{sp,op}} = +\infty$.

Conversely, the span seminorm vanishes exactly on the constant vectors, so it induces a norm, still denoted $\|\cdot\|_{\mathrm{sp}}$, on the quotient spaces $E_p := \mathbb{R}^p/\mathbb{R}\mathbf{1}$ and $E_q := \mathbb{R}^q/\mathbb{R}\mathbf{1}$. Under (ii), the factorization theorem gives that A induces a well-defined linear map $\tilde{A} : E_p \to E_q$, $\tilde{A}[f] = [Af]$. By construction $\|Af\|_{\mathrm{sp}} = \|\tilde{A}[f]\|_{\mathrm{sp}}$ and $\|f\|_{\mathrm{sp}} = \|[f]\|_{\mathrm{sp}}$, so $\|A\|_{\mathrm{sp,op}}$ is the operator norm of $\tilde{A}$ between the normed spaces $(E_p, \|\cdot\|_{\mathrm{sp}})$ and $(E_q, \|\cdot\|_{\mathrm{sp}})$, which is finite since E_p is finite-dimensional.

Finally, the first inequality in (E.5) is the operator-norm inequality $\|\tilde{A}[f]\|_{\mathrm{sp}} \le \|\tilde{A}\|\,\|[f]\|_{\mathrm{sp}}$. For the second, AB also satisfies (ii) and induces the map $\tilde{A}\tilde{B}$, and operator norms are submultiplicative. □

Lemma E.6 (Span seminorm of a row-stochastic matrix)**.** Let $P \in \mathbb{R}^{q\times p}$ be row-stochastic. Then its operator span seminorm equals its Dobrushin coefficient, $\|P\|_{\mathrm{sp,op}} = \alpha(P)$, and consequently $\|Pf\|_{\mathrm{sp}} \le \alpha(P)\|f\|_{\mathrm{sp}}$ for every $f \in \mathbb{R}^p$.

Proof. Since P is row-stochastic, $P\mathbf{1} = \mathbf{1}$, so P satisfies condition (ii) of Lemma E.5: $\|P\|_{\mathrm{sp,op}}$ is finite, and the last claim follows from (E.5) once the equality is proved. For $f \in \mathbb{R}^p$ we have $\|Pf\|_{\mathrm{sp}} = \max_{i,j}\{(Pf)_i - (Pf)_j\} = \max_{i,j}\sum_k (P_{ik} - P_{jk})\, f_k$.

Fix i, j, let $c_k := P_{ik} - P_{jk}$, and let I^+ (resp. I^-) be the set of indices k with $c_k \ge 0$ (resp. $c_k < 0$). Since P is row-stochastic, $\sum_k c_k = 0$, hence $C := \sum_{k\in I^+} c_k = -\sum_{k\in I^-} c_k = \frac{1}{2}\sum_k |c_k|$, and with $\overline{f} := \max_k f_k$ and $\underline{f} := \min_k f_k$, $\sum_k c_k f_k = \sum_{k\in I^+} c_k f_k + \sum_{k\in I^-} c_k f_k \le C\overline{f} - C\underline{f} = C\|f\|_{\mathrm{sp}}$. Maximizing over (i, j) and using Definitions E.1 and E.2 gives $\|Pf\|_{\mathrm{sp}} \le \frac{1}{2}\max_{i,j}\sum_k |P_{ik} - P_{jk}|\,\|f\|_{\mathrm{sp}} = \alpha(P)\|f\|_{\mathrm{sp}}$, i.e. $\|P\|_{\mathrm{sp,op}} \le \alpha(P)$.

If $\alpha(P) = 0$ there is nothing to prove, so assume $\alpha(P) > 0$ and let (i^*, j^*) attain the maximum defining $\alpha(P)$. The coefficients $c_k := P_{i^*k} - P_{j^*k}$ are not all zero and sum to zero, so they take both signs, and the vector f^* defined by $f^*_k := 1$ if $c_k \ge 0$ and $f^*_k := 0$ otherwise is non-constant, with $\|f^*\|_{\mathrm{sp}} = 1$. Then $\|Pf^*\|_{\mathrm{sp}} \ge (Pf^*)_{i^*} - (Pf^*)_{j^*} = \sum_{k\in I^+} c_k = \frac{1}{2}\sum_k |P_{i^*k} - P_{j^*k}| = \alpha(P)$, hence $\|P\|_{\mathrm{sp,op}} \ge \alpha(P)$. □

E.3 A MEAN VALUE INEQUALITY FOR THE SPAN SEMINORM

Under Assumption 2, the regularized value function Ω^* is differentiable with a Lipschitz gradient (Lemma F.1.(i)), and the regularized Bellman operator $\mathcal{T}_\Omega$ is continuously differentiable with Jacobian $K_{\pi_{\Omega,Q}}$ (Corollary F.1). This Jacobian is row-stochastic, hence has constant row sums, which allows a mean value argument in span seminorm.

Proposition E.1 is the classical mean value inequality for curves in a normed space (Dieudonné, 1960, §8.5), applied in the quotient $\mathbb{R}^q/\mathbb{R}\mathbf{1}$, on which $\|\cdot\|_{\mathrm{sp}}$ is a norm; condition (H2) is what makes the Jacobians act on that quotient, and we include a proof for completeness. For linear Markov operators, the contraction rate in this seminorm, known as Hopf's oscillation seminorm, is the Dobrushin coefficient (Lemma E.6); see Gaubert & Qu (2015) for Markov operators on cones and De Pasquale et al. (2024) for the corresponding semicontraction theory.

Throughout this section, for $Q, Q' \in \mathbb{R}^p$ we write $Q(t) := Q' + t(Q - Q')$, $t \in [0, 1]$, for the segment $[Q', Q]$ from Q' to Q, as in Definition G.1.

Proposition E.1 (Mean value inequality in span seminorm)**.** Let $U \subseteq \mathbb{R}^p$ be open, let $\Phi : U \to \mathbb{R}^q$, and let $Q, Q' \in U$ be such that $[Q', Q] \subseteq U$. Assume:

(H1) Φ is differentiable at every point of $[Q', Q]$;
(H2) for every $t \in [0, 1]$, the Jacobian $D\Phi(Q(t))$ has constant row sums, i.e. $D\Phi(Q(t))\mathbf{1} \in \mathbb{R}\mathbf{1}$.

Then

$$\|\Phi(Q) - \Phi(Q')\|_{\mathrm{sp}} \;\le\; \sup_{t\in(0,1)} \|D\Phi(Q(t))\|_{\mathrm{sp,op}}\, \|Q - Q'\|_{\mathrm{sp}}, \tag{E.6}$$

and in particular with the supremum taken over the closed segment.

Proof. Let $\pi : \mathbb{R}^q \to E_q := \mathbb{R}^q/\mathbb{R}\mathbf{1}$ be the quotient map, E_q being normed by $\|\cdot\|_{\mathrm{sp}}$ as in the proof of Lemma E.5, and let $h := Q - Q'$. By (H1) the curve $g(t) := \pi(\Phi(Q(t)))$ is differentiable on $[0,1]$ with $g'(t) = \pi(D\Phi(Q(t))h)$, and by (H2) and (E.5), $\|g'(t)\|_{\mathrm{sp}} = \|D\Phi(Q(t))h\|_{\mathrm{sp}} \le \|D\Phi(Q(t))\|_{\mathrm{sp,op}}\|h\|_{\mathrm{sp}}$. The mean value inequality for differentiable curves in a normed space (Dieudonné, 1960, §8.5) gives $\|g(1)-g(0)\|_{\mathrm{sp}} \le \sup_{t\in(0,1)} \|g'(t)\|_{\mathrm{sp}}$, which is (E.6) since $\|g(1) - g(0)\|_{\mathrm{sp}} = \|\Phi(Q) - \Phi(Q')\|_{\mathrm{sp}}$. □

Corollary E.1 ($\mathcal{C}^1$ case: an exact averaged kernel)**.** If moreover Φ is $\mathcal{C}^1$ on a neighborhood of $[Q', Q]$, then

$$\Phi(Q) - \Phi(Q') = \bar{A}\,(Q - Q'), \qquad \bar{A} := \int_0^1 D\Phi(Q(t))\,\mathrm{d}t.$$

$\bar{A}$ is the average of the Jacobians of Φ along the segment; it has constant row sums and $\|\bar{A}\|_{\mathrm{sp,op}} \le \sup_{t\in(0,1)} \|D\Phi(Q(t))\|_{\mathrm{sp,op}}$.

Proof. The identity is the fundamental theorem of calculus applied coordinatewise to $t \mapsto \Phi(Q(t))$, whose derivative $D\Phi(Q(t))(Q - Q')$ is continuous in t. Integrating (H2) gives $\bar{A}\mathbf{1} = \int_0^1 D\Phi(Q(t))\mathbf{1}\,\mathrm{d}t \in \mathbb{R}\mathbf{1}$. Finally, for every $f \in \mathbb{R}^p$, the triangle inequality gives $\|\bar{A}f\|_{\mathrm{sp}} \le \int_0^1 \|D\Phi(Q(t))f\|_{\mathrm{sp}}\,\mathrm{d}t \le \sup_{t\in(0,1)} \|D\Phi(Q(t))\|_{\mathrm{sp,op}}\,\|f\|_{\mathrm{sp}}$. □

Remark E.1. For $\Phi = \mathcal{T}_\Omega$, which is continuously differentiable with Jacobian $K_{\pi_{\Omega,Q}}$ at Q (Corollary F.1), Corollary E.1 applies on every segment. The state-action kernel is linear in the policy, so $\bar{A} = \int_0^1 K_{\pi_{\Omega,Q(t)}}\,\mathrm{d}t$ is the state-action kernel of the average-path policy $\bar{\pi}_{\Omega,[Q,Q']}$, and the identity is Lemma G.2.(i).

F Proofs for Section 2

Lemma F.1 (Properties of Ω^*)**.** For each $s \in \mathcal{S}$, let $\Omega_s : \Delta(\mathcal{A}) \to \mathbb{R} \cup \{+\infty\}$ be proper, closed and strongly convex, and let $\sigma > 0$ be a common modulus of strong convexity of the Ω_s with respect to $\|\cdot\|_1$, which exists by equivalence of norms on $\mathbb{R}^{\mathcal{A}}$ and finiteness of $\mathcal{S}$; let $\tau \ge 0$ and $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

(i) *Greedy policy*: if $\tau > 0$, $\Omega^*(\cdot)(s)$ is differentiable, $\nabla\Omega^*(Q)(s) = \pi_{\Omega,Q}(\cdot|s)$ is the unique maximizer in its definition, and $\|\pi_{\Omega,Q}(\cdot|s) - \pi_{\Omega,Q'}(\cdot|s)\|_1 \le \frac{1}{2\tau\sigma}\|Q_s - Q'_s\|_{\mathcal{A},\mathrm{sp}}$.
(ii) *Boundedness*: if $L_{\Omega_s} \le \Omega_s \le U_{\Omega_s}$ on $\mathrm{dom}(\Omega_s)$, then $\Omega^*(Q)(s) \le \max_a Q_s(a) - \tau L_{\Omega_s}$; if moreover $\delta_a \in \mathrm{dom}(\Omega_s)$ for every $a \in \mathcal{A}$, then $\Omega^*(Q)(s) \ge \max_a Q_s(a) - \tau U_{\Omega_s}$.
(iii) *Translation equivariance*: $\Omega^*(Q + c\mathbf{1}) = \Omega^*(Q) + c\mathbf{1}$ for every $c \in \mathbb{R}$.
(iv) *Monotonicity*: if $Q_s \le Q'_s$, then $\Omega^*(Q)(s) \le \Omega^*(Q')(s)$.
(v) *Non-expansiveness in* $\|\cdot\|_\infty$: $\|\Omega^*(Q) - \Omega^*(Q')\|_{\mathcal{S},\infty} \le \|Q - Q'\|_{\mathcal{S}\times\mathcal{A},\infty}$.
(vi) *Non-expansiveness in* $\|\cdot\|_{\mathrm{sp}}$: $\|\Omega^*(Q) - \Omega^*(Q')\|_{\mathcal{S},\mathrm{sp}} \le \|Q - Q'\|_{\mathrm{sp}}$.
(vii) *Sub-affinity in* $\|\cdot\|_{\mathrm{sp}}$: $\|\Omega^*(Q)\|_{\mathcal{S},\mathrm{sp}} \le \|Q\|_{\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}}$.

Proof of Lemma F.1. (i). $\Omega^*(\cdot)(s)$ is the convex conjugate of the closed, $\tau\sigma$-strongly convex function $\tau\Omega_s$, so it is differentiable, its gradient is the unique maximizer, and this gradient is $\frac{1}{\tau\sigma}$-Lipschitz from $\|\cdot\|_\infty$ to $\|\cdot\|_1$ (Kakade et al., 2012); see Hiriart-Urruty & Lemaréchal (2001, §E.4) for the Euclidean case and Geist et al. (2019, Prop. 1) for Ω_s finite on $\Delta(\mathcal{A})$. Adding a constant to Q'_s shifts the objective by that constant and leaves the maximizer unchanged, so the bound holds with $\|Q_s - Q'_s - c\mathbf{1}\|_\infty$ for every $c \in \mathbb{R}$, and $\min_{c\in\mathbb{R}} \|Q_s - Q'_s - c\mathbf{1}\|_\infty = \frac{1}{2}\|Q_s - Q'_s\|_{\mathcal{A},\mathrm{sp}}$.

(ii). This refines Geist et al. (2019, Prop. 1), where Ω_s is finite on the whole simplex. For every $\pi_s \in \mathrm{dom}(\Omega_s)$, $\langle \pi_s, Q_s\rangle \le \max_a Q_s(a)$ and $\tau\Omega_s(\pi_s) \ge \tau L_{\Omega_s}$; taking the supremum gives the upper bound. For the lower bound, let $a^\star \in \arg\max_a Q_s(a)$: since $\delta_{a^\star} \in \mathrm{dom}(\Omega_s)$, $\Omega^*(Q)(s) \ge \max_a Q_s(a) - \tau\Omega_s(\delta_{a^\star}) \ge \max_a Q_s(a) - \tau U_{\Omega_s}$.

(iii)–(iv). For every $\pi_s \in \Delta(\mathcal{A})$, $\langle \pi_s, Q_s + c\mathbf{1} \rangle = \langle \pi_s, Q_s \rangle + c$, and $\langle \pi_s, Q_s \rangle \le \langle \pi_s, Q'_s \rangle$ when $Q_s \le Q'_s$ since $\pi_s \ge 0$; take the supremum over π_s.

(v). This is the argument of Crandall & Tartar (1980). With $C := \|Q - Q'\|_{\mathcal{S}\times\mathcal{A},\infty}$, we have $Q' - C\mathbf{1} \le Q \le Q' + C\mathbf{1}$, and (iii)–(iv) give $\Omega^*(Q') - C\mathbf{1} \le \Omega^*(Q) \le \Omega^*(Q') + C\mathbf{1}$.

(vi). The same argument with $c := \min_{s,a}(Q - Q')(s,a)$ and $C := \max_{s,a}(Q - Q')(s,a)$ gives $\Omega^*(Q') + c\mathbf{1} \le \Omega^*(Q) \le \Omega^*(Q') + C\mathbf{1}$, whence $\|\Omega^*(Q) - \Omega^*(Q')\|_{\mathcal{S},\mathrm{sp}} \le C - c = \|Q - Q'\|_{\mathrm{sp}}$. Maps that are monotone and translation equivariant are called topical, and are non-expansive for both norms (Lemmens & Nussbaum, 2012).

(vii). By the triangle inequality and (vi) with $Q' = 0$, $\|\Omega^*(Q)\|_{\mathcal{S},\mathrm{sp}} \le \|\Omega^*(Q) - \Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} \le \|Q\|_{\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}}$. □

Remark F.1 (Strict convexity suffices)**.** We keep strong convexity in Assumption 2, as is standard since Geist et al. (2019), but our results only use part of it. Uniqueness and continuity of the greedy map, and continuous differentiability of Ω^*, i.e. Lemma F.1.(i) without its Lipschitz constant, hold under strict convexity alone: $\tau\Omega_s$ is closed with bounded domain, so $\Omega^*(\cdot)(s)$ is finite and convex, its subdifferential, the set of maximizers, is a singleton (Hiriart-Urruty & Lemaréchal, 2001, §E.4), and a differentiable convex function is continuously differentiable (Rockafellar, 1970, Cor. 25.5.1). This is all that Lemma G.2, the mean value inequality (Proposition E.1 and Corollary E.1), the existence of the fixed point (Theorem G.1) and the finite-time bounds of Section 4 require, so they hold verbatim for strictly convex regularizers; the modulus only enters the Lipschitz constant of Lemma F.1.(i). For instance, the α-divergences with $\alpha > 2$ are strictly but not strongly convex (Theorem K.1.(ii)); they fail (P1), but their ϵ-floored versions satisfy it (Lemma G.10.(iii) only needs a strictly convex base), and our results apply to them.

Remark F.2 (Span convergence and centering)**.** Adding a constant to Q_s shifts the objective defining $\Omega^*(Q)(s)$ by that constant, so $\pi_{\Omega,Q+c\mathbf{1}} = \pi_{\Omega,Q}$: as for the unregularized relative Q-function (Puterman, 1994), the greedy policy depends on Q only modulo constants, which is why the analysis is carried out in $\|\cdot\|_{\mathrm{sp}}$. The centering projection Π turns span convergence into convergence in $\|\cdot\|_\infty$ to the representative $\Pi[Q^*_\Omega]$: since $\Pi[Q] - \Pi[Q']$ has zero mean, $\|\Pi[Q] - \Pi[Q']\|_\infty \le \|\Pi[Q] - \Pi[Q']\|_{\mathrm{sp}} = \|Q - Q'\|_{\mathrm{sp}}$. A centering step plays the same role for asynchronous average-reward Q-learning in Chen & Nanda (2026), where it upgrades span-seminorm convergence to pointwise convergence at the same rate.

Corollary F.1 (Properties of $\mathcal{T}_\Omega$)**.** Under the assumptions of Lemma F.1, $\mathcal{T}_\Omega$ is monotone, translation equivariant, and non-expansive in both $\|\cdot\|_\infty$ and $\|\cdot\|_{\mathrm{sp}}$, and $\|\mathcal{T}_\Omega(Q)\|_{\mathrm{sp}} \le \|Q\|_{\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} + \|r\|_{\mathrm{sp}}$ for every Q, where $\Omega^*(0)(s) = -\tau \min \Omega_s$; in particular $\|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} = 0$ when $\min \Omega_s$ does not depend on s, as for divergences to μ. If $\tau > 0$, $\mathcal{T}_\Omega$ is continuously differentiable, with Jacobian $D\mathcal{T}_\Omega(Q) = K_{\pi_{\Omega,Q}}$.

Proof. $\mathcal{T}_\Omega(Q) = r + P\,\Omega^*(Q)$, with P the transition kernel, row-stochastic from $\mathcal{S}\times\mathcal{A}$ to $\mathcal{S}$. Since $P \ge 0$ and $P\mathbf{1} = \mathbf{1}$, monotonicity and translation equivariance follow from Lemma F.1.(iii)–(iv), and non-expansiveness from (v)–(vi), P being non-expansive in $\|\cdot\|_\infty$ and, as $\alpha(P) \le 1$ (Lemma E.6), in $\|\cdot\|_{\mathrm{sp}}$. The bound follows from (vii) and the triangle inequality, and $\Omega^*(0)(s) = \sup_{\pi_s}\{-\tau\Omega_s(\pi_s)\} = -\tau \min \Omega_s$, the minimum being attained since Ω_s is closed and $\Delta(\mathcal{A})$ is compact. For a divergence to μ, $\min \Omega_s = \Omega_s(\mu_s) = 0$ for every s. Finally, by Lemma F.1.(i), $\partial\mathcal{T}_\Omega(Q)(s,a)/\partial Q(s',a') = p(s'|s,a)\,\pi_{\Omega,Q}(a'|s') = K_{\pi_{\Omega,Q}}(s',a'|s,a)$, which is continuous in Q. □

G Proofs for Section 3

G.1 Proofs for Subsection 3.2

We recall Definition 2 together with the average-path policy used in the proofs.

Definition G.1 (Mixing-regularization objects)**.** Let Ω satisfy Assumption 2, let μ be the mixing policy of Assumption 1, let $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $Q(t) := tQ + (1-t)Q'$ for $t \in [0,1]$, and $R \ge 0$.

(i) *Average-path policy*: $\bar{\pi}_{\Omega,[Q,Q']} := \int_0^1 \pi_{\Omega,Q(t)}\,\mathrm{d}t$ is the mean greedy policy along the segment $[Q', Q]$.

(ii) *Average-path radius*: $R[Q, Q'] := \max_{t\in[0,1]} \|Q(t)\|_{\mathrm{asp}}$ is the largest action span along the segment; by convexity of $\|\cdot\|_{\mathrm{asp}}$, it equals $\max\{\|Q\|_{\mathrm{asp}}, \|Q'\|_{\mathrm{asp}}\}$.

(iii) *Domination profile*: $\gamma_\Omega(R) := \inf\big\{\pi_{\Omega,Q}(a|s)/\mu(a|s) \mid (s,a) \in \mathcal{S}\times\mathcal{A},\ \|Q\|_{\mathrm{asp}} \le R\big\}$ is the largest factor such that $\pi_{\Omega,Q} \ge \gamma_\Omega(R)\,\mu$ for every Q of action span at most R. It lies in $[0,1]$, is non-increasing in R, and governs contraction and exploration (Lemma G.2).

(iv) *Regularization modulus*: $\Lambda_\Omega(R) := \sup\big\{\|g_Q\|_{\mathcal{S},\mathrm{sp}} :\ \|Q\|_{\mathrm{asp}} \le R\big\}$ is the largest variation across states of the *regularization gap* $g_Q(s) := \Omega^*(Q)(s) - \max_a Q(s,a)$ between the regularized and the hard maximum. It is finite, non-decreasing in R (Lemma G.1), and governs the bias (Proposition G.6).

The following lemma gives some basic properties of the regularization modulus defined in Definition G.1.(iv).

Lemma G.1 (Properties of Λ_Ω)**.** Under Assumption 2, the supremum defining $\Lambda_\Omega(R)$ is attained and finite for every $R \ge 0$, Λ_Ω is non-decreasing, and $\Lambda_\Omega(0) = \tau\big(\max_s \underline{\Omega}_s - \underline{\Omega}\big)$, where $\underline{\Omega}_s := \min_{\pi_s\in\Delta(\mathcal{A})} \Omega_s(\pi_s)$ and $\underline{\Omega} := \min_s \underline{\Omega}_s$.

Proof. Since $\Omega^*(Q)(s)$ depends on Q only through Q_s, Lemma F.1.(iii) applied state by state shows that $\Omega^*(Q)(s)$ and $\max_a Q_s(a)$ both increase by c when Q_s is replaced by $Q_s + c\mathbf{1}$: g_Q is invariant under the addition of a constant to each Q_s separately. We may therefore restrict the supremum to the Q with $\max_a Q_s(a) = 0$ for every s, for which $\|Q\|_{\mathrm{asp}} \le R$ reads $Q_s \in [-R, 0]^{\mathcal{A}}$: a compact set. On it, $Q \mapsto \|g_Q\|_{\mathcal{S},\mathrm{sp}}$ is continuous, since Ω^* is non-expansive in $\|\cdot\|_\infty$ (Lemma F.1.(v)) and $Q \mapsto \max_a Q(\cdot,a)$ is 1-Lipschitz; hence the supremum is attained and finite. It is non-decreasing in R because the feasible set grows with R. Finally, for $R = 0$ each Q_s is a constant $c_s\mathbf{1}$, so $\Omega^*(Q)(s) = c_s + \Omega^*(0)(s) = c_s - \tau\underline{\Omega}_s$, the minimum defining $\underline{\Omega}_s$ being attained since Ω_s is closed on the compact simplex. Thus $g_Q = -\tau(\underline{\Omega}_s)_{s\in\mathcal{S}}$, whose span is $\tau(\max_s \underline{\Omega}_s - \underline{\Omega})$. □

Remark G.1 (Boundedness of the regularization modulus)**.** Let $\overline{\Omega} := \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \Omega_s(\delta_a)$, $\underline{\Omega} = \min_s \underline{\Omega}_s$ and $\Delta_\Omega := \overline{\Omega} - \underline{\Omega}$. If $\overline{\Omega} < +\infty$, i.e. if every Dirac lies in the effective domain, then Lemma F.1.(ii) applied state-wise with $L_{\Omega_s} = \underline{\Omega}$ and $U_{\Omega_s} = \max_a \Omega_s(\delta_a)$ gives $\Lambda_\Omega(R) \le \tau\Delta_\Omega$ for every R: the modulus is bounded by a radius-independent constant and all the bounds below reduce to their constant counterparts.

The α-divergences with $\alpha \le 0$, including the forward KL, and the floor-wrapped regularizers of Definition G.4 exclude every Dirac from their effective domain, so $\overline{\Omega} = +\infty$, while $\Lambda_\Omega(R)$ stays finite (Theorem K.1.(v) and Lemma G.10.(v)); this is what makes the radius-dependent form necessary.

Lemma G.2 (Mixing properties of Ω-regularization)**.** Let Assumptions 1 and 2 hold and let $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Then:

(i) *Average-path factorization of the conjugate and of the operator*: for every $s \in \mathcal{S}$, $\Omega^*(Q)(s) - \Omega^*(Q')(s) = \langle \bar\pi_{\Omega,[Q,Q']}(\cdot|s), Q_s - Q'_s\rangle$ and $\mathcal{T}_\Omega(Q) - \mathcal{T}_\Omega(Q') = K_{\bar\pi_{\Omega,[Q,Q']}}(Q - Q')$.

(ii) *Domination of the average-path policy*: $\bar\pi_{\Omega,[Q,Q']}(a|s) \ge \gamma_\Omega(R[Q,Q'])\,\mu(a|s)$ for every $(s,a) \in \mathcal{S}\times\mathcal{A}$.

(iii) *m-step minorization*: if $\pi_1, \dots, \pi_m$ are policies with $\pi_k \ge \gamma_k\mu$ for $k \in [\![1, m]\!]$, the m-step kernel $K_{\pi_1:\pi_m} := K_{\pi_1}\cdots K_{\pi_m}$ satisfies

$$K_{\pi_1:\pi_m}(s', a'|s, a) \ge \beta_m \left(\prod_{k=1}^{m} \gamma_k\right) \eta(s', a'), \qquad \forall (s,a), (s',a') \in \mathcal{S}\times\mathcal{A}.$$

(iv) *Radius-dependent m-step contraction*:

$$\|\mathcal{T}_\Omega^m(Q) - \mathcal{T}_\Omega^m(Q')\|_{\mathrm{sp}} \le \left(1 - \beta_m \prod_{k=1}^{m-1} \gamma_\Omega\left(R[\mathcal{T}_\Omega^k(Q), \mathcal{T}_\Omega^k(Q')]\right)\right) \|Q - Q'\|_{\mathrm{sp}}.$$

(v) *State-action coverage*: let $(Q_n)_{n\ge0}$ be Q-functions and $(s_n, a_n)_{n\ge0}$ a trajectory with $a_n \sim \pi_{\Omega,Q_n}(\cdot|s_n)$ and $s_{n+1} \sim p(\cdot|s_n, a_n)$; let $\mathcal{F}_n := \sigma(Q_0, (s_0,a_0), \dots, (s_n,a_n))$ and assume that Q_n is $\mathcal{F}_{n-1}$-measurable for $n \ge 1$. If $t \ge 0$ and $\gamma_1, \dots, \gamma_m$ are non-negative $\mathcal{F}_t$-measurable random variables with $\pi_{\Omega,Q_{t+k}} \ge \gamma_k\mu$ almost surely for $k \in [\![1, m]\!]$, then almost surely

$$\mathbb{P}\big((s_{t+m}, a_{t+m}) = (s,a) \mid \mathcal{F}_t\big) \ge \beta_m \left(\prod_{k=1}^{m} \gamma_k\right) \eta(s,a), \qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

(vi) *One-step operator gap*: for every $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,

$$\|\mathcal{T}_\Omega(Q) - \mathcal{T}(Q)\|_{\mathrm{sp}} \le \alpha(P)\,\Lambda_\Omega\big(\|Q\|_{\mathrm{asp}}\big) \le \Lambda_\Omega\big(\|Q\|_{\mathrm{asp}}\big),$$

where $\alpha(P)$ is the Dobrushin coefficient (Definition E.2) of the rectangular matrix $(p(s'|s,a))_{(s,a),s'}$. In particular, under the assumptions of Remark G.1, $\|\mathcal{T}_\Omega(Q) - \mathcal{T}(Q)\|_{\mathrm{sp}} \le \alpha(P)\tau\Delta_\Omega \le \tau\Delta_\Omega$ uniformly in Q.

Proof of Lemma G.2. Throughout the proof we write $\bar\pi = \bar\pi_{\Omega,[Q,Q']}$, γ_k for the domination coefficients of each item and $\eta(s',a') = \nu(s')\mu(a'|s')$ for the minorizing measure of Assumption 1. Points (i), (ii), (iv) and (v) use $\tau > 0$ through Lemma F.1.(i), whereas (iii) requires no regularity at all and (vi) only uses Assumption 2 through the finiteness of Λ_Ω (Lemma G.1).

(i). Let $g(t) := \Omega^*(Q(t))(s)$ for $t \in [0,1]$. Since $\Omega^*(Q)(s)$ depends on Q only through Q_s, Lemma F.1.(i) and the chain rule give that g is continuously differentiable, with $g'(t) = \langle \pi_{\Omega,Q(t)}(\cdot|s), Q_s - Q'_s\rangle$. The fundamental theorem of calculus and the linearity of the inner product then give

$$\Omega^*(Q)(s) - \Omega^*(Q')(s) = g(1) - g(0) = \int_0^1 \langle \pi_{\Omega,Q(t)}(\cdot|s), Q_s - Q'_s\rangle\,\mathrm{d}t = \langle \bar\pi(\cdot|s), Q_s - Q'_s\rangle, \quad \text{(G.1)}$$

which is the first equality. Since $\mathcal{T}_\Omega = r + P\,\Omega^*$, the reward terms cancel and applying (G.1) at each next state $s' \in \mathcal{S}$ gives, for every $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}\mathcal{T}_\Omega(Q)(s,a) - \mathcal{T}_\Omega(Q')(s,a) &= \sum_{s'} p(s'|s,a)\big(\Omega^*(Q)(s') - \Omega^*(Q')(s')\big)\\ &= \sum_{s',a'} p(s'|s,a)\,\bar\pi(a'|s')\,(Q-Q')(s',a') = \big(K_{\bar\pi}(Q-Q')\big)(s,a),\end{aligned}$$

which is the second equality.

(ii). By Definition G.1.(ii), $\|Q(t)\|_{\mathrm{asp}} \le R[Q,Q']$ for all $t \in [0,1]$, and Definition G.1.(iii) applied with $R = R[Q,Q']$ then gives $\pi_{\Omega,Q(t)}(a|s) \ge \gamma_\Omega(R[Q,Q'])\mu(a|s)$ for all (s,a) and all $t \in [0,1]$, uniformly in t. Integrating this inequality against the uniform measure on $[0,1]$ and using Definition G.1.(i) yields $\bar\pi(a|s) = \int_0^1 \pi_{\Omega,Q(t)}(a|s)\,\mathrm{d}t \ge \gamma_\Omega(R[Q,Q'])\mu(a|s)$, which is the claim.

(iii). Set $\Gamma_{1:m} = \prod_{k=1}^m \gamma_k$. First, for any policy π with $\pi(a|s) \ge \gamma\mu(a|s)$ we have the transport inequality $P_\pi(s'|s) = \sum_a \pi(a|s)p(s'|s,a) \ge \gamma\sum_a \mu(a|s)p(s'|s,a) = \gamma P_\mu(s'|s)$. By definition of $K_{\pi_1:\pi_m} = K_{\pi_1}\cdots K_{\pi_m}$ and Convention E.1, the first action is the one supplied by the conditioning, the intermediate actions are drawn from $\pi_1,\dots,\pi_{m-1}$ and the terminal action from π_m:

$$K_{\pi_1:\pi_m}(s',a'|s,a) = \pi_m(a'|s') \sum_{s_1,\dots,s_{m-1}} p(s_1|s,a) \prod_{k=1}^{m-1} P_{\pi_k}(s_{k+1}|s_k), \quad \text{(G.2)}$$

with $s_m = s'$, and since all factors are non-negative, the bounds multiply: applying the transport inequality to the $m-1$ intermediate factors and the domination hypothesis to the terminal factor $\pi_m(a'|s') \ge \gamma_m\mu(a'|s')$ gives

$$\begin{aligned}K_{\pi_1:\pi_m}(s',a'|s,a) &\ge \Gamma_{1:m}\,\mu(a'|s') \sum_{s_1,\dots,s_{m-1}} p(s_1|s,a) \prod_{k=1}^{m-1} P_\mu(s_{k+1}|s_k)\\ &= \Gamma_{1:m}\,\mu(a'|s') \sum_{s_1} p(s_1|s,a)\,P_\mu^{m-1}(s'|s_1)\\ &= \Gamma_{1:m}\,\mu(a'|s')\,K_\mu^m(s'|s,a)\\ &\ge \Gamma_{1:m}\,\beta_m\,\nu(s')\mu(a'|s') = \Gamma_{1:m}\,\beta_m\,\eta(s',a'),\end{aligned}$$

the second-to-last line by the definition $K_\mu^m(s'|s,a) = \sum_{s_1} p(s_1|s,a)P_\mu^{m-1}(s'|s_1)$ and the last by Assumption 1.

(iv). For $k \in [\![0, m-1]\!]$ write $Q_k := \mathcal{T}_\Omega^k(Q)$, $Q'_k := \mathcal{T}_\Omega^k(Q')$, $R_k := R[Q_k,Q'_k]$, $\bar\pi^{(k)} = \bar\pi_{\Omega,[Q_k,Q'_k]}$ and $\gamma_k = \gamma_\Omega(R_k)$. (i) applied at level k gives the one-step identity $Q_{k+1} - Q'_{k+1} = \mathcal{T}_\Omega(Q_k) -$

$\mathcal{T}_\Omega(Q'_k) = K_{\bar\pi^{(k)}}(Q_k - Q'_k)$, where $K_{\bar\pi^{(k)}}$ is a row-stochastic state-action kernel. Iterating this identity from $k = m-1$ down to $k = 0$, we obtain

$$\mathcal{T}_\Omega^m(Q) - \mathcal{T}_\Omega^m(Q') = Q_m - Q'_m = K^{(m)}_{Q,Q'}(Q - Q'), \qquad K^{(m)}_{Q,Q'} := K_{\bar\pi^{(m-1)}} K_{\bar\pi^{(m-2)}} \cdots K_{\bar\pi^{(0)}}.$$

Following Convention E.1, $K^{(m)}_{Q,Q'}$ is exactly the m-step state-action kernel $K_{\bar\pi^{(m-1)}:\bar\pi^{(0)}}$ associated with the sequence of policies $\bar\pi^{(m-1)}, \dots, \bar\pi^{(0)}$. By (ii), each of them satisfies $\bar\pi^{(k)} \geq \gamma_k \mu$, so (iii) applies and gives the Doeblin condition $K^{(m)}_{Q,Q'}(s', a'|s, a) \geq \beta_m \left(\prod_{k=0}^{m-1} \gamma_k\right) \eta(s', a')$. Lemma E.4, applied with $\pi_1 = \bar\pi^{(m-1)}, \dots, \pi_m = \bar\pi^{(0)}$, gives $\alpha\big(K^{(m)}_{Q,Q'}\big) \leq 1 - \beta_m \prod_{k=1}^{m-1} \gamma_\Omega(R_k)$, and by Lemma E.6, $\|K^{(m)}_{Q,Q'}\|_{\mathrm{sp,op}} = \alpha\left(K^{(m)}_{Q,Q'}\right)$. Therefore

$$\|\mathcal{T}_\Omega^m(Q) - \mathcal{T}_\Omega^m(Q')\|_{\mathrm{sp}} = \left\|K^{(m)}_{Q,Q'}(Q - Q')\right\|_{\mathrm{sp}} \leq \left(1 - \beta_m \prod_{k=1}^{m-1} \gamma_\Omega(R_k)\right) \|Q - Q'\|_{\mathrm{sp}}.$$

(v). Fix $t \geq 0$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$. For $j \in [\![0, m]\!]$, let $\Gamma_j := \prod_{k=j+1}^{m} \gamma_k$, which is non-negative and $\mathcal{F}_t$-measurable, hence $\mathcal{F}_{t+j}$-measurable. We prove by backward induction on j that, a.s.,

$$\mathbb{E}\left[\mathbb{1}_{\{(s_{t+m}, a_{t+m}) = (s,a)\}} \,\middle|\, \mathcal{F}_{t+j}\right] \;\geq\; \Gamma_j \, K_\mu^{m-j}\big(s, a \mid s_{t+j}, a_{t+j}\big).$$

The case $j = m$ is an equality. Assume the claim at $j+1$. Conditionally on $\mathcal{F}_{t+j}$, $s_{t+j+1} \sim p(\cdot|s_{t+j}, a_{t+j})$ and $a_{t+j+1} \sim \pi_{\Omega, Q_{t+j+1}}(\cdot|s_{t+j+1})$, where Q_{t+j+1} is $\mathcal{F}_{t+j}$-measurable. By the tower property, the induction hypothesis and the $\mathcal{F}_{t+j}$-measurability of Γ_{j+1}, the left-hand side at level j is therefore at least

$$\begin{aligned}
&\Gamma_{j+1} \sum_{s'} p(s'|s_{t+j}, a_{t+j}) \sum_{a'} \pi_{\Omega, Q_{t+j+1}}(a'|s') \, K_\mu^{m-j-1}\big(s, a \mid s', a'\big) \\
&\quad \geq \Gamma_{j+1}\gamma_{j+1} \sum_{s', a'} K_\mu(s', a'|s_{t+j}, a_{t+j}) \, K_\mu^{m-j-1}\big(s, a \mid s', a'\big) = \Gamma_j \, K_\mu^{m-j}\big(s, a \mid s_{t+j}, a_{t+j}\big),
\end{aligned}$$

the inequality by the domination hypothesis and the non-negativity of Γ_{j+1} and of K_μ^{m-j-1}. Taking $j = 0$ and using that the m-step state-action kernel of μ satisfies $K_\mu^m(s', a'|s, a) = K_\mu^m(s'|s, a)\mu(a'|s') \geq \beta_m \nu(s')\mu(a'|s') = \beta_m \eta(s', a')$ by Assumption 1 gives the claim.

(vi). Definition G.1.(iv) applied with $R = \|Q\|_{\mathrm{asp}}$ immediately gives $\|g_Q\|_{\mathcal{S},\mathrm{sp}} \leq \Lambda_\Omega(\|Q\|_{\mathrm{asp}})$. The two operators $\mathcal{T}_\Omega$ and T share the reward term and differ only through the state-value function, so for all (s, a),

$$\mathcal{T}_\Omega(Q)(s, a) - \mathcal{T}(Q)(s, a) = \sum_{s'} p(s'|s, a) \left(\Omega^*(Q)(s') - \max_{a'} Q(s', a')\right) = (P g_Q)(s, a).$$

Lemma E.6 gives $\|P g_Q\|_{\mathrm{sp}} \leq \alpha(P)\|g_Q\|_{\mathcal{S},\mathrm{sp}}$, so $\|\mathcal{T}_\Omega(Q) - \mathcal{T}(Q)\|_{\mathrm{sp}} \leq \alpha(P)\Lambda_\Omega(\|Q\|_{\mathrm{asp}})$. The second inequality follows from $\alpha(P) \leq 1$, and the last claim from Remark G.1. □

Lemma 1 (Mixing properties of Ω-regularization)**.** Under Assumptions 1 and 2 and for any Q, Q':

(i) *Contraction*: with $\mathcal{T}_\Omega^k$ the k-th composition of the regularized Bellman operator,

$$\|\mathcal{T}_\Omega^m(Q) - \mathcal{T}_\Omega^m(Q')\|_{\mathrm{sp}} \leq \left(1 - \beta_m \prod_{k=1}^{m-1} \gamma_\Omega(R[\mathcal{T}_\Omega^k(Q), \mathcal{T}_\Omega^k(Q')])\right) \|Q - Q'\|_{\mathrm{sp}}.$$

(ii) *Coverage*: let $(s_n, a_n)_{n \geq 0}$ be a trajectory with $a_n \sim \pi_{\Omega, Q_n}(\cdot|s_n)$ and $s_{n+1} \sim p(\cdot|s_n, a_n)$, where Q_n is $\mathcal{F}_{n-1}$-measurable for $\mathcal{F}_n := \sigma(Q_0, (s_k, a_k)_{k \leq n})$. If $\gamma_1, \dots, \gamma_m \geq 0$ are $\mathcal{F}_t$-measurable and $\pi_{\Omega, Q_{t+k}} \geq \gamma_k \mu$ almost surely, then almost surely

$$\mathbb{P}\big((s_{t+m}, a_{t+m}) = (s, a) \mid \mathcal{F}_t\big) \geq \beta_m \left(\prod_{k=1}^{m} \gamma_k\right) \eta(s, a), \qquad \text{for every } (s, a).$$

(iii) *Bias*: the one step operator gap is: $\|\mathcal{T}_\Omega(Q) - \mathcal{T}(Q)\|_{\mathrm{sp}} \le \Lambda_\Omega(\|Q\|_{\mathrm{asp}})$. Moreover, if Assumption 3 and (P1) hold, then for every solution (Q^*, ρ^*) of $\mathcal{T}(Q^*) = Q^* + \rho^* \mathbf{1}$,

$$\|Q^*_\Omega - Q^*\|_{\mathrm{sp}} \le \frac{m\,\Lambda_\Omega(R^*)}{\beta_m \gamma_\Omega(R^*)^{m-1}}, \qquad R^* := \max\Big\{\|Q^*_\Omega\|_{\mathrm{asp}}, \max_{0\le k<m} \|\mathcal{T}^k_\Omega(Q^*)\|_{\mathrm{asp}}\Big\},$$

and, if $|\mathcal{S}| \ge 2$ and all the Ω_s have the same minimum $\underline{\Omega}$, as divergences to μ do with $\underline{\Omega} = 0$,

$$-\Lambda_\Omega(\|Q^*\|_{\mathrm{asp}}) \le \rho^*_\Omega - \rho^* + \tau\underline{\Omega} \le 0.$$

Proof. Items (i) and (ii) are Lemma G.2.(iv) and (v). The first claim of (iii) is Lemma G.2.(vi) with $\alpha(P) \le 1$, and the second is Proposition G.6.(i), whose right-hand side is finite under (P1) and the gain bounds are Proposition G.6.(ii), the last one using $|\mathcal{S}| \ge 2$ and a common minimum.. □

Remark G.2 (Only m regularized steps are needed)**.** Let $\pi_1, \dots, \pi_M$ with $M \ge m$ be given, and assume that only the last m of them are known to be regularized, say $\pi_k \ge \gamma_k \mu$ for $k \in [\![M-m+1, M]\!]$, with $\Gamma_{M-m+1:M} := \prod_{k=M-m+1}^{M} \gamma_k$. Then Lemma G.2.(iii) still holds for the full M-step kernel, with the same parameter $\Gamma_{M-m+1:M}\beta_m$ and the same minorizing measure η.

Writing $K_{\mathrm{head}} := K_{\pi_1:\pi_{M-m}}$ for the kernel of the first $M-m$ steps and $K_{\mathrm{tail}} := K_{\pi_{M-m+1}:\pi_M}$ for that of the last m ones, we have, consistently with Convention E.1,

$$K_{\pi_1:\pi_M} = K_{\mathrm{head}} K_{\mathrm{tail}}, \qquad \text{i.e.} \qquad K_{\pi_1:\pi_M}(s', a'|s, a) = \sum_{(\tilde{s},\tilde{a})} K_{\mathrm{head}}(\tilde{s}, \tilde{a}|s, a) K_{\mathrm{tail}}(s', a'|\tilde{s}, \tilde{a}).$$

Lemma G.2.(iii) applies to K_{tail} and gives $K_{\mathrm{tail}}(s', a'|\tilde{s}, \tilde{a}) \ge \Gamma_{M-m+1:M}\beta_m\,\eta(s', a')$ uniformly in the starting pair $(\tilde{s}, \tilde{a})$; the head, which requires no regularization, only contributes a factor 1 by row-stochasticity:

$$K_{\pi_1:\pi_M}(s', a'|s, a) \ge \Gamma_{M-m+1:M}\beta_m \eta(s', a') \sum_{(\tilde{s},\tilde{a})} K_{\mathrm{head}}(\tilde{s}, \tilde{a}|s, a) = \Gamma_{M-m+1:M}\beta_m \eta(s', a').$$

The same holds for the coverage bound of Lemma G.2.(v): if its domination hypothesis holds only for $k \in [\![M-m+1, M]\!]$, then, almost surely,

$$\mathbb{P}\big((s_{t+M}, a_{t+M}) = (s, a) \mid \mathcal{F}_t\big) \ge \Gamma_{M-m+1:M}\beta_m\,\eta(s, a).$$

Indeed, Lemma G.2.(v) applied at time $t + M - m$ gives this lower bound for the probability conditional on $\mathcal{F}_{t+M-m}$, and the tower property transfers it to $\mathcal{F}_t$, the bound being deterministic: the conditional expectation plays the role of the row-stochastic head. Since the proof of (v) only uses the steps after time $t + M - m$, the actions of the head may moreover follow any adapted rule.

G.2 Strict domination

Lemma G.3 ((P1) is full support)**.** Under Assumption 2, Ω satisfies (P1) if and only if $\pi_{\Omega,Q}(a|s) > 0$ for every Q and every (s, a).

Proof. If (P1) holds, $\pi_{\Omega,Q} \ge \gamma_\Omega(\|Q\|_{\mathrm{asp}})\,\mu > 0$. Conversely, $\pi_{\Omega,Q}(\cdot|s) = \nabla\Omega^*_s(Q_s)$ is continuous and invariant under $Q_s \mapsto Q_s + c\mathbf{1}$, so $\gamma_\Omega(R)$ is the minimum of a positive continuous function over the compact set of Q_s with $\max_a Q_s(a) = 0$ and $\min_a Q_s(a) \ge -R$, $s \in \mathcal{S}$. □

Property (P1) is pointwise in R: for the reverse KL, $\gamma_\Omega(R) = \big(\mu_{\min} + (1-\mu_{\min})e^{R/\tau}\big)^{-1}$ vanishes as $R \to \infty$ (Remark K.1). It makes the contraction rate of Lemma G.2.(iv) positive on bounded sets, guarantees that epochs terminate (Proposition H.2) and that $q_R > 0$, and is the full-support requirement of Proposition G.4. It holds for the α-divergences with $\alpha \le 1$ and fails for $\alpha > 1$ (Theorem K.1.(iv)); the floor-wrapped regularizers satisfy it with $\gamma_{\Omega_\epsilon} \ge \epsilon$ (Lemma G.10.(iii)).

G.3 Existence and uniqueness of the fixed point of $\mathcal{T}_\Omega$

Theorem G.1 (Existence of a fixed point for $\mathcal{T}_\Omega$)**.** Under Assumptions 1 to 3, there exist $Q^*_\Omega \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\rho^*_\Omega \in \mathbb{R}$ such that $\mathcal{T}_\Omega(Q^*_\Omega) = Q^*_\Omega + \rho^*_\Omega \mathbf{1}$, with

$$\|Q^*_\Omega\|_{\mathrm{sp}} \le \|r\|_{\mathrm{sp}} + \frac{4mR_{\max}}{\beta_m \nu_{\min}}, \qquad |\rho^*_\Omega| \le 2R_{\max},$$

where $\nu_{\min} = \min_s \nu(s) > 0$ and $R_{\max} = \|r\|_\infty + \tau \max_s \max\{|\underline{\Omega}_s|, |\Omega_s(\mu_s)|\}$, both finite under Assumption 3. Moreover ρ^*_Ω is unique and is the optimal regularized gain (Propositions G.3 and G.5), and under (P1) Q^*_Ω is unique up to an additive constant (Proposition G.4).

This subsection proves Theorem G.1 as a corollary of a more general statement, Theorem G.2, which rests on a condition that is both necessary and sufficient and does not presuppose Assumption 1. The argument is the regularized counterpart of the *vanishing discount* approach. In the unregularized case, the equivalence it establishes is classical (Schweitzer, 1984), as is the bound on the span of the relative discounted values under weak communication, which underlies regret analysis (Jaksch et al., 2010; Bartlett & Tewari, 2009) (see also Puterman, 1994, §8.4.4; Bertsekas, 2012, §4.1–4.3; Hernández-Lerma & Lasserre, 1996, §5.5; and the survey of Arapostathis et al., 1993). We nevertheless give a self-contained argument, since the classical statements for compact action sets require the reward to be continuous in the action (Schweitzer, 1985), which fails for $\tilde{r}_\pi$ as soon as $\mathrm{dom}(\Omega_s) \subsetneq \Delta(\mathcal{A})$, as for the α-divergences with $\alpha \le 0$. Special cases are known by other routes: entropy regularization by convex duality (Neu et al., 2017), and the reverse KL as a linearly solvable problem, whose average-cost equation is a Perron–Frobenius eigenproblem (Todorov, 2007).

Throughout Ω satisfies Assumption 2 and we also assume the following.

Assumption G.1 (Admissible policy)**.** There exists a full-support policy $\tilde{\pi}$ such that $\tilde{\pi}_s \in \mathrm{dom}(\Omega_s)$ for every $s \in \mathcal{S}$ and $\max_{s\in\mathcal{S}} |\Omega_s(\tilde{\pi}_s)| < +\infty$.

Notation and constants. Since Ω_s is proper and closed and $\Delta(\mathcal{A})$ is compact, $\underline{\Omega}_s := \min_{\pi_s \in \Delta(\mathcal{A})} \Omega_s(\pi_s)$ is finite. Under Assumption G.1, the following constants are well-defined and finite:

$$\delta_\Omega := \max_{s\in\mathcal{S}} \max\{|\underline{\Omega}_s|,\ |\Omega_s(\tilde{\pi}_s)|\} < +\infty, \qquad R_{\max} := \|r\|_\infty + \tau\delta_\Omega. \tag{G.3}$$

For an *admissible* policy π (i.e. $\pi_s \in \mathrm{dom}(\Omega_s)$ for every s) and discount factor $\lambda \in (0,1)$ define the *regularized reward* and the discounted evaluation and optimality operators on $\mathbb{R}^{\mathcal{S}}$,

$$\tilde{r}_\pi(s) := \langle \pi_s, r_s\rangle - \tau\,\Omega_s(\pi_s), \qquad \mathcal{E}^\lambda_\pi V := \tilde{r}_\pi + \lambda P_\pi V, \qquad \mathcal{T}^\lambda_\Omega V := \max_{\pi \text{ admissible}} \mathcal{E}^\lambda_\pi V, \tag{G.4}$$

the maximum in $\mathcal{T}^\lambda_\Omega$ being taken state-wise. By (G.3), $\tilde{r}_\pi(s) \le \|r\|_\infty - \tau\underline{\Omega}_s \le R_{\max}$ for every admissible π, and $|\tilde{r}_{\tilde{\pi}}(s)| \le R_{\max}$.

Remark G.3 (Sufficient conditions for Assumption G.1)**.** Under Assumption 1, Assumption 3 gives Assumption G.1 with $\tilde{\pi} = \mu$. More generally, if every Dirac lies in the domain, as in Remark G.1, convexity gives $\Omega_s(\pi_s) \le \max_a \Omega_s(\delta_a) < +\infty$ on $\Delta(\mathcal{A})$, so any full-support $\tilde{\pi}$ works; this covers the negative entropy, the Tsallis negentropies, the quadratic regularizer and the α-divergences to μ with $\alpha > 0$ (Theorem K.1.(i)). For $\alpha \le 0$, including the forward KL, Ω_s is unbounded on the relative interior of $\Delta(\mathcal{A})$, but $\tilde{\pi} = \mu$ still gives $\Omega_s(\mu_s) = 0$; for the floor-wrapped regularizers, $\Omega_{\epsilon,s}(\mu_s) = (1-\epsilon)\Omega_{0,s}(\mu_s)$ (Lemma G.10). The assumption says nothing about the domination profile: regularizers with sparse greedy policies satisfy it while failing (P1) (Remark G.6).

G.3.1 The discounted regularized problem

Lemma G.4 (Discounted fixed point)**.** Let $\lambda \in (0,1)$ and $\mathcal{T}^\lambda_\Omega(Q) := r + \lambda P\,\Omega^*(Q)$. Then $\mathcal{T}^\lambda_\Omega$ is a λ-contraction for $\|\cdot\|_\infty$ with unique fixed point Q^λ, and $V^\lambda := \Omega^*(Q^\lambda)$ satisfies

$$Q^\lambda(s,a) = r(s,a) + \lambda \sum_{s'} p(s'|s,a)\, V^\lambda(s'), \qquad V^\lambda = \mathcal{T}^\lambda_\Omega V^\lambda. \tag{G.5}$$

The operator $\mathcal{T}^\lambda_\Omega$ is monotone and λ-translation equivariant, and the maximum defining $\mathcal{T}^\lambda_\Omega V^\lambda$ is attained at the unique $\pi^\lambda := \pi_{\Omega,Q^\lambda} = \nabla\Omega^*(Q^\lambda)$; in particular $V^\lambda = \mathcal{E}^\lambda_{\pi^\lambda} V^\lambda \ge \mathcal{E}^\lambda_\pi V^\lambda$ for every admissible π.

Proof. We have $\mathcal{T}_\Omega^\lambda(Q) - \mathcal{T}_\Omega^\lambda(Q') = \lambda P(\Omega^*(Q) - \Omega^*(Q'))$ with P row-stochastic and Ω^* non-expansive for $\|\cdot\|_\infty$ (Lemma F.1.(v)), so $\mathcal{T}_\Omega^\lambda$ is a λ-contraction on the complete space $(\mathbb{R}^{\mathcal{S}\times\mathcal{A}}, \|\cdot\|_\infty)$ and Banach's theorem applies; this is the regularized discounted operator of (Geist et al., 2019, Thm. 1). Substituting the first identity of (G.5) into the definition of Ω^* gives, for every s,

$$V^\lambda(s) = \max_{\pi_s \in \mathrm{dom}(\Omega_s)} \left\{\tilde{r}_\pi(s) + \lambda \left(P_\pi V^\lambda\right)(s)\right\} = \left(\mathcal{T}_\Omega^\lambda V^\lambda\right)(s), \tag{G.6}$$

the maximizer being unique and equal to $\nabla\Omega^*(Q^\lambda)$ by Lemma F.1.(i), since $\tau > 0$. Monotonicity and translation equivariance of $\mathcal{T}_\Omega^\lambda$ are immediate from (G.4), each $\mathcal{E}_\pi^\lambda$ having them and the state-wise maximum preserving them. □

Lemma G.5 (Variational representation of the discounted values)**.** For an admissible stationary policy π define the discounted regularized return $J_\pi^\lambda(s) := \mathbb{E}_\pi\left[\sum_{t\geq 0} \lambda^t \tilde{r}_\pi(s_t) | s_0 = s\right]$, which is finite since $\tilde{r}_\pi$ is a finite vector. Then, for every $\lambda \in (0,1)$ and every $s \in \mathcal{S}$,

$$V^\lambda(s) = \max_{\pi \text{ admissible stationary}} J_\pi^\lambda(s), \tag{G.7}$$

the maximum being attained at π^λ.

Proof. Let π be an admissible stationary policy. Iterating $V^\lambda \geq \mathcal{E}_\pi^\lambda V^\lambda$ (Lemma G.4) K times, the operator $\mathcal{E}_\pi^\lambda$ being monotone, and using the Markov property we obtain $V^\lambda(s) \geq \left(\mathcal{E}_\pi^\lambda\right)^K V^\lambda(s) = \mathbb{E}_\pi\left[\sum_{t<K} \lambda^t \tilde{r}_\pi(s_t) | s_0 = s\right] + \lambda^K P_\pi^K V^\lambda(s)$. Since V^λ is a finite vector and P_π^K is row-stochastic, $\left\|\lambda^K P_\pi^K V^\lambda\right\|_{\mathcal{S},\infty} \leq \lambda^K \|V^\lambda\|_{\mathcal{S},\infty} \to 0$ as $K \to +\infty$ at fixed λ, so $V^\lambda \geq J_\pi^\lambda$. The same computation with $\pi = \pi^\lambda$, for which $V^\lambda = \mathcal{E}_{\pi^\lambda}^\lambda V^\lambda$ holds with equality, gives $V^\lambda = J_{\pi^\lambda}^\lambda$. □

Lemma G.6 (Normalized values are uniformly bounded)**.** For every $\lambda \in (0,1)$,

$$(1-\lambda)\left\|V^\lambda\right\|_{\mathcal{S},\infty} \leq R_{\max}, \qquad (1-\lambda)\left\|Q^\lambda\right\|_\infty \leq 2R_{\max}. \tag{G.8}$$

Proof. By (G.3), every admissible π satisfies the one-sided bound $\tilde{r}_\pi \leq \|r\|_\infty - \tau\underline{\Omega}_s \leq R_{\max}$, while the policy $\tilde{\pi}$ of Assumption G.1 satisfies the two-sided bound $|\tilde{r}_{\tilde{\pi}}(s)| \leq \|r\|_\infty + \tau|\Omega_s(\tilde{\pi}_s)| \leq R_{\max}$. Bounding the summand of J_π^λ termwise and summing the geometric series therefore gives $J_\pi^\lambda \leq \frac{R_{\max}}{1-\lambda}$ for every admissible π, and $J_{\tilde{\pi}}^\lambda \geq -\frac{R_{\max}}{1-\lambda}$. Since V^λ is the maximum of J_π^λ over admissible π (Lemma G.5), the upper bound propagates to V^λ and the lower bound is inherited from the single policy $\tilde{\pi}$, whence $\|V^\lambda\|_{\mathcal{S},\infty} \leq \frac{R_{\max}}{1-\lambda}$. The claim on Q^λ follows from (G.5): $(1-\lambda)\|Q^\lambda\|_\infty \leq (1-\lambda)\|r\|_\infty + \lambda(1-\lambda)\|V^\lambda\|_{\mathcal{S},\infty} \leq 2R_{\max}$. □

G.3.2 A NECESSARY AND SUFFICIENT CONDITION FOR THE EXISTENCE OF THE REGULARIZED FIXED POINT

Assumption G.2 (Uniformly bounded relative values)**.** There exist $\lambda_0 \in (0,1)$ and $C_V < +\infty$ such that $\left\|V^\lambda\right\|_{\mathcal{S},\mathrm{sp}} \leq C_V$ for every $\lambda \in [\lambda_0, 1)$.

Theorem G.2 (Existence of a fixed point for $\mathcal{T}_\Omega$)**.** Let Assumptions 2, G.1 and G.2 hold. Then there exist $Q_\Omega^* \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\rho_\Omega^* \in \mathbb{R}$ such that

$$\mathcal{T}_\Omega\left(Q_\Omega^*\right) = Q_\Omega^* + \rho_\Omega^* \mathbf{1}, \tag{G.9}$$

with $\left\|Q_\Omega^*\right\|_{\mathrm{sp}} \leq \|r\|_{\mathrm{sp}} + C_V$ and $|\rho_\Omega^*| \leq 2R_{\max}$.

Proof. Fix $(s_0, a_0) \in \mathcal{S} \times \mathcal{A}$ and set, for $\lambda \in [\lambda_0, 1)$, $\bar{Q}^\lambda := Q^\lambda - Q^\lambda(s_0,a_0)\mathbf{1}$ and $\rho^\lambda := (1-\lambda)Q^\lambda(s_0,a_0)$.

By translation equivariance of Ω^* (Lemma F.1.(iii)) and $P\mathbf{1} = \mathbf{1}$, substituting $Q^\lambda = \bar{Q}^\lambda + Q^\lambda(s_0,a_0)\mathbf{1}$ into (G.5) and subtracting $\lambda Q^\lambda(s_0,a_0)$ gives

$$\bar{Q}^\lambda(s,a) + \rho^\lambda = r(s,a) + \lambda \sum_{s'} p(s'|s,a)\, \Omega^*\left(\bar{Q}^\lambda\right)(s'). \tag{G.10}$$

By Lemma G.6, $|\rho^\lambda| \le (1-\lambda)\|Q^\lambda\|_\infty \le 2R_{\max}$. By (G.5) and Lemma E.6, $\|Q^\lambda\|_{\mathrm{sp}} \le \|r\|_{\mathrm{sp}} + \lambda\|V^\lambda\|_{\mathcal{S},\mathrm{sp}} \le \|r\|_{\mathrm{sp}} + C_V$ under Assumption G.2, hence $\|\bar{Q}^\lambda\|_\infty = \max_{s,a}|Q^\lambda(s,a) - Q^\lambda(s_0,a_0)| \le \|Q^\lambda\|_{\mathrm{sp}} \le \|r\|_{\mathrm{sp}} + C_V$. Both bounds are uniform in $\lambda \in [\lambda_0, 1)$.

Let $\lambda_n \uparrow 1$ with $\lambda_n \ge \lambda_0$. The pairs $(\bar{Q}^{\lambda_n}, \rho^{\lambda_n})$ lie in a compact subset of $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}\times\mathbb{R}$, so by Bolzano–Weierstrass some subsequence $\lambda_{\varphi(n)} \uparrow 1$ satisfies $\bar{Q}^{\lambda_{\varphi(n)}} \to Q^*_\Omega$ and $\rho^{\lambda_{\varphi(n)}} \to \rho^*_\Omega$. Since Ω^* is non-expansive hence continuous, letting $n \to +\infty$ in (G.10) gives $Q^*_\Omega + \rho^*_\Omega \mathbf{1} = r + P\,\Omega^*(Q^*_\Omega) = \mathcal{T}_\Omega(Q^*_\Omega)$. The bounds are preserved in the limit. □

Proposition G.1 (Assumption G.2 is also necessary)**.** Conversely, if (G.9) admits a solution $(Q^*_\Omega, \rho^*_\Omega)$, then, with $V^* := \Omega^*(Q^*_\Omega)$,

$$\left\|V^\lambda\right\|_{\mathcal{S},\mathrm{sp}} \;\le\; 2\,\|V^*\|_{\mathcal{S},\mathrm{sp}} \qquad \text{for every } \lambda \in (0,1),$$

so Assumption G.2 holds. It therefore characterizes exactly the solvability of the regularized average-reward optimality equation, as in the unregularized case (Schweitzer & Federgruen, 1978; Puterman, 1994, §8.10).

Proof. Replacing Q^*_Ω by $Q^*_\Omega + c\mathbf{1}$ changes V^* by c (Lemma F.1.(iii)) and leaves (G.9) unchanged, so we may center V^* so that $\|V^*\|_{\mathcal{S},\infty} = \frac{1}{2}\|V^*\|_{\mathcal{S},\mathrm{sp}} =: \kappa$. Taking the conjugate in (G.9) as in (G.6) gives the regularized average-reward optimality equation

$$V^* + \rho^*_\Omega \mathbf{1} = \max_{\pi \text{ admissible}} \{\tilde{r}_\pi + P_\pi V^*\}. \tag{G.11}$$

Set $\tilde{V} := V^* + \frac{\rho^*_\Omega}{1-\lambda}\mathbf{1}$. Using (G.11) and $P_\pi V^* \ge -\kappa\mathbf{1}$ termwise,

$$\begin{aligned}\mathcal{T}^\lambda_\Omega \tilde{V} &= \max_\pi \{\tilde{r}_\pi + P_\pi V^* - (1-\lambda)P_\pi V^*\} + \frac{\lambda\rho^*_\Omega}{1-\lambda}\mathbf{1} \\ &\le V^* + \rho^*_\Omega \mathbf{1} + (1-\lambda)\kappa\mathbf{1} + \frac{\lambda\rho^*_\Omega}{1-\lambda}\mathbf{1} = \tilde{V} + (1-\lambda)\kappa\mathbf{1}.\end{aligned}$$

Hence $\mathcal{T}^\lambda_\Omega(\tilde{V} + \kappa\mathbf{1}) \le \tilde{V} + (1-\lambda)\kappa\mathbf{1} + \lambda\kappa\mathbf{1} = \tilde{V} + \kappa\mathbf{1}$ by λ-translation equivariance, and monotonicity of $\mathcal{T}^\lambda_\Omega$ (Lemma G.4) makes the iterates from $\tilde{V} + \kappa\mathbf{1}$ non-increasing, so their limit V^λ satisfies $V^\lambda \le \tilde{V} + \kappa\mathbf{1}$. The symmetric argument gives $V^\lambda \ge \tilde{V} - \kappa\mathbf{1}$, whence $\|V^\lambda - \tilde{V}\|_{\mathcal{S},\infty} \le \kappa$ and $\|V^\lambda\|_{\mathcal{S},\mathrm{sp}} \le \|V^*\|_{\mathcal{S},\mathrm{sp}} + 2\kappa = 2\|V^*\|_{\mathcal{S},\mathrm{sp}}$. □

G.3.3 Sufficient structural conditions

We now record the classical setting in which Assumption G.2 holds, and show that Assumption 1 is a special case of it with explicit constants.

Assumption G.3 (Weakly communicating MDP)**.** The state space splits as $\mathcal{S} = \mathcal{C} \sqcup \mathcal{S}_{\mathrm{tr}}$, where

(i) $\mathcal{C}$ is *closed*: $s \in \mathcal{C}$, $a \in \mathcal{A}$ and $p(s'|s,a) > 0$ imply $s' \in \mathcal{C}$;
(ii) $\mathcal{C}$ is *communicating*: for all $s, s' \in \mathcal{C}$ there exist a stationary policy π and $k \ge 0$ with $P^k_\pi(s'|s) > 0$;
(iii) every state of $\mathcal{S}_{\mathrm{tr}}$ is transient under every stationary policy.

This is the standard definition (Puterman, 1994, §8.3.1); $\mathcal{S}_{\mathrm{tr}}$ may be empty, in which case the MDP is communicating.

Lemma G.7 (Stopped evaluation)**.** Let $\lambda \in (0,1)$, $\mathcal{B} \subseteq \mathcal{S}$, $\sigma := \inf\{t \ge 0 | s_t \in \mathcal{B}\} \in \mathbb{N} \cup \{+\infty\}$, and let π be an admissible stationary policy. For every $K \ge 0$ and $s \in \mathcal{S}$,

$$V^\lambda(s) \;\ge\; \mathbb{E}_\pi\left[\sum_{t=0}^{\sigma\wedge K - 1} \lambda^t \tilde{r}_\pi(s_t) + \lambda^{\sigma\wedge K} V^\lambda\left(s_{\sigma\wedge K}\right) \mid s_0 = s\right], \tag{G.12}$$

with equality when $\pi = \pi^\lambda$. If moreover $\sigma < +\infty$ $\mathbb{P}_\pi(\cdot|s_0 = s)$-almost surely, the same holds with $\sigma \wedge K$ replaced by σ.

Proof. Let $G_K(s)$ denote the right-hand side of (G.12), empty sums being 0. We show $V^\lambda \geq G_K$ pointwise by induction on K, simultaneously over all states.

For $K = 0$ we have $\sigma \wedge 0 = 0$, so $G_0 = V^\lambda$ for every π. Assume $V^\lambda \geq G_K$. If $s \in \mathcal{B}$ then $\sigma = 0$ and $G_{K+1}(s) = V^\lambda(s)$. If $s \notin \mathcal{B}$ then $\sigma \geq 1$, so splitting off the first step and applying the Markov property at s_1 gives $G_{K+1}(s) = \tilde{r}_\pi(s) + \lambda \left(P_\pi G_K\right)(s)$; the induction hypothesis, P_π being non-negative, and Lemma G.4 then give

$$G_{K+1}(s) \;\leq\; \tilde{r}_\pi(s) + \lambda \left(P_\pi V^\lambda\right)(s) \;=\; \left(\mathcal{E}_\pi^\lambda V^\lambda\right)(s) \;\leq\; V^\lambda(s).$$

Both inequalities are equalities at π^λ, which attains the state-wise maximum of Lemma G.4, whence the equality case.

For the last claim, $\mathcal{S}$ is finite and π is admissible, so $\max_s |\tilde{r}_\pi(s)|$ and $\|V^\lambda\|_{\mathcal{S},\infty}$ are finite and the integrand of (G.12) is bounded, uniformly in K, by $\frac{\max_s |\tilde{r}_\pi(s)|}{1-\lambda} + \|V^\lambda\|_{\mathcal{S},\infty}$. On $\{\sigma < +\infty\}$ we have $\sigma \wedge K = \sigma$ for K large enough, so the integrand converges almost surely to its value at σ; conclude by dominated convergence. □

Lemma G.8 (Uniformly bounded absorption time)**.** Under Assumption G.3 there is $L < +\infty$ with $\mathbb{E}_\pi[\sigma_\mathcal{C} | s_0 = s] \leq L$ for every stationary policy π and every $s \in \mathcal{S}$, where $\sigma_\mathcal{C} := \inf\{t \geq 0 | s_t \in \mathcal{C}\}$.

Proof. This is the uniform-transience estimate underlying stochastic shortest path problems (Bertsekas, 2012, §3.1); we give the short argument. Let $n := |\mathcal{S}_{\mathrm{tr}}|$, and fix a stationary π and $s \in \mathcal{S}_{\mathrm{tr}}$. If $\mathbb{P}_\pi(\forall t, s_t \in \mathcal{S}_{\mathrm{tr}} | s_0 = s) > 0$, then with positive probability some state of the finite set $\mathcal{S}_{\mathrm{tr}}$ is visited infinitely often, hence is recurrent for P_π, contradicting Assumption G.3.(iii). So $\sigma_\mathcal{C} < +\infty$ a.s., and deleting cycles from a path reaching $\mathcal{C}$ gives $\mathbb{P}_\pi(\sigma_\mathcal{C} > n | s_0 = s) < 1$. The map $\pi \mapsto \mathbb{P}_\pi(\sigma_\mathcal{C} > n | s_0 = s)$ is polynomial, hence continuous on the compact $\prod_s \Delta(\mathcal{A})$, so $\theta := \max_{s \in \mathcal{S}_{\mathrm{tr}}} \max_\pi \mathbb{P}_\pi(\sigma_\mathcal{C} > n | s_0 = s) < 1$; the Markov property gives $\mathbb{P}_\pi(\sigma_\mathcal{C} > kn | s_0 = s) \leq \theta^k$ and $\mathbb{E}_\pi[\sigma_\mathcal{C} | s_0 = s] = \sum_{j \geq 0} \mathbb{P}_\pi(\sigma_\mathcal{C} > j | s_0 = s) \leq n \sum_{k \geq 0} \mathbb{P}_\pi(\sigma_\mathcal{C} > kn | s_0 = s) \leq \frac{n}{1-\theta} =: L$. If $s \in \mathcal{C}$ then $\sigma_\mathcal{C} = 0$, so the bound holds for every $s \in \mathcal{S}$. □

Proposition G.2 (Weakly communicating implies Assumption G.2)**.** Let Assumptions G.1 and G.3 hold. Let $K^\star \geq 1$ and $p^\star > 0$ be such that for every ordered pair (s, s') of states of $\mathcal{C}$ there is $k(s, s') \leq K^\star$ with $P_{\tilde{\pi}}^{k(s,s')}(s' | s) \geq p^\star$. Such constants exist, with $K^\star \leq \max\{1, |\mathcal{C}| - 1\} \leq |\mathcal{S}|$, and they depend only on p and $\tilde{\pi}$, not on λ. For any such pair, Assumption G.2 holds with

$$\lambda_0 := 1 - \frac{1}{2K^\star}, \qquad C_V := \frac{4K^\star R_{\max}}{p^\star} + 4LR_{\max}. \tag{G.13}$$

Proof. Let $s, s' \in \mathcal{C}$. By Assumption G.3.(ii) there are π and k with $P_\pi^k(s' | s) > 0$, hence a path $s = s_0, \ldots, s_k = s'$ with $p(s_{i+1} | s_i, a_i) > 0$ for some actions a_i; it stays in $\mathcal{C}$, which is closed, and deleting its cycles leaves a simple path, so $k(s, s') \leq \max\{1, |\mathcal{C}| - 1\}$. Since $\tilde{\pi}$ has full support, $P_{\tilde{\pi}}(s_{i+1} | s_i) \geq \tilde{\pi}(a_i | s_i) p(s_{i+1} | s_i, a_i) > 0$ along that path, so $P_{\tilde{\pi}}^{k(s,s')}(s' | s) > 0$; taking for $K^\star$ the largest of these lengths and for $p^\star$ the smallest of these finitely many positive numbers proves that admissible constants exist, and they involve only p and $\tilde{\pi}$. Note also that $\tilde{\pi}$ is admissible by Assumption G.1 and $\pi^\lambda = \nabla\Omega^*(Q^\lambda)$ is admissible as a maximizer, so Lemma G.7 applies to both, with $|\tilde{r}_{\tilde{\pi}}| \leq R_{\max}$ and $\tilde{r}_\pi \leq R_{\max}$ for every admissible π by (G.3).

We first bound the span on $\mathcal{C}$. Fix $\lambda \in [\lambda_0, 1)$ and let $s_{\max}$ and $s_{\min}$ maximize and minimize V^λ over $\mathcal{C}$; if $s_{\max} = s_{\min}$ that span vanishes and there is nothing to prove, so assume $s_{\max} \neq s_{\min}$ and set $K := k(s_{\min}, s_{\max}) \geq 1$ and $p := P_{\tilde{\pi}}^K(s_{\max} | s_{\min}) \geq p^\star$. Applying Lemma G.7 at $s_{\min}$ with $\mathcal{B} = \emptyset$, so that $\sigma \wedge K = K$, with $\pi = \tilde{\pi}$ and $\sum_{t<K} \lambda^t \leq K$, gives $V^\lambda(s_{\min}) \geq -KR_{\max} + \lambda^K \mathbb{E}_{\tilde{\pi}}[V^\lambda(s_K) \mid s_0 = s_{\min}]$. As $\mathcal{C}$ is closed, $s_K \in \mathcal{C}$ almost surely, so $V^\lambda(s_K) \geq V^\lambda(s_{\min})$ pointwise and, isolating the event $\{s_K = s_{\max}\}$ of probability p, $\mathbb{E}_{\tilde{\pi}}[V^\lambda(s_K) \mid s_0 = s_{\min}] \geq pV^\lambda(s_{\max}) + (1-p)V^\lambda(s_{\min})$. Rearranging, the coefficient of $V^\lambda(s_{\min})$ being $1 - \lambda^K$, and using $1 - \lambda^K \leq K(1 - \lambda)$ with Lemma G.6,

$$p\,\lambda^K \left(V^\lambda(s_{\max}) - V^\lambda(s_{\min})\right) \;\leq\; KR_{\max} + \left(1 - \lambda^K\right) V^\lambda(s_{\min}) \;\leq\; 2KR_{\max}.$$

Since $K \leq K^\star$ and $\lambda^K \geq \lambda_0^{K^\star} = \left(1 - \frac{1}{2K^\star}\right)^{K^\star} \geq \frac{1}{2}$, we get $\left\|V^\lambda|_\mathcal{C}\right\|_{\mathcal{C},\mathrm{sp}} \leq 4K^\star R_{\max}/p^\star =: C_\mathcal{C}$.

It remains to pass from $\mathcal{C}$ to $\mathcal{S}$. Write $\sigma := \sigma_{\mathcal{C}}$ and fix $s \in \mathcal{S}$. By Lemma G.8, $\mathbb{E}_\pi[\sigma \mid s_0 = s] \le L$ for every stationary π, so $\sigma < +\infty$ almost surely and Lemma G.7 applies with $\mathcal{B} = \mathcal{C}$ and the stopping time σ itself. Its two terms are each bounded by $LR_{\max}$: the reward term because $\sum_{t<\sigma} \lambda^t \le \sigma$, and the boundary term, once written as $\mathbb{E}_\pi[V^\lambda(s_\sigma) \mid s_0 = s] + \mathbb{E}_\pi[(\lambda^\sigma - 1)V^\lambda(s_\sigma) \mid s_0 = s]$, because $0 \le 1 - \lambda^\sigma \le \sigma(1-\lambda)$ and $(1-\lambda)\|V^\lambda\|_{\mathcal{S},\infty} \le R_{\max}$ by Lemma G.6. Since $s_\sigma \in \mathcal{C}$, the previous paragraph gives $0 \le V^\lambda(s_\sigma) - V^\lambda(s_{\min}) \le C_{\mathcal{C}}$. Taking $\pi = \pi^\lambda$, for which Lemma G.7 holds with equality, and $\tilde{r}_{\pi^\lambda} \le R_{\max}$, yields $V^\lambda(s) \le V^\lambda(s_{\min}) + C_{\mathcal{C}} + 2LR_{\max}$; taking $\pi = \tilde{\pi}$ and $\tilde{r}_{\tilde{\pi}} \ge -R_{\max}$ yields $V^\lambda(s) \ge V^\lambda(s_{\min}) - 2LR_{\max}$. Both hold for every s, so $\left\|V^\lambda\right\|_{\mathcal{S},\mathrm{sp}} \le C_{\mathcal{C}} + 4LR_{\max} = C_V$. □

Corollary G.1 (Assumption 1 implies Assumption G.2, hence Theorem G.1)**.** Let Assumptions 1 to 3 hold. Then the MDP is communicating ($\mathcal{C} = \mathcal{S}$, $\mathcal{S}_{\mathrm{tr}} = \emptyset$, $L = 0$), Proposition G.2 applies with $\tilde{\pi} = \mu$, $K^\star = m$ and $p^\star = \beta_m \nu_{\min}$, where $\nu_{\min} := \min_s \nu(s) > 0$, so Assumption G.2 holds with

$$\lambda_0 = 1 - \frac{1}{2m}, \qquad C_V = \frac{4mR_{\max}}{\beta_m \nu_{\min}},$$

and Theorem G.2 gives Theorem G.1.

Proof. Assumption 3 makes Assumption G.1 hold with $\tilde{\pi} = \mu$, which has full support, so that $\delta_\Omega = \max_s \max\{|\underline{\Omega}_s|, |\Omega_s(\mu_s)|\}$ and $R_{\max}$ are the constants of Theorem G.1. For all s, s', $P_\mu^m(s'|s) = \sum_a \mu(a|s) K_\mu^m(s'|s,a) \ge \beta_m \nu(s') \ge \beta_m \nu_{\min} > 0$, so every ordered pair of states is joined by a path of length m under μ: Assumption G.3 holds with $\mathcal{C} = \mathcal{S}$, which is closed and communicating, and $\mathcal{S}_{\mathrm{tr}} = \emptyset$, so $\sigma_{\mathcal{C}} \equiv 0$ and $L = 0$ in Lemma G.8. Proposition G.2 then applies with $K^\star = m$ and $p^\star = \beta_m \nu_{\min}$, its last paragraph being vacuous, and gives Assumption G.2 with the stated constants. Finally Theorem G.2 yields (G.9) with $\|Q_\Omega^*\|_{\mathrm{sp}} \le \|r\|_{\mathrm{sp}} + C_V$ and $|\rho_\Omega^*| \le 2R_{\max}$. The gain ρ_Ω^* is unique and optimal by Propositions G.3 and G.5, and under (P1) Q_Ω^* is unique up to an additive constant by Proposition G.4, whose Assumption G.4 holds since the MDP is communicating (Remark G.5). □

Remark G.4 (On the structural hypothesis)**.** Among the classical ergodic classes of finite MDPs (Puterman, 1994, §8.3.1), weak communication is the weakest: both *unichain* models — in which every stationary policy induces a chain with a single recurrent class plus a possibly empty set of transient states — and *communicating* models are weakly communicating, while unichain and communicating are incomparable, neither class containing the other. Proposition G.2 therefore covers all these models at once.

Assumption G.2 is nonetheless strictly weaker, being stated for the reward at hand rather than for p alone: two isolated states carrying the same reward do not form a weakly communicating MDP, yet V^λ is constant, its span vanishes and the fixed point exists. It is moreover necessary as well as sufficient (Proposition G.1), stated directly for the regularized model.

G.3.4 Uniqueness and interpretation of the gain

Proposition G.3 (Uniqueness of the gain)**.** Let Assumption 2 hold. If $\mathcal{T}_\Omega(Q_i) = Q_i + \rho_i \mathbf{1}$ for $i = 1, 2$, then $\rho_1 = \rho_2$. No structural assumption on the MDP is needed here, unlike for the uniqueness of Q_Ω^* (Proposition G.4).

Proof. Let $c := \max_{s,a}(Q_1 - Q_2)(s,a)$, so $Q_1 \le Q_2 + c\mathbf{1}$. By monotonicity and translation equivariance of $\mathcal{T}_\Omega$ (Corollary F.1), $\mathcal{T}_\Omega(Q_1) \le \mathcal{T}_\Omega(Q_2) + c\mathbf{1}$, i.e. $Q_1 - Q_2 \le (\rho_2 - \rho_1 + c)\mathbf{1}$. Evaluating at a maximizer of $Q_1 - Q_2$ gives $c \le \rho_2 - \rho_1 + c$, i.e. $\rho_1 \le \rho_2$; exchanging the roles of Q_1 and Q_2 gives the reverse inequality. □

Assumption G.4 (Unique terminal class)**.** The union graph G of Definition D.1 has a unique terminal strongly connected component, i.e. a unique strongly connected component $\mathcal{C}$ with no outgoing edge.

Remark G.5. Assumption G.3 implies Assumption G.4: its set $\mathcal{C}$ is terminal by closedness and connectedness, and a second terminal component $\mathcal{C}' \subseteq \mathcal{S}_{\mathrm{tr}}$ would be closed, so every stationary policy then has a recurrent state inside, contradicting (iii). The converse fails: with $\mathcal{S} = \{1, 2, 3\}$, a self-loop and an edge $1 \to 2$ at state 1, and $2 \to 3$, $3 \to 3$, the terminal component $\{3\}$ is unique, yet state 1 is recurrent under the self-loop policy, so the MDP is not weakly communicating.

Proposition G.4 (Uniqueness of Q^*_Ω up to an additive constant)**.** Let Assumption G.4 hold and assume (P1), i.e. that the greedy policies have full support (Lemma G.3). If $\mathcal{T}_\Omega(Q_i) = Q_i + \rho_i \mathbf{1}$ for $i = 1, 2$, then $Q_1 - Q_2 \in \mathbb{R}\mathbf{1}$. Conversely $Q^*_\Omega + c\mathbf{1}$ solves (G.9) with the same ρ^*_Ω, for every $c \in \mathbb{R}$.

Proof. The converse is translation equivariance (Corollary F.1). Directly, $\rho_1 = \rho_2$ by Proposition G.3, so $h := Q_1 - Q_2$ satisfies $h = \mathcal{T}_\Omega(Q_1) - \mathcal{T}_\Omega(Q_2) = K_{\bar\pi} h$ by Lemma G.2.(i), where $\bar\pi := \bar\pi_{\Omega,[Q_1,Q_2]}$: the function h is harmonic for the state-action kernel $K_{\bar\pi}$.

By Definition G.1.(i), $\bar\pi$ is a mixture of greedy policies, hence has full support, so $P_{\bar\pi}(s'|s) > 0$ exactly when $s \to s'$ in G. The recurrent classes of the state chain are therefore the terminal strongly connected components of G, and there is only one, $\mathcal{C}$, by Assumption G.4; since $\bar\pi(a|s) > 0$ for every a, the state-action chain has $\mathcal{C} \times \mathcal{A}$ as its unique recurrent class. A bounded harmonic function of a finite Markov chain is constant on each recurrent class (Puterman, 1994, App. A.4), and on a transient s we have $h(s,a) = \mathbb{E}[h(s_n, a_n)|s_0 = s, a_0 = a]$ for every n, which converges by absorption to the value of h on that unique class. Hence h is constant. □

Remark G.6 (Where the full-support requirement bites)**.** The full-support requirement of Proposition G.4 is exactly (P1) (Lemma G.3). It holds for the α-divergences with $\alpha \le 1$ (Theorem K.1.(iv)), hence for both KL divergences and for the Hellinger regularizer, for the negative entropy, and for every ϵ-floor-wrapped regularizer, where $\pi_{\Omega_\epsilon,Q} \ge \epsilon\mu > 0$ by Lemma G.10.(ii). It fails for regularizers whose greedy policy may be sparse, that is for $\alpha > 1$ and in particular for the χ^2: there ρ^*_Ω is still unique by Proposition G.3, but Q^*_Ω need not be, exactly as at $\tau = 0$ (Puterman, 1994, §9.1).

Proposition G.5 (ρ^*_Ω is the optimal regularized gain)**.** Let $(Q^*_\Omega, \rho^*_\Omega)$ solve (G.9) and set $\pi^* := \pi_{\Omega,Q^*_\Omega}$. For an admissible stationary policy π define the regularized gain $\tilde\rho_\pi(s) := \liminf_{n\to+\infty} \frac{1}{n}\mathbb{E}_\pi\left[\sum_{t=0}^{n-1} \tilde r_\pi(s_t)|s_0 = s\right]$. Then $\tilde\rho_\pi(s) \le \rho^*_\Omega = \tilde\rho_{\pi^*}(s)$ for every s and every admissible stationary π: the gain is constant in s, ρ^*_Ω is optimal, and π^* is gain-optimal.

Proof. Equation (G.11) holds with $V^* = \Omega^*(Q^*_\Omega)$, the maximum being attained state-wise at π^*. For any admissible π it gives $\tilde r_\pi \le V^* + \rho^*_\Omega \mathbf{1} - P_\pi V^*$; applying P^t_π and summing over $t < n$ telescopes to $\mathbb{E}_\pi\left[\sum_{t<n} \tilde r_\pi(s_t)|s_0 = s\right] \le n\rho^*_\Omega + V^*(s) - (P^n_\pi V^*)(s)$. Divide by n and let $n \to +\infty$, V^* being bounded. Equality holds throughout for $\pi = \pi^*$. This is the standard verification argument (Puterman, 1994, §8.4; Bertsekas, 2012, §4.2). □

G.4 Control of the bias

We compare the regularized solution with the solutions (Q^*, ρ^*) of the average-reward optimality equation $\mathcal{T}(Q^*) = Q^* + \rho^*\mathbf{1}$. Since Assumption 1 makes the MDP communicating, such solutions exist (Schweitzer & Federgruen, 1978, Theorem 3.1(b)) and ρ^* is the optimal gain, but Q^* need not be unique up to an additive constant (Wan et al., 2024). The comparison uses two tools: a pointwise version of the regularization modulus, and the classical bounds on the gain given by an approximate fixed point.

Definition G.2 (Pointwise regularization modulus)**.** For a radius $R \ge 0$, with g_Q the regularization gap of Definition G.1.(iv),

$$\Lambda^{\mathrm{pt}}_\Omega(R) := \max_{Q\,:\,\|Q\|_{\mathrm{asp}} \le R} \|g_Q\|_{\mathcal{S},\infty}\,.$$

The maximum is attained and finite by the argument of the proof of Lemma G.1, and $\Lambda_\Omega(R) \le 2\Lambda^{\mathrm{pt}}_\Omega(R)$.

Lemma G.9 (Gain bounds from an approximate fixed point)**.** Assume that $\mathcal{T}_\Omega(Q^*_\Omega) = Q^*_\Omega + \rho^*_\Omega\mathbf{1}$. If $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $c_-, c_+ \in \mathbb{R}$ satisfy $Q + c_-\mathbf{1} \le \mathcal{T}_\Omega(Q) \le Q + c_+\mathbf{1}$, then $c_- \le \rho^*_\Omega \le c_+$.

Proof. This is the regularized form of the classical value-iteration bounds on the gain (Puterman, 1994, §8.5.5), (Bertsekas, 2012, §4.3). By monotonicity and translation equivariance of $\mathcal{T}_\Omega$ (Corollary F.1), $\mathcal{T}^{n+1}_\Omega(Q) \le \mathcal{T}^n_\Omega(Q + c_+\mathbf{1}) = \mathcal{T}^n_\Omega(Q) + c_+\mathbf{1}$, so $\mathcal{T}^n_\Omega(Q) \le Q + nc_+\mathbf{1}$ by induction. Since $\mathcal{T}_\Omega$ is non-expansive for $\|\cdot\|_\infty$ (Corollary F.1), with $d := \|Q - Q^*_\Omega\|_\infty$ we get $Q^*_\Omega + n\rho^*_\Omega\mathbf{1} = \mathcal{T}^n_\Omega(Q^*_\Omega) \le \mathcal{T}^n_\Omega(Q) + d\mathbf{1} \le Q + nc_+\mathbf{1} + d\mathbf{1}$. Reading any coordinate, dividing by n and letting $n \to +\infty$ gives $\rho^*_\Omega \le c_+$; the other inequality is symmetric. □

Proposition G.6 (Cost of Ω-regularization)**.** Let Assumptions 1 to 3 hold, and let (Q^*, ρ^*) solve $\mathcal{T}(Q^*) = Q^* + \rho^* \mathbf{1}$.

(i) *Value gap*: with $R^* := \max\{\|Q^*_\Omega\|_{\mathrm{asp}}, \max_{0\le k<m} \|\mathcal{T}^k_\Omega(Q^*)\|_{\mathrm{asp}}\}$ and $\alpha(P) \le 1$ the Dobrushin coefficient of the transition kernel P (Definition E.2),

$$\|Q^*_\Omega - Q^*\|_{\mathrm{sp}} \le \frac{m\,\alpha(P)\,\Lambda_\Omega(R^*)}{\beta_m \gamma_\Omega(R^*)^{m-1}} \le \frac{m\,\Lambda_\Omega(R^*)}{\beta_m \gamma_\Omega(R^*)^{m-1}},$$

the right-hand sides being $+\infty$ if $\gamma_\Omega(R^*) = 0$.

(ii) *Gain gap*: with $\underline{\Omega} := \min_{s\in\mathcal{S}} \min_{\pi_s \in \Delta(\mathcal{A})} \Omega_s(\pi_s)$,

$$\min_{s\in\mathcal{S}} g_{Q^*}(s) \;\le\; \rho^*_\Omega - \rho^* \;\le\; \max_{s\in\mathcal{S}} g_{Q^*}(s) \;\le\; -\tau\underline{\Omega}, \tag{G.14}$$

and consequently

$$|\rho^*_\Omega - \rho^*| \;\le\; \|g_{Q^*}\|_{\mathcal{S},\infty} \;\le\; \Lambda^{\mathrm{pt}}_\Omega(\|Q^*\|_{\mathrm{asp}}). \tag{G.15}$$

If moreover $|\mathcal{S}| \ge 2$ and $\min \Omega_s = \underline{\Omega}$ for every s, then $-\Lambda_\Omega(\|Q^*\|_{\mathrm{asp}}) \le \rho^*_\Omega - \rho^* + \tau\underline{\Omega} \le 0$.

Proof. The regularized fixed point $(Q^*_\Omega, \rho^*_\Omega)$ exists by Theorem G.1, and both $\mathcal{T}$ and $\mathcal{T}_\Omega$ are translation equivariant (Corollary F.1), so that $\mathcal{T}^k(Q^*) = Q^* + k\rho^*\mathbf{1}$ and $\mathcal{T}^k_\Omega(Q^*_\Omega) = Q^*_\Omega + k\rho^*_\Omega \mathbf{1}$ for every $k \ge 0$.

(i) The span seminorm being invariant under constants,

$$\|Q^*_\Omega - Q^*\|_{\mathrm{sp}} = \|\mathcal{T}^m_\Omega(Q^*_\Omega) - \mathcal{T}^m(Q^*)\|_{\mathrm{sp}} \le \underbrace{\|\mathcal{T}^m_\Omega(Q^*_\Omega) - \mathcal{T}^m_\Omega(Q^*)\|_{\mathrm{sp}}}_{\text{(a)}} + \underbrace{\|\mathcal{T}^m_\Omega(Q^*) - \mathcal{T}^m(Q^*)\|_{\mathrm{sp}}}_{\text{(b)}}. \tag{G.16}$$

Term (a). Since $\|\cdot\|_{\mathrm{asp}}$ is invariant under constants, $\|\mathcal{T}^k_\Omega(Q^*_\Omega)\|_{\mathrm{asp}} = \|Q^*_\Omega\|_{\mathrm{asp}}$, so that for every $k \in [\![1, m-1]\!]$,

$$R\big[\mathcal{T}^k_\Omega(Q^*_\Omega), \mathcal{T}^k_\Omega(Q^*)\big] = \max\{\|Q^*_\Omega\|_{\mathrm{asp}}, \|\mathcal{T}^k_\Omega(Q^*)\|_{\mathrm{asp}}\} \le R^*.$$

As γ_Ω is non-increasing (Definition G.1.(iii)), every factor of the product in Lemma G.2.(iv), applied with $Q = Q^*_\Omega$ and $Q' = Q^*$, is at least $\gamma_\Omega(R^*)$, hence (a) $\le \big(1 - \beta_m \gamma_\Omega(R^*)^{m-1}\big)\|Q^*_\Omega - Q^*\|_{\mathrm{sp}}$.

Term (b). Let $A_k := \mathcal{T}^k_\Omega\big(\mathcal{T}^{m-k}(Q^*)\big)$ for $k \in [\![0, m]\!]$, so that $A_0 = \mathcal{T}^m(Q^*)$ and $A_m = \mathcal{T}^m_\Omega(Q^*)$. Since $\mathcal{T}^{m-k-1}(Q^*) = Q^* + (m-k-1)\rho^*\mathbf{1}$, translation equivariance makes the constants cancel, and k applications of the non-expansiveness of $\mathcal{T}_\Omega$ in $\|\cdot\|_{\mathrm{sp}}$ give

$$\begin{aligned}\|A_{k+1} - A_k\|_{\mathrm{sp}} &= \big\|\mathcal{T}^k_\Omega\big(\mathcal{T}_\Omega(Q^*)\big) - \mathcal{T}^k_\Omega\big(\mathcal{T}(Q^*)\big)\big\|_{\mathrm{sp}} \le \|\mathcal{T}_\Omega(Q^*) - \mathcal{T}(Q^*)\|_{\mathrm{sp}} \\ &\le \alpha(P)\Lambda_\Omega(\|Q^*\|_{\mathrm{asp}}) \le \alpha(P)\Lambda_\Omega(R^*),\end{aligned}$$

by Lemma G.2.(vi), $\|Q^*\|_{\mathrm{asp}} \le R^*$ and the monotonicity of Λ_Ω (Lemma G.1). Summing over k, (b) $\le m\,\alpha(P)\Lambda_\Omega(R^*)$.

Injecting both bounds into (G.16) yields $\beta_m\gamma_\Omega(R^*)^{m-1}\|Q^*_\Omega - Q^*\|_{\mathrm{sp}} \le m\,\alpha(P)\Lambda_\Omega(R^*)$, which is the claim when $\gamma_\Omega(R^*) > 0$; otherwise the bound is trivial.

(ii) By Lemma G.2.(vi) and its proof, $\mathcal{T}_\Omega(Q) - \mathcal{T}(Q) = Pg_Q$ for every Q, so that $\mathcal{T}_\Omega(Q^*) = Q^* + \rho^*\mathbf{1} + Pg_{Q^*}$. Each entry of Pg_{Q^*} is a convex combination of the entries of g_{Q^*}, and Lemma G.9 with $c_- := \rho^* + \min_s g_{Q^*}(s)$ and $c_+ := \rho^* + \max_s g_{Q^*}(s)$ gives the first two inequalities of (G.14). The last one is the upper part of Lemma F.1.(ii), which needs no assumption on the Diracs: $g_{Q^*}(s) \le -\tau\underline{\Omega}_s \le -\tau\underline{\Omega}$. Finally, (G.15) follows from (G.14) and Definition G.2. For the last claim, fix s and $s_0 \ne s$, and let $\tilde{Q}$ equal Q^* except at s_0, where it is constant in the action. Then $\|\tilde{Q}\|_{\mathrm{asp}} \le \|Q^*\|_{\mathrm{asp}}$, and since $g_Q(s)$ only depends on $Q(s,\cdot)$ and equals $-\tau \min \Omega_{s_0}$ at a constant, $g_{\tilde{Q}}(s) = g_{Q^*}(s)$ and $g_{\tilde{Q}}(s_0) = -\tau\underline{\Omega}$; hence $-\tau\underline{\Omega} - g_{Q^*}(s) \le \|g_{\tilde{Q}}\|_{\mathcal{S},\mathrm{sp}} \le \Lambda_\Omega(\|Q^*\|_{\mathrm{asp}})$. Together with $g_{Q^*} \le -\tau\underline{\Omega}$ and (G.14), this gives the claim. □

Item (i) bounds the gap by the regularization modulus times the effective horizon $m/(\beta_m\gamma_\Omega(R^*)^{m-1})$, the average-reward counterpart of the discounted bias bound $\tau(\overline{\Omega} - \underline{\Omega})/(1-\gamma)$

of Geist et al. (2019), where $\overline{\Omega} := \max_{s,a} \Omega_s(\delta_a)$. It exhibits the trade-off that runs through the paper: a larger temperature raises γ_Ω, hence the contraction rates of Section 4, but also inflates Λ_Ω. Unlike $1/(1-\gamma)$, this effective horizon depends on τ, so (i) is informative only when $\gamma_\Omega(R^*)$ is not too small; for the reverse KL the bound grows as $e^{(m-1)R^*/\tau}$ when $\tau \to 0$. Item (ii) has no such limitation, as the next corollary shows. Regularization also removes a degeneracy of the unregularized problem: under (P1), Q^*_Ω is unique up to an additive constant (Proposition G.4), and (i) bounds its distance to every solution Q^*.

Corollary G.2 (The gain gap vanishes with τ)**.** Under the assumptions of Remark G.1, $-\tau\overline{\Omega} \le \rho^*_\Omega - \rho^* \le -\tau\underline{\Omega}$, so $\rho^*_\Omega \to \rho^*$ linearly as $\tau \to 0$, with constants that are free of β_m, of γ_Ω and of the radius.

Proof. When every Dirac lies in the effective domain, Lemma F.1.(ii) gives $g_{Q^*}(s) \ge -\tau\overline{\Omega}$ for every s, and (G.14) concludes. □

Example G.1. For the reverse KL (Corollary K.1, $\alpha = 1$), $\underline{\Omega}_s = 0$ and $\overline{\Omega} = \log(1/\mu_{\min})$ with $\mu_{\min} := \min_{s,a} \mu(a|s)$, so Corollary G.2 reads $-\tau \log(1/\mu_{\min}) \le \rho^*_\Omega - \rho^* \le 0$. For the forward KL (Corollary K.1, $\alpha = 0$), the Diracs lie outside the effective domain, but $\underline{\Omega}_s = 0$ still gives $\rho^*_\Omega \le \rho^*$, and (G.15) and Theorem K.1.(v) give $|\rho^*_\Omega - \rho^*| \le \tau\big(1 + \log_+(\|Q^*\|_{\mathrm{asp}}/\tau)\big) = O\big(\tau \log(1/\tau)\big)$ as $\tau \to 0$.

The gain gap also controls the performance of the policies returned by our algorithms: the greedy policy of any Q close to Q^*_Ω in span is near-optimal.

Corollary G.3 (Performance of the greedy policy)**.** Let Assumptions 1 to 3 hold, let $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\pi := \pi_{\Omega,Q}$. Then, for every $s \in \mathcal{S}$, the regularized gain of Proposition G.5 satisfies $\rho^*_\Omega - \|Q - Q^*_\Omega\|_{\mathrm{sp}} \le \tilde{\rho}_\pi(s) \le \rho^*_\Omega$, and the gain of Subsection 2.1 satisfies $\rho^\pi(s) \ge \tilde{\rho}_\pi(s) + \tau\underline{\Omega}$. In particular, under the assumptions of Remark G.1, $\rho^* - \rho^\pi(s) \le \|Q - Q^*_\Omega\|_{\mathrm{sp}} + \tau\Delta_\Omega$.

Proof. The upper bound is Proposition G.5. For the lower bound, let $\mathcal{T}^\pi(U)(s,a) := r(s,a) + \sum_{s'} p(s'|s,a)\big(\langle \pi_{s'}, U_{s'}\rangle - \tau\Omega_{s'}(\pi_{s'})\big)$ be the regularized evaluation operator of π; it is monotone and translation equivariant, and $\mathcal{T}^\pi(Q) = \mathcal{T}_\Omega(Q)$ since π is greedy for Q. With $u := Q - Q^*_\Omega$, Lemma G.2.(i) gives $\mathcal{T}_\Omega(Q) - Q - \rho^*_\Omega \mathbf{1} = K_{\bar{\pi}} u - u \ge -\|u\|_{\mathrm{sp}}\mathbf{1}$, the entries of $K_{\bar{\pi}}u$ lying in $[\min u, \max u]$. Hence $\mathcal{T}^\pi(Q) \ge Q + (\rho^*_\Omega - \|u\|_{\mathrm{sp}})\mathbf{1}$ and, as in the proof of Lemma G.9, $(\mathcal{T}^\pi)^n(Q) \ge Q + n(\rho^*_\Omega - \|u\|_{\mathrm{sp}})\mathbf{1}$. Since $(\mathcal{T}^\pi)^n(Q)(s,a)$ is $r(s,a)$ plus the expected sum of $\tilde{r}_\pi(s_t)$, $1 \le t < n$, along π from (s,a), up to a bounded term, dividing by n gives $\tilde{\rho}_\pi(s) \ge \rho^*_\Omega - \|u\|_{\mathrm{sp}}$. Next, $\langle \pi_s, r_s\rangle = \tilde{r}_\pi(s) + \tau\Omega_s(\pi_s) \ge \tilde{r}_\pi(s) + \tau\underline{\Omega}$, whence $\rho^\pi \ge \tilde{\rho}_\pi + \tau\underline{\Omega}$. Finally, Corollary G.2 gives $\rho^* - \rho^*_\Omega \le \tau\overline{\Omega}$, and $\rho^* - \rho^\pi \le (\rho^* - \rho^*_\Omega) + \|u\|_{\mathrm{sp}} - \tau\underline{\Omega} \le \|u\|_{\mathrm{sp}} + \tau\Delta_\Omega$. □

G.5 ϵ-Floor-Wrapped Regularization

Regularizers with sparse greedy policies, such as the α-divergences with $\alpha > 1$, violate (P1), and even when (P1) holds the domination profile $\gamma_\Omega(R)$ may go to zero as R grows, making contraction and exploration results local. We wrap the base regularization in a hard policy floor. This restores (P1) with $\gamma_{\Omega_\epsilon} \ge \epsilon$ uniformly in R, which makes the results global (Lemma G.11), at the price of an additional bias of order ϵR^* (Remark G.7). The resulting greedy policy $\epsilon\mu + (1-\epsilon)\pi_{\Omega_0,Q}$ (Lemma G.10.(ii)) is, for a uniform μ and the negative entropy as Ω_0, the ϵ-softmax learning policy of Nanda & Chen (2026), where the mixing weight ϵ plays the same role: a uniform lower bound on the action probabilities.

Definition G.3 (ϵ-floored simplex)**.** Let $\epsilon \in [0,1)$ and let μ be the full-support mixing policy of Assumption 1. For $s \in \mathcal{S}$ define $\Delta_\epsilon(s) := \{\pi_s \in \Delta(\mathcal{A}) : \pi_s \ge \epsilon\mu_s\}$ and the affine bijection $\Phi_{\epsilon,s} : \Delta(\mathcal{A}) \to \Delta_\epsilon(s)$, $\Phi_{\epsilon,s}(p) = \epsilon\mu_s + (1-\epsilon)p$, whose inverse is $\Phi^{-1}_{\epsilon,s}(\pi_s) = (\pi_s - \epsilon\mu_s)/(1-\epsilon)$.

Definition G.4 (ϵ-floor-wrapped regularization)**.** Let Ω_0 be a base regularization satisfying Assumption 2 and let $\epsilon \in [0,1)$. The ϵ-floor-wrapped regularization Ω_ϵ is defined state-wise by

$$\Omega_{\epsilon,s}(\pi_s) = \begin{cases} (1-\epsilon)\,\Omega_{0,s}\left(\frac{\pi_s - \epsilon\mu_s}{1-\epsilon}\right) & \text{if } \pi_s \in \Delta_\epsilon(s), \\ +\infty & \text{otherwise.} \end{cases}$$

Equivalently, $\Omega_{\epsilon,s} = (1-\epsilon)\left(\Omega_{0,s} \circ \Phi_{\epsilon,s}^{-1}\right) + \iota_{\Delta_\epsilon(s)}$, where ι_C denotes the convex indicator of the set C, equal to 0 on C and $+\infty$ elsewhere. For $\epsilon = 0$ we recover Ω_0.

Lemma G.10 (Properties of Ω_ϵ)**.** Let Ω_0 satisfy Assumption 2 with strong convexity modulus σ, and let $\epsilon \in [0,1)$. Then $\Omega_{\epsilon,s}$ is proper, closed, and $\frac{\sigma}{1-\epsilon}$-strongly convex, with effective domain $\Phi_{\epsilon,s}(\mathrm{dom}(\Omega_{0,s}))$; in particular Ω_ϵ satisfies Assumption 2, and it satisfies Assumption 3 whenever Ω_0 does. Moreover:

(i) *Conjugate*: $\Omega_\epsilon^*(Q)(s) = \epsilon\langle \mu_s, Q_s\rangle + (1-\epsilon)\Omega_0^*(Q)(s)$; in particular Ω_ϵ^* is differentiable with Lipschitz gradient.
(ii) *Greedy policy*: $\pi_{\Omega_\epsilon,Q}(a|s) = \epsilon\mu(a|s) + (1-\epsilon)\pi_{\Omega_0,Q}(a|s)$.
(iii) *Domination profile*: $\gamma_{\Omega_\epsilon}(R) = \epsilon + (1-\epsilon)\gamma_{\Omega_0}(R) \geq \epsilon$ for every $R \geq 0$.
(iv) *Span of the conjugate at the origin*: $\|\Omega_\epsilon^*(0)\|_{\mathcal{S},\mathrm{sp}} = (1-\epsilon)\|\Omega_0^*(0)\|_{\mathcal{S},\mathrm{sp}}$.
(v) *Regularization modulus*: $\Lambda_{\Omega_\epsilon}(R) \leq \epsilon R + (1-\epsilon)\Lambda_{\Omega_0}(R)$ for every $R \geq 0$. In particular, if Ω_0 satisfies the assumptions of Remark G.1, then $\Lambda_{\Omega_\epsilon}(R) \leq \epsilon R + (1-\epsilon)\tau\Delta_{\Omega_0}$.

Proof of Lemma G.10. Fix $s \in \mathcal{S}$ and write $\mu_s = \mu(\cdot|s)$, $\Phi = \Phi_{\epsilon,s}$ for the affine bijection of Definition G.3.

Effective domain and properness. Since $\Omega_{\epsilon,s}(\Phi(p)) = (1-\epsilon)\Omega_{0,s}(p)$ for $p \in \Delta(\mathcal{A})$ and $\Omega_{\epsilon,s} = +\infty$ outside $\Delta_\epsilon(s) = \Phi(\Delta(\mathcal{A}))$, the effective domain of $\Omega_{\epsilon,s}$ is $\Phi(\mathrm{dom}(\Omega_{0,s}))$. It is non-empty since $\Omega_{0,s}$ is proper, and $\Omega_{\epsilon,s}$ never takes the value $-\infty$ since $\Omega_{0,s}$ does not; hence $\Omega_{\epsilon,s}$ is proper. Moreover $\Phi(\mu_s) = \mu_s$, so $\Omega_{\epsilon,s}(\mu_s) = (1-\epsilon)\Omega_{0,s}(\mu_s)$, which gives Assumption 3 for Ω_ϵ whenever it holds for Ω_0.

Closedness. $\Delta_\epsilon(s)$ is the intersection of the simplex with the closed half-spaces $\{\pi_s : \pi_s(a) \geq \epsilon\mu(a|s)\}$, hence closed and convex. The map $\Omega_{0,s}\circ\Phi^{-1}$ is proper, convex and lower semicontinuous as the composition of a proper, closed convex function with a continuous affine map, hence closed, a proper convex function being closed if and only if it is lower semicontinuous (Rockafellar, 1970, Theorem 7.1); and $\iota_{\Delta_\epsilon(s)}$ is closed since $\Delta_\epsilon(s)$ is closed. A sum of closed functions is closed.

Strong convexity. Let $\pi_s, \pi_s' \in \mathrm{dom}(\Omega_{\epsilon,s})$, $t \in [0,1]$, and set $p = \Phi^{-1}(\pi_s)$, $p' = \Phi^{-1}(\pi_s')$, so that $\pi_s - \pi_s' = (1-\epsilon)(p-p')$. Since Φ^{-1} is affine, $\Phi^{-1}(t\pi_s + (1-t)\pi_s') = tp + (1-t)p'$, and the σ-strong convexity of $\Omega_{0,s}$ gives

$$\begin{aligned}\Omega_{\epsilon,s}(t\pi_s + (1-t)\pi_s') &= (1-\epsilon)\Omega_{0,s}(tp + (1-t)p') \\ &\leq (1-\epsilon)\left[t\,\Omega_{0,s}(p) + (1-t)\,\Omega_{0,s}(p') - \frac{\sigma}{2}t(1-t)\|p-p'\|^2\right] \\ &= t\,\Omega_{\epsilon,s}(\pi_s) + (1-t)\,\Omega_{\epsilon,s}(\pi_s') - \frac{\sigma}{2(1-\epsilon)}t(1-t)\|\pi_s - \pi_s'\|^2,\end{aligned}$$

using $\|p-p'\|^2 = \|\pi_s - \pi_s'\|^2/(1-\epsilon)^2$ in the last line. Hence $\Omega_{\epsilon,s}$ is $\frac{\sigma}{1-\epsilon}$-strongly convex, and together with $\tau > 0$ this shows that Ω_ϵ satisfies Assumption 2.

(i). The constraint $\pi_s \in \Delta_\epsilon(s)$ is exactly the range of Φ, so the change of variable $\pi_s = \Phi(p) = \epsilon\mu_s + (1-\epsilon)p$ is a bijection from $\Delta(\mathcal{A})$ onto $\Delta_\epsilon(s)$. Since $\Omega_{\epsilon,s}(\Phi(p)) = (1-\epsilon)\Omega_{0,s}(p)$ and $\langle\Phi(p), Q_s\rangle = \epsilon\langle\mu_s, Q_s\rangle + (1-\epsilon)\langle p, Q_s\rangle$, we get

$$\begin{aligned}\Omega_\epsilon^*(Q)(s) &= \max_{\pi_s\in\Delta(\mathcal{A})}\{\langle\pi_s, Q_s\rangle - \tau\Omega_{\epsilon,s}(\pi_s)\} \\ &= \max_{\pi_s\in\Delta_\epsilon(s)}\{\langle\pi_s, Q_s\rangle - \tau\Omega_{\epsilon,s}(\pi_s)\} \\ &= \max_{p\in\Delta(\mathcal{A})}\{\epsilon\langle\mu_s, Q_s\rangle + (1-\epsilon)\langle p, Q_s\rangle - \tau(1-\epsilon)\,\Omega_{0,s}(p)\} \\ &= \epsilon\langle\mu_s, Q_s\rangle + (1-\epsilon)\max_{p\in\Delta(\mathcal{A})}\{\langle p, Q_s\rangle - \tau\Omega_{0,s}(p)\} \\ &= \epsilon\langle\mu_s, Q_s\rangle + (1-\epsilon)\Omega_0^*(Q)(s),\end{aligned}$$

where the second equality holds because $\Omega_{\epsilon,s} = +\infty$ outside $\Delta_\epsilon(s)$, the third is the change of variable, the fourth uses $1-\epsilon > 0$, and the last is the definition of Ω_0^*. The map $Q_s \mapsto \epsilon\langle\mu_s, Q_s\rangle$ is linear with constant gradient $\epsilon\mu_s$, so Ω_ϵ^* is differentiable as soon as Ω_0^* is (which is the case by Lemma F.1.(i) since $\tau > 0$), with $\nabla\Omega_\epsilon^* = \epsilon\mu_s + (1-\epsilon)\nabla\Omega_0^*$; if $\nabla\Omega_0^*$ is L_0-Lipschitz then $\nabla\Omega_\epsilon^*$ is $(1-\epsilon)L_0$-Lipschitz.

(ii). The maximum over p in the third line above is attained at $p^\star = \arg\max_{p\in\Delta(\mathcal{A})}\{\langle p, Q_s\rangle - \tau\Omega_{0,s}(p)\} = \pi_{\Omega_0,Q}(\cdot|s)$, which is unique by Lemma F.1.(i) applied to Ω_0. Since Φ is a bijection, the maximizer of the original problem is $\pi_{\Omega_\epsilon,Q}(\cdot|s) = \Phi(p^\star) = \epsilon\mu_s + (1-\epsilon)\pi_{\Omega_0,Q}(\cdot|s)$. This is consistent with (i) and Lemma F.1.(i), which identify the greedy policy with $\nabla\Omega^*_\epsilon(Q)(s)$.

(iii). Let $(s,a) \in \mathcal{S}\times\mathcal{A}$ and Q with $\|Q\|_{\mathrm{asp}} \le R$. Dividing (ii) by $\mu(a|s) > 0$ gives

$$\frac{\pi_{\Omega_\epsilon,Q}(a|s)}{\mu(a|s)} = \epsilon + (1-\epsilon)\frac{\pi_{\Omega_0,Q}(a|s)}{\mu(a|s)}.$$

Taking the infimum over (s,a) and over $\{Q : \|Q\|_{\mathrm{asp}} \le R\}$, we obtain $\gamma_{\Omega_\epsilon}(R) = \epsilon + (1-\epsilon)\gamma_{\Omega_0}(R)$ from Definition G.1.(iii). Since $\gamma_{\Omega_0}(R) \ge 0$ and $\epsilon < 1$, we get $\gamma_{\Omega_\epsilon}(R) \ge \epsilon$ for every $R \ge 0$.

(iv). Trivial from (i).

(v). Let Q satisfy $\|Q\|_{\mathrm{asp}} \le R$ and write g^ϵ_Q and g^0_Q for the regularization gaps of Ω_ϵ and Ω_0. Subtracting $\max_a Q_s(a)$ from (i) gives, for every $s \in \mathcal{S}$,

$$g^\epsilon_Q(s) = \epsilon\, h_Q(s) + (1-\epsilon)\, g^0_Q(s), \qquad h_Q(s) := \langle \mu_s, Q_s\rangle - \max_a Q_s(a).$$

Since $\mu_s \in \Delta(\mathcal{A})$, we have $h_Q(s) \in [-\|Q_s\|_{\mathcal{A},\mathrm{sp}}, 0] \subseteq [-R, 0]$ for every s, hence $\|h_Q\|_{\mathcal{S},\mathrm{sp}} \le R$. The triangle inequality and Definition G.1.(iv) applied to Ω_0 give

$$\|g^\epsilon_Q\|_{\mathcal{S},\mathrm{sp}} \le \epsilon\|h_Q\|_{\mathcal{S},\mathrm{sp}} + (1-\epsilon)\|g^0_Q\|_{\mathcal{S},\mathrm{sp}} \le \epsilon R + (1-\epsilon)\Lambda_{\Omega_0}(R).$$

Taking the maximum over $\{Q : \|Q\|_{\mathrm{asp}} \le R\}$ yields $\Lambda_{\Omega_\epsilon}(R) \le \epsilon R + (1-\epsilon)\Lambda_{\Omega_0}(R)$. The last claim follows from Remark G.1 applied to Ω_0. □

Lemma G.11 (Global properties of Ω_ϵ)**.** Let Assumption 1 hold, let Ω_0 satisfy Assumption 2, let $\epsilon \in (0,1)$, and let Q, Q' be two Q-value functions. Then:

(i) *Global m-step contraction*:

$$\begin{aligned}\|\mathcal{T}^m_{\Omega_\epsilon}(Q) - \mathcal{T}^m_{\Omega_\epsilon}(Q')\|_{\mathrm{sp}} &\le \Big(1 - \beta_m \prod_{k=1}^{m-1} \gamma_{\Omega_\epsilon}\big(R[\mathcal{T}^k_{\Omega_\epsilon}(Q), \mathcal{T}^k_{\Omega_\epsilon}(Q')]\big)\Big)\|Q - Q'\|_{\mathrm{sp}} \\ &\le \big(1 - \beta_m\epsilon^{m-1}\big)\,\|Q - Q'\|_{\mathrm{sp}}.\end{aligned}$$

(ii) *Global coverage*: let $(Q_n)_{n\ge0}$ be Q-functions and let $(s_n, a_n)_{n\ge0}$ be a trajectory with $a_n \sim \pi_{\Omega_\epsilon,Q_n}(\cdot|s_n)$ and $s_{n+1} \sim p(\cdot|s_n, a_n)$; let $\mathcal{F}_n := \sigma(Q_0, (s_0,a_0), \dots, (s_n,a_n))$ and assume that Q_n is $\mathcal{F}_{n-1}$-measurable for $n \ge 1$. Then for all $t \ge 0$ and all (s,a):

$$\mathbb{P}\left((s_{t+m}, a_{t+m}) = (s,a) \mid \mathcal{F}_t\right) \ge \beta_m\epsilon^m\eta(s,a) \ge \beta_m\epsilon^m\eta_{\min}. \tag{G.17}$$

(iii) *Fixed point up to a constant*: there exists $Q^*_{\Omega_\epsilon} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, unique up to an additive constant, and a unique $\rho^*_{\Omega_\epsilon} \in \mathbb{R}$, such that $\mathcal{T}_{\Omega_\epsilon}(Q^*_{\Omega_\epsilon}) = Q^*_{\Omega_\epsilon} + \rho^*_{\Omega_\epsilon}\mathbf{1}$.

Proof of Lemma G.11. By Lemma G.10, Ω_ϵ satisfies Assumption 2, so Lemma G.2 applies to Ω_ϵ.

(i). The first inequality is Lemma G.2.(iv) applied to Ω_ϵ. By Lemma G.10.(iii), $\gamma_{\Omega_\epsilon}(R) \ge \epsilon$ for every $R \ge 0$, so $\prod_{k=1}^{m-1}\gamma_{\Omega_\epsilon}(R[\mathcal{T}^k_{\Omega_\epsilon}(Q), \mathcal{T}^k_{\Omega_\epsilon}(Q')]) \ge \epsilon^{m-1}$ regardless of the radii, and the second inequality follows.

(ii). By Lemma G.10.(ii), $\pi_{\Omega_\epsilon,Q_n} \ge \epsilon\mu$ for every n, so Lemma G.2.(v) applies to Ω_ϵ with the deterministic constants $\gamma_1 = \cdots = \gamma_m = \epsilon$.

(iii). By translation equivariance (Corollary F.1), $\mathcal{T}_{\Omega_\epsilon}$ induces a well-defined map $\bar{\mathcal{T}}$ on the quotient space $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbb{R}\mathbf{1}$, which is complete for the norm induced by $\|\cdot\|_{\mathrm{sp}}$, and (i) makes $\bar{\mathcal{T}}^m$ a contraction with constant $1 - \beta_m\epsilon^{m-1} < 1$. Banach's theorem gives a unique fixed point $\bar{Q}^*$ of $\bar{\mathcal{T}}^m$; since $\bar{\mathcal{T}}$ commutes with $\bar{\mathcal{T}}^m$, $\bar{\mathcal{T}}(\bar{Q}^*)$ is also a fixed point of $\bar{\mathcal{T}}^m$, hence equals $\bar{Q}^*$. Any representative $Q^*_{\Omega_\epsilon}$ of $\bar{Q}^*$ therefore satisfies $\mathcal{T}_{\Omega_\epsilon}(Q^*_{\Omega_\epsilon}) = Q^*_{\Omega_\epsilon} + \rho^*_{\Omega_\epsilon}\mathbf{1}$ for some $\rho^*_{\Omega_\epsilon} \in \mathbb{R}$, unique by Proposition G.3, and uniqueness of $Q^*_{\Omega_\epsilon}$ up to an additive constant is that of $\bar{Q}^*$. □

Remark G.7 (Bias of the floor)**.** Applying Proposition G.6 to Ω_ϵ, which satisfies its assumptions by Lemma G.10, and using Lemma G.10.(iii) and (v) gives

$$\|Q^*_{\Omega_\epsilon} - Q^*\|_{\mathrm{sp}} \le \frac{m\,(\epsilon R^* + (1-\epsilon)\Lambda_{\Omega_0}(R^*))}{\beta_m(\epsilon + (1-\epsilon)\gamma_{\Omega_0}(R^*))^{m-1}},$$

where R^* is the radius of Proposition G.6 associated with Ω_ϵ. The floor therefore trades a *global* contraction rate $\beta_m\epsilon^{m-1}$ against a bias term of order ϵR^* which does not vanish as $\tau \to 0$: small ϵ keeps the fixed point close to Q^* but degrades the rate, large ϵ does the opposite.

H Proofs for the learning algorithms of Section 4

Proof techniques. Three ingredients drive the analysis. First, since $\nabla\Omega^*$ is the greedy policy, differences of regularized Bellman operators are kernels of average-path policies (Lemma G.2); unlike the single randomized policy that represents differences of hard maxima (Chen et al., 2026, Lemma B.6), this policy inherits the domination of the greedy policies, which turns the Doeblin condition on μ into a Doeblin condition on every product of m such kernels, much as lower bounds on learning policies transfer irreducibility in discounted on-policy Q-learning (Nanda & Chen, 2026). Second, we analyze the unmodified one-step update although $\mathcal{T}_\Omega$ only contracts over m steps: expanding the Jacobian of a block of relaxed updates as an expectation over Bernoulli selections shows that the step size acts as a lazy transformation of the state-action kernels, the counterpart of the aperiodicity transformation of Puterman (1994, §8.5.4), so that blocks with at least m selected updates inherit the m-step contraction. This avoids both the multi-step estimators of Zhang et al. (2021) and the lazy transformation of the dynamics of Chen et al. (2026). Third, since the contraction rate degrades with the radius, a stopping-time argument keeps the iterates, with high probability, in a ball on which the rate holds, with a self-consistent choice of the radius.

Throughout this section, Assumptions 1 to 3 and (P1) hold, and Q^*_Ω denotes a solution of (G.9), unique up to an additive constant (Theorem G.1); every quantity below is invariant under the choice of that constant. The Q-iteration bound is the regularized counterpart of the span-contraction analysis of relative value iteration (Puterman, 1994, §8.5). For Q-learning, we first analyze the noiseless relaxed flow, which contracts over blocks of updates (Appendix H.2.1), then track the stochastic iterates along this flow up to their exit time from a span ball (Appendix H.2.2); the epoch algorithm follows the same scheme, with batch averaging and a cover-time bound (Appendix H.3).

The results for general step-size schedules use the following assumption.

Assumption H.1 (Step sizes)**.** The sequence of step sizes $(\alpha_t)_{t\ge0}$ is non-increasing, satisfies $0 < \alpha_t \le 1$ for all $t \ge 0$, and $\sum_{t\ge0}\alpha_t = +\infty$.

H.1 Proof of Corollary 1

Corollary 1 (Synchronous Ω-regularized Q-iteration)**.** Let Assumptions 1 to 3 and (P1) hold. Then $(\|Q_t - Q^*_\Omega\|_{\mathrm{sp}})_{t\ge0}$ is non-increasing and, for every $t \ge 0$,

$$\|Q_{t+m} - Q^*_\Omega\|_{\mathrm{sp}} \le \Big(1 - \beta_m\,\gamma_\Omega\big(\|Q^*_\Omega\|_{\mathrm{asp}} + \|Q_t - Q^*_\Omega\|_{\mathrm{sp}}\big)^{m-1}\Big)\|Q_t - Q^*_\Omega\|_{\mathrm{sp}}. \tag{1}$$

Consequently, for $\varepsilon \in (0, e_0)$, $\|Q_T - Q^*_\Omega\|_{\mathrm{sp}} \le \varepsilon$ for every $T \ge m\lceil\tilde\beta^{-1}\log(e_0/\varepsilon)\rceil$, i.e. after $O\big(|\mathcal{S}|^2|\mathcal{A}|\,m\,\tilde\beta^{-1}\log(e_0/\varepsilon)\big)$ model queries, one query being an evaluation of $p(s'|s,a)$.

Proof of Corollary 1. Let Q^*_Ω be given by Theorem G.1. By translation equivariance (Corollary F.1), $\mathcal{T}^k_\Omega(Q^*_\Omega) = Q^*_\Omega + k\rho^*_\Omega\mathbf{1}$ for every $k \ge 0$ and, since the projection Π subtracts a constant, $Q_{j+k} = \mathcal{T}^k_\Omega(Q_j) - c\mathbf{1}$ for some $c \in \mathbb{R}$. Since $\|\cdot\|_{\mathrm{sp}}$ annihilates constants, $e_{j+k} = \|\mathcal{T}^k_\Omega(Q_j) - \mathcal{T}^k_\Omega(Q^*_\Omega)\|_{\mathrm{sp}}$ for every $j, k \ge 0$. With $k = 1$, the non-expansiveness of $\mathcal{T}_\Omega$ in $\|\cdot\|_{\mathrm{sp}}$ (Corollary F.1) shows that $(e_j)_j$ is non-increasing.

With $k = m$, Lemma G.2.(iv) applied to (Q_j, Q^*_Ω) gives (1), since γ_Ω is non-increasing and $R[\mathcal{T}^k_\Omega(Q_j), \mathcal{T}^k_\Omega(Q^*_\Omega)] \le \|Q^*_\Omega\|_{\mathrm{asp}} + e_j$ for $k \in [\![0, m-1]\!]$. Indeed, $\|\cdot\|_{\mathrm{asp}}$ is invariant under constants, so $\|\mathcal{T}^k_\Omega(Q^*_\Omega)\|_{\mathrm{asp}} = \|Q^*_\Omega\|_{\mathrm{asp}}$, and the triangle inequality together with $\|\cdot\|_{\mathrm{asp}} \le \|\cdot\|_{\mathrm{sp}}$ gives $\|\mathcal{T}^k_\Omega(Q_j)\|_{\mathrm{asp}} \le \|Q^*_\Omega\|_{\mathrm{asp}} + e_{j+k} \le \|Q^*_\Omega\|_{\mathrm{asp}} + e_j$.

Finally, $e_j \le e_0$ for every j and γ_Ω is non-increasing, so each factor in (1) is at most $1 - \beta_m \gamma_\Omega(\|Q^*_\Omega\|_{\mathrm{asp}} + e_0)^{m-1} \le 1 - \tilde{\beta}$. Hence $e_{km} \le (1-\tilde{\beta})^k e_0 \le e^{-\tilde{\beta}k} e_0$, which is at most ε for $k \ge \tilde{\beta}^{-1} \log(e_0/\varepsilon)$, and monotonicity extends the bound to every $j \ge km$. Each sweep evaluates $p(s'|s,a)$ for all (s,a,s'), i.e. uses $|\mathcal{S}|^2|\mathcal{A}|$ queries, which gives the announced count. □

H.2 Proofs for Subsection 4.2 (Synchronous Ω-regularized Q-learning)

H.2.1 Synchronous full-knowledge updates

We let $(Q_t)_{t\ge 0}$ denote the sequence produced from an arbitrary initialization Q_0 by the *full-knowledge synchronous update*

$$Q_{t+1} = F_t(Q_t), \qquad \text{where } F_t(Q) := (1-\alpha_t)Q + \alpha_t \mathcal{T}_\Omega(Q), \tag{H.1}$$

and we measure the error in span seminorm, $e_t := \|Q_t - Q^*_\Omega\|_{\mathrm{sp}}$.

Without noise, (H.1) is a damped value iteration, whose linearization $(1-\alpha_t)I + \alpha_t K_\pi$ is a lazy version of the greedy state-action kernel: the step size plays the role of the aperiodicity transformation of Puterman (1994, §8.5.4), which acts on the state kernel and which Chen et al. (2026) impose on the dynamics through lazy sampling. The analysis below exploits this laziness at the level of the Jacobians (Lemma H.7): it turns the m-step minorization of Assumption 1 into a contraction over blocks, and it is this deterministic flow that the stochastic analysis of Appendix H.2.2 tracks.

The fixed point is stationary only up to a constant, so it is not a fixed point of F_t; the next lemma records that it nonetheless generates a trajectory of (H.1) along which the error is exactly e_t. This is what allows us to dispense with a second sequence.

Lemma H.1 (Reference trajectory)**.** Let $Q^*_t := Q^*_\Omega + c_t \mathbf{1}$ with $c_t := \rho^*_\Omega \sum_{k<t} \alpha_k$. Then $Q^*_{t+1} = F_t(Q^*_t)$ for every $t \ge 0$, and $\|Q_t - Q^*_t\|_{\mathrm{sp}} = e_t$.

Proof. By translation equivariance of $\mathcal{T}_\Omega$ (Corollary F.1) and the fixed-point equation, $\mathcal{T}_\Omega(Q^*_\Omega + c\mathbf{1}) = Q^*_\Omega + (c + \rho^*_\Omega)\mathbf{1}$, hence $F_t(Q^*_\Omega + c\mathbf{1}) = Q^*_\Omega + (c + \alpha_t\rho^*_\Omega)\mathbf{1}$, which is the announced recursion. The second claim follows since the span seminorm annihilates constants. □

Lemma H.2 (Monotonicity of the error)**.** For any Q_0, the sequence $(e_t)_{t\ge 0}$ is non-increasing.

Proof. Using $\mathcal{T}_\Omega(Q^*_\Omega) = Q^*_\Omega + \rho^*_\Omega \mathbf{1}$, we have $Q_{t+1} - Q^*_\Omega = (1-\alpha_t)(Q_t - Q^*_\Omega) + \alpha_t\left(\mathcal{T}_\Omega(Q_t) - \mathcal{T}_\Omega(Q^*_\Omega)\right) + \alpha_t\rho^*_\Omega\mathbf{1}$, and the constant term vanishes in the span seminorm. The triangle inequality and the non-expansiveness of $\mathcal{T}_\Omega$ in $\|\cdot\|_{\mathrm{sp}}$ (Corollary F.1) then give $e_{t+1} \le (1-\alpha_t)e_t + \alpha_t e_t = e_t$. □

Definition H.1 (Sub-blocks and global blocks)**.** Let $\kappa_0 > 0$ be a fixed sub-block budget. We define recursively the sub-blocks $J_n = \{t_n, \dots, t_{n+1} - 1\}$ by $t_0 = 0$ and

$$t_{n+1} = \min\left\{ t > t_n \,\middle|\, \sum_{k=t_n}^{t-1} \alpha_k \ge \kappa_0 \right\}, \tag{H.2}$$

and the i-th global block as the chronological concatenation of m consecutive sub-blocks,

$$\mathcal{B}_i = \bigcup_{\ell=0}^{m-1} J_{im+\ell}, \qquad b_i := \min \mathcal{B}_i = t_{im}. \tag{H.3}$$

Note that these definitions are valid under Assumption H.1.

We write $\Phi^{(i)} := F_{b_{i+1}-1} \circ \dots \circ F_{b_i}$, so that $Q_{b_{i+1}} = \Phi^{(i)}(Q_{b_i})$ and, by Lemma H.1, $Q^*_{b_{i+1}} = \Phi^{(i)}(Q^*_{b_i})$.

Lemma H.3. Under Assumption H.1, for any sub-block J_n we have $\sum_{k\in J_n} \alpha_k < \kappa_0 + \alpha_0$, and consequently, for any global block $\mathcal{B}_i$, $\sum_{k\in\mathcal{B}_i} \alpha_k < m(\kappa_0 + \alpha_0)$.

Proof. This follows from the minimality in the definition of a sub-block (H.2) and Assumption H.1 on the step sizes. A global block is the chronological concatenation of m consecutive sub-blocks, which gives the second result. □

Lemma H.4 (Boundedness along the frozen flow)**.** Let $\bar{R}_i := \|Q^*_\Omega\|_{\mathrm{asp}} + e_{b_i}$. The sequence $(\bar{R}_i)_{i\geq 0}$ is non-increasing, and for every $\tilde{Q} \in [Q_{b_i}, Q^*_{b_i}]$ and every $k \in \mathcal{B}_i$, the frozen-flow iterate $\tilde{Q}_k := F_{k-1} \circ \ldots \circ F_{b_i}(\tilde{Q})$ (with $\tilde{Q}_{b_i} := \tilde{Q}$) satisfies $\|\tilde{Q}_k\|_{\mathrm{asp}} \leq \bar{R}_i$.

Proof. The monotonicity of $(\bar{R}_i)_i$ is that of $(e_t)_t$ (Lemma H.2).

Let now $\tilde{Q} = (1-u)Q_{b_i} + uQ^*_{b_i}$ for some $u \in [0,1]$. Since $Q^*_{b_i} = Q^*_\Omega + c_{b_i}\mathbf{1}$, we get $\tilde{Q} - Q^*_\Omega = (1-u)(Q_{b_i} - Q^*_\Omega) + uc_{b_i}\mathbf{1}$, and the constant is annihilated by the span seminorm, so $\|\tilde{Q} - Q^*_\Omega\|_{\mathrm{sp}} = (1-u)e_{b_i} \leq e_{b_i}$. The sequence $(\tilde{Q}_k)_{k\geq b_i}$ follows the update (H.1) started at $\tilde{Q}$, so Lemma H.2 applied to it, whose proof holds for any starting time, gives $\|\tilde{Q}_k - Q^*_\Omega\|_{\mathrm{sp}} \leq \|\tilde{Q} - Q^*_\Omega\|_{\mathrm{sp}} \leq e_{b_i}$, whence $\|\tilde{Q}_k\|_{\mathrm{asp}} \leq \|Q^*_\Omega\|_{\mathrm{asp}} + \|\tilde{Q}_k - Q^*_\Omega\|_{\mathrm{sp}} \leq \bar{R}_i$. □

Lemma H.5 (Contraction of products of at least m regularized greedy kernels)**.** Let $R > 0$, let $(U_k)_{k\in\mathcal{K}}$ be a finite family of Q-functions with $\|U_k\|_{\mathrm{asp}} \leq R$ for every k, and let $K_k := K_{\pi_{\Omega,U_k}}$, i.e. $K_k(s',a'|s,a) = p(s'|s,a)\,\pi_{\Omega,U_k}(a'|s')$. For a subset $\mathcal{C} = \{k_1 < \ldots < k_n\} \subseteq \mathcal{K}$, let $K_\mathcal{C} := K_{k_n} \cdots K_{k_1}$ be the associated ordered product. Then $K_\mathcal{C}$ is row-stochastic and:

(i) if $n \geq m$, then $\|K_\mathcal{C}\|_{\mathrm{sp,op}} \leq 1 - \tilde{\beta}_m(R)$, where $\tilde{\beta}_m(R) := \beta_m\,\gamma_\Omega(R)^{m-1}$;
(ii) if $n < m$, then $\|K_\mathcal{C}\|_{\mathrm{sp,op}} \leq 1$.

Proof. Each K_k is row-stochastic, hence so is $K_\mathcal{C}$, and (ii) follows from Lemma E.6.

Assume $n \geq m$. With Convention E.1, the Markov chain associated with $K_\mathcal{C} = K_{k_n} \cdots K_{k_1}$ applies the factors in the order $K_{k_n}, K_{k_{n-1}}, \ldots, K_{k_1}$. We therefore split off its *last* m steps in chain order, i.e. the m *oldest* algorithmic indices: $K_\mathcal{C} = HB$ with $H := K_{k_n} \cdots K_{k_{m+1}}$ and $B := K_{k_m} \cdots K_{k_1}$, and $H = I$ when $n = m$. Since $\|U_{k_\ell}\|_{\mathrm{asp}} \leq R$ for each ℓ, Definition G.1.(iii) gives $\pi_{\Omega,U_{k_\ell}} \geq \gamma_\Omega(R)\mu$, so Lemma E.4, applied to B, whose $m-1$ intermediate policies are $\pi_{\Omega,U_{k_m}}, \ldots, \pi_{\Omega,U_{k_2}}$, gives $\alpha(B) \leq 1 - \beta_m\gamma_\Omega(R)^{m-1} = 1 - \tilde{\beta}_m(R)$. Finally, H is row-stochastic, so Lemma E.6 and the submultiplicativity (E.5) give $\|K_\mathcal{C}\|_{\mathrm{sp,op}} \leq \|H\|_{\mathrm{sp,op}}\,\|B\|_{\mathrm{sp,op}} \leq \alpha(B)$. □

It remains to count the effective updates within a global block. This is where the sub-block structure is used, and a plain independence argument suffices: no concentration inequality is required (a Chernoff bound gives a sharper estimate for large m, see Remark H.2).

Lemma H.6 (Effective updates within a global block)**.** Let $\mathcal{B}_i$ be the i-th global block of Definition H.1, let $(X_k)_{k\in\mathcal{B}_i}$ be auxiliary independent Bernoulli random variables with $X_k \sim \mathcal{B}(\alpha_k)$ and let $N_i := \sum_{k\in\mathcal{B}_i} X_k$. Under Assumption H.1,

$$\mathbb{P}(N_i \geq m) \geq p_m := \left(1 - e^{-\kappa_0}\right)^m. \tag{H.4}$$

Proof. For each of the m sub-blocks $J_n \subseteq \mathcal{B}_i$, let $\mathcal{U}_n := \bigcup_{k\in J_n}\{X_k = 1\}$ be the event that at least one effective update occurs in J_n. By independence of the X_k and $1 - u \leq e^{-u}$,

$$\mathbb{P}(\mathcal{U}_n) = 1 - \prod_{k\in J_n}(1-\alpha_k) \geq 1 - \exp\left(-\sum_{k\in J_n}\alpha_k\right) \geq 1 - e^{-\kappa_0},$$

the last inequality by the definition (H.2) of the sub-block. The sub-blocks of $\mathcal{B}_i$ have disjoint index sets, so the events $\mathcal{U}_{im}, \ldots, \mathcal{U}_{im+m-1}$ are independent, and $\bigcap_{\ell=0}^{m-1}\mathcal{U}_{im+\ell} \subseteq \{N_i \geq m\}$. Hence $\mathbb{P}(N_i \geq m) \geq \prod_{\ell=0}^{m-1}\mathbb{P}(\mathcal{U}_{im+\ell}) \geq (1 - e^{-\kappa_0})^m$. □

Lemma H.7 (Jacobian bound over a global block)**.** Under Assumption H.1, for every global block $\mathcal{B}_i$ and every $\tilde{Q} \in [Q_{b_i}, Q^*_{b_i}]$,

$$\|D\Phi^{(i)}(\tilde{Q})\|_{\mathrm{sp,op}} \leq 1 - p_m\,\tilde{\beta}_m(\bar{R}_i).$$

Proof. By the chain rule and Corollary F.1, the Jacobian of $\Phi^{(i)}$ at $\tilde{Q}$ is

$$D\Phi^{(i)}(\tilde{Q}) = \prod_{k\in\mathcal{B}_i}\left((1-\alpha_k)I + \alpha_k K_k\right), \qquad K_k := K_{\pi_{\Omega,\tilde{Q}_k}}, \quad \tilde{Q}_k := F_{k-1} \circ \ldots \circ F_{b_i}(\tilde{Q}),$$

the product being written in decreasing order of k. By Lemma H.4, $\|\tilde{Q}_k\|_{\mathrm{asp}} \le \bar{R}_i$ for every $k \in \mathcal{B}_i$, so the family $(K_k)_{k\in\mathcal{B}_i}$ satisfies the hypotheses of Lemma H.5 with $U_k = \tilde{Q}_k$ and $R = \bar{R}_i$.

We introduce a probabilistic argument to bound the Jacobian. Let $(X_k)_{k\in\mathcal{B}_i}$ be the auxiliary Bernoulli variables of Lemma H.6, realized as the coordinate maps of the finite product space $\{0,1\}^{\mathcal{B}_i}$ endowed with $\bigotimes_{k\in\mathcal{B}_i} \mathcal{B}(\alpha_k)$, which is well defined since $\alpha_k \in (0,1]$ by Assumption H.1. The flow (H.1) being deterministic, this finite space carries all the randomness of the argument. We define $Y_k = (1-X_k)I + X_k K_k$, so that $(1-\alpha_k)I + \alpha_k K_k = \mathbb{E}[Y_k]$. By independence of the Y_k, we have $\prod_{k\in\mathcal{B}_i} \mathbb{E}[Y_k] = \mathbb{E}\left[\prod_{k\in\mathcal{B}_i} Y_k\right]$, and the triangle inequality gives

$$\|D\Phi^{(i)}(\tilde{Q})\|_{\mathrm{sp,op}} = \left\| \mathbb{E}\left[\prod_{k\in\mathcal{B}_i} Y_k\right] \right\|_{\mathrm{sp,op}} \le \mathbb{E}\left[\left\| \prod_{k\in\mathcal{B}_i} Y_k \right\|_{\mathrm{sp,op}} \right].$$

Let $\mathcal{C} := \{k \in \mathcal{B}_i \mid X_k = 1\}$, so that $N_i = |\mathcal{C}|$. Since the factors equal to I can be removed from the product, we have the pathwise identity $\prod_{k\in\mathcal{B}_i} Y_k = K_{\mathcal{C}}$ with the notation of Lemma H.5.

On the event $\{N_i \ge m\}$, point (i) of Lemma H.5 gives $\|K_{\mathcal{C}}\|_{\mathrm{sp,op}} \le 1 - \tilde{\beta}_m(\bar{R}_i)$, and on the event $\{N_i < m\}$, point (ii) gives $\|K_{\mathcal{C}}\|_{\mathrm{sp,op}} \le 1$; hence $\|K_{\mathcal{C}}\|_{\mathrm{sp,op}} \le 1 - \tilde{\beta}_m(\bar{R}_i)\mathbb{1}_{\{N_i\ge m\}}$ pointwise. Taking expectations,

$$\|D\Phi^{(i)}(\tilde{Q})\|_{\mathrm{sp,op}} \le 1 - \mathbb{P}(N_i \ge m)\,\tilde{\beta}_m(\bar{R}_i) \le 1 - p_m \tilde{\beta}_m(\bar{R}_i),$$

the last inequality by Lemma H.6. □

Lemma H.8 (Per-block contraction)**.** Under Assumption H.1, for every global block $\mathcal{B}_i$,

$$e_{b_{i+1}} \le \left(1 - p_m\, \tilde{\beta}_m(\bar{R}_i)\right) e_{b_i}.$$

Proof. By Lemma H.1, $e_{b_{i+1}} = \|\Phi^{(i)}(Q_{b_i}) - \Phi^{(i)}(Q^*_{b_i})\|_{\mathrm{sp}}$. The map $\Phi^{(i)}$ is continuously differentiable (Corollary F.1) and all its Jacobians $D\Phi^{(i)}(\tilde{Q})$, $\tilde{Q} \in [Q_{b_i}, Q^*_{b_i}]$, are products of row-stochastic matrices, hence row-stochastic, so the mean value inequality in the span seminorm (Proposition E.1) applies and gives $e_{b_{i+1}} \le \sup_{\tilde{Q}\in[Q_{b_i},Q^*_{b_i}]} \|D\Phi^{(i)}(\tilde{Q})\|_{\mathrm{sp,op}}\, \|Q_{b_i} - Q^*_{b_i}\|_{\mathrm{sp}}$. Lemma H.7 bounds the supremum by $1 - p_m\tilde{\beta}_m(\bar{R}_i)$, while $\|Q_{b_i} - Q^*_{b_i}\|_{\mathrm{sp}} = e_{b_i}$. □

Corollary H.1 (Full-knowledge updates for $m = 1$)**.** Assume $m = 1$ and $\alpha_t \in (0,1]$ for every t. For any initialization Q_0, the iterates of (H.1) satisfy

$$e_T \le \prod_{k=0}^{T-1} (1 - \alpha_k\beta_1)\, e_0.$$

In particular, Q-iteration ($\alpha_t \equiv 1$) contracts at rate $1-\beta_1$ per sweep, and reaches $e_T \le \varepsilon$ after $O\big(|\mathcal{S}|^2|\mathcal{A}|\,\beta_1^{-1}\log(e_0/\varepsilon)\big)$ model queries.

Proof. Let $\Psi_T := F_{T-1} \circ \ldots \circ F_0$ and $\tilde{Q} \in [Q_0, Q_0^*]$. As in Lemma H.7, $D\Psi_T(\tilde{Q}) = \prod_{k=0}^{T-1} ((1-\alpha_k)I + \alpha_k P_k)$. By Lemma E.4 with $n = 1$, each factor satisfies

$$((1-\alpha_k)I + \alpha_k K_k)\,(s',a'|s,a) \ge \alpha_k K_k(s',a'|s,a) \ge \alpha_k\beta_1\, \nu(s')\, \pi_{\Omega,\tilde{Q}_k}(a'|s'),$$

which is a Doeblin condition with parameter $\alpha_k\beta_1$. Lemmas E.3 and E.6 give $\|(1-\alpha_k)I + \alpha_k P_k\|_{\mathrm{sp,op}} \le 1 - \alpha_k\beta_1$, and the submultiplicativity of the operator span seminorm together with Proposition E.1 and Lemma H.1 conclude. For $\alpha_t \equiv 1$, the update (H.1) is Q-iteration without the centering projection, which is invisible to $\|\cdot\|_{\mathrm{sp}}$, so $e_T \le (1-\beta_1)^T e_0 \le \varepsilon$ for $T \ge \beta_1^{-1}\log(e_0/\varepsilon)$, each sweep using $|\mathcal{S}|^2|\mathcal{A}|$ queries. □

Lemma H.9 (Number of completed blocks)**.** Under Assumption H.1, let $I(T) := \max\{i \mid b_i \le T\}$ be the number of global blocks completed within the first T updates. Then $I(T) \ge \left\lfloor \frac{\sum_{t=0}^{T-1}\alpha_t}{m(\kappa_0+\alpha_0)} \right\rfloor$.

Proof. By definition of $I(T)$ we have $b_{I(T)} \le T < b_{I(T)+1}$, and by Lemma H.3 each global block consumes at most $m(\kappa_0 + \alpha_0)$. The step sizes being non-negative,

$$\sum_{k=0}^{T-1} \alpha_k \le \sum_{k=0}^{b_{I(T)+1}-1} \alpha_k < (I(T)+1)\, m(\kappa_0+\alpha_0), \qquad \text{i.e.} \qquad I(T) > \frac{\sum_{k=0}^{T-1} \alpha_k}{m(\kappa_0+\alpha_0)} - 1,$$

and $I(T)$ being an integer gives the claim. □

Remark H.1 (Radius-adaptive rate)**.** Chaining Lemma H.8 over the completed blocks gives the radius-adaptive bound $e_{b_{I(T)}} \le \prod_{i<I(T)} \left(1 - p_m \beta_m \gamma_\Omega (\bar{R}_i)^{m-1}\right) e_0$, in which the per-block rate improves as the error decreases, since $\bar{R}_i = \|Q^*_\Omega\|_{\mathrm{asp}} + e_{b_i}$ is non-increasing. The uniform constant $\tilde{\beta}_m(\bar{R}_0)$ is its worst case, attained at the first block. Asymptotically, $\bar{R}_i \downarrow \|Q^*_\Omega\|_{\mathrm{asp}}$, so the rate approaches $1 - p_m \beta_m \gamma_\Omega (\|Q^*_\Omega\|_{\mathrm{asp}})^{m-1}$: the local geometry of the fixed point, not the initialization, governs the asymptotic speed. The bounds of Theorems H.1 and H.2 use instead the worst-case constant $\bar{\beta}(R)$ of Definition H.2 on the whole ball of radius R; making them radius-adaptive would require shrinking the radius along the run.

Remark H.2 (An alternative bound on $\mathbb{P}(N_i \ge m)$ through a Chernoff inequality)**.** Each global block has a total budget $\bar{\alpha}_i := \sum_{k \in \mathcal{B}_i} \alpha_k \ge m\kappa_0$, so for $\kappa_0 > 1$ the multiplicative Chernoff inequality for the lower tail (Mitzenmacher & Upfal, 2005, Theorem 4.5) gives

$$\mathbb{P}(N_i < m) \le \exp\left(-\frac{(\bar{\alpha}_i - m)^2}{2\bar{\alpha}_i}\right) \le \exp\left(-\frac{(\kappa_0-1)^2}{2\kappa_0} m\right),$$

the last inequality because $u \mapsto \frac{(u-m)^2}{2u}$ has derivative $\frac{1}{2}\left(1 - \frac{m^2}{u^2}\right) > 0$ on (m, ∞) and is therefore minimized at $u = m\kappa_0$. This variant requires $\kappa_0 > 1$ but is much better for large m, since it tends to 1 whereas $(1 - e^{-\kappa_0})^m$ decays exponentially in m. Both bounds being valid, Lemma H.6 holds with

$$p_m = \max\left\{ \left(1 - e^{-\kappa_0}\right)^m,\ 1 - \exp\left(-\frac{(\kappa_0-1)^2}{2\kappa_0} m\right) \mathbb{1}\{\kappa_0 > 1\} \right\}.$$

The proof of Theorem 1 uses $\kappa_0 = 4$, for which $p_m \ge 1 - e^{-9/8} \ge \frac{1}{2}$.

H.2.2 Setting and localized constants

We now study Algorithm 3, the synchronous Ω-regularized Q-learning algorithm, and write $e_t := \|Q_t - Q^*_\Omega\|_{\mathrm{sp}}$ for its error. We define the filtration $\mathcal{F}_t := \sigma\left\{Q_0,\ (s'_k(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}} : k < t\right\}$ and write the stochastic update as

$$Q_{t+1} = Q_t + \alpha_t \left(\mathcal{T}_\Omega(Q_t) - Q_t + W_{t+1}\right), \tag{H.5}$$

where for every pair (s,a) the next state $s'_t(s,a) \sim p(\cdot|s,a)$ is sampled independently and

$$W_{t+1}(s,a) := \Omega^*(Q_t)\left(s'_t(s,a)\right) - \sum_{s'} p(s'|s,a)\, \Omega^*(Q_t)(s'). \tag{H.6}$$

By construction W_{t+1} is $\mathcal{F}_{t+1}$-measurable and $\mathbb{E}[W_{t+1} \mid \mathcal{F}_t] = 0$. By translation equivariance of Ω^* (Lemma F.1.(iii)), the centering projection of Algorithm 3 only shifts the iterates by a constant, which $\|\cdot\|_{\mathrm{sp}}$ and $\|\cdot\|_{\mathrm{asp}}$ annihilate, so we omit it; its only role is to upgrade span convergence to $\|\cdot\|_\infty$ convergence (Remark F.2). We reuse the sub-blocks J_n and global blocks $\mathcal{B}_i$ of Definition H.1; the step sizes being deterministic, the block boundaries b_i, the number $I(T)$ of completed blocks, the bound $I_{\max} := \left\lfloor \frac{1}{m\kappa_0} \sum_{t<T} \alpha_t \right\rfloor$ and the last incomplete block $\mathcal{B}_{\mathrm{res}} := [\![b_{I(T)}, T-1]\!]$ are all deterministic.

Definition H.2 (Localized radii, exit time and localized constants)**.** For a radius $R \ge e_0$, define the exit time

$$\sigma_R := \inf\{t \in \mathbb{N} \mid e_t > R\} \in \mathbb{N} \cup \{+\infty\}, \tag{H.7}$$

the *domination radius* and the *noise radius*

$$A_R := \|Q^*_\Omega\|_{\mathrm{asp}} + R, \qquad C_R := \|Q^*_\Omega\|_{\mathrm{sp}} + R,$$

and the localized constants

$$\tilde{\beta}_m(A_R) := \beta_m\,\gamma_\Omega(A_R)^{m-1}, \quad \bar{\beta}(R) := p_m\,\tilde{\beta}_m(A_R) \in (0,1), \quad \bar{w}(R) := C_R + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}}\,, \tag{H.8}$$

where p_m is the constant of Lemma H.6 (see also Remark H.2).

Remark H.3. (i) σ_R is an $(\mathcal{F}_t)$-stopping time and $\sigma_R \geq 1$ since $e_0 \leq R$. (ii) By minimality, $t < \sigma_R$ implies $e_t \leq R$, hence $\|Q_t\|_{\mathrm{sp}} \leq C_R$ and, since $\|\cdot\|_{\mathrm{asp}} \leq \|\cdot\|_{\mathrm{sp}}$, also $\|Q_t\|_{\mathrm{asp}} \leq \|Q^*_\Omega\|_{\mathrm{asp}} + \|Q_t - Q^*_\Omega\|_{\mathrm{asp}} \leq A_R$. (iii) γ_Ω is non-increasing (Definition G.1.(iii)), so $R \mapsto \bar{\beta}(R)$ is non-increasing and any contraction estimate valid at a radius $A \leq A_R$ holds a fortiori with $\bar{\beta}(R)$; moreover $\bar{\beta}(R) > 0$ by (P1). (iv) The localization serves two purposes: it bounds the noise scale by $\bar{w}(R)$, which grows linearly in R through the sub-affinity of Ω^* (Lemma F.1.(vii)), and, since γ_Ω may vanish at infinity, it keeps the contraction rate at least $\bar{\beta}(R)$.

H.2.3 The noiseless flow before the exit time

For each global block $\mathcal{B}_i$ we define the *virtual sequence* $(\tilde{Q}_k)_{k \geq b_i}$ by

$$\tilde{Q}_{b_i} := Q_{b_i}, \qquad \tilde{Q}_{k+1} := \tilde{Q}_k + \alpha_k \left(\mathcal{T}_\Omega(\tilde{Q}_k) - \tilde{Q}_k\right), \tag{H.9}$$

i.e. the full-knowledge synchronous update (H.1) started at Q_{b_i}; for $i = I(T)$, it is defined on the last incomplete block $\mathcal{B}_{\mathrm{res}}$. It is a deterministic function of Q_{b_i}, hence $\mathcal{F}_{b_i}$-measurable.

Lemma H.10 (Contraction of the virtual flow at radius R)**.** Let $i \in \mathbb{N}$ with $e_{b_i} \leq R$. Then, pathwise:

(i) $\|\tilde{Q}_k - Q^*_\Omega\|_{\mathrm{sp}} \leq e_{b_i}$ for every $k \geq b_i$;
(ii) $\|\tilde{Q}_{b_{i+1}} - Q^*_\Omega\|_{\mathrm{sp}} \leq \big(1 - \bar{\beta}(R)\big)\, e_{b_i}$.

Proof. The virtual sequence is exactly the full-knowledge update (H.1) started at Q_{b_i}, so (i) is Lemma H.2 applied to that sequence, and (ii) is Lemma H.8 applied within the single block $\mathcal{B}_i$, which gives the factor $1 - p_m\tilde{\beta}_m(\bar{R}_i)$ where $\bar{R}_i = \|Q^*_\Omega\|_{\mathrm{asp}} + e_{b_i}$ is the block radius of Lemma H.4. It remains to check $\bar{R}_i \leq A_R$, which follows from $e_{b_i} \leq R$; γ_Ω being non-increasing, $\tilde{\beta}_m(\bar{R}_i) \geq \tilde{\beta}_m(A_R)$, i.e. $1 - p_m\tilde{\beta}_m(\bar{R}_i) \leq 1 - \bar{\beta}(R)$. □

H.2.4 Localized noise control

Lemma H.11 (Stopped noise)**.** Define $\widehat{W}_{t+1} := \mathbb{1}_{\{t<\sigma_R\}} W_{t+1}$. Then $\widehat{W}_{t+1}$ is $\mathcal{F}_{t+1}$-measurable, $\mathbb{E}\left[\widehat{W}_{t+1} \mid \mathcal{F}_t\right] = 0$, and each entry $\widehat{W}_{t+1}(s,a)$ lies in an $\mathcal{F}_t$-measurable interval containing 0 and of length at most $\bar{w}(R)$; in particular $\left\|\widehat{W}_{t+1}\right\|_\infty \leq \bar{w}(R)$. Moreover $\widehat{W}_{t+1} = W_{t+1}$ on $\{t < \sigma_R\}$.

Proof. Since σ_R is a stopping time, $\{t < \sigma_R\} = \{\sigma_R \leq t\}^c \in \mathcal{F}_t$, so $\mathbb{E}[\widehat{W}_{t+1} \mid \mathcal{F}_t] = \mathbb{1}_{\{t<\sigma_R\}}\mathbb{E}[W_{t+1} \mid \mathcal{F}_t] = 0$. Conditionally on $\mathcal{F}_t$, $\Omega^*(Q_t)$ is fixed, and $W_{t+1}(s,a)$ is a value of $\Omega^*(Q_t)$ minus the $\mathcal{F}_t$-measurable convex combination $\sum_{s'} p(s'|s,a)\,\Omega^*(Q_t)(s')$; it therefore lies in an $\mathcal{F}_t$-measurable interval containing 0 and of length $\|\Omega^*(Q_t)\|_{\mathcal{S},\mathrm{sp}}$, and so does $\widehat{W}_{t+1}(s,a)$, with length $\mathbb{1}_{\{t<\sigma_R\}}\|\Omega^*(Q_t)\|_{\mathcal{S},\mathrm{sp}}$. On $\{t < \sigma_R\}$ we have $\|Q_t\|_{\mathrm{sp}} \leq C_R$ by Remark H.3.(ii), and Lemma F.1.(vii) gives $\|\Omega^*(Q_t)\|_{\mathcal{S},\mathrm{sp}} \leq \|Q_t\|_{\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} \leq C_R + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} = \bar{w}(R)$. Off the event, $\widehat{W}_{t+1} = 0$. □

For an interval $\mathcal{K} = [\![t_a, t_b]\!]$ we define the stopped noise martingale $\widehat{U}^{(\mathcal{K})}_{t_a} := 0$, $\widehat{U}^{(\mathcal{K})}_{k+1} := \widehat{U}^{(\mathcal{K})}_k + \alpha_k\widehat{W}_{k+1}$ for $k \in \mathcal{K}$, we write $v_\mathcal{K} := \sum_{k\in\mathcal{K}} \alpha_k^2$, and we set

$$\bar{U}_\mathcal{K}(\delta') := \bar{w}(R)\sqrt{\frac{1}{2}v_\mathcal{K}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta'}\right)}, \tag{H.10}$$

$$\varepsilon_\mathcal{K}(\delta') := 2\left(1 + 2\sum_{k\in\mathcal{K}}\alpha_k\right)\bar{U}_\mathcal{K}(\delta') \;\leq\; 2\,(1 + 2m(\kappa_0 + \alpha_0))\,\bar{U}_\mathcal{K}(\delta'), \tag{H.11}$$

the last inequality by Lemma H.3 when $\mathcal{K}$ is contained in a global block.

Lemma H.12 (Good event)**.** Let $\delta \in (0,1)$ and $\delta' := \frac{\delta}{I_{\max}+1}$. The event

$$G_R(\delta) := \bigcap_{i=0}^{I(T)-1} \left\{ \max_{b_i \le k \le b_{i+1}} \left\| \widehat{U}_k^{(\mathcal{B}_i)} \right\|_\infty \le \bar{U}_{\mathcal{B}_i}(\delta') \right\} \cap \left\{ \max_{b_{I(T)} \le k \le T} \left\| \widehat{U}_k^{(\mathcal{B}_{\mathrm{res}})} \right\|_\infty \le \bar{U}_{\mathcal{B}_{\mathrm{res}}}(\delta') \right\} \tag{H.12}$$

satisfies $\mathbb{P}\left(G_R(\delta)\right) \ge 1-\delta$.

Proof. Fix an interval $\mathcal{K}$ among $\mathcal{B}_0, \dots, \mathcal{B}_{I(T)-1}, \mathcal{B}_{\mathrm{res}}$ and a pair (s,a). By Lemma H.11, the scalar process $\left(\widehat{U}_k^{(\mathcal{K})}(s,a)\right)_k$ is an $\mathcal{F}_k$-martingale whose increment $\alpha_k \widehat{W}_{k+1}(s,a)$ lies in an $\mathcal{F}_k$-measurable interval of length at most $\alpha_k \bar{w}(R)$, so $\sum_{k\in\mathcal{K}} \left(\alpha_k \bar{w}(R)\right)^2 = \bar{w}(R)^2 v_{\mathcal{K}}$ and the maximal Azuma–Hoeffding inequality (Lemma E.2) bounds the failure probability of each brace, for each pair, by $\frac{\delta'}{|\mathcal{S}||\mathcal{A}|}$. A union bound over the $|\mathcal{S}||\mathcal{A}|$ pairs and the $I(T)+1 \le I_{\max}+1$ intervals concludes. □

H.2.5 Tracking before the exit time

Lemma H.13 (Tracking up to the exit time)**.** Work on the event $G_R(\delta)$. Let $i \le I(T)$, let $\mathcal{K} := \mathcal{B}_i$ if $i < I(T)$ and $\mathcal{K} := \mathcal{B}_{\mathrm{res}}$ if $i = I(T)$, write $t_a := b_i$ and $t_b + 1 := b_{i+1}$ (resp. T), and let $\tilde{Q}$ be the virtual sequence started at Q_{t_a}. Then, for every k with $t_a \le k \le \min\{t_b+1, \sigma_R\}$, the tracking error $E_k := Q_k - \tilde{Q}_k$ satisfies $\|E_k\|_{\mathrm{sp}} \le \varepsilon_{\mathcal{K}}(\delta')$.

Proof. If $\sigma_R \le t_a$ the range reduces to $k = t_a$, where $E_{t_a} = 0$. Otherwise, for $t_a \le k < \min\{t_b+1, \sigma_R\}$ the true update (H.5) occurs and $W_{k+1} = \widehat{W}_{k+1}$ by Lemma H.11. Subtracting (H.9) from (H.5) and using the average-path identity of Lemma G.2.(i) at the pair $(Q_k, \tilde{Q}_k)$, i.e. $\mathcal{T}_\Omega(Q_k) - \mathcal{T}_\Omega(\tilde{Q}_k) = \bar{K}_k E_k$ with $\bar{K}_k := K_{\bar{\pi}_{\Omega,[Q_k,\tilde{Q}_k]}}$,

$$E_{k+1} = (1-\alpha_k)E_k + \alpha_k \bar{K}_k E_k + \alpha_k \widehat{W}_{k+1} = M_k E_k + \alpha_k \widehat{W}_{k+1}, \tag{H.13}$$

where $M_k := (1-\alpha_k)I + \alpha_k \bar{K}_k$ is row-stochastic as a convex combination of row-stochastic matrices.

Decompose $E_k = \widehat{R}_k + \widehat{U}_k^{(\mathcal{K})}$ with $\widehat{R}_{t_a} = 0$. The noise terms in (H.13) cancel with the definition of $\widehat{U}_{k+1}^{(\mathcal{K})}$, so that $\widehat{R}_{k+1} = M_k \widehat{R}_k + \alpha_k \left(\bar{K}_k - I\right) \widehat{U}_k^{(\mathcal{K})}$. Since $\bar{K}_k$ is row-stochastic, every entry of $\bar{K}_k \widehat{U}_k^{(\mathcal{K})}$ is a convex combination of entries of $\widehat{U}_k^{(\mathcal{K})}$, hence lies in $\left[\min \widehat{U}_k^{(\mathcal{K})}, \max \widehat{U}_k^{(\mathcal{K})}\right]$, and so does every entry of $\widehat{U}_k^{(\mathcal{K})}$; therefore every entry of $\left(\bar{K}_k - I\right) \widehat{U}_k^{(\mathcal{K})}$ lies in $\left[-\|\widehat{U}_k^{(\mathcal{K})}\|_{\mathrm{sp}}, \|\widehat{U}_k^{(\mathcal{K})}\|_{\mathrm{sp}}\right]$ and $\left\|\alpha_k \left(\bar{K}_k - I\right) \widehat{U}_k^{(\mathcal{K})}\right\|_{\mathrm{sp}} \le 2\alpha_k \left\|\widehat{U}_k^{(\mathcal{K})}\right\|_{\mathrm{sp}} \le 4\alpha_k \left\|\widehat{U}_k^{(\mathcal{K})}\right\|_\infty$. By Lemma E.6 the row-stochastic matrix M_k is span non-expansive, so $\|\widehat{R}_{k+1}\|_{\mathrm{sp}} \le \|\widehat{R}_k\|_{\mathrm{sp}} + 4\alpha_k \|\widehat{U}_k^{(\mathcal{K})}\|_\infty$. Summing from t_a and using the definition of $G_R(\delta)$, for every $k \in [\![t_a, \min\{t_b+1, \sigma_R\}]\!]$, $\|\widehat{R}_k\|_{\mathrm{sp}} \le 4\left(\sum_{k'\in\mathcal{K}} \alpha_{k'}\right) \bar{U}_{\mathcal{K}}(\delta')$ and

$$\|E_k\|_{\mathrm{sp}} \le \|\widehat{R}_k\|_{\mathrm{sp}} + 2\left\|\widehat{U}_k^{(\mathcal{K})}\right\|_\infty \le 2\left(1 + 2\sum_{k'\in\mathcal{K}} \alpha_{k'}\right) \bar{U}_{\mathcal{K}}(\delta') = \varepsilon_{\mathcal{K}}(\delta').$$

□

H.2.6 No escape before the horizon

Proposition H.1 (No escape)**.** Let $\delta \in (0,1)$, $R \ge e_0$, $\delta' = \frac{\delta}{I_{\max}+1}$ and

$$\bar{\varepsilon}_R(\delta) := \max\left\{ \max_{0 \le i < I(T)} \varepsilon_{\mathcal{B}_i}(\delta'),\ \varepsilon_{\mathcal{B}_{\mathrm{res}}}(\delta') \right\}.$$

Assume the *radius admissibility condition*

$$\bar{\varepsilon}_R(\delta) \le \frac{\bar{\beta}(R)}{\bar{\beta}(R)+1}\,(R-e_0)\,. \tag{H.14}$$

Then, on the event $G_R(\delta)$, we have $\sigma_R > T$, in particular $\sup_{0\le t\le T} e_t \le R$, and for every $0 \le i \le I(T)$,

$$e_{b_i} \le (1-\bar{\beta}(R))^i e_0 + \sum_{\ell=0}^{i-1}(1-\bar{\beta}(R))^{i-1-\ell}\varepsilon_{\mathcal{B}_\ell}(\delta'). \tag{H.15}$$

Proof. Work on $G_R(\delta)$. We first record that (H.15) implies, bounding the geometric sum and using (H.14),

$$e_{b_i} \le e_0 + \frac{\bar{\varepsilon}_R(\delta)}{\bar{\beta}(R)}, \qquad \text{hence} \qquad e_{b_i} + \bar{\varepsilon}_R(\delta) \le e_0 + \frac{\bar{\beta}(R)+1}{\bar{\beta}(R)}\bar{\varepsilon}_R(\delta) \le R. \tag{H.16}$$

We prove by induction on $i \in [\![0, I(T)]\!]$ that $\sigma_R > b_i$ and (H.15) holds at rank i.

Base case. $b_0 = 0$ and $\sigma_R \ge 1$ by Remark H.3.(i); (H.15) at $i = 0$ reads $e_0 \le e_0$.

Inductive step. Assume the claim at rank $i < I(T)$ and let $\tilde{Q}$ be the virtual sequence started at Q_{b_i}; by (H.16), $e_{b_i} \le R$, so Lemma H.10 applies to $\mathcal{B}_i$. Suppose $\sigma_R \le b_{i+1}$; since $\sigma_R > b_i$, we would have $t^\star := \sigma_R \in [\![b_i+1, b_{i+1}]\!]$, and as $t^\star \le \min\{b_{i+1}, \sigma_R\}$, Lemma H.13 gives $\|Q_{t^\star} - \tilde{Q}_{t^\star}\|_{\mathrm{sp}} \le \varepsilon_{\mathcal{B}_i}(\delta') \le \bar{\varepsilon}_R(\delta)$ while Lemma H.10.(i) gives $\|\tilde{Q}_{t^\star} - Q^*_\Omega\|_{\mathrm{sp}} \le e_{b_i}$. Hence, by (H.16), $e_{\sigma_R} \le e_{b_i} + \bar{\varepsilon}_R(\delta) \le R$, contradicting the definition (H.7) of σ_R. Therefore $\sigma_R > b_{i+1}$; in particular $b_{i+1} \le \min\{b_{i+1}, \sigma_R\}$, so Lemma H.13 at $k = b_{i+1}$ combined with Lemma H.10.(ii) yields $e_{b_{i+1}} \le \|\tilde{Q}_{b_{i+1}} - Q^*_\Omega\|_{\mathrm{sp}} + \|Q_{b_{i+1}} - \tilde{Q}_{b_{i+1}}\|_{\mathrm{sp}} \le (1-\bar{\beta}(R))e_{b_i} + \varepsilon_{\mathcal{B}_i}(\delta')$, which, injected into (H.15) at rank i, gives it at rank $i+1$.

Residual interval. The last step gives $\sigma_R > b_{I(T)}$. If $\sigma_R \le T$, the same argument on $\mathcal{B}_{\mathrm{res}}$ — Lemma H.13 with $\mathcal{K} = \mathcal{B}_{\mathrm{res}}$ and Lemma H.10.(i) — gives $e_{\sigma_R} \le e_{b_{I(T)}} + \bar{\varepsilon}_R(\delta) \le R$, again a contradiction. Hence $\sigma_R > T$ and the containment follows from the definition of σ_R. □

H.2.7 Main result

Theorem H.1 (Finite-time bound for Ω-regularized synchronous Q-learning)**.** Let $\delta \in (0,1)$ and $T \ge 1$. Under Assumption H.1, let the sub-block budget $\kappa_0 > 0$ be arbitrary and let the radius $R \ge e_0$ satisfy the admissibility condition (H.14). Then, with probability at least $1-\delta$, $\sup_{0\le t\le T} e_t \le R$ and

$$e_T \le (1-\bar{\beta}(R))^{I(T)} e_0 + 2\,(1+2m(\kappa_0+\alpha_0))\,\bar{w}(R)\sqrt{\frac{1}{2}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|(I_{\max}+1)}{\delta}\right)} \times \left(\sum_{i=0}^{I(T)-1}(1-\bar{\beta}(R))^{I(T)-1-i}\sqrt{\sum_{k\in\mathcal{B}_i}\alpha_k^2} + \sqrt{\sum_{k\in\mathcal{B}_{\mathrm{res}}}\alpha_k^2}\right),$$

where $\bar{\beta}(R)$ and $\bar{w}(R)$ are the localized constants of Definition H.2, $I(T) \ge \left\lfloor \frac{\sum_{t<T}\alpha_t}{m(\kappa_0+\alpha_0)} \right\rfloor$ is the number of completed blocks and $I_{\max} \ge I(T)$.

Proof. Work on $G_R(\delta)$, which holds with probability at least $1-\delta$ by Lemma H.12. Proposition H.1 gives $\sigma_R > T$, so the ranges of Lemma H.13 cover every block entirely, as well as the residual block. Applying Lemma H.13 with $\mathcal{K} = \mathcal{B}_{\mathrm{res}}$ at $k = T$, and Lemma H.10.(i) to the virtual sequence started at $Q_{b_{I(T)}}$, the triangle inequality gives $e_T \le e_{b_{I(T)}} + \varepsilon_{\mathcal{B}_{\mathrm{res}}}(\delta')$. Injecting (H.15) at $i = I(T)$ and the explicit form (H.11) of the $\varepsilon_{\mathcal{K}}(\delta')$ yields the announced bound; the lower bound on $I(T)$ is Lemma H.9. □

Corollary H.2 (Constant step size)**.** Under the assumptions of Theorem H.1, let the step size be constant, $\alpha_t = \alpha \in (0,1)$, and set $C_\kappa := \sqrt{2}\,(1 + 2m(\kappa_0 + \alpha))$. If the radius $R \geq e_0$ satisfies

$$C_\kappa\, \bar{w}(R) \sqrt{m(\kappa_0+\alpha)\,\alpha\,\log\left(\frac{2|\mathcal{S}||\mathcal{A}|\left(\frac{\alpha T}{m\kappa_0}+1\right)}{\delta}\right)} \leq \frac{\bar{\beta}(R)}{\bar{\beta}(R)+1}\,(R - e_0)\,, \tag{H.17}$$

then (H.14) holds and, with probability at least $1-\delta$,

$$e_T \leq \frac{1}{1-\bar{\beta}(R)}\left(1-\bar{\beta}(R)\right)^{\frac{\alpha T}{m(\kappa_0+\alpha)}} e_0 + \frac{(\bar{\beta}(R)+1)C_\kappa \bar{w}(R)}{\bar{\beta}(R)} \sqrt{m(\kappa_0+\alpha)\,\alpha\,\log\left(\frac{2|\mathcal{S}||\mathcal{A}|\left(\frac{\alpha T}{m\kappa_0}+1\right)}{\delta}\right)}.$$

Proof. With a constant step size, every sub-block has exactly $\lceil \kappa_0/\alpha \rceil$ indices, so $|\mathcal{B}_i| = m\lceil \kappa_0/\alpha \rceil$ and, using $\lceil \kappa_0/\alpha \rceil < \frac{\kappa_0+\alpha}{\alpha}$ and $\lfloor x \rfloor > x - 1$, $I(T) > \frac{\alpha T}{m(\kappa_0+\alpha)} - 1$; moreover $v_{\mathcal{B}_i} = |\mathcal{B}_i|\alpha^2 \leq m(\kappa_0+\alpha)\alpha$ for every i, and the same bound holds for $v_{\mathcal{B}_{\mathrm{res}}}$ since $\mathcal{B}_{\mathrm{res}} \subseteq \mathcal{B}_{I(T)}$. Hence, for every interval $\mathcal{K}$ among the $\mathcal{B}_i$ and $\mathcal{B}_{\mathrm{res}}$,

$$\varepsilon_{\mathcal{K}}(\delta') \leq 2\,(1 + 2m(\kappa_0+\alpha))\,\bar{w}(R) \sqrt{\frac{1}{2} m(\kappa_0+\alpha)\alpha \log\left(\frac{2|\mathcal{S}||\mathcal{A}|\left(\frac{\alpha T}{m\kappa_0}+1\right)}{\delta}\right)},$$

which is the left-hand side of (H.17), so (H.17) implies (H.14). Theorem H.1 then applies, and the noise term is summed using $\sum_{i=0}^{I(T)-1}(1-\bar{\beta}(R))^{I(T)-1-i} \leq \frac{1}{\bar{\beta}(R)}$, so that the bracket is at most $\left(\frac{1}{\bar{\beta}(R)}+1\right)\sqrt{m(\kappa_0+\alpha)\alpha}$. Finally $1-\bar{\beta}(R) \in (0,1)$ and $I(T) > \frac{\alpha T}{m(\kappa_0+\alpha)} - 1$ give the prefactor $\frac{1}{1-\bar{\beta}(R)}$ on the contraction term. □

H.2.8 Proof of Theorem 1

Theorem 1 (Synchronous Ω-regularized Q-learning)**.** Let Assumptions 1 to 3 and (P1) hold, and set $\alpha(u) := (\tilde{\beta}^2 u^2)(9m(2+\tilde{\beta})^2(1+9m)^2\bar{w}^2\iota)^{-1}$ for $u > 0$. Run Algorithm 3 for T iterations with a constant step size $\alpha \leq \alpha(e_0)$. Then, with probability at least $1-\delta$, $\sup_{t \leq T} \|Q_t - Q^*_\Omega\|_{\mathrm{sp}} \leq 2e_0$ and

$$\|Q_T - Q^*_\Omega\|_{\mathrm{sp}} \leq \frac{(1-\tilde{\beta}/2)^{\frac{2\alpha T}{9m}}}{1-\tilde{\beta}/2}\,\|Q_0 - Q^*_\Omega\|_{\mathrm{sp}} + \frac{3(2+\tilde{\beta})(1+9m)}{\tilde{\beta}}\,\bar{w}\sqrt{m\alpha\iota}. \tag{2}$$

In particular, for $\varepsilon \in (0, e_0)$, the choices $\alpha = \alpha(\varepsilon/2)$ and $T \geq \frac{9m}{\alpha\tilde{\beta}} \log\left(\frac{2e_0}{(1-\tilde{\beta}/2)\varepsilon}\right)$ give $\|Q_T - Q^*_\Omega\|_{\mathrm{sp}} \leq \varepsilon$, i.e. $\tilde{O}\big(|\mathcal{S}||\mathcal{A}|\, m^4\bar{w}^2/(\tilde{\beta}^3\varepsilon^2)\big)$ sampled transitions.

Proof of Theorem 1. Take $\kappa_0 = 4$ and the Chernoff bound of Remark H.2, so that $p_m \geq 1 - e^{-9/8} \geq \frac{1}{2}$; for $\alpha \leq \frac{1}{2}$, we also have $m(\kappa_0+\alpha) \leq \frac{9}{2}m$ and $C_\kappa \leq \sqrt{2}(1+9m)$, and $\alpha(e_0) \leq \frac{1}{2}$ since $\bar{w} \geq e_0$. We apply Corollary H.2 with $R = 2e_0$: then $A_R = \|Q^*_\Omega\|_{\mathrm{asp}} + 2e_0$, so that $\bar{\beta}(R) = p_m\tilde{\beta}_m(A_R) \geq \tilde{\beta}/2$ and $\bar{w}(R) = \bar{w}$, and the logarithm in (H.17) is at most ι since $\frac{\alpha T}{m\kappa_0} + 1 \leq T + 1$.

Admissibility. The left-hand side of (H.17) is at most $\sqrt{2}(1+9m)\bar{w}\sqrt{\frac{9}{2}m\alpha\iota} = 3(1+9m)\bar{w}\sqrt{m\alpha\iota}$, which is at most $\frac{\tilde{\beta}}{2+\tilde{\beta}}e_0$ when $\alpha \leq \alpha(e_0)$. Since $x \mapsto \frac{x}{1+x}$ is increasing and $\bar{\beta}(R) \geq \tilde{\beta}/2$, this is at most $\frac{\bar{\beta}(R)}{1+\bar{\beta}(R)}(R - e_0)$, which is (H.17).

Bound. With probability at least $1-\delta$, Theorem H.1 thus gives $\sup_{t \leq T} e_t \leq 2e_0$. Since $\bar{\beta}(R) \geq \tilde{\beta}/2$, $I(T) \geq 0$, and $I(T) > \frac{\alpha T}{m(\kappa_0+\alpha)} - 1 \geq \frac{2\alpha T}{9m} - 1$, its contraction term satisfies $(1-\bar{\beta}(R))^{I(T)}e_0 \leq$

$(1-\tilde{\beta}/2)^{I(T)}e_0 \le (1-\tilde{\beta}/2)^{\frac{2\alpha T}{9m}-1}e_0$. Since $x \mapsto (1+x)/x$ is non-increasing, the noise term of Corollary H.2 is at most $\frac{2+\tilde{\beta}}{\tilde{\beta}}\, 3(1+9m)\bar{w}\sqrt{m\alpha\iota}$. This is (2).

Sample complexity. For $\varepsilon \in (0, e_0)$, the choice $\alpha = \alpha(\varepsilon/2) \le \alpha(e_0)$ makes the noise term equal to $\varepsilon/2$, and $(1-\tilde{\beta}/2)^x \le e^{-\tilde{\beta}x/2}$ makes the contraction term at most $\varepsilon/2$ as soon as $T \ge \frac{9m}{\alpha\tilde{\beta}} \log\Big(\frac{2e_0}{(1-\tilde{\beta}/2)\varepsilon}\Big)$, which is possible since ι depends on T only logarithmically. With $\alpha(\varepsilon/2) = \tilde{\Theta}\big(\tilde{\beta}^2\varepsilon^2/(m^3\bar{w}^2)\big)$, this gives $\tilde{O}\big(m^4\bar{w}^2/(\tilde{\beta}^3\varepsilon^2)\big)$ iterations, each drawing $|\mathcal{S}||\mathcal{A}|$ transitions. Since the radius is $2e_0$, the constants $\tilde{\beta}$ and $\bar{w}$ do not depend on ε. □

Corollary H.3 (Synchronous Ω-regularized Q-learning for $m = 1$)**.** Let $m = 1$. Then $\tilde{\beta} = \beta_1$, and Theorem 1 holds even without (P1). Moreover $\mathcal{T}_\Omega$ is a $(1-\beta_1)$-contraction for $\|\cdot\|_{\mathrm{sp}}$ and $\|Q^*_\Omega\|_{\mathrm{sp}} \le (\|r\|_{\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}})/\beta_1$; in particular, for $Q_0 = 0$ and a regularizer whose minimal value does not depend on the state, $\bar{w} \le 3\|r\|_{\mathrm{sp}}/\beta_1$, and the sample complexity of Theorem 1 is $\tilde{O}\big(|\mathcal{S}||\mathcal{A}|\,\|r\|^2_{\mathrm{sp}}/(\beta_1^5\varepsilon^2)\big)$.

Proof. For $m = 1$, $\tilde{\beta}_1(R) = \beta_1$ for every R, so the localized constants of Definition H.2 do not depend on the radius and are positive without (P1), which the proof of Theorem 1 only uses to ensure $\tilde{\beta} > 0$; that proof then applies verbatim. For the second claim, Lemma G.2.(iv) with $m = 1$, whose product is empty, gives $\|\mathcal{T}_\Omega(Q) - \mathcal{T}_\Omega(Q')\|_{\mathrm{sp}} \le (1-\beta_1)\|Q - Q'\|_{\mathrm{sp}}$. Since $\|\cdot\|_{\mathrm{sp}}$ annihilates constants, $\|Q^*_\Omega\|_{\mathrm{sp}} = \|\mathcal{T}_\Omega(Q^*_\Omega)\|_{\mathrm{sp}} \le (1-\beta_1)\|Q^*_\Omega\|_{\mathrm{sp}} + \|\mathcal{T}_\Omega(0)\|_{\mathrm{sp}}$, and $\|\mathcal{T}_\Omega(0)\|_{\mathrm{sp}} = \|r + P\,\Omega^*(0)\|_{\mathrm{sp}} \le \|r\|_{\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}}$. Finally, for $Q_0 = 0$, $e_0 = \|Q^*_\Omega\|_{\mathrm{sp}}$, and for such a regularizer $\|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} = 0$ (Corollary F.1), so $\bar{w} = 3\|Q^*_\Omega\|_{\mathrm{sp}} \le 3\|r\|_{\mathrm{sp}}/\beta_1$. □

H.3 Proofs for Subsection 4.3 (Epoch Ω-regularized Q-learning)

Every constant below is a function of the domination profile γ_Ω (Definition G.1.(iii)) evaluated at the radius actually reached by the iterates, which drives both the contraction rate and the exploration rate.

Algorithm 1 has two noise handling mechanisms: the step sizes α_j, which averages across epochs, and the batch size B_j, which averages within an epoch. We therefore control the noise in two ways, both valid at all parameters:

- a *batch bound* (Lemma H.21.(i)), which bounds the per-epoch noise by $O(\bar{w}(R)/\sqrt{B_j})$ and sums it over a block by the triangle inequality;
- a *block-martingale bound* (Lemma H.21.(ii)), which concatenates all the transition-level increments of a block into a single martingale, as in Appendix H.2.

Neither dominates: for constant parameters their ratio is $(1+2N\alpha)/\sqrt{N}$, with N the number of epochs of a block, so the batch bound is tighter at $\alpha_j \equiv 1$ (Corollary H.4), the setting of Theorem 2, where the block-martingale bound loses a factor $\Theta(\sqrt{m})$ on the noise term, while the block-martingale bound wins for small step sizes and handles small batches, down to $B_j \equiv 1$, where it recovers Theorem H.1 (Remark H.6).

H.3.1 Setting, terminology and the epoch update

We write $e_j := \|Q_j - Q^*_\Omega\|_{\mathrm{sp}}$ for the error after j epochs and, for a radius $R \ge e_0$, we use the localized constants A_R, C_R, $\tilde{\beta}_m(A_R)$, $\bar{\beta}(R)$ and $\bar{w}(R)$ of Definition H.2; under (P1), $\tilde{\beta}_m(A_R)$ and $\bar{\beta}(R)$ are positive. We reserve the index j and the notation E_j for algorithmic epochs, and the index i for the blocks $\mathcal{B}_i$ introduced by the analysis; B_j is always the batch size. Three time scales coexist: transitions (n), epochs (j) and blocks (i).

The j-th epoch is $E_j := \{n_j, \dots, n_{j+1} - 1\}$ with $n_0 := 0$ and

$$n_{j+1} := \min\left\{ n > n_j \;\middle|\; \min_{(s,a)} \sum_{k=n_j}^{n-1} \mathbb{1}_{\{(s_k,a_k)=(s,a)\}} \ge B_j \right\}.$$

We reuse the sub-blocks and blocks $\mathcal{B}_i$ of Definition H.1 on the epoch scale, i.e. with the iteration index replaced by the epoch index, so that $m\kappa_0 \le \sum_{j\in\mathcal{B}_i}\alpha_j < m(\kappa_0+\alpha_0)$ (Definition H.1 and Lemma H.3), and we write $b_i := \min \mathcal{B}_i$. The step sizes and batch sizes being deterministic, so are the blocks and the b_i.

We write $\ell_j := n_{j+1} - n_j \in \mathbb{N}\cup\{+\infty\}$ for the (random) length of epoch j, with the convention

$$Q_{j+1} := Q_j \qquad \text{on } \{\ell_j = +\infty\}. \tag{H.18}$$

Let $\mathcal{G}_j := \sigma\big(Q_0, (s_0,a_0), \dots, (s_{n_j-1}, a_{n_j-1}), s_{n_j}\big)$; then Q_j is $\mathcal{G}_j$-measurable and, by construction of Algorithm 1, the behavior policy $\pi_j := \pi_{\Omega,Q_j}$ is frozen throughout epoch j. Set $\mathcal{H}_n^{(j)} := \sigma\big(\mathcal{G}_j, (s_{n_j}, a_{n_j}), \dots, (s_{n_j+n}, a_{n_j+n})\big)$. As in Appendix H.2.2, the centering projection Π only shifts Q_{j+1} by a constant, which leaves the greedy policies, hence the law of the trajectory, as well as $\|\cdot\|_{\mathrm{sp}}$ and $\|\cdot\|_{\mathrm{asp}}$ unchanged (Corollary F.1 and Lemma F.1.(i),(iii)), so we omit it; its only role is to upgrade span convergence to $\|\cdot\|_\infty$ convergence (Remark F.2).

Convention H.1 (Extended trajectory)**.** Within epoch j we let the trajectory be defined for *all* $n \ge 0$ under π_j: $\tilde{s}_0^{(j)} := s_{n_j}$, $\tilde{a}_n^{(j)} \sim \pi_j(\cdot|\tilde{s}_n^{(j)})$, $\tilde{s}_{n+1}^{(j)} \sim p(\cdot|\tilde{s}_n^{(j)}, \tilde{a}_n^{(j)})$, coupled with the algorithm for $n < \ell_j$. All visit times below refer to this extended trajectory, which conditionally on $\mathcal{G}_j$ is a time-homogeneous Markov chain with kernel K_{π_j}; its filtration $\widetilde{\mathcal{H}}_n^{(j)}$ coincides with $\mathcal{H}_n^{(j)}$ on $\{n < \ell_j\}$. The extension is a definition and requires no probabilistic input; Proposition H.2 shows a posteriori that it is almost surely never used.

Lemma H.14 (Exact form of the epoch update)**.** (i) On $\{\ell_j < +\infty\}$ we have $N_j(s,a) = B_j$ for every pair at exit, and

$$Q_{j+1} = F_j(Q_j) + \alpha_j W_{j+1}, \qquad F_j(Q) := (1-\alpha_j)Q + \alpha_j \mathcal{T}_\Omega(Q), \tag{H.19}$$

where, with $t_1(s,a) < \dots < t_{B_j}(s,a)$ the first B_j visit times to (s,a) in the extended trajectory of epoch j, counted from the beginning n_j of the epoch,

$$\begin{aligned} W_{j+1}(s,a) &= \frac{1}{B_j}\sum_{l=1}^{B_j} Z_l^{(j)}(s,a), \\ Z_l^{(j)}(s,a) &:= \Omega^*(Q_j)\big(\tilde{s}^{(j)}_{t_l(s,a)+1}\big) - \mathbb{E}_{s'\sim p(\cdot|s,a)}\big[\Omega^*(Q_j)(s')\big]. \end{aligned} \tag{H.20}$$

(ii) Every Q_j is a well-defined $\mathcal{G}_j$-measurable random vector of $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, with $\|Q_j\|_\infty < +\infty$ everywhere. Consequently π_j is a well-defined policy and the extended trajectory of Convention H.1 is a well-defined Markov chain, for every j and on every path.

Proof. (i). The inner loop takes into account a transition at (s,a) only while $N_j(s,a) < B_j$, so it accumulates exactly the first B_j visits and exits when the last pair reaches B_j; hence $N_j \equiv B_j$ at exit and $t_l(s,a) < \ell_j$ for $l \le B_j$, so the extended trajectory coincides with the algorithm at every visit time used and at the next state. After normalization the estimate is $\widehat{\mathcal{T}_\Omega Q_j}(s,a) = r(s,a) + \frac{1}{B_j}\sum_l \Omega^*(Q_j)(\tilde{s}^{(j)}_{t_l(s,a)+1})$, the reward being deterministic; subtracting $\mathcal{T}_\Omega(Q_j)(s,a)$ gives (H.20), and injecting into the update gives (H.19).

(ii). By induction on j; the argument is deterministic. The initialization Q_0 is a fixed finite vector. Assume Q_j finite. On $\{\ell_j = +\infty\}$, (H.18) gives $Q_{j+1} = Q_j$. On $\{\ell_j < +\infty\}$, (i) gives $Q_{j+1} = F_j(Q_j) + \alpha_j W_{j+1}$, and each coordinate of W_{j+1} is a difference between a value of $\Omega^*(Q_j)$ and a convex combination of such values, so $\|W_{j+1}\|_\infty \le \|\Omega^*(Q_j)\|_{\mathcal{S},\mathrm{sp}} \le \|Q_j\|_{\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} < +\infty$ by Lemma F.1.(vii); hence Q_{j+1} is finite. Measurability is that of the update, and $\pi_j = \nabla\Omega^*(Q_j)$ is well defined for any finite argument by Lemma F.1.(i). □

The length ℓ_j is random, but because of the cap the number of terms in (H.20) is deterministic, equal to B_j. The martingale arguments of Lemmas H.19 and H.20 rely on this, the second one through the deterministic length $\sum_{j\in\mathcal{B}_i} B_j$ of the concatenated martingale.

H.3.2 EXIT TIME, WELL-POSEDNESS AND EPOCH LENGTH

We define the exit time of the iterates in the epoch time scale, at fixed radius $R \geq e_0$, by

$$\sigma_R := \inf\{j \geq 0 \mid e_j > R\} \in \mathbb{N} \cup \{+\infty\}. \tag{H.21}$$

Remark H.4 (Exit time)**.** The iterates are frozen within an epoch, so σ_R is a stopping time for $(\mathcal{G}_j)_{j\geq 0}$ and $\{j < \sigma_R\} \in \mathcal{G}_j$. Since $e_0 \leq R$, $\sigma_R \geq 1$. As in Remark H.3.(ii), $j < \sigma_R$ implies $\|Q_j\|_{\mathrm{sp}} \leq C_R$ and $\|Q_j\|_{\mathrm{asp}} \leq A_R$. As in Remark H.3.(iv), the localization bounds the noise and, since γ_Ω may vanish at infinity, keeps both the contraction and the coverage rates at least at their values at radius A_R.

Lemma H.15 (Conditional cover time at a deterministic rate)**.** Let $j \geq 0$, let $u \geq 0$ be a radius and let $\delta' \in (0,1)$. Set

$$\mathcal{D}_j(u) := \{\|Q_j\|_{\mathrm{asp}} \leq u\} \in \mathcal{G}_j, \qquad q(u) := \beta_m\, \gamma_\Omega(u)^m\, \eta_{\min}, \tag{H.22}$$

$$K(u,\delta') := \left\lceil \frac{\left(\sqrt{B_j} + \sqrt{\iota(\delta')}\right)^2}{q(u)} \right\rceil, \qquad L(u,\delta') := m\,K(u,\delta')+1, \qquad \iota(\delta') := \log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta'}\right). \tag{H.23}$$

Under (P1), $q(u) > 0$ and, almost surely,

$$\mathbb{P}\left(\{\ell_j > L(u,\delta')\} \cap \mathcal{D}_j(u) \mid \mathcal{G}_j\right) \;\leq\; \delta'\, \mathbb{1}_{\mathcal{D}_j(u)}. \tag{H.24}$$

Moreover $L(u,\delta') \geq m|\mathcal{S}||\mathcal{A}|B_j$. In particular, for $R \geq e_0$, the coverage constant and the localized length

$$q_R := q(A_R) = \gamma_\Omega(A_R)\,\tilde{\beta}_m(A_R)\,\eta_{\min}, \qquad L_j(\delta') := L(A_R,\delta') = m\left\lceil \frac{\left(\sqrt{B_j} + \sqrt{\iota(\delta')}\right)^2}{q_R} \right\rceil + 1 \tag{H.25}$$

satisfy $\mathbb{P}\left(\{\ell_j > L_j(\delta')\} \cap \{j < \sigma_R\} \mid \mathcal{G}_j\right) \leq \delta'\mathbb{1}_{\{j<\sigma_R\}}$ almost surely.

Proof. Since $\mathcal{D}_j(u) \in \mathcal{G}_j$, it suffices to prove $\mathbb{P}(\ell_j > L(u,\delta') \mid \mathcal{G}_j) \leq \delta'$ almost surely on $\mathcal{D}_j(u)$; we work under a regular conditional probability given $\mathcal{G}_j$ and fix an outcome in $\mathcal{D}_j(u)$. All the constants $q(u) > 0$, $K := K(u,\delta')$ and $L(u,\delta')$ are deterministic, u being deterministic; positivity of $q(u)$ is (P1) at the deterministic radius u.

By Lemma H.14.(ii), the extended trajectory of epoch j is well defined and, conditionally on $\mathcal{G}_j$, is a time-homogeneous Markov chain with kernel K_{π_j}; the conditioning is legitimate because $\mathcal{D}_j(u) \in \mathcal{G}_j \subseteq \widetilde{\mathcal{H}}_n^{(j)}$. On $\mathcal{D}_j(u)$, Definition G.1.(iii) applied at radius u gives $\pi_j(a|s) \geq \gamma_\Omega(u)\mu(a|s)$ for all (s,a), so Lemma G.2.(iii) applied with $\pi_1 = \cdots = \pi_m = \pi_j$ gives, for every $n \geq 0$ and every $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\mathbb{P}\left(\left(\tilde{s}_{n+m}^{(j)}, \tilde{a}_{n+m}^{(j)}\right) = (s,a) \;\middle|\; \widetilde{\mathcal{H}}_n^{(j)}\right) = K_{\pi_j}^m\left(s,a \,|\, \tilde{s}_n^{(j)}, \tilde{a}_n^{(j)}\right) \;\geq\; \beta_m\,\gamma_\Omega(u)^m\,\eta(s,a) \;\geq\; q(u). \tag{H.26}$$

Fix (s,a) and set $Z_l := \mathbb{1}_{\{(\tilde{s}_{lm}^{(j)}, \tilde{a}_{lm}^{(j)}) = (s,a)\}} \in [0,1]$ for $l \in [\![1, K]\!]$, adapted to $\mathcal{G}_l^{\mathrm{sub}} := \widetilde{\mathcal{H}}_{lm}^{(j)}$. Consecutive sub-sampling instants being exactly m steps apart, (H.26) gives $\mathbb{E}[Z_l \mid \mathcal{G}_{l-1}^{\mathrm{sub}}] \geq q(u)$, so the predictable mean satisfies $\bar{S} \geq Kq(u)$ deterministically. The exact multiplicative-Chernoff form (E.1) of Lemma E.1, applied with increment bound $b = 1$, $a = Kq(u)$ and $c = B_j$, all deterministic, gives

$$\mathbb{P}\left(\sum_{l=1}^{K} Z_l < B_j \;\middle|\; \mathcal{G}_j\right) \leq \exp\left(-B_j\, h(\theta)\right), \qquad \theta := \frac{Kq(u)}{B_j}, \quad h(\theta) := \theta - 1 - \log\theta, \tag{H.27}$$

the hypothesis $c < a$ holding since $Kq(u) \geq (\sqrt{B_j} + \sqrt{\iota(\delta')})^2 > B_j$.

Now $h(\theta) \geq (\sqrt{\theta} - 1)^2$ for $\theta \geq 1$: both sides vanish at $\theta = 1$ and $h'(\theta) = 1 - \theta^{-1} \geq 1 - \theta^{-1/2}$ there, the right-hand side being the derivative of $\theta \mapsto (\sqrt{\theta} - 1)^2$. By the choice of K, $\theta \geq (1 +$

$\sqrt{\iota(\delta')/B_j})^2$, hence $B_j h(\theta) \geq B_j(\sqrt{\theta}-1)^2 \geq \iota(\delta')$ and the right-hand side of (H.27) is at most $\delta'/(|\mathcal{S}||\mathcal{A}|)$.

A union bound over the $|\mathcal{S}||\mathcal{A}|$ pairs shows that, with conditional probability at least $1-\delta'$, every pair is visited at least B_j times among the instants $\{m, 2m, \dots, Km\} \subseteq [\![0, mK]\!]$, so $\ell_j \leq mK+1 = L(u,\delta')$. This is (H.24).

Finally $q(u) \leq \eta_{\min} \leq (|\mathcal{S}||\mathcal{A}|)^{-1}$, so $K(u,\delta') \geq |\mathcal{S}||\mathcal{A}|B_j$ and $L(u,\delta') \geq m|\mathcal{S}||\mathcal{A}|B_j$. For the last claim, Remark H.4 gives $\{j < \sigma_R\} \subseteq \mathcal{D}_j(A_R)$, both events belonging to $\mathcal{G}_j$; apply (H.24) with the deterministic radius $u = A_R$ and intersect with $\{j < \sigma_R\}$. □

Proposition H.2 (Well-posedness)**.** For every $j \geq 0$, $\ell_j < +\infty$ almost surely. Consequently convention (H.18) is vacuous, the extension of Convention H.1 is almost surely never used, and Algorithm 1 produces the update (H.19) at every epoch almost surely.

Proof. Fix j. For every deterministic $u \geq 0$ and every $\delta' \in (0,1)$, integrating (H.24) gives

$$\mathbb{P}(\ell_j = +\infty) \;\leq\; \mathbb{P}(\{\ell_j > L(u,\delta')\} \cap \mathcal{D}_j(u)) + \mathbb{P}(\mathcal{D}_j(u)^c) \;\leq\; \delta' + \mathbb{P}(\|Q_j\|_{\mathrm{asp}} > u)\,.$$

Letting $\delta' \downarrow 0$ at fixed u gives $\mathbb{P}(\ell_j = +\infty) \leq \mathbb{P}(\|Q_j\|_{\mathrm{asp}} > u)$ for every u, and letting $u \uparrow +\infty$ gives $\mathbb{P}(\ell_j = +\infty) = 0$, since $\|Q_j\|_{\mathrm{asp}} \leq 2\|Q_j\|_\infty < +\infty$ everywhere by Lemma H.14.(ii). □

Given a transition budget $T \geq 0$ we write

$$J(T) := \max\{j \mid n_j \leq T\}, \qquad I(T) := \max\{i \mid b_i \leq J(T)\}, \tag{H.28}$$

and $\mathcal{B}_{\mathrm{res}} := [\![b_{I(T)}, J(T)-1]\!] \subseteq \mathcal{B}_{I(T)}$ for the last incomplete block.

Lemma H.16 (Epoch and block counts)**.** Let $T \geq 0$, $\underline{B} := \min_j B_j \geq 1$. Then:

(i) *Deterministic upper bounds*: $\ell_j \geq |\mathcal{S}||\mathcal{A}|B_j$, hence $J(T) \leq J_{\max}(T) := \left\lfloor \frac{T}{|\mathcal{S}||\mathcal{A}|\underline{B}} \right\rfloor$ and $I(T) \leq I_{\max}(T) := \left\lfloor \frac{J_{\max}(T)}{m} \right\rfloor$.

(ii) *From epochs to blocks*: $I(T) > \frac{\sum_{j<J(T)}\alpha_j}{m(\kappa_0+\alpha_0)} - 1$, i.e. $I(T) \geq \left\lfloor \frac{\sum_{j<J(T)}\alpha_j}{m(\kappa_0+\alpha_0)} \right\rfloor$. If $\alpha_j \equiv \alpha$, every block contains exactly $m\lceil\kappa_0/\alpha\rceil$ epochs and $I(T) = \left\lfloor \frac{J(T)}{m\lceil\kappa_0/\alpha\rceil} \right\rfloor$.

(iii) *From transitions to epochs*: set $J^-(T) := \max\left\{J \mid \sum_{j<J} L_j(\delta') \leq T\right\}$. If $\ell_j \leq L_j(\delta')$ for every $j < \min\{\sigma_R, J_{\max}(T)\}$ and $\sigma_R > J(T)$, then $J(T) \geq J^-(T)$. In particular for $\alpha_j \equiv \alpha$, $B_j \equiv B$, writing $L(\delta')$ for the common value,

$$J(T) \geq \left\lfloor \frac{T}{L(\delta')} \right\rfloor, \qquad I(T) \geq \left\lfloor \frac{\alpha T}{m(\kappa_0+\alpha)L(\delta')} - \frac{\alpha}{m(\kappa_0+\alpha)} \right\rfloor. \tag{H.29}$$

Proof. (i). Epoch j terminates only when each of the $|\mathcal{S}||\mathcal{A}|$ pairs has been visited at least B_j times, and each transition visits one pair, so $\ell_j \geq |\mathcal{S}||\mathcal{A}|B_j$; summing, $J(T)|\mathcal{S}||\mathcal{A}|\underline{B} \leq n_{J(T)} \leq T$. Each sub-block is non-empty, so each block has at least m epochs.

(ii). By definition of $I(T)$, $b_{I(T)+1} > J(T)$, and each block consumes at most $m(\kappa_0+\alpha_0)$ (Lemma H.3), so $\sum_{j<J(T)}\alpha_j \leq \sum_{j<b_{I(T)+1}}\alpha_j < (I(T)+1)m(\kappa_0+\alpha_0)$; the floor form follows since $I(T)$ is an integer. For $\alpha_j \equiv \alpha$ the stopping rule fires after exactly $\lceil\kappa_0/\alpha\rceil$ epochs.

(iii). If $J(T) = J_{\max}(T)$ there is nothing to prove, since $J^-(T) \leq J_{\max}(T)$ by the bound $L_j(\delta') \geq m|\mathcal{S}||\mathcal{A}|B_j$ of Lemma H.15. Otherwise suppose $J(T) < J^-(T)$. Every $j \leq J(T)$ satisfies $j < \sigma_R$ and $j < J_{\max}(T)$, hence $\ell_j \leq L_j(\delta')$, so $n_{J(T)+1} = \sum_{j\leq J(T)}\ell_j \leq \sum_{j<J(T)+1} L_j(\delta') \leq T$, the last inequality because $J(T)+1 \leq J^-(T)$; this contradicts the maximality of $J(T)$. The constant case follows by injecting $J(T) \geq \lfloor T/L(\delta')\rfloor$ into the exact identity of (ii) and using $\lfloor x\rfloor > x-1$ along with $\lceil\kappa_0/\alpha\rceil < (\kappa_0+\alpha)/\alpha$. □

Remark H.5. Only (i) is used for the union bounds, and it is deterministic: this is what lets δ' be fixed in advance although $J(T)$ and $I(T)$ are random. There is no circularity in (iii): Proposition H.3 establishes $\sigma_R > J(T)$ using only the almost sure finiteness of Proposition H.2, and only then are the length bounds available.

H.3.3 THE NOISELESS FLOW BEFORE THE EXIT TIME

For each block $\mathcal{B}_i$ define the *virtual sequence* by $\tilde{Q}_{b_i} := Q_{b_i}$ and $\tilde{Q}_{j+1} := F_j(\tilde{Q}_j)$ for $j \geq b_i$, i.e. the full-knowledge synchronous update (H.1) started at Q_{b_i}; it is $\mathcal{G}_{b_i}$-measurable.

Lemma H.17 (Contraction of the virtual flow at radius R)**.** Let i be such that $e_{b_i} \leq R$. Then, pathwise:

(i) $\|\tilde{Q}_j - Q^*_\Omega\|_{\mathrm{sp}} \leq e_{b_i}$ for every $j \geq b_i$;
(ii) $\|\tilde{Q}_{b_{i+1}} - Q^*_\Omega\|_{\mathrm{sp}} \leq (1 - \bar{\beta}(R))\, e_{b_i}$.

Proof. This is Lemma H.10 with the step index replaced by the epoch index: the statement concerns the full-knowledge update (H.1) only, and $\mathcal{B}_i$ is a concatenation of m sub-blocks in the sense of Definition H.1. □

H.3.4 LOCALIZED NOISE CONTROL

The noise increments below lie in intervals whose endpoints are known one step ahead, which is the setting of the maximal Azuma–Hoeffding inequality of Lemma E.2. We write $\widehat{W}_{j+1} := \mathbb{1}_{\{j<\sigma_R\}} W_{j+1}$ and $\widehat{Z}_l^{(j)} := \mathbb{1}_{\{j<\sigma_R\}} Z_l^{(j)}$ for the stopped noise, so that $\widehat{W}_{j+1} = \frac{1}{B_j}\sum_l \widehat{Z}_l^{(j)}$.

Lemma H.18 (Epoch noise: martingale structure and range)**.** Fix $x = (s,a)$ and let $\mathcal{K}_l^{(j)} := \sigma\left(\widetilde{\mathcal{H}}_{t_l(x)}^{(j)}, \tilde{s}_{t_l(x)+1}^{(j)}\right)$ for $l \in [\![1, B_j]\!]$, $\mathcal{K}_0^{(j)} := \mathcal{G}_j$. Then:

(i) $\mathcal{K}_0^{(j)} \subseteq \cdots \subseteq \mathcal{K}_{B_j}^{(j)} \subseteq \mathcal{G}_{j+1} \subseteq \mathcal{K}_0^{(j+1)}$, so the concatenation over j is a filtration at fixed x; in particular W_{j+1} is $\mathcal{G}_{j+1}$-measurable;
(ii) $\mathbb{E}[\widehat{Z}_l^{(j)}(x) \mid \mathcal{K}_{l-1}^{(j)}] = 0$; consequently $\mathbb{E}[\widehat{W}_{j+1} \mid \mathcal{G}_j] = 0$;
(iii) $\widehat{Z}_l^{(j)}(x)$ lies in a $\mathcal{G}_j$-measurable interval of length $\mathbb{1}_{\{j<\sigma_R\}}\|\Omega^*(Q_j)\|_{\mathcal{S},\mathrm{sp}} \leq \bar{w}(R)$.

Proof. (i). The $t_l(x)$ are stopping times for $(\widetilde{\mathcal{H}}_n^{(j)})_n$, a.s. finite by Proposition H.2, and $t_{l-1}(x)+1 \leq t_l(x)$, so $\mathcal{K}_{l-1}^{(j)} \subseteq \widetilde{\mathcal{H}}_{t_l(x)}^{(j)} \subseteq \mathcal{K}_l^{(j)}$; and $t_{B_j}(x) + 1 \leq \ell_j$, the extended trajectory coinciding with the algorithm before ℓ_j, gives $\mathcal{K}_{B_j}^{(j)} \subseteq \mathcal{G}_{j+1}$.

(ii). On $\{t_l(x) = n\}$ we have $(\tilde{s}_n^{(j)}, \tilde{a}_n^{(j)}) = x$ and the conditional law of $\tilde{s}_{n+1}^{(j)}$ given $\widetilde{\mathcal{H}}_n^{(j)}$ is $p(\cdot \mid x)$, so $\mathbb{E}[Z_l^{(j)}(x) \mid \widetilde{\mathcal{H}}_{t_l(x)}^{(j)}] = 0$. Since $\mathbb{1}_{\{j<\sigma_R\}}$ is $\mathcal{G}_j$-, hence $\mathcal{K}_{l-1}^{(j)}$-measurable and $\mathcal{K}_{l-1}^{(j)} \subseteq \widetilde{\mathcal{H}}_{t_l(x)}^{(j)}$, the tower property gives the claim.

(iii). Conditionally on $\mathcal{G}_j$, $\Omega^*(Q_j)$ is fixed, so $Z_l^{(j)}(x)$ lies in the $\mathcal{G}_j$-measurable interval of length exactly $\|\Omega^*(Q_j)\|_{\mathcal{S},\mathrm{sp}}$, translated by $-(P\Omega^*(Q_j))(x)$. On $\{j < \sigma_R\}$, Lemma F.1.(vii) and Remark H.4 give $\|\Omega^*(Q_j)\|_{\mathcal{S},\mathrm{sp}} \leq \|Q_j\|_{\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} \leq C_R + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} = \bar{w}(R)$; off the event $\widehat{Z}_l^{(j)} = 0$. □

Batch bound.

Lemma H.19 (Per-epoch noise)**.** For every j and $\delta' \in (0,1)$, with probability at least $1 - \delta'$,

$$\left\|\widehat{W}_{j+1}\right\|_\infty \leq \bar{U}_j(\delta') := \bar{w}(R)\sqrt{\frac{1}{2B_j}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta'}\right)}, \qquad \text{hence} \qquad \left\|\widehat{W}_{j+1}\right\|_{\mathrm{sp}} \leq 2\bar{U}_j(\delta'). \tag{H.30}$$

Proof. Fix x and apply Lemma E.2 to $M_k := \sum_{l\leq k} \widehat{Z}_l^{(j)}(x)$ with $n = B_j$ deterministic (Lemma H.14) and $c_l \equiv \bar{w}(R)$, legitimate by Lemma H.18. With $u = B_j\bar{U}_j(\delta')$ and a union bound over the pairs this is (H.30); the last inequality is $\|\cdot\|_{\mathrm{sp}} \leq 2\|\cdot\|_\infty$. □

Block-martingale bound.

Lemma H.20 (Block martingale)**.** Fix i and define $\widehat{U}^{(i)}_{b_i} := 0$, $\widehat{U}^{(i)}_{j+1} := \widehat{U}^{(i)}_j + \alpha_j \widehat{W}_{j+1}$. Then for every $\delta' \in (0,1)$, with probability at least $1-\delta'$,

$$\max_{b_i \le j \le b_{i+1}} \left\| \widehat{U}^{(i)}_j \right\|_\infty \le \bar{V}_i(\delta') := \bar{w}(R) \sqrt{\frac{1}{2} \left(\sum_{j \in \mathcal{B}_i} \frac{\alpha_j^2}{B_j} \right) \log \left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta'} \right)}. \tag{H.31}$$

The same holds with $\mathcal{B}_i$ replaced by any prefix, in particular by $\mathcal{B}_{\mathrm{res}}$.

Proof. Fix x and concatenate chronologically the increments $\left(\frac{\alpha_j}{B_j} \widehat{Z}^{(j)}_l(x) \right)_{j \in \mathcal{B}_i,\, l \le B_j}$, adapted to the filtration of Lemma H.18.(i). They are martingale differences by (ii), and by (iii) the l-th increment of epoch j has predictable range of deterministic length $\frac{\alpha_j}{B_j} \bar{w}(R)$. The total number of increments $\sum_{j \in \mathcal{B}_i} B_j$ is deterministic (Lemma H.14) and

$$\sum_{j \in \mathcal{B}_i} \sum_{l=1}^{B_j} \left(\frac{\alpha_j}{B_j} \bar{w}(R) \right)^2 = \bar{w}(R)^2 \sum_{j \in \mathcal{B}_i} \frac{\alpha_j^2}{B_j}.$$

The partial sums at epoch boundaries are exactly $\widehat{U}^{(i)}_j(x)$, so Lemma E.2 with $u = \bar{V}_i(\delta')$ and a union bound over the pairs give (H.31). □

The tracking lemma.

Lemma H.21 (Tracking within a block)**.** Let i be such that $e_{b_i} \le R$, let $\tilde{Q}$ be the virtual sequence started at Q_{b_i} and $E_j := Q_j - \tilde{Q}_j$. Then for every $j \in \mathcal{B}_i \cup \{b_{i+1}\}$ with $j \le \sigma_R$:

(i) on $\bigcap_{k \in \mathcal{B}_i} \{ \|\widehat{W}_{k+1}\|_\infty \le \bar{U}_k(\delta') \}$: $\|E_j\|_{\mathrm{sp}} \le \varepsilon_i^{\mathrm{bat}}(\delta') := 2 \sum_{k \in \mathcal{B}_i} \alpha_k \bar{U}_k(\delta')$;
(ii) on $\{ \max_k \|\widehat{U}^{(i)}_k\|_\infty \le \bar{V}_i(\delta') \}$: $\|E_j\|_{\mathrm{sp}} \le \varepsilon_i^{\mathrm{mg}}(\delta') := 2 \left(1 + 2 \sum_{k \in \mathcal{B}_i} \alpha_k \right) \bar{V}_i(\delta')$.

Both hold verbatim on any prefix, in particular on $\mathcal{B}_{\mathrm{res}}$.

Proof. For $k \in \mathcal{B}_i$ with $k < \sigma_R$, $\widehat{W}_{k+1} = W_{k+1}$ and $\ell_k < +\infty$ (Proposition H.2), so (H.19) applies. Subtracting the virtual recursion and using the average-path identity of Lemma G.2.(i), i.e. $\mathcal{T}_\Omega(Q_k) - \mathcal{T}_\Omega(\tilde{Q}_k) = \bar{K}_k E_k$ with $\bar{K}_k := K_{\bar{\pi}_{\Omega,[Q_k,\tilde{Q}_k]}}$,

$$E_{k+1} = M_k E_k + \alpha_k \widehat{W}_{k+1}, \qquad M_k := (1-\alpha_k) I + \alpha_k \bar{K}_k, \tag{H.32}$$

with M_k row-stochastic, hence span non-expansive (Lemma E.6).

(i). $\|E_{k+1}\|_{\mathrm{sp}} \le \|E_k\|_{\mathrm{sp}} + 2\alpha_k \bar{U}_k(\delta')$ by (H.30), and an immediate induction from $E_{b_i} = 0$ concludes.

(ii). The recursion (H.32) has the same form as (H.13), so the decomposition $E_k = \widehat{R}_k + \widehat{U}^{(i)}_k$ in the proof of Lemma H.13, with $\widehat{U}^{(i)}$ and $\bar{V}_i(\delta')$ in place of $\widehat{U}^{(\mathcal{K})}$ and $\bar{U}_{\mathcal{K}}(\delta')$, gives $\|E_j\|_{\mathrm{sp}} \le \varepsilon_i^{\mathrm{mg}}(\delta')$. □

H.3.5 No escape before the horizon

Let $\delta \in (0,1)$, $T \ge 0$ and $\delta' := \frac{\delta}{3(J_{\max}(T)+1)}$. We set

$$\varepsilon_i(\delta') := \min \left\{ \varepsilon_i^{\mathrm{bat}}(\delta'), \varepsilon_i^{\mathrm{mg}}(\delta') \right\}, \qquad \bar{\varepsilon}_R(\delta) := \max_{0 \le i \le I_{\max}(T)} \varepsilon_i(\delta'), \tag{H.33}$$

and write $\varepsilon_{\mathrm{res}}(\delta')$ for the same quantity with $\mathcal{B}_i$ replaced by $\mathcal{B}_{\mathrm{res}}$. Both bounds of Lemma H.21 are non-decreasing when a prefix is extended, so $\varepsilon_{\mathrm{res}}(\delta') \le \varepsilon_{I(T)}(\delta') \le \bar{\varepsilon}_R(\delta)$.

Lemma H.22 (Good event)**.** The event

$$\mathcal{E}_R(\delta) := \bigcap_{j<J_{\max}(T)} \left\{ \left\| \widehat{W}_{j+1} \right\|_\infty \le \bar{U}_j(\delta') \right\} \cap \bigcap_{i\le I_{\max}(T)} \left\{ \max_{b_i\le j\le b_{i+1}} \left\| \widehat{U}_j^{(i)} \right\|_\infty \le \bar{V}_i(\delta') \right\} \cap \bigcap_{j<J_{\max}(T)} (\{j \ge \sigma_R\} \cup \{\ell_j \le L_j(\delta')\}) \tag{H.34}$$

satisfies $\mathbb{P}(\mathcal{E}_R(\delta)) \ge 1-\delta$.

Proof. The three families contain at most $J_{\max}(T)$, $I_{\max}(T)+1 \le J_{\max}(T)+1$ and $J_{\max}(T)$ events, hence at most $3(J_{\max}(T)+1)$ in total, each of failure probability at most δ' by Lemmas H.19 and H.20 and Lemma H.15, the latter applied conditionally on $\mathcal{G}_j$ and integrated. □

Proposition H.3 (No escape)**.** Let $R \ge e_0$ satisfy the *radius admissibility condition*

$$\bar{\varepsilon}_R(\delta) \le \frac{\bar{\beta}(R)}{\bar{\beta}(R)+1}\,(R-e_0)\,. \tag{H.35}$$

Then, on $\mathcal{E}_R(\delta)$, $\sigma_R > J(T)$; in particular $\sup_{0\le j\le J(T)} e_j \le R$ and $\ell_j \le L_j(\delta')$ for every $j < J(T)$, and for every $0 \le i \le I(T)$,

$$e_{b_i} \le (1-\bar{\beta}(R))^i e_0 + \sum_{\ell=0}^{i-1} (1-\bar{\beta}(R))^{i-1-\ell} \varepsilon_\ell(\delta'). \tag{H.36}$$

Proof. Work on $\mathcal{E}_R(\delta)$. The proof of Proposition H.1 applies verbatim on the epoch scale, with the exit time (H.21), Remark H.4 for $\sigma_R \ge 1$, Lemmas H.17 and H.21 in place of Lemmas H.10 and H.13, and $\varepsilon_i(\delta')$ and $\varepsilon_{\mathrm{res}}(\delta') \le \bar{\varepsilon}_R(\delta)$ in place of $\varepsilon_{\mathcal{B}_i}(\delta')$ and $\varepsilon_{\mathcal{B}_{\mathrm{res}}}(\delta')$; it gives $\sigma_R > J(T)$ and (H.36). The length bound is then the third family of (H.34), since every $j < J(T)$ satisfies $j < \sigma_R$ and $j < J_{\max}(T)$. □

H.3.6 Main result

Theorem H.2 (Finite-time bound for Algorithm 1)**.** Let $\delta \in (0,1)$ and $T \ge 0$. Under Assumption H.1, let $\kappa_0 > 0$ be arbitrary and let $R \ge e_0$ satisfy the admissibility condition (H.35). Then, with probability at least $1-\delta$, $\sup_{0\le j\le J(T)} e_j \le R$ and

$$e_{J(T)} \le (1-\bar{\beta}(R))^{I(T)} e_0 + \sum_{i=0}^{I(T)-1} (1-\bar{\beta}(R))^{I(T)-1-i} \varepsilon_i(\delta') + \varepsilon_{\mathrm{res}}(\delta'), \tag{H.37}$$

where $J(T)$ and $I(T)$ are bounded below in Lemma H.16.(iii). Explicitly, with $\iota_1 := \log\left(\frac{6|\mathcal{S}||\mathcal{A}|(J_{\max}(T)+1)}{\delta}\right)$,

$$\varepsilon_i(\delta') = \bar{w}(R)\sqrt{2\iota_1}\,\min\left\{ \sum_{j\in\mathcal{B}_i} \frac{\alpha_j}{\sqrt{B_j}},\ \left(1+2\sum_{j\in\mathcal{B}_i}\alpha_j\right)\sqrt{\sum_{j\in\mathcal{B}_i}\frac{\alpha_j^2}{B_j}} \right\}. \tag{H.38}$$

Proof. Work on $\mathcal{E}_R(\delta)$, of probability at least $1-\delta$ (Lemma H.22). As in the proof of Theorem H.1, Proposition H.3, Lemma H.21 on $\mathcal{B}_{\mathrm{res}}$ and Lemma H.17.(i) give the containment and $e_{J(T)} \le e_{b_{I(T)}} + \varepsilon_{\mathrm{res}}(\delta')$, and (H.36) at rank $I(T)$ yields (H.37). For (H.38), $2\bar{U}_j(\delta') = \bar{w}(R)B_j^{-1/2}\sqrt{2\iota_1}$ and $2\bar{V}_i(\delta') = \bar{w}(R)\sqrt{2\iota_1\sum_{j\in\mathcal{B}_i}\alpha_j^2/B_j}$. □

Remark H.6 (Recovering Theorem H.1)**.** At $B_j \equiv 1$, the block-martingale bound in (H.38) is exactly the quantity $\varepsilon_{\mathcal{B}_i}(\delta')$ of (H.11), with a different value of δ'. Theorems H.1 and H.2 are then the same statement up to this value; the only difference is the cost model, $|\mathcal{S}||\mathcal{A}|$ sampled transitions per iteration there against $L_j(\delta')$ per epoch here.

H.3.7 THE CASE $\alpha_j \equiv 1$ AND PROOF OF THEOREM 2

Corollary H.4 ($\alpha_j \equiv 1$: empirical Ω-regularized value iteration)**.** Take $\alpha_j \equiv 1$, $B_j \equiv B$, and define the blocks directly as $\mathcal{B}_i := [\![im, (i+1)m - 1]\!]$, so that $I(T) = \lfloor J(T)/m \rfloor$. Then $F_j = \mathcal{T}_\Omega$, the block map is $\mathcal{T}_\Omega^m$, and Lemma H.17 holds with $\bar\beta(R)$ replaced by $\tilde\beta_m(A_R) = \beta_m \gamma_\Omega(A_R)^{m-1}$, with no sub-block budget κ_0 and no probability p_m. With $\bar U := \bar w(R)\sqrt{\iota_1/(2B)}$, the batch bound gives $\varepsilon_i(\delta') \le 2m\bar U$ and $\varepsilon_{\mathrm{res}}(\delta') \le 2m\bar U$, condition (H.35) becomes

$$2m\bar U \le \frac{\tilde\beta_m(A_R)}{1 + \tilde\beta_m(A_R)}\,(R - e_0)\,, \tag{H.39}$$

and with probability at least $1 - \delta$,

$$\sup_{0 \le j \le J(T)} e_j \le R, \qquad e_{J(T)} \le \Big(1 - \tilde\beta_m(A_R)\Big)^{I(T)} e_0 + \frac{(1 + \tilde\beta_m(A_R))2m\bar U}{\tilde\beta_m(A_R)}. \tag{H.40}$$

Proof. With $\alpha_j = 1$, $\tilde Q_j = \mathcal{T}_\Omega^{j-b_i}(Q_{b_i})$. By Corollary F.1, $\mathcal{T}_\Omega^k(Q^*_\Omega) = Q^*_\Omega + k\rho^*_\Omega \mathbf{1}$, of action-span $\|Q^*_\Omega\|_{\mathrm{asp}}$. If $e_{b_i} \le R$, span non-expansiveness gives $\|\tilde Q_j - Q^*_\Omega\|_{\mathrm{sp}} \le e_{b_i} \le R$, which is Lemma H.17.(i), hence $\|\tilde Q_j\|_{\mathrm{asp}} \le A_R$, i.e. $R[\mathcal{T}_\Omega^k(Q_{b_i}), \mathcal{T}_\Omega^k(Q^*_\Omega)] \le A_R$ for $k \in [\![0, m-1]\!]$ (Definition G.1.(ii)). Since γ_Ω is non-increasing, Lemma G.2.(iv) gives

$$\|\mathcal{T}_\Omega^m(Q_{b_i}) - \mathcal{T}_\Omega^m(Q^*_\Omega)\|_{\mathrm{sp}} \le \big(1 - \beta_m \gamma_\Omega(A_R)^{m-1}\big)\, e_{b_i},$$

which is Lemma H.17.(ii) with $\tilde\beta_m(A_R)$ in place of $\bar\beta(R)$; neither Lemma H.6 nor the probabilistic argument of Lemma H.7 is needed. Proposition H.3 and Theorem H.2 then apply verbatim with $\tilde\beta_m(A_R)$ in place of $\bar\beta(R)$, the geometric sum being bounded by $1/\tilde\beta_m(A_R)$ and the residual block adding one copy of $2m\bar U$. □

Theorem 2 (Epoch Ω-regularized Q-learning with $\alpha_j \equiv 1$)**.** Let Assumptions 1 to 3 and (P1) hold, and set $B(u) := \frac{8m^2\bar w^2 \iota}{(\tilde\beta u)^2}$ for $u > 0$. Run Algorithm 1 with $\alpha_j \equiv 1$ and a constant batch size $B \ge B(e_0)$, and let $J(T)$ be the number of epochs completed within T transitions. Then, with probability at least $1 - \delta$, $\sup_{j \le J(T)} \|Q_j - Q^*_\Omega\|_{\mathrm{sp}} \le 2e_0$, each epoch completed within the budget lasts at most $L := m\lceil(\sqrt{B} + \sqrt{\iota})^2/(\bar\gamma\tilde\beta\eta_{\min})\rceil + 1$ transitions, so that $J(T) \ge \lfloor T/L \rfloor$, and

$$\|Q_{J(T)} - Q^*_\Omega\|_{\mathrm{sp}} \le (1 - \tilde\beta)^{\lfloor J(T)/m \rfloor}\|Q_0 - Q^*_\Omega\|_{\mathrm{sp}} + \frac{2\sqrt{2}\, m\, \bar w}{\tilde\beta}\sqrt{\frac{\iota}{B}}. \tag{3}$$

In particular, for $\varepsilon \in (0, e_0)$, the choices $B = \lceil B(\varepsilon/2)\rceil$ and $T \ge mL\lceil\tilde\beta^{-1}\log(2e_0/\varepsilon)\rceil$ give $\|Q_{J(T)} - Q^*_\Omega\|_{\mathrm{sp}} \le \varepsilon$, i.e. $\tilde O\big(m^4\bar w^2/(\tilde\beta^4\bar\gamma\eta_{\min}\varepsilon^2)\big)$ sampled transitions.

Proof of Theorem 2. We apply Corollary H.4 with $R = 2e_0$. Then $A_R = \|Q^*_\Omega\|_{\mathrm{asp}} + 2e_0$, so that $\tilde\beta_m(A_R) = \tilde\beta$, $\bar w(R) = \bar w$ and $q_R = \bar\gamma\tilde\beta\eta_{\min}$; moreover $\iota(\delta') \le \iota_1 \le \iota$ since $J_{\max}(T) \le T$.

Admissibility. We have $2m\bar U \le \sqrt{2}\, m\bar w\sqrt{\iota/B}$ and $1 + \tilde\beta \le 2$, so (H.39) holds as soon as $\sqrt{2}\, m\bar w\sqrt{\iota/B} \le \tilde\beta e_0/2$, i.e. $B \ge B(e_0)$.

Bound. With probability at least $1 - \delta$, Corollary H.4 gives $\sup_{j \le J(T)} e_j \le 2e_0$, and its residual term is at most $\frac{2\cdot\sqrt{2}\, m\bar w\sqrt{\iota/B}}{\tilde\beta}$, which gives (3). On the same event, Proposition H.3 gives $\ell_j \le L_j(\delta') \le L$ for every $j < J(T)$, since $\iota(\delta') \le \iota$, so Lemma H.16.(iii) gives $J(T) \ge \lfloor T/L \rfloor$, and $I(T) = \lfloor J(T)/m \rfloor$.

Sample complexity. For $\varepsilon \in (0, e_0)$, the choice $B = \lceil B(\varepsilon/2)\rceil \ge B(e_0)$ makes the residual term at most $\varepsilon/2$. For $T \ge mL\lceil\tilde\beta^{-1}\log(2e_0/\varepsilon)\rceil$, which is possible since ι depends on T only logarithmically, we get $I(T) \ge \tilde\beta^{-1}\log(2e_0/\varepsilon)$, and $1 - u \le e^{-u}$ makes the contraction term at most $\varepsilon/2$. Since $L = \tilde\Theta\big(mB/(\bar\gamma\tilde\beta\eta_{\min})\big)$ and $B = \tilde\Theta\big(m^2\bar w^2/(\tilde\beta\varepsilon)^2\big)$, this gives $\tilde O\big(m^4\bar w^2/(\tilde\beta^3 \cdot \bar\gamma\tilde\beta\eta_{\min} \cdot \varepsilon^2)\big)$ sampled transitions. □

Corollary H.5 (Epoch Ω-regularized Q-learning for $m = 1$)**.** Let $m = 1$. Then $\tilde{\beta} = \beta_1$, so that in Theorem 2 the domination profile only enters through the coverage factor $\bar{\gamma}\beta_1\eta_{\min}$ of the epoch length L, and the sample complexity is $\tilde{O}\big(\bar{w}^2/(\beta_1^3 \cdot \bar{\gamma}\beta_1\eta_{\min} \cdot \varepsilon^2)\big)$.

Proof. For $m = 1$, $\tilde{\beta}_1(A_R) = \beta_1$ for every R (Definition H.2), so the contraction constants of the proof of Theorem 2 do not depend on the radius, whereas the coverage constant $q_R = \gamma_\Omega(A_R)\,\tilde{\beta}_1(A_R)\,\eta_{\min}$ of Lemma H.15 does. At $R = 2e_0$, it equals $\bar{\gamma}\beta_1\eta_{\min}$, and the claim is Theorem 2 with $m = 1$. □

Remark H.7 (Acting with μ or with the greedy policies)**.** The update of Algorithm 1 is off-policy: its target $r(s,a) + \Omega^*(Q_j)(s')$ is an expectation under the greedy policy and does not sample the next action, so the noise analysis only uses that the successor of a visited pair (s,a) is distributed as $p(\cdot|s,a)$ (Lemma H.18.(ii)), whatever the law of the action. Running the algorithm with actions $a_n \sim \mu(\cdot|s_n)$ is therefore covered by the same analysis, with γ_Ω replaced by 1 in Lemma H.15: q_R becomes $\beta_m\eta_{\min}$, L becomes $m\lceil(\sqrt{B} + \sqrt{\iota})^2/(\beta_m\eta_{\min})\rceil + 1$ in Theorem 2, and the sample complexity becomes $\tilde{O}\big(m^4\bar{w}^2/(\tilde{\beta}^3 \cdot \beta_m\eta_{\min} \cdot \varepsilon^2)\big)$, the factor $\tilde{\beta}^3$ of the contraction being unchanged. Acting with the greedy policies thus costs at most a factor $\bar{\gamma}^m$ in coverage.

Our criterion, uniform accuracy on all pairs of a tabular MDP, favors a fixed well-mixing behavior, and cannot credit what acting with the learned policy brings: the reward collected during learning, data concentrated where a good policy goes and, beyond the tabular setting, the stability of on-policy data distributions under function approximation (Sutton & Barto, 2018, §11.3). What the analysis does show is that regularization makes this choice safe: with a hard maximum, the greedy policy is deterministic and needs an added exploration mechanism, whereas under (P1) it explores by itself, at a quantified rate.

I Planning: Setting and Main Results

This section sets up the objects of Subsection 4.4 and states the planning results, which are proved in Appendix J. From a reset distribution $\xi \in \Delta(\mathcal{S})$, planning computes the regularized H-step solution bootstrapped at depth H by a boundary value $Q_B \in \mathbb{R}^{\mathcal{S}'\times\mathcal{A}}$, e.g. a value estimate or a heuristic, and uses its root level to approximate Q^*_Ω. It is learning localized in space, since only the states reachable from ξ are visited, and in time, since only the levels closest to the root matter.

Depth-dependent objects. Let $F(X) := \{s' \in \mathcal{S} \mid \exists (s,a) \in X \times \mathcal{A},\ p(s'|s,a) > 0\}$ for $X \subseteq \mathcal{S}$. The *reachable layers* are $\mathcal{S}_0 := \mathrm{supp}(\xi)$ and $\mathcal{S}_{h+1} := F(\mathcal{S}_h)$; we write $\mathcal{X}_h := \mathcal{S}_h \times \mathcal{A}$, $\mathcal{S}' := \bigcup_{h\geq 0}\mathcal{S}_h$ for the reachable set, which satisfies $F(\mathcal{S}') \subseteq \mathcal{S}'$, so that the MDP restricted to $\mathcal{S}'$ is well defined and $\mathcal{T}_\Omega$ acts on $\mathbb{R}^{\mathcal{S}'\times\mathcal{A}}$, and $\bar{S}_H := \max_{h\leq H}|\mathcal{S}_h|$. A *profile* is a family $(Q_h)_{h=0}^H$ of depth-dependent Q-functions $Q_h \in \mathbb{R}^{\mathcal{X}_h}$. Since the successors of $\mathcal{S}_h$ lie in $\mathcal{S}_{h+1}$, the *depth-dependent* operator $\mathcal{T}_{\Omega,h} : \mathbb{R}^{\mathcal{X}_{h+1}} \to \mathbb{R}^{\mathcal{X}_h}$, $\mathcal{T}_{\Omega,h}(Q)(s,a) := r(s,a) + \sum_{s'\in\mathcal{S}_{h+1}} p(s'|s,a)\,\Omega^*(Q)(s')$, is well defined, and the *planning solution* is the unique profile defined backward by $Q^*_{\Omega,H} := Q_B|_{\mathcal{X}_H}$ and $Q^*_{\Omega,h} := \mathcal{T}_{\Omega,h}(Q^*_{\Omega,h+1})$. It solves this triangular system exactly, whereas Q^*_Ω only solves $\mathcal{T}_\Omega(Q^*_\Omega) = Q^*_\Omega + \rho^*_\Omega\mathbf{1}$, so the planners need no centering projection.

Mixing and target. Throughout the planning appendices, Assumptions 2 to 4 and (P1) hold, Assumption 1 is not assumed, and μ, m, β_m, ν denote the objects of Assumption 4, with respect to which Assumption 3 and (P1) are understood; we write $\eta(s,a) := \nu(s)\mu(a|s)$ and $\eta_{\min} := \min_{\mathcal{S}_\infty\times\mathcal{A}}\eta$.

Remark I.1 (Reading Assumption 4)**.**

(i) *The core $\mathcal{S}_\infty$ is closed*: by (i), $\mathcal{S}_\infty = \mathcal{S}_{h_1+1} = F(\mathcal{S}_{h_1}) = F(\mathcal{S}_\infty)$. On the MDP restricted to $\mathcal{S}_\infty$, (ii) is Assumption 1 with a full-support ν, so Theorem G.1 gives $(Q^*_\Omega, \rho^*_\Omega)$ on $\mathcal{S}_\infty \times \mathcal{A}$, with Q^*_Ω unique up to an additive constant under (P1).

(ii) *Global case*: if Assumption 1 holds, then $K^m_\mu(s'|s,a) > 0$ for all (s,a,s'), so $\mathcal{S}_h = \mathcal{S}$ for $h \geq m$ and Assumption 4 holds with $\mathcal{S}_\infty = \mathcal{S}$ and $h_1 = m$. The spatial localization is then vacuous, and only the temporal one remains.

(iii) *Transient layers*: nothing is assumed on $\mathcal{S}_0, \dots, \mathcal{S}_{h_1-1}$, which enter the analysis only through $\bar{S}_H$ and the coverage of the shallow levels (Lemma J.9).

(iv) *When (i) holds*: $\mathcal{S}_h$ is the set of states reachable in exactly h steps from $\mathrm{supp}(\xi)$ in the union graph G (Definition D.1), so $(\mathcal{S}_h)_h$ is eventually periodic, and (i) holds for some finite h_1 as soon as every strongly connected component of G that contains a cycle and is reachable from $\mathrm{supp}(\xi)$ is aperiodic, e.g. contains a state s with $p(s|s,a) > 0$ for some a: closed walks in such a component have all large enough lengths, so a state reached at arbitrarily large depths is reached at every large enough depth.

By (i), Q^*_Ω is defined on $\mathcal{S}_\infty \times \mathcal{A}$; it extends uniquely to $\mathcal{S}' \times \mathcal{A}$ through the fixed-point equation (Lemma J.2), and the norms of Q^*_Ω are taken over $\mathcal{S}' \times \mathcal{A}$.

Errors and constants. We measure the quality of the boundary by $\Delta_B := \|Q_B - Q^*_\Omega\|_{\mathrm{sp},\mathcal{X}_H}$ and that of the initial profile by $D_0 := \max_{h<H} \|Q_{0,h} - Q^*_{\Omega,h}\|_{\mathrm{sp}}$, and use $\tilde{\beta}$, $\bar{w}$ and H_{eff} of Section 4 with $e_0 := D_0 + \Delta_B$. We assume $H \geq h_1 + m$. With $\eta^{\mathrm{res}}_h(s,a) := (\xi P^h_\mu)(s)\,\mu(a|s)$, the probability that a rollout started from ξ and following μ visits (s,a) at depth h, which is positive exactly on $\mathcal{X}_h$ (Lemma J.1), the coverage constants under μ and under a behavior dominating $\gamma_\Omega(A)\mu$ are

$$q_\mu := \min\Big\{ \min_{h<h_1+m} \min_{(s,a)\in\mathcal{X}_h} \eta^{\mathrm{res}}_h(s,a),\ \beta_m \eta_{\min} \Big\}, \qquad q_\pi(A) := \gamma_\Omega(A)^{h_1+m}\, q_\mu \quad (A \geq 0), \tag{I.1}$$

both positive, $q_\pi(A)$ by (P1).

Algorithms and results. With a model, the backward sweep $Q_h \leftarrow \mathcal{T}_{\Omega,h}(Q_{h+1})$ (Algorithm 4) returns the planning solution exactly, and Proposition J.1 bounds its horizon bias and its cost. Without a model, Algorithm 5 estimates each depth operator from rollouts started from ξ. Whereas the learning algorithms iterate the stationary update $Q_{j+1} = \mathcal{T}_\Omega(Q_j)$, it applies, up to sampling noise, the depth-shifted update $Q_{j+1,h} = \mathcal{T}_{\Omega,h}(Q_{j,h+1})$: level h at epoch $j+1$ is computed from level $h+1$ at epoch j, which is frozen during epoch j. Without noise, this recursion reaches the planning solution after $H - h$ epochs at level h (Lemma J.10), so the initialization is forgotten exactly rather than geometrically. As in Algorithm 1, a cap makes the number of samples per pair and per level deterministic.

Theorem I.1 (Epoch Ω-regularized Q-planning)**.** Let $q := q_\pi(\|Q^*_\Omega\|_{\mathrm{asp}} + 2e_0)$, or $q := q_\mu$ if the actions are drawn from μ instead of the greedy policies (Remark J.1). Let $J \geq H$, $\iota_H := \log(4JH\bar{S}_H|\mathcal{A}|/\delta)$, $M := h_1 + H_{\mathrm{eff}}$ and $B(u) := 2M^2\bar{w}^2\iota_H/u^2$. Run Algorithm 5 for J epochs with a batch size $B \geq B(e_0)$. Then, with probability at least $1-\delta$, each epoch lasts at most $H\lceil(\sqrt{B} + \sqrt{\iota_H})^2/q\rceil$ transitions and

$$\big\|Q_{J,0} - Q^*_\Omega\big\|_{\mathrm{sp},\mathcal{X}_0} \leq (1-\tilde{\beta})^{\lfloor (H-h_1)/m \rfloor}\Delta_B + 2M\bar{w}\sqrt{\frac{\iota_H}{2B}}. \tag{I.2}$$

In particular, for $\varepsilon \in (0, e_0)$, the choices $H \geq h_1 + m + H_{\mathrm{eff}}\log_+(2\Delta_B/\varepsilon)$, $J = H$ and $B = \lceil B(\varepsilon/2)\rceil$ give $\|Q_{J,0} - Q^*_\Omega\|_{\mathrm{sp},\mathcal{X}_0} \leq \varepsilon$, i.e. $\tilde{O}\big((h_1 + H_{\mathrm{eff}})^4\,\bar{w}^2/(q\,\varepsilon^2)\big)$ sampled transitions.

The bound (I.2) has the structure of (3), with the horizon bias in place of the initialization term. Finally, Algorithm 6 merges the H tables into a single table learned from resets, which removes the extra factor $1/\tilde{\beta}$ (Theorem J.2).

J Analysis of the Planning Algorithms

This section proves the results of Appendix I and analyses Algorithm 6, which merges the depth-indexed tables of the epoch planner into a single one.

The analysis rests on three facts. First, the planning solution solves its equations exactly (Appendix I), so the planners forget their initialization after H epochs. Second, a single lemma (Lemma J.8) bounds the error at depth h by the Bellman residuals of the deeper levels, damped at the rate of Lemma G.2.(iv) every m levels: it gives the planning depth and shows that noise does not accumulate across the horizon. Third, resets provide the exploration of the shallow levels, and the mixing of Assumption 4 that of the deep ones (Lemma J.9). The model-based sweep is treated in Appendix J.5, epoch planning in Appendix J.6 and epoch learning with resets in Appendix J.7.

J.1 Reachability and the extended target

Lemma J.1 (Reachability under μ)**.** For every $h \geq 0$ and $s \in \mathcal{S}$, $(\xi P_\mu^h)(s) > 0$ if and only if $s \in \mathcal{S}_h$.

Proof. Induction on h, the case $h = 0$ being the definition of $\mathcal{S}_0$. Since μ has full support, $P_\mu(s'|s) > 0$ if and only if $p(s'|s,a) > 0$ for some a, so $(\xi P_\mu^{h+1})(s') = \sum_s (\xi P_\mu^h)(s) P_\mu(s'|s) > 0$ if and only if $s' \in F(\mathcal{S}_h) = \mathcal{S}_{h+1}$. □

Lemma J.2 (Extension of the target to the transient states)**.** Let $\mathcal{V} := \mathcal{S}' \setminus \mathcal{S}_\infty$. The accessibility relation restricted to $\mathcal{V}$ has no directed cycle, and Q_Ω^* has a unique extension to $\mathcal{S}' \times \mathcal{A}$ such that

$$\mathcal{T}_\Omega(Q_\Omega^*) = Q_\Omega^* + \rho_\Omega^* \mathbf{1} \qquad \text{on } \mathcal{S}' \times \mathcal{A}. \tag{J.1}$$

Shifting Q_Ω^* by a constant on $\mathcal{S}_\infty$ shifts its extension by the same constant.

Proof. If $s \in \mathcal{V} \cap \mathcal{S}_j$ lay on a cycle of length $\ell \geq 1$ within $\mathcal{V}$, then $s \in \mathcal{S}_{j+\ell i}$ for every $i \geq 0$, hence $s \in \mathcal{S}_h = \mathcal{S}_\infty$ for some $h \geq h_1$, a contradiction. The subgraph induced on $\mathcal{V}$ is therefore acyclic, and its vertices admit a topological order in which every successor of $s \in \mathcal{V}$ lies either in $\mathcal{S}_\infty$ or later in the order. Defining, along the reverse of that order,

$$Q_\Omega^*(s,a) := r(s,a) - \rho_\Omega^* + \sum_{s'} p(s'|s,a)\, \Omega^*(Q_\Omega^*)(s'), \qquad s \in \mathcal{V},$$

each right-hand side only involves values already defined, which gives existence and uniqueness. The last claim follows from the translation equivariance of Ω^* (Lemma F.1.(iii)). □

We keep the notation Q_Ω^* for the extension. By (J.1) and Corollary F.1, $\mathcal{T}_\Omega^k(Q_\Omega^*) = Q_\Omega^* + k\rho_\Omega^* \mathbf{1}$ on $\mathcal{S}' \times \mathcal{A}$ for every $k \geq 0$; this is the only property of the extension used below, and it allows us to compare a profile to Q_Ω^* at the depths $h < h_1$.

J.2 Depth-dependent objects

Lemma J.3 (Depth-indexed average-path representation)**.** Let $Q, Q' \in \mathbb{R}^{\mathcal{X}_{h+1}}$ and let $\bar{\pi} := \bar{\pi}_{\Omega,[Q,Q']}$ be the average-path policy of Definition G.1.(i). Then

$$\mathcal{T}_{\Omega,h}(Q) - \mathcal{T}_{\Omega,h}(Q') = K_{\bar{\pi}}^{(h)}\,(Q - Q')\,, \qquad K_{\bar{\pi}}^{(h)}(s',a'|s,a) := p(s'|s,a)\,\bar{\pi}(a'|s'),$$

where $K_{\bar{\pi}}^{(h)}$, the restriction of the state-action kernel $K_{\bar{\pi}}$ to $\mathcal{X}_h \times \mathcal{X}_{h+1}$, is row-stochastic from $\mathcal{X}_h$ to $\mathcal{X}_{h+1}$, and $\bar{\pi} \geq \gamma_\Omega(R[Q,Q'])\,\mu$ on $\mathcal{X}_{h+1}$.

Proof. This is the computation of Lemma G.2.(i) restricted to $(s,a) \in \mathcal{X}_h$, all successors lying in $\mathcal{S}_{h+1}$; domination is Lemma G.2.(ii), the greedy policy at s' depending on Q only through $Q_{s'}$. □

For policies $\pi^{(h)}, \dots, \pi^{(h+k-1)}$, we write $K_{h:h+k} := K_{\pi^{(h)}}^{(h)} K_{\pi^{(h+1)}}^{(h+1)} \cdots K_{\pi^{(h+k-1)}}^{(h+k-1)}$, with $K_{h:h} := I$, for the product of depth kernels from $\mathcal{X}_h$ to $\mathcal{X}_{h+k}$ in the sense of Convention E.1; it is row-stochastic.

Lemma J.4 (The planning solution is a restriction)**.** For every $Q \in \mathbb{R}^{\mathcal{S}' \times \mathcal{A}}$ and $h \in [\![0, H]\!]$, $\big(\mathcal{T}_\Omega^{H-h} Q\big)\big|_{\mathcal{X}_h} = \mathcal{T}_{\Omega,h} \circ \cdots \circ \mathcal{T}_{\Omega,H-1}\big(Q|_{\mathcal{X}_H}\big)$. In particular, $Q_{\Omega,h}^* = \big(\mathcal{T}_\Omega^{H-h} Q_B\big)\big|_{\mathcal{X}_h}$ and $\big(\mathcal{T}_\Omega^{H-h} Q_\Omega^*\big)\big|_{\mathcal{X}_h} = Q_\Omega^*|_{\mathcal{X}_h} + (H-h)\rho_\Omega^* \mathbf{1}$.

Proof. Backward induction on h: at $(s,a) \in \mathcal{X}_h$, $\mathcal{T}_\Omega$ only involves the values of Q at the states $s' \in \mathcal{S}_{h+1}$, which coincide with the depth-$(h+1)$ restriction by the induction hypothesis. The last identity is (J.1) iterated. □

Constants. For a radius $R \geq 0$, we use the localized constants $A_R = \|Q^*_\Omega\|_{\mathrm{asp}} + R$, $\tilde\beta_m(A_R) = \beta_m \gamma_\Omega(A_R)^{m-1}$ and $\bar w(R) = \|Q^*_\Omega\|_{\mathrm{sp}} + R + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}}$ of Definition H.2; here R bounds the span distance to Q^*_Ω at every depth. With $e_0 = D_0 + \Delta_B$ as in Appendix I, the constants of the main text are $\tilde\beta = \tilde\beta_m(A_{2e_0})$ and $\bar w = \bar w(2e_0)$. We also write

$$H_{\mathrm{eff}}(A) := \frac{m}{\tilde\beta_m(A)}, \qquad M(h,A) := (h_1 - h)_+ + H_{\mathrm{eff}}(A), \qquad M(A) := M(0,A) = h_1 + H_{\mathrm{eff}}(A), \tag{J.2}$$

so that $H_{\mathrm{eff}} = H_{\mathrm{eff}}(A_{2e_0})$. By (P1), $\tilde\beta_m(A) > 0$ for every finite A.

Lemma J.5 (Distance between the planning solution and the target)**.** For every $h \in [\![0, H]\!]$, $\|Q^*_{\Omega,h} - Q^*_\Omega\|_{\mathrm{sp},\mathcal{X}_h} \leq \Delta_B$. Consequently $\|Q^*_{\Omega,h}\|_{\mathrm{asp}} \leq A_{\Delta_B}$ and $\|Q^*_{\Omega,h}\|_{\mathrm{sp}} \leq \|Q^*_\Omega\|_{\mathrm{sp}} + \Delta_B$, uniformly in h and H.

Proof. By Lemma J.4, $Q^*_{\Omega,l} = \mathcal{T}_{\Omega,l}(Q^*_{\Omega,l+1})$ and $V_l := Q^*_\Omega|_{\mathcal{X}_l} + (H-l)\rho^*_\Omega \mathbf{1}$ satisfies $V_l = \mathcal{T}_{\Omega,l}(V_{l+1})$ for every $l < H$, with $V_H = Q^*_\Omega|_{\mathcal{X}_H}$. Lemma J.3 gives $Q^*_{\Omega,l} - V_l = K^{(l)}_{\pi^{(l)}}(Q^*_{\Omega,l+1} - V_{l+1})$ with $\pi^{(l)} := \bar\pi_{\Omega,[Q^*_{\Omega,l+1}, Q^*_\Omega]}$, the greedy policies being invariant under constants. Unrolling from $l = h$ to $l = H$,

$$Q^*_{\Omega,h} - Q^*_\Omega|_{\mathcal{X}_h} - (H-h)\rho^*_\Omega \mathbf{1} = K_{h:H}\big(Q_B - Q^*_\Omega\big)\big|_{\mathcal{X}_H}.$$

$K_{h:H}$ is row-stochastic, hence non-expansive in span (Lemma E.6), and the span annihilates constants, whence the first claim. The other two follow from $\|\cdot\|_{\mathrm{asp}} \leq \|\cdot\|_{\mathrm{sp}}$, the invariance of $\|\cdot\|_{\mathrm{asp}}$ under constants and the triangle inequality. □

J.3 The master lemma

For $0 \leq h \leq h + k \leq H$, the *contracting block count*

$$n(h,k) := \left\lfloor \frac{\big(h + k - \max\{h, h_1\}\big)_+}{m} \right\rfloor \tag{J.3}$$

is the number of disjoint blocks of m consecutive depths inside $[\![\max\{h, h_1\}, h + k - 1]\!]$; when $h \geq h_1$, $n(h,k) = \lfloor k/m \rfloor$.

Lemma J.6 (Contraction of a product of depth kernels)**.** Let $A \geq 0$, $0 \leq h \leq h + k \leq H$, and let $\pi^{(h)}, \ldots, \pi^{(h+k-1)}$ be policies with $\pi^{(l)} \geq \gamma_\Omega(A)\,\mu$ for every l. Then $\|K_{h:h+k}\|_{\mathrm{sp,op}} \leq (1 - \tilde\beta_m(A))^{n(h,k)}$.

Proof. Group the factors of depths $l \geq \max\{h, h_1\}$ into $n(h,k)$ consecutive blocks of exactly m factors, leaving ungrouped the factors of depths $l < h_1$ and the fewer than m remaining ones. A block starting at depth $l \geq h_1$ is the product of m depth kernels from $\mathcal{X}_l = \mathcal{S}_\infty \times \mathcal{A}$ to $\mathcal{X}_{l+m} = \mathcal{S}_\infty \times \mathcal{A}$, whose policies dominate $\gamma_\Omega(A)\mu$; as $\mathcal{S}_\infty$ is closed, these kernels coincide with the unrestricted ones on $\mathcal{S}_\infty \times \mathcal{A}$. The proof of Lemma E.4 only uses the minorization $K^m_\mu(\cdot|s,a) \geq \beta_m \nu$ at the starting pairs, which Assumption 4.(ii) provides on $\mathcal{S}_\infty \times \mathcal{A}$, so it bounds the Dobrushin coefficient of each block by $1 - \beta_m \gamma_\Omega(A)^{m-1} = 1 - \tilde\beta_m(A)$, and Lemma E.6 bounds its operator span seminorm by the same quantity. The ungrouped factors are row-stochastic, hence of operator span seminorm at most 1, and we conclude by submultiplicativity (E.5). □

Lemma J.7 (Geometric sum)**.** For every $h \geq 0$ and $A \geq 0$, $\sum_{k \geq 0}(1 - \tilde\beta_m(A))^{n(h,k)} \leq M(h,A)$.

Proof. For $k < (h_1 - h)_+$, $n(h,k) = 0$, and these terms contribute $(h_1 - h)_+$. For $k \geq (h_1 - h)_+$, $n(h,k) = \lfloor (k - (h_1 - h)_+)/m \rfloor$ takes each value $q \geq 0$ exactly m times, and these terms contribute $m \sum_{q \geq 0}(1 - \tilde\beta_m(A))^q = H_{\mathrm{eff}}(A)$. □

The next lemma drives every planning bound. It compares a partial profile to the planning solution through its Bellman residuals: the model-based sweep has no residual, and the epoch planner is a profile whose residuals are the noise of successive epochs (Appendix J.6).

Lemma J.8 (Master lemma)**.** Let $0 \le h_0 \le h_e \le H$, let $Q_l \in \mathbb{R}^{\mathcal{X}_l}$ for $l \in [\![h_0, h_e]\!]$, with residuals

$$r_l := Q_l - \mathcal{T}_{\Omega,l}(Q_{l+1}), \qquad l \in [\![h_0, h_e - 1]\!], \tag{J.4}$$

and assume $\|Q_l\|_{\mathrm{asp}} \le A$ for $l \in [\![h_0 + 1, h_e]\!]$, with $A \ge A_{\Delta_B}$. Then, writing $\tilde{\beta} := \tilde{\beta}_m(A)$,

$$\begin{aligned} \big\|Q_{h_0} - Q^*_{\Omega,h_0}\big\|_{\mathrm{sp}} &\le (1-\tilde{\beta})^{n(h_0,h_e-h_0)} \big\|Q_{h_e} - Q^*_{\Omega,h_e}\big\|_{\mathrm{sp}} + \sum_{l=h_0}^{h_e-1} (1-\tilde{\beta})^{n(h_0,l-h_0)} \|r_l\|_{\mathrm{sp}} \\ &\le (1-\tilde{\beta})^{n(h_0,h_e-h_0)} \big\|Q_{h_e} - Q^*_{\Omega,h_e}\big\|_{\mathrm{sp}} + 2M(h_0,A) \max_{h_0 \le l < h_e} \|r_l\|_\infty. \end{aligned} \tag{J.5}$$

Proof. Let $e_l := Q_l - Q^*_{\Omega,l}$. By (J.4), the definition of the planning solution and Lemma J.3,

$$e_l = r_l + \mathcal{T}_{\Omega,l}(Q_{l+1}) - \mathcal{T}_{\Omega,l}(Q^*_{\Omega,l+1}) = r_l + K^{(l)}_{\pi^{(l)}} e_{l+1}, \qquad \pi^{(l)} := \bar{\pi}_{\Omega,[Q_{l+1},Q^*_{\Omega,l+1}]}.$$

Unrolling from h_0 to h_e gives $e_{h_0} = K_{h_0:h_e} e_{h_e} + \sum_{l=h_0}^{h_e-1} K_{h_0:l}\, r_l$. By Lemma J.5, $R[Q_{l+1}, Q^*_{\Omega,l+1}] \le \max\{A, A_{\Delta_B}\} = A$, so every $\pi^{(l)}$ dominates $\gamma_\Omega(A)\mu$ (Lemma J.3) and Lemma J.6 applies to each product. The triangle inequality gives the first bound, and $\|\cdot\|_{\mathrm{sp}} \le 2\|\cdot\|_\infty$ together with Lemma J.7 gives the second. □

Corollary J.1 (Horizon bias and planning depth)**.** For every $h \in [\![0, H]\!]$,

$$\big\|Q^*_{\Omega,h} - Q^*_\Omega\big\|_{\mathrm{sp},\mathcal{X}_h} \le \Big(1 - \tilde{\beta}_m(A_{\Delta_B})\Big)^{n(h,H-h)} \Delta_B, \tag{J.6}$$

and the right-hand side is at most ε as soon as

$$H \ge \max\{h, h_1\} + m + H_{\mathrm{eff}}(A_{\Delta_B}) \log_+\left(\frac{\Delta_B}{\varepsilon}\right). \tag{J.7}$$

Proof. By Lemma J.4, the profile $V_l := Q^*_\Omega|_{\mathcal{X}_l} + (H-l)\rho^*_\Omega \mathbf{1}$ satisfies $V_l = \mathcal{T}_{\Omega,l}(V_{l+1})$ for $l < H$, and $\|V_l\|_{\mathrm{asp}} = \|Q^*_\Omega\|_{\mathrm{asp}} \le A_{\Delta_B}$. Lemma J.8 with $h_0 = h$, $h_e = H$, zero residuals and $A = A_{\Delta_B}$ gives (J.6), since $\|V_H - Q^*_{\Omega,H}\|_{\mathrm{sp}} = \Delta_B$ and the span annihilates constants. Finally, $(1-\tilde{\beta})^n \le e^{-\tilde{\beta}n}$ and $n(h, H-h) \ge (H - \max\{h, h_1\})/m - 1$ give (J.7). □

Lemma J.8 has two consequences. First, noise does not accumulate across the horizon. Without contraction, the errors of the H levels of a backward induction add up. Here, the residual at depth l is damped by $(1-\tilde{\beta})^{n(h_0,l-h_0)}$ before it reaches depth h_0, so only the $M(h_0, A)$ levels closest to h_0 matter. Second, $H_{\mathrm{eff}}(A)$ plays the role of the discounted horizon $1/(1-\gamma)$ of discounted planning, but it is set by the regularizer through γ_Ω. The boundary value enters the required depth (J.7) only through $\log(\Delta_B/\varepsilon)$: a better bootstrap allows a shallower lookahead.

J.4 Coverage

Two mechanisms explore in planning: resets cover the shallow levels, and the mixing of Assumption 4.(ii) covers the deep ones. The following lemma gives the coverage constants (I.1).

Lemma J.9 (Level coverage)**.** Let $\mathcal{G}$ be a σ-algebra and let $(s_h, a_h)_{h<H}$ be a trajectory with $s_0 \sim \xi$, $a_h \sim \pi_h(\cdot|s_h)$ and $s_{h+1} \sim p(\cdot|s_h, a_h)$, where the policies π_h are $\mathcal{G}$-measurable and the trajectory is otherwise independent of $\mathcal{G}$. Let $E \in \mathcal{G}$ be an event on which $\pi_h \ge \gamma_h \mu$ for every h, with deterministic $\gamma_h \in [0,1]$, and write $\mathcal{F}_n := \sigma(\mathcal{G}, s_0, a_0, \dots, s_n, a_n)$. Then, almost surely on E, for every $h < H$ and $(s,a) \in \mathcal{X}_h$:

(i) if $h \ge h_1 + m$, $\mathbb{P}\big((s_h, a_h) = (s,a) \mid \mathcal{F}_{h-m}\big) \ge \big(\prod_{l=h-m+1}^{h} \gamma_l\big) \beta_m\, \eta(s,a)$;
(ii) if $h < h_1 + m$, $\mathbb{P}\big((s_h, a_h) = (s,a) \mid \mathcal{G}\big) \ge \big(\prod_{l=0}^{h} \gamma_l\big) \eta^{\mathrm{res}}_h(s,a)$.

In particular, almost surely on E, $\mathbb{P}((s_h, a_h) = (s,a) \mid \mathcal{G}) \ge q_\mu$ if $\gamma_h \equiv 1$, and $\ge q_\pi(A)$ if $\gamma_h \equiv \gamma_\Omega(A)$.

Proof. (i). Since $h - m \geq h_1$, $s_{h-m} \in \mathcal{S}_{h-m} = \mathcal{S}_\infty$, so $K_\mu^m(\cdot|s_{h-m}, a_{h-m}) \geq \beta_m \nu$ by Assumption 4.(ii). The proof of Lemma G.2.(v) only uses this minorization at the conditioning pair and the domination of the m following policies, and applies verbatim.

(ii). Conditionally on $\mathcal{G}$, (s_h) is a Markov chain with kernels P_{π_l}, and the transport inequality $P_{\pi_l} \geq \gamma_l P_\mu$ of the proof of Lemma G.2.(iii) holds entrywise on E. Hence $\mathbb{P}(s_h = s \mid \mathcal{G}) \geq \big(\prod_{l<h} \gamma_l\big)(\xi P_\mu^h)(s)$, and $\pi_h(a|s) \geq \gamma_h \mu(a|s)$ concludes.

For the last claim, integrate (i) and use $\gamma_\Omega(A)^{h+1} \geq \gamma_\Omega(A)^{h_1+m}$ in (ii) and $\gamma_\Omega(A)^m \geq \gamma_\Omega(A)^{h_1+m}$ in (i). □

Remark J.1 (Which behavior to use)**.** The sample $r(s,a)+\Omega^*(Q)(s')$ is unbiased for $\mathcal{T}_{\Omega,h}(Q)(s,a)$ whatever the law of the action a, so the behavior policy only affects which pairs are visited. The factor γ_Ω^{m-1} of the contraction rate comes from the average-path policy of Lemma J.3, internal to the operator, whereas the factor $\gamma_\Omega(A)^{h_1+m}$ of $q_\pi(A)$ is the price of exploring with the greedy policies, and disappears under the behavior μ, exactly as on a single trajectory (Remark H.7). The shallow levels may however be covered slowly, since $(\xi P_\mu^h)(s)$ is small for states reachable only through a narrow path; if ξ and p are bounded below by $\xi_{\min}$ and $p_{\min}$ on their supports, then $\eta_h^{\mathrm{res}} \geq \xi_{\min}(\mu_{\min} p_{\min})^h \mu_{\min}$ on $\mathcal{X}_h$.

J.5 Model-based planning

Proposition J.1 (Horizon bias and model-based planning)**.** For every $h \leq H$,

$$\big\|Q^*_{\Omega,h} - Q^*_\Omega\big\|_{\mathrm{sp},\mathcal{X}_h} \leq (1-\tilde\beta)^{\lfloor (H-\max\{h,h_1\})/m \rfloor}\, \Delta_B. \tag{J.8}$$

In particular, Algorithm 4 returns $(Q^*_{\Omega,h})_{h=0}^H$ exactly after $\sum_{h<H} |\mathcal{S}_h||\mathcal{A}||\mathcal{S}_{h+1}| \leq H \bar S_H^2 |\mathcal{A}|$ model queries, one query being an evaluation of $p(s'|s,a)$, and $Q^*_{\Omega,0}$ is ε-accurate for Q^*_Ω on $\mathcal{X}_0$ as soon as $H \geq h_1 + m + H_{\mathrm{eff}} \log_+(\Delta_B/\varepsilon)$.

Proof. With a model, $D_0 = 0$, so $e_0 = \Delta_B$ and $\tilde\beta = \tilde\beta_m(A_{2\Delta_B}) \leq \tilde\beta_m(A_{\Delta_B})$, γ_Ω being non-increasing. Since $n(h, H-h) = \lfloor (H - \max\{h,h_1\})/m \rfloor$, (J.8) follows from Corollary J.1. The sweep computes $Q_h = \mathcal{T}_{\Omega,h}(Q_{h+1})$ from $Q_H = Q_B|_{\mathcal{X}_H}$, which is the definition of the planning solution, and each value $\mathcal{T}_{\Omega,h}(Q_{h+1})(s,a)$ uses the $|\mathcal{S}_{h+1}|$ probabilities $p(s'|s,a)$, $s' \in \mathcal{S}_{h+1}$, whence the count. Finally, $H_{\mathrm{eff}} \geq H_{\mathrm{eff}}(A_{\Delta_B})$, so the condition on H implies (J.7) at $h = 0$. □

Compared with Corollary 1, model-based planning is Q-iteration restricted to the reachable set, with the boundary mismatch Δ_B in place of the initial error: it saves a factor $(\bar S_H/|\mathcal{S}|)^2$ in model queries, which is non-trivial only when $\mathcal{S}_\infty \subsetneq \mathcal{S}$ (Remark I.1.(ii)).

The same exactness holds for the depth-shifted recursion $Q_{j+1,h} = \mathcal{T}_{\Omega,h}(Q_{j,h+1})$, which is the noiseless version of epoch planning (Algorithm 5): each level is computed from the previous iterate of the level below, and the initialization is forgotten after exactly H iterations.

Lemma J.10 (Exact forgetting of the initialization)**.** Let $(Q_{j,h})_{j\geq 0,\, h\leq H}$ satisfy $Q_{j+1,h} = \mathcal{T}_{\Omega,h}(Q_{j,h+1})$ for $h < H$ and $Q_{j,H} = Q_B|_{\mathcal{X}_H}$ for every j. Then $Q_{j,h} = Q^*_{\Omega,h}$ for every $h \leq H$ and $j \geq H - h$.

Proof. Backward induction on h. The case $h = H$ holds by assumption. If $Q_{j,h+1} = Q^*_{\Omega,h+1}$ for every $j \geq H - h - 1$, then $Q_{j+1,h} = \mathcal{T}_{\Omega,h}(Q^*_{\Omega,h+1}) = Q^*_{\Omega,h}$ for every such j, i.e. $Q_{j,h} = Q^*_{\Omega,h}$ for every $j \geq H - h$. □

J.6 Epoch planning

Algorithm 5 is the depth-indexed analogue of Corollary H.4: its step size is $\alpha_j \equiv 1$, which makes the recursion over epochs nilpotent (Lemma J.10), and its batch sizes B_j are deterministic. Let $\mathcal{G}_j$ be the σ-algebra generated by Q_0 and the rollouts of epochs $0, \ldots, j-1$; the analysis covers Algorithm 5 as written, with actions drawn from $\pi_{\Omega,Q_{j,h}}$ at level h, and its variant with actions drawn from μ

(Remark J.1); in both cases the behavior of epoch j is $\mathcal{G}_j$-measurable. As in Lemma H.14, on the event that epoch j terminates, for $h < H$,

$$Q_{j+1,h} = \mathcal{T}_{\Omega,h}(Q_{j,h+1}) + w_{j,h},$$
$$w_{j,h}(s,a) := \frac{1}{B_j}\sum_{l=1}^{B_j}\Big(\Omega^*(Q_{j,h+1})(s'_l) - \mathbb{E}_{s'\sim p(\cdot|s,a)}\big[\Omega^*(Q_{j,h+1})(s')\big]\Big), \tag{J.9}$$

where s'_l is the successor of the l-th level-h visit of (s,a) during epoch j, and $Q_{j,H} = Q_B|_{\mathcal{X}_H}$. We set $Q_{j+1,\cdot} := Q_{j,\cdot}$ if epoch j does not terminate, which happens with probability zero (Proposition J.2).

Remark J.2 (The outer loop is sample splitting)**.** Level h at epoch $j+1$ is computed from level $h+1$ at epoch j, which is $\mathcal{G}_j$-measurable, hence independent of the data of epoch j. A single backward sweep over stored transitions would instead evaluate $\Omega^*(\widehat{Q}_{h+1})$ at the level-$(h+1)$ states of the very rollouts on which $\widehat{Q}_{h+1}$ depends, and lose the martingale structure of the noise. Restoring independence would require H disjoint groups of rollouts, one per level, which costs the same factor H as the H epochs of Lemma J.10.

Proposition J.2 (Well-posedness)**.** Every epoch terminates almost surely, and every $Q_{j,h}$ is a finite, $\mathcal{G}_j$-measurable random vector.

Proof. Finiteness and measurability follow by induction on j as in Lemma H.14.(ii), each coordinate of $w_{j,h}$ being bounded by $\|Q_{j,h+1}\|_{\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}}$ (Lemma F.1.(vii)). For termination, fix j and a deterministic $u \geq 0$. Conditionally on $\mathcal{G}_j$, the rollouts of epoch j are i.i.d. and, on $\{\max_{h<H}\|Q_{j,h}\|_{\mathrm{asp}} \leq u\} \in \mathcal{G}_j$, each of them visits a given (h,x), $x \in \mathcal{X}_h$, with probability at least $\min\{q_\mu, q_\pi(u)\} > 0$ by Lemma J.9. Each pair is thus visited B times after finitely many rollouts, almost surely on that event, and letting $u \uparrow \infty$ concludes. □

For $R \geq e_0$, let

$$\sigma_R := \inf\Big\{j \geq 0 \;\Big|\; \max_{h<H}\|Q_{j,h} - Q^*_\Omega\|_{\mathrm{sp},\mathcal{X}_h} > R\Big\},$$

a $(\mathcal{G}_j)$-stopping time with $\sigma_R \geq 1$, since $\|Q_{0,h} - Q^*_\Omega\|_{\mathrm{sp},\mathcal{X}_h} \leq D_0 + \Delta_B = e_0$ by Lemma J.5, and write $\widehat{w}_{j,h} := \mathbb{1}_{\{j<\sigma_R\}}w_{j,h}$. On $\{j < \sigma_R\}$, $\|Q_{j,h}\|_{\mathrm{asp}} \leq A_R$ and $\|\Omega^*(Q_{j,h})\|_{\mathcal{S},\mathrm{sp}} \leq \|Q_{j,h}\|_{\mathrm{sp}} + \|\Omega^*(0)\|_{\mathcal{S},\mathrm{sp}} \leq \bar{w}(R)$ for every $h \leq H$, the level H being covered by $\Delta_B \leq e_0 \leq R$. We write $q := q_\pi(A_R)$, or $q := q_\mu$ for the variant with actions drawn from μ, and, for $J \geq 1$ and $\delta \in (0,1)$, $\iota_H := \log(4JH\bar{S}_H|\mathcal{A}|/\delta)$.

Lemma J.11 (Noise)**.** With probability at least $1 - \delta/2$, $\|\widehat{w}_{j,h}\|_\infty \leq \bar{u}_j(R) := \bar{w}(R)\sqrt{\iota_H/(2B_j)}$ for all $j < J$ and $h < H$.

Proof. Fix (j,h,x) with $x \in \mathcal{X}_h$, and order the B terms of $\widehat{w}_{j,h}(x)$ by the successive level-h visits of x during epoch j, the rollouts being taken in chronological order. Given $\mathcal{G}_j$ and the past of these visits, the successor of a visit is distributed as $p(\cdot|x)$, while $\Omega^*(Q_{j,h+1})$ is $\mathcal{G}_j$-measurable. The terms are therefore martingale differences, lying in $\mathcal{G}_j$-measurable intervals of length at most $\bar{w}(R)/B_j$ on $\{j < \sigma_R\}$ and vanishing off it, and their number is B_j by the cap (Lemma H.14). Lemma E.2 with $u = \bar{u}_j(R)$ gives a failure probability $2e^{-\iota_H}$, and a union bound over the at most $JH\bar{S}_H|\mathcal{A}|$ triples concludes. □

Lemma J.12 (Epoch length)**.** With probability at least $1 - \delta/4$, every epoch $j < J$ with $j < \sigma_R$ lasts at most N_jH transitions, where $N_j := \lceil(\sqrt{B_j} + \sqrt{\iota_H})^2/q\rceil$.

Proof. Conditionally on $\mathcal{G}_j$, the rollouts of epoch j, extended beyond its end if needed, are i.i.d., and on $\{j < \sigma_R\} \in \mathcal{G}_j$, Lemma J.9 with $\mathcal{G} = \mathcal{G}_j$ shows that each of them visits a given (h,x), $x \in \mathcal{X}_h$, with probability at least q. As in the proof of Lemma H.15, the exact Chernoff bound (E.1) with $b = 1$, $a = N_jq$ and $c = B_j$ bounds by $e^{-\iota_H}$ the conditional probability that fewer than B_j of the first N_j rollouts visit (h,x), and a union bound over the at most $JH\bar{S}_H|\mathcal{A}|$ triples concludes. Independence across resets replaces the m-step sub-sampling of Lemma H.15. □

Theorem J.1 (Finite-time bound for epoch planning)**.** Let $R \geq e_0$ satisfy the *radius admissibility condition*

$$e_0 + 2M(A_R) \max_{j<J} \bar{u}_j(R) \leq R. \tag{J.10}$$

Then, with probability at least $1-\delta$, every epoch $j < J$ lasts at most $N_j H$ transitions, $\max_{h<H} \|Q_{j,h} - Q^*_\Omega\|_{\mathrm{sp},\mathcal{X}_h} \leq R$ for every $j \leq J$, and, for every $j \leq J$ and $h < H$,

$$\begin{aligned}\left\|Q_{j,h} - Q^*_{\Omega,h}\right\|_{\mathrm{sp}} \leq\ & \mathbb{1}_{\{j<H-h\}} \left(1-\tilde{\beta}_m(A_R)\right)^{n(h,j)} D_0 \\ & + 2 \sum_{k=0}^{\min\{j,H-h\}-1} \left(1-\tilde{\beta}_m(A_R)\right)^{n(h,k)} \bar{u}_{j-1-k}(R).\end{aligned} \tag{J.11}$$

In particular, the initialization has vanished exactly for $j \geq H-h$.

Proof. Work on the intersection of the events of Lemmas J.11 and J.12, of probability at least $1-\delta$, on which all epochs terminate (Proposition J.2). We prove by induction on $j \leq J$ that $\sigma_R > j$ and that (J.11) holds at rank j; the length bound then follows from Lemma J.12. At $j = 0$, both claims were checked above. Assume $\sigma_R > j-1$, fix $h < H$, let $\ell := \min\{j, H-h\}$ and consider the *diagonal* profile $P_l := Q_{j-(l-h),l}$, $l \in [\![h, h+\ell]\!]$. For $l < h+\ell$, the epoch $i := j-(l-h)-1$ satisfies $0 \leq i < j \leq \sigma_R$, so (J.9) gives $P_l = \mathcal{T}_{\Omega,l}(P_{l+1}) + \widehat{w}_{i,l}$: the residuals of the diagonal are the stopped noises, and Lemma J.11 and $\|\cdot\|_{\mathrm{sp}} \leq 2\|\cdot\|_\infty$ bound their span by $2\bar{u}_i(R)$. For $l > h$, P_l is either computed at epoch $j-(l-h) < j \leq \sigma_R$ or equal to $Q_B|_{\mathcal{X}_H}$, whose action span is at most A_{Δ_B} by Lemma J.5, so $\|P_l\|_{\mathrm{asp}} \leq A_R$, and $A_R \geq A_{\Delta_B}$. The last element is $P_{h+\ell} = Q_B|_{\mathcal{X}_H} = Q^*_{\Omega,H}$ if $\ell = H-h$, and $P_{h+\ell} = Q_{0,h+j}$, within D_0 of $Q^*_{\Omega,h+j}$, if $\ell = j < H-h$. The first inequality of Lemma J.8 with $h_0 = h$ and $h_e = h+\ell$ gives (J.11) at rank j, since $n(h,\ell) = n(h,j)$ when $\ell = j$. Finally, Lemma J.7 and $M(h, A_R) \leq M(A_R)$ bound the sum in (J.11) by $2M(A_R)\max_{i<J}\bar{u}_i(R)$, so Lemma J.5 and (J.10) give $\|Q_{j,h} - Q^*_\Omega\|_{\mathrm{sp},\mathcal{X}_h} \leq D_0 + 2M(A_R)\max_{i<J}\bar{u}_i(R) + \Delta_B \leq R$ for every $h < H$, i.e. $\sigma_R > j$. □

Unlike (H.14) and (H.35), condition (J.10) has no factor $\bar{\beta}(R)/(\bar{\beta}(R)+1)$: the profile at epoch j is an explicit function of the data of epochs $i < j$ along finite diagonals, and the initialization is forgotten exactly, not geometrically.

Proof of Theorem I.1. We apply Theorem J.1 with $B_j \equiv B$ and $R = 2e_0$, for which (J.11) and Lemma J.7 give $2M(h, A_R)\bar{u}(R)$. Then $\tilde{\beta}_m(A_R) = \tilde{\beta}$, $\bar{w}(R) = \bar{w}$, $M(A_R) = M$ and $q_\pi(A_R)$ is the constant q of the greedy behavior, so that $2M\bar{u}(R) = 2M\bar{w}\sqrt{\iota_H/(2B)}$, which is at most e_0 if and only if $B \geq B(e_0)$: condition (J.10) holds. With probability at least $1-\delta$, every epoch lasts at most $NH = H\lceil(\sqrt{B}+\sqrt{\iota_H})^2/q\rceil$ transitions and, since $J \geq H$, (J.11) at $h = 0$ gives $\|Q_{J,0} - Q^*_{\Omega,0}\|_{\mathrm{sp}} \leq 2M\bar{w}\sqrt{\iota_H/(2B)}$. Corollary J.1 at $h = 0$, with $n(0,H) = \lfloor (H-h_1)/m \rfloor$ and $\tilde{\beta} \leq \tilde{\beta}_m(A_{\Delta_B})$, bounds $\|Q^*_{\Omega,0} - Q^*_\Omega\|_{\mathrm{sp},\mathcal{X}_0}$ by $(1-\tilde{\beta})^{\lfloor (H-h_1)/m \rfloor}\Delta_B$, and the triangle inequality gives (I.2).

For $\varepsilon \in (0, e_0)$, the choice $B = \lceil B(\varepsilon/2)\rceil \geq B(e_0)$ makes the noise term at most $\varepsilon/2$, and the choice of H gives $\lfloor (H-h_1)/m \rfloor \geq \tilde{\beta}^{-1}\log(2\Delta_B/\varepsilon)$, so that $1-u \leq e^{-u}$ makes the bias term at most $\varepsilon/2$. The $J = H$ epochs use at most $H \cdot NH$ transitions, with $N = \tilde{\Theta}(B/q)$, $B = \tilde{\Theta}(M^2\bar{w}^2/\varepsilon^2)$ and $H = \tilde{\Theta}(h_1 + H_{\mathrm{eff}})$, which gives $\tilde{O}\big((h_1+H_{\mathrm{eff}})^4\bar{w}^2/(q\varepsilon^2)\big)$; the dependence of ι_H on $J = H$ is logarithmic. □

Remark J.3 (Planning and learning)**.** For $h_1 = 0$, Theorem I.1 costs $\tilde{O}(m^4\bar{w}^2/(\tilde{\beta}^4 q\varepsilon^2))$ transitions, against $\tilde{O}(m^4\bar{w}^2/(\tilde{\beta}^3 q_R\varepsilon^2))$ for Theorem 2: at equal coverage constant, epoch planning is a factor $1/\tilde{\beta}$ more expensive. Propagation is paid twice, once in depth, since a rollout gives one sample per level to H distinct tables, and once in epochs, since $J = H$ iterations are needed to propagate the boundary to the root. Per table, planning is m times more efficient, the independence across resets removing the sub-sampling of Lemma H.15. Merging the tables removes the extra factor (Appendix J.7).

J.7 MERGING THE TABLES: EPOCH Q-LEARNING WITH RESETS

The factor $1/\tilde{\beta}$ of Remark J.3 is the price of maintaining H depth-indexed tables. By Corollary J.1, the planning solution is within ε of Q^*_Ω, up to constants, at every depth $h_1 \le h \le H - m - H_{\text{eff}}(A_{\Delta_B}) \log_+(\Delta_B/\varepsilon)$, so these tables are redundant on most of the horizon. Merging them gives Algorithm 6, which is Algorithm 1 run on a reset process: the transient prefix of each rollout is played with μ and discarded, and the transitions retained at depths $h \ge h_1$ are transitions of the MDP restricted to $\mathcal{S}_\infty$, since $s_h, s_{h+1} \in \mathcal{S}_\infty$. This MDP satisfies Assumption 1 (Remark I.1.(i)). We keep the notation of Appendix H.3, every constant being computed on $\mathcal{S}_\infty$: $e_0 = \|Q_0 - Q^*_\Omega\|_{\text{sp}}$, the localized constants of Definition H.2, the exit time (H.21), the blocks $\mathcal{B}_i$ and the σ-algebras $\mathcal{G}_j$. Only the coverage constant and the epoch length change.

Lemma J.13 (Cover time with resets)**.** Assume $H \ge h_1 + m + 1$ and let $K_0 := \lfloor (H - 1 - h_1)/m \rfloor \ge 1$. For $j \ge 0$, a deterministic radius $u \ge 0$ and $\delta' \in (0,1)$, let $\mathcal{D}_j(u) := \{\|Q_j\|_{\text{asp}} \le u\}$, let $q^{\text{res}}(u) := \beta_m \gamma_\Omega(u)^m \eta_{\min}$, or $q^{\text{res}}(u) := \beta_m \eta_{\min}$ if the actions at depths $h \ge h_1$ are drawn from μ instead of π_{Ω,Q_j}, let $\iota(\delta') := \log(|\mathcal{S}_\infty||\mathcal{A}|/\delta')$ and

$$L_j^{\text{res}}(u, \delta') := H \left\lceil \frac{1}{K_0} \left\lceil \frac{\left(\sqrt{B_j} + \sqrt{\iota(\delta')}\right)^2}{q^{\text{res}}(u)} \right\rceil \right\rceil. \tag{J.12}$$

Then, almost surely, $\mathbb{P}\big(\{\ell_j > L_j^{\text{res}}(u,\delta')\} \cap \mathcal{D}_j(u) \mid \mathcal{G}_j\big) \le \delta' \, \mathbb{1}_{\mathcal{D}_j(u)}$. If moreover $H \ge 2(h_1 + m + 1)$, then $L_j^{\text{res}}(u,\delta') \le 2m\lceil (\sqrt{B_j} + \sqrt{\iota(\delta')})^2 / q^{\text{res}}(u) \rceil + H$.

Proof. Conditionally on $\mathcal{G}_j$, the rollouts of epoch j, extended beyond its end if needed, are i.i.d. Consider the instants of depths $h_1 + km$, $k \in [\![1, K_0]\!]$, of the successive rollouts, in chronological order, fix $(s,a) \in \mathcal{S}_\infty \times \mathcal{A}$, and let Z_l indicate that the l-th instant visits (s,a). On $\mathcal{D}_j(u) \in \mathcal{G}_j$, the actions at depths $h \ge h_1$ are drawn from μ or from a policy dominating $\gamma_\Omega(u)\mu$, and the m steps before each instant start at a depth $h \ge h_1$, hence in $\mathcal{S}_\infty$. Lemma J.9.(i) with $\mathcal{G} = \mathcal{G}_j$, conditional on the previous instant of the same rollout or integrated over a fresh reset, therefore gives $\mathbb{E}[Z_l \mid Z_1, \dots, Z_{l-1}, \mathcal{G}_j] \ge q^{\text{res}}(u)$. The rest is the proof of Lemma H.15: the exact Chernoff bound (E.1) with $b = 1$, $a = K q^{\text{res}}(u)$ and $c = B_j$, where $K := \lceil (\sqrt{B_j} + \sqrt{\iota(\delta')})^2 / q^{\text{res}}(u) \rceil$, and a union bound over $\mathcal{S}_\infty \times \mathcal{A}$ show that, with conditional probability at least $1 - \delta'$, every pair is visited B_j times within K instants, which $\lceil K/K_0 \rceil$ rollouts of H transitions provide. Finally, $K_0 \ge (H - 1 - h_1 - m)/m \ge H/(2m)$ when $H \ge 2(h_1 + m + 1)$. □

Theorem J.2 (Epoch Q-learning with resets)**.** Let $H \ge h_1 + m + 1$. Theorem H.2, under Assumption H.1, and Corollary H.4 hold for Algorithm 6, with every constant computed on the MDP restricted to $\mathcal{S}_\infty$ and the epoch length $L_j(\delta')$ of Lemma H.15 replaced by $L_j^{\text{res}}(A_R, \delta')$. In particular, Theorem 2 holds with $L := H\lceil \frac{1}{K_0} \lceil (\sqrt{B} + \sqrt{\iota})^2 / q^{\text{res}} \rceil \rceil$, where $q^{\text{res}} := \bar{\gamma}\tilde{\beta}\eta_{\min}$, or $q^{\text{res}} := \beta_m \eta_{\min}$ if the actions at depths $h \ge h_1$ are drawn from μ. For $H \ge 2(h_1 + m + 1)$, this gives the sample complexity

$$\tilde{O}\left(\frac{m^4 \, \bar{w}^2}{\tilde{\beta}^3 \, q^{\text{res}} \, \varepsilon^2} + \frac{mH}{\tilde{\beta}} \right), \tag{J.13}$$

whose first term is $\tilde{O}\big(m^4 \bar{w}^2 / (\tilde{\beta}^3 \cdot \beta_m \eta_{\min} \cdot \varepsilon^2)\big)$ under the behavior μ and $\tilde{O}\big(m^4 \bar{w}^2 / (\tilde{\beta}^3 \cdot \bar{\gamma}\tilde{\beta}\eta_{\min} \cdot \varepsilon^2)\big)$ under the greedy behavior.

Proof. The proofs of Appendix H.3 use the trajectory through three properties only: the behavior during epoch j is $\mathcal{G}_j$-measurable; given the past, the successor of a retained visit of (s,a) is distributed as $p(\cdot|s,a)$ (Lemma H.18.(ii)); and the cap makes the number of retained samples deterministic (Lemma H.14). The reset process satisfies all three, by the Markov property within each rollout and the independence of the rollouts given $\mathcal{G}_j$. The virtual flow, tracking and no-escape arguments (Lemmas H.17 and H.21, Proposition H.3) only concern the update (H.19), and the trajectory structure only enters through Lemma H.15, which Lemma J.13 replaces, and Proposition H.2, whose proof applies verbatim with it. The count follows as in the proof of Theorem 2, with $L \le 2m\lceil (\sqrt{B} + \sqrt{\iota})^2 / q^{\text{res}} \rceil + H$ by Lemma J.13. □

Remark J.4 (What the merge buys and what it loses)**.**

(i) *What the resets bring.* Relative to Theorem 2, (J.13) takes $\eta_{\min}$ over $\mathcal{S}_\infty \times \mathcal{A}$, with the minorizing measure of Assumption 4, which can be much larger when $\mathcal{S}_\infty \subsetneq \mathcal{S}$. The factor γ_Ω^m removed under the behavior μ is not due to the resets: it is removed on a single trajectory as well (Remark H.7). Relative to epoch planning, merging the tables saves the factor $1/\tilde{\beta}$ of Remark J.3, and the rollout length only enters through $H \geq 2(h_1 + m + 1)$ and the second term of (J.13).

(ii) *The boundary becomes the initialization.* There is no horizon bias: the target is Q_Ω^* itself, and the boundary value Q_B of the planners plays the role of Q_0. The horizon bias $(1-\tilde{\beta})^{\lfloor (H-h_1)/m \rfloor}\Delta_B$ of Theorem I.1 and the initialization term $(1-\tilde{\beta})^{\lfloor J(T)/m \rfloor}e_0$ of Theorem 2 are the same term under two names.

(iii) *The exact fixed point is lost.* The single table only solves $\mathcal{T}_\Omega(Q_\Omega^*) = Q_\Omega^* + \rho_\Omega^* \mathbf{1}$, which is why Algorithm 6 needs the centering projection Π, and the depth profile of Theorem J.1 has no counterpart.

(iv) *A single parameterized algorithm.* Without resets, Algorithm 6 is Algorithm 1, and Theorems 2 and J.2 are one statement in which (H, ξ) only enters through the coverage constant and the epoch length.

K THE α-DIVERGENCE REGULARIZERS

Table 1 summarizes Theorem K.1 and Corollary K.1 for $\Omega_{\alpha,s} := D_\alpha(\cdot\|\mu_s)$. In it, $Q_a := Q(s,a)$, $\mu_a := \mu(a|s)$, $\exp_q(x) := [1 + (1-q)x]_+^{1/(1-q)}$ with $\exp_1 = \exp$, and ζ_s is the normalizer making $\pi_{\Omega,Q}(\cdot|s)$ sum to one; strong convexity is with respect to $\|\cdot\|_1$ and fails for $\alpha > 2$, and the profile bounds are tight as $\mu_{\min} \to 0$ (Remark K.1).

On the simplex $D_\alpha(p\|q) = D_{1-\alpha}(q\|p)$ (Cichocki & Amari, 2010), so divergences to μ in either argument belong to the family. For a general reference μ_s, the greedy map of $\Omega_{\alpha,s}$ is the f-softargmax of Roulet et al. (2025) with reference measure μ_s, and Belousov & Peters (2019) give its closed form in MDPs. For μ uniform, $\Omega_{\alpha,s}$ is the Tsallis negentropy of index α up to the factor $|\mathcal{A}|^{\alpha-1}$ on τ and an additive constant, and for $\alpha \geq 1$ its greedy map is the α-entmax of Peters et al. (2019). Tsallis-KL regularization toward a reference policy, which is this family up to a reparametrization of the index and a rescaling of τ, is used in reinforcement learning by Zhu et al. (2023), and Labbi et al. (2026) use the same divergences for policy gradient. Parts (i)–(iii) of Theorem K.1 collect known facts in our notation; our contribution is the domination profile (iv) and the regularization modulus (v). Since $\Omega_{\alpha,s} \geq 0$ vanishes at μ_s, Lemma F.1.(ii) gives $g_Q \leq 0$, while $g_Q \geq -\Lambda_{\Omega_\alpha}^{\mathrm{pt}}(R)$ for $\|Q\|_{\mathrm{asp}} \leq R$ by definition; hence $\Lambda_{\Omega_\alpha} \leq \Lambda_{\Omega_\alpha}^{\mathrm{pt}}$, and we bound the latter.

Theorem K.1 (The α-divergence regularizer)**.** Let $\alpha \in \mathbb{R}$, $\tau > 0$ and μ have full support.

(i) *Domain.* $\mathrm{dom}(\Omega_{\alpha,s}) = \Delta(\mathcal{A})$ for $\alpha > 0$ and $\{\pi_s > 0\}$ for $\alpha \leq 0$; $\Omega_{\alpha,s}$ is proper and closed, and Assumption G.1 holds with $\tilde{\pi} = \mu$.

(ii) *Strong convexity* (Pinsker's inequality at $\alpha = 1$). $\Omega_{\alpha,s}$ is $\mu_{\min}^{(1-\alpha)_+}$-strongly convex with respect to $\|\cdot\|_1$ for $\alpha \leq 2$, and strictly but not strongly convex for $\alpha > 2$.

(iii) *Greedy policy and conjugate* (Roulet et al., 2025, Proposition 1). For $\alpha \neq 1$,

$$\pi_{\Omega_\alpha,Q}(a|s) = \mu(a|s)\left[1 + \frac{(\alpha-1)\big(Q(s,a) - \zeta_s(Q_s)\big)}{\tau}\right]_+^{\frac{1}{\alpha-1}}, \tag{K.1}$$

$$\Omega_\alpha^*(Q)(s) = \frac{(\alpha-1)\langle \pi_{\Omega_\alpha,Q}(\cdot|s), Q_s\rangle + \zeta_s(Q_s)}{\alpha} \quad (\alpha \neq 0), \tag{K.2}$$

where $\zeta_s(Q_s)$ is the unique normalizer making (K.1) sum to one; the truncation is active only for $\alpha > 1$. For any $a^* \in \arg\max_a Q(s,a)$,

$$\max_a Q(s,a) - \tau\,\frac{\mu(a^*|s)^{1-\alpha} - 1}{\alpha - 1} \;\leq\; \zeta_s(Q_s) \;\leq\; \max_a Q(s,a),$$

so that ζ_s is computable by bisection.

(iv) *Domination profile.* $\gamma_{\Omega_\alpha}(R) \geq \exp_{2-\alpha}(-R/\tau)$. If $\alpha > 1$ and $|\mathcal{A}| \geq 2$, then $\gamma_{\Omega_\alpha}(R) = 0$ for $R \geq \tau\mu_{\max}^{1-\alpha}/(\alpha-1)$, with $\mu_{\max} := \max_{s,a}\mu(a|s)$. Hence (P1) holds if and only if $\alpha \leq 1$.

(v) *Regularization modulus.* For $\alpha > 0$, $\Lambda^{\mathrm{pt}}_{\Omega_\alpha}(R) \le \tau\Delta_\alpha$ with $\Delta_\alpha := \frac{\mu_{\min}^{1-\alpha}-1}{\alpha(\alpha-1)}$, and $\Delta_\alpha \le \frac{1}{\alpha(1-\alpha)}$ for $0 < \alpha < 1$. For $\alpha \le 0$,

$$\Lambda^{\mathrm{pt}}_{\Omega_\alpha}(R) \;\le\; \inf_{0<\varepsilon\le 1}\Big\{\varepsilon R + \tau\,\frac{1-\varepsilon^\alpha}{\alpha(1-\alpha)}\Big\} = \begin{cases}\Theta\big(\tau^{\frac{1}{1-\alpha}} R^{\frac{-\alpha}{1-\alpha}}\big) & \alpha < 0,\\ \Theta\big(\tau\log(R/\tau)\big) & \alpha = 0,\end{cases} \tag{K.3}$$

the fraction read as $\log(1/\varepsilon)$ at $\alpha = 0$, and these rates are attained. Hence the modulus is bounded in R if and only if $\alpha > 0$.

Proof. Fix s and write π_a, μ_a, Q_a for $\pi_s(a)$, $\mu(a|s)$, $Q(s,a)$.

(i) The term $\pi_a^\alpha \mu_a^{1-\alpha}$ is finite for every $\pi_a \ge 0$ when $\alpha > 0$, and infinite at $\pi_a = 0$ when $\alpha < 0$, as is $\log(\mu_a/\pi_a)$ at $\alpha = 0$. $\Omega_{\alpha,s}$ is convex and continuous on its domain and tends to $+\infty$ at the boundary when $\alpha \le 0$, hence proper and closed; $D_\alpha(\mu_s\|\mu_s) = 0$ gives Assumption G.1 with $\tilde\pi = \mu$.

(ii) On the relative interior $\nabla^2\Omega_{\alpha,s}(\pi) = \mathrm{diag}(\pi_a^{\alpha-2}\mu_a^{1-\alpha})$, and Cauchy–Schwarz gives $\langle\nabla^2\Omega_{\alpha,s}(\pi)h,h\rangle \ge \|h\|_1^2/\sum_a \pi_a^{2-\alpha}\mu_a^{\alpha-1}$. For $1 \le \alpha \le 2$, Hölder bounds the sum by $(\sum_a \pi_a)^{2-\alpha}(\sum_a \mu_a)^{\alpha-1} = 1$; for $\alpha < 1$, $\pi_a^{2-\alpha} \le \pi_a$ bounds it by $\mu_{\min}^{\alpha-1}$. Integrating along segments and closing by lower semicontinuity gives the modulus. For $\alpha > 2$ the Hessian is positive definite on the interior but its entries vanish as $\pi_a \to 0$.

(iii) The form (K.1) of the maximizer, with a unique normalizer, is Roulet et al. (2025, Proposition 1) with reference measure μ_s and logits Q_s/τ. For the bracket, $S(\zeta) := \sum_a \mu_a[1 + (\alpha-1)(Q_a - \zeta)/\tau]_+^{1/(\alpha-1)}$ is continuous and non-increasing on its domain, $\mathbb{R}$ if $\alpha > 1$ and $(\max_a Q_a - \tau/(1-\alpha), +\infty)$ if $\alpha < 1$; $S(\max_a Q_a) \le 1$, every ratio π_a/μ_a being then at most 1; and $S \ge 1$ at $\zeta = \max_a Q_a - \tau(\mu_{a^*}^{1-\alpha} - 1)/(\alpha-1)$, which lies in the domain and where the term of a^* alone equals 1. Finally, for $\alpha \ne 0$, with $u_a := 1 + (\alpha-1)(Q_a - \zeta_s)/\tau$, we have $\pi_a^\alpha \mu_a^{1-\alpha} = \pi_a u_a$ and $\sum_a \pi_a u_a = 1 + (\alpha-1)(\langle\pi, Q_s\rangle - \zeta_s)/\tau$, hence $\tau\Omega_{\alpha,s}(\pi) = (\langle\pi, Q_s\rangle - \zeta_s)/\alpha$, and $\Omega^*_\alpha(Q)(s) = \langle\pi, Q_s\rangle - \tau\Omega_{\alpha,s}(\pi)$ is (K.2).

(iv) If $\|Q\|_{\mathrm{asp}} \le R$ then $Q_a - \zeta_s \ge Q_a - \max_b Q_b \ge -R$ by (iii), and $x \mapsto [1+(\alpha-1)x/\tau]_+^{1/(\alpha-1)}$ is non-decreasing for every $\alpha \ne 1$, which gives the bound; $\alpha = 1$ is the limit. For $\alpha > 1$, take a^* with $\mu_{a^*} = \mu_{\max}$, $Q_{a^*} = 0$ and $Q_b = -R$ otherwise. The candidate $\pi = \delta_{a^*}$ has normalizer $\zeta = \tau(1-\mu_{\max}^{1-\alpha})/(\alpha-1)$ and $u_b = \mu_{\max}^{1-\alpha} - (\alpha-1)R/\tau \le 0$ for $b \ne a^*$ as soon as $R \ge \tau\mu_{\max}^{1-\alpha}/(\alpha-1)$; by uniqueness it is the greedy policy, so $\gamma_{\Omega_\alpha}(R) = 0$. That the greedy map can be sparse for $\alpha > 1$ is known (Peters et al., 2019; Roulet et al., 2025); what (P1) requires, and what (iv) provides, is the uniform lower bound and the explicit threshold.

(v) Let $a^* \in \arg\max_a Q_a$. For $\alpha > 0$, $\Omega_{\alpha,s}$ is bounded on $\Delta(\mathcal{A})$ by $0 \le \Omega_{\alpha,s} \le \max_a \Omega_{\alpha,s}(\delta_a) = \Delta_\alpha$, the maximum of a convex function on the simplex being attained at a vertex, and the boundedness property of the conjugate (Geist et al., 2019, Proposition 1(ii)) gives $\max_a Q_a - \tau\Delta_\alpha \le \Omega^*_\alpha(Q)(s) \le \max_a Q_a$; testing the maximum defining $\Omega^*_\alpha(Q)(s)$ against δ_{a^*} gives the same bound with μ_{a^*} in place of $\mu_{\min}$. For $0 < \alpha < 1$, $\Delta_\alpha = \frac{1-\mu_{\min}^{1-\alpha}}{\alpha(1-\alpha)} \le \frac{1}{\alpha(1-\alpha)}$. For $\alpha \le 0$, $\Omega_{\alpha,s}$ is unbounded and we test instead $\pi^\varepsilon := (1-\varepsilon)\delta_{a^*} + \varepsilon\mu_s$: $\langle\pi^\varepsilon, Q_s\rangle \ge \max_a Q_a - \varepsilon R$, and $\pi^\varepsilon \ge \varepsilon\mu_s$ with $x \mapsto x^\alpha$ non-increasing gives $\Omega_{\alpha,s}(\pi^\varepsilon) \le \frac{\varepsilon^\alpha - 1}{\alpha(\alpha-1)}$, with limit $\log(1/\varepsilon)$ at $\alpha = 0$. Minimizing over ε, at $\varepsilon = \min\{1, (\tau/((1-\alpha)R))^{1/(1-\alpha)}\}$, gives (K.3). The rates are attained with $Q_{a^*} = R$ and $Q_b = 0$: writing $\pi_b = \delta$, the gap is $\min_\delta\{R\delta + \tau\Omega_{\alpha,s}(\pi)\}$ with $\Omega_{\alpha,s}(\pi) \sim \mu_b^{1-\alpha}\delta^\alpha/(\alpha(\alpha-1))$ for $\alpha < 0$ and $\sim \mu_b\log(1/\delta)$ for $\alpha = 0$, whose minimizer gives the same powers of R. □

Corollary K.1 (Four landmarks)**.** Theorem K.1 at $\alpha \in \{1, 0, \frac{1}{2}, 2\}$ gives the columns of Table 1, and the conjugates:

α	$\Omega^*(Q)(s)$	$\zeta_s \in$
1	$\tau\log\sum_a \mu_a e^{Q_a/\tau}$	–
0	$\zeta_s - \tau\sum_a \mu_a \log\big(1 + \frac{\zeta_s - Q_a}{\tau}\big)$	$(\max_a Q_a - \tau, \max_a Q_a]$
$\frac{1}{2}$	$2\zeta_s - \langle\pi_{\Omega,Q}(\cdot\|s), Q_s\rangle$	$(\max_a Q_a - 2\tau, \max_a Q_a]$
2	$\frac{1}{2}\big(\langle\pi_{\Omega,Q}(\cdot\|s), Q_s\rangle + \zeta_s\big)$	$[\min_a Q_a, \max_a Q_a]$

Proof. At $\alpha = 1$ the normalizer factors out of the exponential. At $\alpha = 0$, (K.1) reads $\pi_a = \tau\mu_a/(\tau + \zeta_s - Q_a)$, so $\sum_a \pi_a(\tau + \zeta_s - Q_a) = \tau$ gives $\langle \pi, Q_s\rangle = \zeta_s$, and $\Omega^*(Q)(s) = \langle \pi, Q_s\rangle - \tau\mathrm{KL}(\mu_s\|\pi)$ gives the conjugate; the modulus bound is (K.3) at $\varepsilon = \min\{1, \tau/R\}$. At $\alpha = \frac{1}{2}$ and $\alpha = 2$ the conjugates are those of Theorem K.1.(iii), and $\Omega_s(\delta_a) = 4(1 - \sqrt{\mu_a})$ and $\frac{1}{2}(1/\mu_a - 1)$ give the modulus bounds. The brackets at $\alpha \in \{0, \frac{1}{2}\}$ follow from Theorem K.1.(iii); at $\alpha = 2$, the sum in (K.1) is at least 1 at $\zeta = \min_a Q_a$ and at most 1 at $\zeta = \max_a Q_a$. □

Remark K.1 (Sharpness)**.** For a fixed μ, ζ_s is non-decreasing in each Q_a while π_b/μ_b is non-increasing in $\zeta_s - Q_b$, so the infimum defining $\gamma_\Omega(R)$ is attained with the least likely action at the bottom and every other one at $+R$; symmetrically, $\Lambda^{\mathrm{pt}}_\Omega(R)$ is attained with the least likely action alone at the top and every other one at $-R$. Both depend on μ only through $\mu_{\min}$, and the explicit normalizers give, for instance, $\gamma_\Omega(R) = \big(\mu_{\min} + (1 - \mu_{\min})e^{R/\tau}\big)^{-1}$ for the reverse KL and $\gamma_\Omega(R) = [1 - (1 - \mu_{\min})R/\tau]_+$ for the χ^2, whose limits as $\mu_{\min} \to 0$ are the bounds of Table 1. Inside the window $0 < \alpha \leq 1$ the comparison is quantitative and depends on how τ is matched across α. At fixed τ, the profile decays as $(R/\tau)^{-1/(1-\alpha)}$, more slowly for smaller α, while the modulus constant $\Delta_\alpha = (1 - \mu_{\min}^{1-\alpha})/(\alpha(1 - \alpha))$ tends, as $\mu_{\min} \to 0$, to $1/(\alpha(1 - \alpha))$ for $\alpha < 1$, minimal at the Hellinger value 4, and to $+\infty$ at $\alpha = 1$: the Hellinger regularizer is the most robust to a small $\mu_{\min}$, whereas for a balanced μ the reverse KL can have the smaller constant: for a uniform μ, $\log|\mathcal{A}| < 4(1 - |\mathcal{A}|^{-1/2})$ as long as $|\mathcal{A}| \leq 24$. The lower bound of Theorem K.1.(iv) equals $(f')^{-1}(-R/\tau)$ for the generator $f(u) = (u^\alpha - \alpha u + \alpha - 1)/(\alpha(\alpha - 1))$ of D_α, the inverse link that also drives the policy-gradient analysis of Labbi et al. (2026): their probability floor (Labbi et al., 2026, Theorem 3.5) is proportional to $(f')^{-1}$ evaluated at an argument of order $-1/\lambda$, where λ is their temperature, up to factors depending on the discount and the initial distribution, and their exponential-versus-polynomial dichotomy between the KL and $\alpha < 1$ is the one displayed by our profile bounds. They also find that the best α lies strictly inside $(0, 1)$ (Labbi et al., 2026, Corollary 4.8), for a different criterion.