A SMOOTH POLYNOMIAL LYAPUNOV CERTIFICATE FOR CONVERGENCE OF Q-LEARNING AND ITS SMOOTH VARIANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Classical convergence analyses of Q-learning rely on the ∞-norm contraction of Bellman operators, and existing ordinary differential equation (ODE) arguments often use the non-differentiable ∞-norm directly. This paper develops a smooth polynomial Lyapunov-function-based stability certificate for convergence of Q-learning by transferring ∞-norm contraction to a weighted degree-$2p$ polynomial Lyapunov function induced by a finite $2p$-norm. The framework is conceptual and structural: it avoids non-differentiability, handles preconditioned dynamics arising in Q-learning and its variants, and gives a unified stability argument for standard Q-learning and smooth variants based on log-sum-exp (LSE), mellowmax, and Boltzmann softmax operators. For contractive operators, including the max, LSE, and mellowmax cases, the associated ODEs are globally exponentially stable and, under the stated independent and identically distributed (i.i.d.) sampling model, the stochastic approximation iterates converge almost surely. For the Boltzmann operator, which need not be contractive, the same framework yields convergence to an explicit invariant error set around the optimal Q-function. The resulting theory is not intended as a finite-time bound, but as a clean ODE foundation that unifies and simplifies asymptotic analyses of Q-learning and its smooth variants.

1 INTRODUCTION

Reinforcement learning (RL) provides a mathematical framework for sequential decision-making in unknown environments, where an agent learns policies by interacting with the environment and observing rewards (Sutton & Barto, 1998). Its recent empirical successes, including human-level or superhuman performance in challenging domains (Mnih et al., 2015), have intensified interest in understanding the stability and convergence properties of RL algorithms. Among these algorithms, Q-learning (Watkins & Dayan, 1992) remains one of the most fundamental: it is simple, model-free, and directly targets the optimal action-value function. Its asymptotic convergence has been studied extensively (Jaakkola et al., 1994; Borkar & Meyn, 2000; Lee & He, 2020; Tsitsiklis, 1994; Singh et al., 2000), and more recent work has developed finite-time guarantees that quantify the sample complexity of convergence (Szepesvári, 1998; Even-Dar & Mansour, 2003; Beck & Srikant, 2012; Wainwright, 2019; Qu & Wierman, 2020; Li et al., 2022; Chen et al., 2024; 2020; Lee, 2024).

The ordinary differential equation (ODE) method (Borkar & Meyn, 2000) is a particularly useful tool for studying stochastic-approximation algorithms in RL. It replaces the noisy discrete-time recursion with a limiting continuous-time dynamical system and derives convergence from stability of the associated ODE. Although finite-time analyses provide sharper quantitative rates, the ODE viewpoint remains valuable because it often yields cleaner proofs, exposes the underlying dynamical mechanism, and provides a useful first stability check before pursuing more refined non-asymptotic results. This perspective has been applied to a broad range of RL and stochastic-approximation schemes (Borkar & Meyn, 2000; Lee & He, 2020; Sutton et al., 2008; 2009; Ghiassian et al., 2020; Lee et al., 2023; Melo et al., 2008; Bhatnagar et al., 2013).

This paper develops an ODE-based convergence framework for standard Q-learning and several smooth Q-learning variants (Haarnoja et al., 2017; Song et al., 2019; Pan et al., 2020; Asadi & Littman, 2017; Barber, 2023). Smooth variants replace the hard max operator in the Bellman update

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.

with a differentiable approximation, such as log-sum-exp (LSE) (Haarnoja et al., 2017), mellowmax (Asadi & Littman, 2017), or Boltzmann softmax (Pan et al., 2020). These operators are often used to improve empirical behavior, encourage smoother action selection, and reduce overestimation effects (Haarnoja et al., 2017; Pan et al., 2020; Song et al., 2019; Barber, 2023). However, a unified ODE-based convergence theory that covers both standard Q-learning and such smooth variants is still limited.

Our main idea is to replace the classical non-smooth ∞-norm Lyapunov argument with a smooth weighted degree-$2p$ polynomial Lyapunov function. The framework transfers ∞-norm contraction of Bellman-type operators to a differentiable Lyapunov certificate, while also handling preconditioned dynamics of the form $\dot{x} = D(F(x) - x)$. Here, $x \in \mathbb{R}^n$ is the state of the limiting ODE, $F : \mathbb{R}^n \to \mathbb{R}^n$ is the operator whose fixed point is the target solution, and D is a positive diagonal matrix that scales the coordinatewise dynamics; in asynchronous Q-learning, its diagonal entries represent coordinate-update frequencies. At the ODE level, this allows a single proof template to cover the separately defined synchronous full-update mean field with $D = I$ and asynchronous mean fields with a positive diagonal D induced by coordinate-update frequencies. For the max, LSE, and mellowmax operators, the induced Bellman operators are contractive and the associated ODEs are globally exponentially stable. For the Boltzmann operator, which is generally not contractive, the same framework yields convergence to an explicit invariant error set around the optimal Q-function.

The main contributions are summarized as follows:

1. **A smooth Lyapunov principle for preconditioned operator dynamics.** We establish stability results for ODEs of the form $\dot{x}_t = DF(x_t) - Dx_t$ using a weighted degree-$2p$ polynomial Lyapunov function. When F is an ∞-norm contraction, the framework transfers this contraction to a differentiable Lyapunov certificate and avoids the non-differentiability of the ∞-norm.
2. **A unified ODE stability analysis for standard and smooth Q-learning.** Applying the general result to Bellman operators, we prove global exponential stability of the ODE models associated with standard Q-learning and smooth variants based on the LSE and mellowmax operators. The analysis covers both the separately defined synchronous full-update mean field with $D = I$ and asynchronous mean fields with a positive diagonal D induced by coordinate-update frequencies.
3. **A non-contractive stability result for Boltzmann smooth Q-learning.** For the Boltzmann softmax operator, which need not be non-expansive, we prove that the corresponding ODE converges to an explicit error set around $Q^*_{\max}$. The radius of this set scales as $\gamma\lambda \ln(|\mathcal{A}|)w_{\max}/((1-\gamma)w_{\min})$, showing how the approximation error shrinks as the temperature parameter λ decreases to zero.
4. **A unified asymptotic convergence theorem for the stochastic algorithms.** Using the ODE stability results, we derive almost-sure convergence guarantees for the corresponding discrete-time Q-learning variants under independent and identically distributed (i.i.d.) sampling and standard diminishing step-sizes. The resulting analysis gives a single ODE-based proof architecture for standard Q-learning, contractive smooth Q-learning variants, and the non-contractive Boltzmann case.

Overall, the paper provides a smooth Lyapunov-based ODE foundation for Q-learning and its smooth variants. The framework is intentionally asymptotic rather than finite-time: its purpose is to clarify the stability mechanism and to offer a compact route for proving convergence of both classical and smoothed Bellman updates.

1.1 Related work

Despite a growing body of work on smooth variants of Q-learning, an ODE-based convergence analysis has not yet been fully established. In what follows, we briefly review prior studies on smooth Q-learning and clarify how our contributions differ from and extend existing results.

Smooth Q-learning algorithms: Smooth Bellman updates have been studied in several settings. Haarnoja et al. (Haarnoja et al., 2017) developed deep energy-based RL with an LSE structure, while Song et al. (Song et al., 2019) analyzed theoretical properties of the Boltzmann softmax Bellman operator and evaluated it with deep Q-learning. These works did not establish almost-sure convergence

of the tabular stochastic-approximation recursions considered here. Asadi and Littman (Asadi & Littman, 2017) introduced the mellowmax operator, proved its non-expansiveness, and established convergence for generalized value iteration and a corresponding state–action–reward–state–action (SARSA) scheme. Pan et al. (Pan et al., 2020) later analyzed tabular learning with dynamic Boltzmann softmax updates, whereas Barber (Barber, 2023) proposed a different convergent off-policy rule that replaces maximization with averaging. These studies do not establish stability of the associated ODE models. The direct finite-time analysis in Jeong & Lee (2025) covers the LSE, Boltzmann, and mellowmax variants, but likewise does not provide a unified ODE stability theory. In contrast, our analysis develops one ODE framework for the standard max operator and all three smooth operators considered here.

ODE-based analysis of Q-learning: Recent work (Lee & He, 2020) developed a switching-system model (Liberzon, 2003) for asynchronous Q-learning and employed the ODE method (Borkar & Meyn, 2000), leveraging tools from switching-system theory to establish global asymptotic stability without explicit Lyapunov-function constructions. A key limitation, however, is that its global stability result requires restrictive conditions (e.g., quasi-monotonicity) on the induced switching systems, which makes the approach difficult to extend to other RL algorithms such as smooth Q-learning variants. In contrast, the proposed method applies more broadly and yields a unified treatment of both standard Q-learning and a variety of smooth Q-learning algorithms. Consequently, it provides an alternative ODE-based route for analyzing asynchronous Q-learning beyond Lee & He (2020).

Prior ODE analyses (Borkar & Meyn, 2000; Borkar & Soumyanath, 1997) established stability of Q-learning mean fields using the ∞-norm as a Lyapunov function. The analysis in Borkar & Soumyanath (1997), however, addressed only synchronous Q-learning, whereas the ODE models studied here include both synchronous and asynchronous Q-learning and their smooth variants. Moreover, ∞-norm–based arguments require careful treatment because the ∞-norm is nondifferentiable. By using a smooth polynomial Lyapunov function induced by a weighted $2p$-norm, our approach avoids this difficulty while retaining a direct stability argument. Thus, relative to Borkar & Meyn (2000); Borkar & Soumyanath (1997), we cover positive-diagonally scaled mean fields arising from asynchronous updates and obtain concise stability proofs through a smooth polynomial certificate.

Finite-time analysis of smooth Q-learning: In Jeong & Lee (2025), a finite-time analysis was developed for smooth Q-learning algorithms based on the LSE, Boltzmann softmax, and mellowmax operators. That work analyzed the discrete-time algorithms directly without invoking an ODE framework. In contrast, this paper establishes asymptotic convergence through ODE analysis. Beyond the discrete-time iterates, we study a positive-diagonally scaled ODE family that includes the mean fields induced by asynchronous updates.

2 Preliminaries

2.1 Markov decision process

We consider a finite discounted Markov decision process (MDP) with state space $\mathcal{S} = \{1, 2, \dots, |\mathcal{S}|\}$, action space $\mathcal{A} = \{1, 2, \dots, |\mathcal{A}|\}$, transition probabilities $P(s' \mid s, a)$, deterministic rewards $r(s, a, s') \in \mathbb{R}$, and discount factor $\gamma \in (0, 1)$. At time k, the agent observes s_k, selects a_k, receives $r_{k+1} := r(s_k, a_k, s_{k+1})$, and moves to s_{k+1} according to $P(\cdot \mid s_k, a_k)$.

A deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ induces the action-value function

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} \,\middle|\, s_0 = s, a_0 = a, \pi\right].$$

The optimal action-value function is $Q^*(s, a) = Q^{\pi^*}(s, a)$, where π^* maximizes the expected discounted return over deterministic policies. Once Q^* is known, an optimal policy is obtained by the greedy rule $\pi^*(s) \in \arg\max_{a \in \mathcal{A}} Q^*(s, a)$. Whenever a state-sampling distribution p is used below, it is treated as a fixed distribution. It may be chosen as the stationary distribution of an ergodic behavior chain, but ergodicity alone does not make successive observations independent; the convergence theorems separately assume independent resampling from p.

2.2 DEFINITIONS

This subsection introduces the definitions and auxiliary results used below. Throughout the paper, we use the following notation for compact matrix–vector representations:

$$P := \begin{bmatrix} P_1 \\ \vdots \\ P_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}, \; R := \begin{bmatrix} R_1 \\ \vdots \\ R_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}, \; Q := \begin{bmatrix} Q(\cdot,1) \\ \vdots \\ Q(\cdot,|\mathcal{A}|) \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}, \quad (1)$$

where $P_a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is the transition probability matrix under the action $a \in \mathcal{A}$, $Q(\cdot,a) \in \mathbb{R}^{|\mathcal{S}|}$ for $a \in \mathcal{A}$, and $R_a(s) := \mathbb{E}[r(s,a,s') \mid s,a]$. In this notation, the Q-function is encoded as a single vector $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ that enumerates $Q(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. In particular, $Q(s,a) = (e_a \otimes e_s)^\top Q$, where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ and $e_a \in \mathbb{R}^{|\mathcal{A}|}$ are the s-th and a-th standard basis vectors, respectively. Similarly, $R(s,a) = (e_a \otimes e_s)^\top R$ and $P(s' \mid s,a) = (e_a \otimes e_s)^\top P e_{s'}$.

We focus on the weighted $2p$-norm and its associated polynomial Lyapunov function. For $p \in \mathbb{N} := \{1,2,\ldots\}$, define

$$\|x\|_{2p,w} := \left(\sum_{i=1}^n w_i x_i^{2p}\right)^{1/(2p)}, \qquad W_p(x) := \frac{1}{2p}\|x\|_{2p,w}^{2p}. \quad (2)$$

The real numbers $w_i > 0, i \in \{1,2,\ldots,n\}$, are fixed weights. The weighted $2p$-norm reduces to the standard $2p$-norm when $w_i = 1$ for every i. Because $2p$ is even, W_p is a polynomial and is therefore infinitely differentiable on all of $\mathbb{R}^n$, including the origin. Moreover, $\lim_{p\to\infty}\|x\|_{2p,w} = \|x\|_\infty$. This limiting relation is crucial in our analysis. The following norm comparisons will be used in both the polynomial and norm-based estimates.

Lemma 1 (Norm comparisons)**.** *For any $x \in \mathbb{R}^n$ and any $p \in \mathbb{N}$, we have*

1. $w_{\min}^{1/(2p)}\|x\|_\infty \le \|x\|_{2p,w} \le n^{1/(2p)} w_{\max}^{1/(2p)}\|x\|_\infty$
2. $w_{\min}^{1/(2p)}\|x\|_{2p} \le \|x\|_{2p,w} \le w_{\max}^{1/(2p)}\|x\|_{2p}$
3. $\|x\|_\infty \le \|x\|_{2p} \le n^{1/(2p)}\|x\|_\infty$

where $w_{\min} := \min_{i\in\{1,2,\ldots,n\}} w_i$ and $w_{\max} := \max_{i\in\{1,2,\ldots,n\}} w_i$.

The proof of Lemma 1 is provided in Section A.

Definition 1 (Monotone operator)**.** *A mapping $h : \mathbb{R}^n \to \mathbb{R}$ is monotone with respect to the componentwise order if $u_i \le v_i$ for every i implies $h(u) \le h(v)$.*

Define the mappings $h_{\max} : \mathbb{R}^n \to \mathbb{R}$, $h_{\mathrm{mm}} : \mathbb{R}^n \to \mathbb{R}$, $h_{\mathrm{lse}} : \mathbb{R}^n \to \mathbb{R}$, and $h_{\mathrm{bz}} : \mathbb{R}^n \to \mathbb{R}$ by

$$h_{\max}(x) := \max_{i\in\{1,2,\ldots,n\}} x_i, \qquad h_{\mathrm{lse}}^\lambda(x) := \lambda \ln\left(\sum_{i=1}^n e^{x_i/\lambda}\right),$$

$$h_{\mathrm{mm}}^\lambda(x) := \lambda \ln\left(\frac{1}{n}\sum_{i=1}^n e^{x_i/\lambda}\right), \qquad h_{\mathrm{bz}}^\lambda(x) := \frac{\sum_{i=1}^n x_i e^{x_i/\lambda}}{\sum_{i=1}^n e^{x_i/\lambda}}.$$

Here, $\lambda > 0$ is the temperature parameter for the log-sum-exp (LSE), mellowmax, and Boltzmann operators, and each operator uses the factor $e^{x_i/\lambda}$. The mapping $h_{\max}$ is the standard max operator. The mapping h_{lse}^λ is the LSE, or smooth-max, operator used in machine learning and reinforcement learning (RL) (Haarnoja et al., 2017; Dai et al., 2018). The mellowmax operator h_{mm}^λ was introduced in Asadi & Littman (2017) to address limitations of the Boltzmann softmax operator h_{bz}^λ. The latter forms a smooth approximation of the max operator through softmax weights that are widely used in RL (Pan et al., 2020; Gao & Pavel, 2017). For each choice of $h \in \{h_{\max}, h_{\mathrm{lse}}^\lambda, h_{\mathrm{mm}}^\lambda, h_{\mathrm{bz}}^\lambda\}$, define $H : \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|}$ by

$$H(Q) := \begin{bmatrix} h(Q(1,\cdot)) \\ h(Q(2,\cdot)) \\ \vdots \\ h(Q(|\mathcal{S}|,\cdot)) \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|}.$$

Accordingly, $H_{\max}$, $H^{\lambda}_{\mathrm{lse}}$, $H^{\lambda}_{\mathrm{mm}}$, and $H^{\lambda}_{\mathrm{bz}}$ denote the mappings obtained from $h_{\max}$, $h^{\lambda}_{\mathrm{lse}}$, $h^{\lambda}_{\mathrm{mm}}$, and $h^{\lambda}_{\mathrm{bz}}$, respectively. In this paper, we consider smooth variants of Q-learning based on these approximations of the max operator and analyze their convergence in a unified manner using ordinary differential equation (ODE) analysis. The proofs of the two main polynomial Lyapunov-function-based stability theorems, together with the norm-based and stochastic-approximation results, are provided in the Appendix.

3 Stability of nonlinear ODE models under contraction

We consider the following ODE:

$$\frac{d}{dt}x_t = DF(x_t) - Dx_t, \quad \forall t \geq 0, \quad x_0 \in \mathbb{R}^n \tag{3}$$

where $t \geq 0$ is the continuous time, $x_t \in \mathbb{R}^n$ is the state at time t, $F : \mathbb{R}^n \to \mathbb{R}^n$ is a mapping that will be specified later, and $D \in \mathbb{R}^{n\times n}$ is a positive definite diagonal matrix with strictly positive diagonal elements $d_i > 0$, $i \in \{1, 2, \ldots, n\}$. This ODE will be used to model Q-learning and its variants in the remainder of the paper. A similar ODE was originally considered in Borkar & Soumyanath (1997). The distinction here is the presence of D: when $D = I_n$, where I_n is the n-dimensional identity matrix, (3) reduces to the ODE in Borkar & Soumyanath (1997). When D is later induced by the update probabilities of the one-coordinate recursion in Algorithm 1, its diagonal entries sum to one, so $D = I_n$ does not arise from that sampler unless $n = 1$. Instead, $D = I_n$ represents the mean field of a separately defined synchronous full-update recursion in which every coordinate is updated at each iteration. To account for the diagonal scaling, we choose $w_i = 1/d_i$ and use the polynomial Lyapunov function W_p in (2). Because $2p$ is even, this function is smooth on all of $\mathbb{R}^n$, including at the origin. The following result is the first main stability result for (3).

Theorem 1 (Polynomial contractive certificate)**.** *Let x_t be the unique solution of (3). Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is an ∞-norm contraction: for some $\alpha \in (0, 1)$,*

$$\|F(x) - F(y)\|_\infty \leq \alpha\|x - y\|_\infty, \qquad x, y \in \mathbb{R}^n.$$

Let x^* be its unique fixed point, and choose $w_i = 1/d_i$. For any $p \in \mathbb{N}$ satisfying

$$\left\lceil \frac{\ln n}{\ln(\alpha^{-1})} \right\rceil < 2p,$$

we have

$$\begin{aligned} \frac{d}{dt}W_p(x_t - x^*) &\leq -\left(1 - \alpha n^{1/(2p)}\right)\|x_t - x^*\|_{2p}^{2p} \\ &\leq -\frac{2p\left(1 - \alpha n^{1/(2p)}\right)}{w_{\max}} W_p(x_t - x^*), \end{aligned} \tag{4}$$

and hence

$$W_p(x_t - x^*) \leq \exp\left(-\frac{2p\left(1 - \alpha n^{1/(2p)}\right)}{w_{\max}}t\right) W_p(x_0 - x^*), \qquad t \geq 0, \tag{5}$$

where $w_{\max} := \max_{i\in\{1,2,\ldots,n\}} w_i$. Therefore, W_p is a globally smooth polynomial Lyapunov function and x^* is the unique globally exponentially stable equilibrium.

The proof of Theorem 1 is provided in Section B. Theorem 1 uses W_p itself as the Lyapunov function. The condition on p guarantees that $1 - \alpha n^{1/(2p)} > 0$. For any fixed n and $\alpha \in (0, 1)$, a sufficiently large integer p satisfies this condition, while W_p remains a finite-degree polynomial. As $p \to \infty$, the contraction margin $1 - \alpha n^{1/(2p)}$ approaches $1 - \alpha$.

Weighted p-norm Lyapunov functions have previously been used to analyze the contractive dynamics

$$\dot{x}_t = F(x_t) - x_t$$

in Borkar & Soumyanath (1997); Borkar (2008). In those works, the Lyapunov function is matched to a mapping that is contractive in the same weighted p-norm. In contrast, in Theorem 1, F is assumed to be contractive only in the ∞-norm, while the polynomial W_p is induced by a weighted

finite $2p$-norm. The comparison between the finite $2p$-norm and the ∞-norm transfers the contraction property to the smooth polynomial certificate. This approach avoids the directional-derivative or subgradient arguments required when the non-differentiable ∞-norm is used directly, accommodates the preconditioned dynamics in (3), and keeps the stability proof concise.

Next, we consider a mapping that is uniformly close to an ∞-norm contraction.

Theorem 2 (Polynomial perturbation certificate)**.** *Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is globally Lipschitz continuous, and let x_t be the unique global solution of (3). Suppose that $F' : \mathbb{R}^n \to \mathbb{R}^n$ is an ∞-norm contraction with factor $\alpha \in (0,1)$ and fixed point x^*:*

$$\|F'(x) - F'(y)\|_\infty \le \alpha \|x - y\|_\infty, \qquad x, y \in \mathbb{R}^n.$$

Suppose also that F satisfies

$$\|F(x) - F'(x)\|_\infty \le \eta, \qquad x \in \mathbb{R}^n,$$

for some $\eta > 0$. Choose $w_i = 1/d_i$, and let $w_{\min} := \min_i w_i$ and $w_{\max} := \max_i w_i$. For any $p \in \mathbb{N}$ satisfying

$$\left\lceil \frac{\ln n}{\ln(\alpha^{-1})} \right\rceil < 2p,$$

we have

$$\frac{d}{dt} W_p(x_t - x^*) \le -\left(1 - \alpha n^{1/(2p)}\right) \|x_t - x^*\|_{2p}^{2p} + n^{1/(2p)} \eta \|x_t - x^*\|_{2p}^{2p-1}. \tag{6}$$

Moreover,

$$\|x_t - x^*\|_{2p,w} \le \|x_0 - x^*\|_{2p,w} \exp\left(\frac{\alpha n^{1/(2p)} - 1}{w_{\max}} t\right) + \frac{n^{1/(2p)} \eta w_{\max}}{\left(1 - \alpha n^{1/(2p)}\right) w_{\min}^{(2p-1)/(2p)}}, \qquad t \ge 0. \tag{7}$$

In particular, $W_p(x_t - x^*)$ decreases whenever $\|x_t - x^*\|_{2p} > n^{1/(2p)} \eta / (1 - \alpha n^{1/(2p)})$.

The proof of Theorem 2 is provided in Section C. Theorem 2 shows that the polynomial Lyapunov function is strictly decreasing outside an explicit neighborhood of x^*. Thus, even when F is not itself contractive, the solution of (3) is ultimately bounded under the stated uniform approximation condition. This result will be used for the Boltzmann Q-learning ODE.

4 ODE MODEL OF Q-LEARNING AND ITS SMOOTH VARIANTS

In this section, we consider the following specific ODE model associated with Q-learning and analyze its convergence properties:

$$\frac{d}{dt} Q_t = DR + \gamma DPH(Q_t) - DQ_t, \quad \forall t \ge 0, \quad Q_0 \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}. \tag{8}$$

Here, $H \in \{H_{\max}, H_{\mathrm{lse}}^\lambda, H_{\mathrm{mm}}^\lambda, H_{\mathrm{bz}}^\lambda\}$, and $D \in \mathbb{R}^{n\times n}$ is a positive definite diagonal matrix with strictly positive diagonal elements $d_i > 0$, $i \in \{1, 2, \ldots, n\}$. The ODE is a special case of (3) and is the continuous-time model commonly associated with Q-learning and its smooth variants. Define the Bellman operator

$$F(Q) := R + \gamma PH(Q).$$

Then (8) can be written as

$$\frac{d}{dt} Q_t = DF(Q_t) - DQ_t, \quad \forall t \ge 0, \quad Q_0 \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}, \tag{9}$$

which has the form of (3). The four Bellman operators considered here are

$$F_{\max}(Q) := R + \gamma PH_{\max}(Q), \qquad F_{\mathrm{bz}}^\lambda(Q) := R + \gamma PH_{\mathrm{bz}}^\lambda(Q),$$
$$F_{\mathrm{lse}}^\lambda(Q) := R + \gamma PH_{\mathrm{lse}}^\lambda(Q), \qquad F_{\mathrm{mm}}^\lambda(Q) := R + \gamma PH_{\mathrm{mm}}^\lambda(Q).$$

The max, LSE, and mellowmax operators are non-expansive (Asadi & Littman, 2017; Dai et al., 2018); consequently, their Bellman operators are γ-contractions by Lemma 11. Their unique fixed points are defined by

$$Q^*_{\max} = F_{\max}(Q^*_{\max}), \qquad Q^\lambda_{\mathrm{lse}} = F^\lambda_{\mathrm{lse}}(Q^\lambda_{\mathrm{lse}}), \qquad Q^\lambda_{\mathrm{mm}} = F^\lambda_{\mathrm{mm}}(Q^\lambda_{\mathrm{mm}}). \tag{10}$$

The Boltzmann softmax operator is not non-expansive in general (Asadi & Littman, 2017). We therefore first treat the three contractive cases and then use the perturbation result for the Boltzmann case.

Theorem 3. *Consider (8) with $H \in \{H_{\max}, H^\lambda_{\mathrm{lse}}, H^\lambda_{\mathrm{mm}}\}$, and let Q_t be its solution. Let Q_e denote the corresponding fixed point in (10), set $n = |\mathcal{S} \times \mathcal{A}|$, choose $w_i = 1/d_i$, and let $w_{\max} := \max_i w_i$. For any $p \in \mathbb{N}$ satisfying*

$$\left\lceil \frac{\ln n}{\ln(\gamma^{-1})} \right\rceil < 2p,$$

we have

$$W_p(Q_t - Q_e) \le \exp\left(-\frac{2p\left(1 - \gamma n^{1/(2p)}\right)}{w_{\max}} t \right) W_p(Q_0 - Q_e), \qquad t \ge 0. \tag{11}$$

Therefore, Q_e is the unique globally exponentially stable equilibrium of the corresponding ODE.

The proof of Theorem 3 is provided in Section D.

Theorem 3 states that the ODE associated with the max, LSE, or mellowmax operator converges exponentially to its respective fixed point, with the smooth polynomial W_p serving as a Lyapunov function. Because the Boltzmann softmax operator is not contractive in general, we instead combine its uniform approximation to the max operator with the polynomial perturbation certificate in Theorem 2 to obtain the following result.

Theorem 4. *Consider (8) with $H = H^\lambda_{\mathrm{bz}}$, and let Q_t be its solution. Set $n = |\mathcal{S} \times \mathcal{A}|$, choose $w_i = 1/d_i$, and let $w_{\min} := \min_i w_i$ and $w_{\max} := \max_i w_i$. For any $p \in \mathbb{N}$ satisfying*

$$\left\lceil \frac{\ln n}{\ln(\gamma^{-1})} \right\rceil < 2p,$$

we have

$$\begin{aligned} \frac{d}{dt} W_p(Q_t - Q^*_{\max}) \le & -\left(1 - \gamma n^{1/(2p)}\right) \|Q_t - Q^*_{\max}\|_{2p}^{2p} \\ & + n^{1/(2p)} \gamma \lambda \ln(|\mathcal{A}|) \|Q_t - Q^*_{\max}\|_{2p}^{2p-1}, \end{aligned} \tag{12}$$

and

$$\begin{aligned} \|Q_t - Q^*_{\max}\|_{2p,w} \le & \|Q_0 - Q^*_{\max}\|_{2p,w} \exp\left(\frac{\gamma n^{1/(2p)} - 1}{w_{\max}} t \right) \\ & + \frac{n^{1/(2p)} w_{\max}}{\left(1 - \gamma n^{1/(2p)}\right) w_{\min}^{(2p-1)/(2p)}} \gamma \lambda \ln(|\mathcal{A}|), \qquad t \ge 0. \end{aligned} \tag{13}$$

The proof of Theorem 4 is provided in Section E. Theorem 4 shows that W_p decreases outside an explicit neighborhood of $Q^*_{\max}$ and gives a quantitative ultimate bound in the weighted $2p$-norm. The exact ∞-norm invariant-set counterpart and its numerical illustration are collected in the appendix.

5 Q-LEARNING AND ITS SMOOTH VARIANTS

We consider the recursion in Algorithm 1, where

$$Q_k(s_k', \cdot) := \begin{bmatrix} Q_k(s_k', 1) \\ Q_k(s_k', 2) \\ \vdots \\ Q_k(s_k', |\mathcal{A}|) \end{bmatrix} \in \mathbb{R}^{|\mathcal{A}|},$$

Algorithm 1 Asynchronous Q-learning variants under independent sampling

1: Initialize $Q_0 \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$
2: **for** $k = 0, 1, \ldots$ **do**
3: Sample $s_k \sim p(\cdot)$ and $a_k \sim \beta(\cdot \mid s_k)$
4: Sample $s'_k \sim P(\cdot \mid s_k, a_k)$ and $r(s_k, a_k, s'_k)$
5: Update $Q_{k+1}(s_k, a_k) = Q_k(s_k, a_k) + \alpha_k\big\{r(s_k, a_k, s'_k) + \gamma\, h(Q_k(s'_k, \cdot)) - Q_k(s_k, a_k)\big\}$
6: **end for**

and $h : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}$ is chosen from $\{h_{\max}, h^{\lambda}_{\mathrm{lse}}, h^{\lambda}_{\mathrm{mm}}, h^{\lambda}_{\mathrm{bz}}\}$. The choice $h = h_{\max}$ gives standard Q-learning (Watkins & Dayan, 1992); $h = h^{\lambda}_{\mathrm{lse}}$ gives the LSE variant (Haarnoja et al., 2017); $h = h^{\lambda}_{\mathrm{mm}}$ gives the mellowmax variant (Asadi & Littman, 2017); and $h = h^{\lambda}_{\mathrm{bz}}$ gives the Boltzmann softmax variant (Pan et al., 2020; Song et al., 2019).

Algorithm 1 is the asynchronous single-coordinate recursion. Its mean field contains the diagonal matrix with entries $d(s, a) = p(s)\beta(a \mid s)$. A synchronous full-update recursion is a separate algorithm that updates every state–action coordinate at each iteration using one Bellman sample for each coordinate; its mean field is obtained by setting $D = I$ in (9). Thus, $D = I$ is not obtained by assigning sampling probabilities in Algorithm 1.

In this paper, $\{(s_k, a_k, s'_k)\}_{k=0}^{\infty}$ are independent and identically distributed (i.i.d.) samples generated as follows: $s_k \sim p$, $a_k \sim \beta(\cdot \mid s_k)$, and $s'_k \sim P(\cdot \mid s_k, a_k)$. This is a generative-model or independent replay-sampling assumption, not the distribution of successive observations along a single Markov trajectory. In particular, ergodicity of a behavior chain may justify the choice of a stationary distribution p, but it does not imply the independence required here. Note that s'_k denotes the next-state sample associated with the independently sampled pair (s_k, a_k) at iteration k. The current state used at iteration $k+1$ is resampled from p and is therefore independent of s_k. Under this assumption, the joint state–action distribution at every iteration is

$$d(s, a) = p(s)\beta(a \mid s), \quad (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{14}$$

Throughout the paper, we adopt the following assumption for convenience.

Assumption 1. *$d(s, a) > 0$ holds for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.*

Under the i.i.d. sampling model, Assumption 1 guarantees that every state–action pair is sampled infinitely often with probability one. Similar assumptions have been adopted in Li et al. (2022) and Chen et al. (2024). In contrast, Beck & Srikant (2012) imposes an alternative exploration requirement, known as the *cover time* condition, which posits that within a prescribed time horizon every state–action pair is expected to be visited at least once. Related variants of cover-time conditions have also been used in Even-Dar & Mansour (2003) and Li et al. (2022) to derive convergence-rate guarantees.

In practice, replay sampling can reduce temporal dependence, but samples drawn from a finite replay buffer are not generally exactly i.i.d. The convergence results below require the i.i.d. model stated above. The analysis can in principle be extended to Markovian observation settings using recently developed tools, albeit at the cost of substantially more involved arguments (Lim & Lee, 2024; Qu & Wierman, 2020). Such extensions, however, are not the main focus of this paper. Our primary goal is to develop a useful framework that interfaces naturally with the ODE method and the Borkar–Meyn theorem, two widely used tools for establishing convergence in RL. Many subsequent refinements build on these tools. The corresponding numerical illustration of smooth Q-learning with the Boltzmann operator is provided in Figure 3.

6 Convergence Analysis

This section summarizes the convergence consequences of the ODE stability results. The detailed stochastic-approximation verification and the proofs of the two convergence claims are provided in Section K. Under the i.i.d. sampling model in Algorithm 1 and the step-size conditions $\alpha_k > 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, the update can be written as a stochastic approximation recursion

$$Q_{k+1} = Q_k + \alpha_k\{f(Q_k) + \varepsilon_{k+1}\},$$

where $f(Q) = DR + \gamma DPH(Q) - DQ$ and ε_{k+1} is a martingale-difference noise term. Thus convergence follows by combining the ODE stability results in Theorems 3 and 4 with standard ODE-based stochastic approximation arguments. For $h \in \{h_{\max}, h^{\lambda}_{\mathrm{lse}}, h^{\lambda}_{\mathrm{mm}}\}$, the corresponding Bellman operator is an ∞-norm contraction. Hence the associated ODE is globally exponentially stable, and the stochastic iterates converge almost surely to the corresponding fixed point $Q^*_{\max}$, $Q^{\lambda}_{\mathrm{lse}}$, or $Q^{\lambda}_{\mathrm{mm}}$; see Section K and Theorem 11. For the Boltzmann softmax operator, the induced Bellman operator need not be contractive. In this case, the ODE and the stochastic iterates are guaranteed to approach the invariant error set

$$\mathcal{H} := \left\{ Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : \|Q - Q^*_{\max}\|_\infty \le \frac{\gamma\lambda \ln(|\mathcal{A}|) w_{\max}}{(1-\gamma) w_{\min}} \right\}.$$

Consequently, the limiting error shrinks as the temperature parameter λ decreases to zero; see Section K and Theorem 12.

Conclusion

This paper presents a unified ODE framework for analyzing convergence of Q-learning and its smooth variants. A smooth polynomial Lyapunov function treats positive-diagonal mean fields arising from asynchronous updates, while $D = I$ represents the separately defined synchronous full-update mean field. The framework also complements switching-system analyses of asynchronous Q-learning by avoiding restrictive structural conditions on the ODE model.

Reproducibility Statement

This paper primarily makes theoretical contributions. The assumptions and main results are stated in the main text, and detailed proofs and supporting analyses are provided in Sections A to E and Section K. The numerical examples and simulations are simple illustrations of the theoretical results rather than large-scale empirical evaluations. They can be readily reconstructed from the small finite-state MDP and the simulation settings described in the appendix entitled "Examples."

AI Use Statement

Large language model (LLM) tools were used only to aid or polish the writing and for retrieval and discovery (e.g., finding related work).

References

Kavosh Asadi and Michael L. Littman. An alternative softmax operator for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 243–252. PMLR, 2017.

David Barber. Smoothed Q-learning, 2023. arXiv preprint arXiv:2303.08631.

Carolyn L. Beck and Rayadurgam Srikant. Error bounds for constant step-size Q-learning. *Systems & Control Letters*, 61(12):1203–1208, 2012. doi: 10.1016/j.sysconle.2012.08.014.

Shalabh Bhatnagar, H. L. Prasad, and L. A. Prashanth. *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods*, volume 434 of *Lecture Notes in Control and Information Sciences*. Springer London, London, 2013. doi: 10.1007/978-1-4471-4285-0.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48 of *Texts and Readings in Mathematics*. Hindustan Book Agency, Gurgaon, 2008. doi: 10.1007/978-93-86279-38-5.

Vivek S. Borkar and Sean P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000. doi: 10.1137/S0363012997331639.

Vivek S. Borkar and K. Soumyanath. An analog scheme for fixed-point computation. I. theory. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44(4):351–355, 1997. doi: 10.1109/81.563625.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. In *Advances in Neural Information Processing Systems 33*, pp. 8223–8234. Curran Associates, Inc., 2020.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A Lyapunov theory for finite-sample guarantees of markovian stochastic approximation. *Operations Research*, 72(4):1352–1367, 2024. doi: 10.1287/opre.2022.0249.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1125–1134. PMLR, 2018.

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

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning, 2017. arXiv preprint arXiv:1704.00805.

Sina Ghiassian, Andrew Patterson, Shivam Garg, Dhawal Gupta, Adam White, and Martha White. Gradient temporal-difference learning with regularized corrections. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3524–3534. PMLR, 2020.

Thomas Hakon Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919. doi: 10.2307/1967124.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 2017.

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994. doi: 10.1162/neco.1994.6.6.1185.

Narim Jeong and Donghwan Lee. Unified finite-time error analysis of soft Q-learning. *Neurocomputing*, 626:129582, 2025. doi: 10.1016/j.neucom.2025.129582.

Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3 edition, 2002.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Stochastic Modelling and Applied Probability*. Springer New York, New York, 2 edition, 2003. doi: 10.1007/b97441.

Donghwan Lee. Final iteration convergence bound of Q-learning: Switching system approach. *IEEE Transactions on Automatic Control*, 69(7):4765–4772, 2024. doi: 10.1109/TAC.2024.3355326.

Donghwan Lee and Niao He. A unified switching system perspective and convergence analysis of Q-learning algorithms. In *Advances in Neural Information Processing Systems 33*, pp. 15556–15567. Curran Associates, Inc., 2020.

Donghwan Lee, Han-Dong Lim, Jihoon Park, and Okyong Choi. New versions of gradient temporal-difference learning. *IEEE Transactions on Automatic Control*, 68(8):5006–5013, 2023. doi: 10.1109/TAC.2022.3213763.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473, 2022. doi: 10.1109/TIT.2021.3120096.

Daniel Liberzon. *Switching in Systems and Control*. Systems & Control: Foundations & Applications. Birkhäuser, Boston, MA, 2003. doi: 10.1007/978-1-4612-0017-8.

Han-Dong Lim and Donghwan Lee. Finite-time analysis of asynchronous Q-learning under diminishing step-size from control-theoretic view. *IEEE Access*, 12:149916–149939, 2024. doi: 10.1109/ACCESS.2024.3476564.

Francisco S. Melo, Sean P. Meyn, and M. Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 664–671. ACM, 2008. doi: 10.1145/1390156.1390240.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236.

Ling Pan, Qingpeng Cai, Qi Meng, Wei Chen, and Longbo Huang. Reinforcement learning with dynamic Boltzmann softmax updates. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 1992–1998. International Joint Conferences on Artificial Intelligence Organization, 2020. doi: 10.24963/ijcai.2020/276.

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3185–3205. PMLR, 2020.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

Satinder Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3):287–308, 2000. doi: 10.1023/A:1007678930559.

Zhao Song, Ron Parr, and Lawrence Carin. Revisiting the softmax Bellman operator: New benefits and new perspective. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5916–5925. PMLR, 2019.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Richard S. Sutton, Csaba Szepesvári, and Hamid Reza Maei. A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximation. In *Advances in Neural Information Processing Systems 21*, pp. 1609–1616. Curran Associates, Inc., 2008.

Richard S. Sutton, Hamid R. Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 993–1000. ACM, 2009. doi: 10.1145/1553374.1553501.

Csaba Szepesvári. The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems 10*, pp. 1064–1070. MIT Press, 1998.

John N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16 (3):185–202, 1994. doi: 10.1023/A:1022689125041.

Martin J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning, 2019. arXiv preprint arXiv:1905.06265.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992. doi: 10.1007/BF00992698.

Appendix

A Proof of Lemma 1

Restatement of Lemma 1 (Norm comparisons).

For any $x \in \mathbb{R}^n$ and any $p \in \mathbb{N}$, we have

1. $w_{\min}^{1/(2p)} \|x\|_\infty \le \|x\|_{2p,w} \le n^{1/(2p)} w_{\max}^{1/(2p)} \|x\|_\infty$;
2. $w_{\min}^{1/(2p)} \|x\|_{2p} \le \|x\|_{2p,w} \le w_{\max}^{1/(2p)} \|x\|_{2p}$;
3. $\|x\|_\infty \le \|x\|_{2p} \le n^{1/(2p)} \|x\|_\infty$,

where $w_{\min} := \min_i w_i$ and $w_{\max} := \max_i w_i$.

Proof of Lemma 1. For the first statement, the lower bound follows from

$$\|x\|_{2p,w} \ge (w_j {x_j}^{2p})^{1/(2p)} = w_j^{1/(2p)} |x_j| \ge w_{\min}^{1/(2p)} |x_j|,$$

for any $j \in \{1, 2, \dots, n\}$. By taking the maximum over j, we have $\|x\|_{2p,w} \ge w_{\min}^{1/(2p)} \|x\|_\infty$. For the upper bound, we have

$$\begin{aligned}\|x\|_{2p,w} &= \left(\sum_{i=1}^n w_i {x_i}^{2p}\right)^{1/(2p)} \le \left(\sum_{i=1}^n w_i \|x\|_\infty^{2p}\right)^{1/(2p)} = \|x\|_\infty \left(\sum_{i=1}^n w_i\right)^{1/(2p)} \\ &\le \|x\|_\infty (n w_{\max})^{1/(2p)} = n^{1/(2p)} w_{\max}^{1/(2p)} \|x\|_\infty.\end{aligned}$$

The second statement follows from the inequalities

$$\begin{aligned}\|x\|_{2p,w} &= \left(\sum_{i=1}^n w_i x_i^{2p}\right)^{1/(2p)} \\ &\le \left(w_{\max} \sum_{i=1}^n x_i^{2p}\right)^{1/(2p)} \\ &= w_{\max}^{1/(2p)} \left(\sum_{i=1}^n x_i^{2p}\right)^{1/(2p)} = w_{\max}^{1/(2p)} \|x\|_{2p}\end{aligned}$$

and

$$\begin{aligned}\|x\|_{2p,w} &= \left(\sum_{i=1}^n w_i x_i^{2p}\right)^{1/(2p)} \\ &\ge \left(w_{\min} \sum_{i=1}^n x_i^{2p}\right)^{1/(2p)} \\ &= w_{\min}^{1/(2p)} \left(\sum_{i=1}^n x_i^{2p}\right)^{1/(2p)} = w_{\min}^{1/(2p)} \|x\|_{2p}.\end{aligned}$$

For the third statement, the largest coordinate satisfies

$$\|x\|_\infty^{2p} = \max_{1 \le i \le n} x_i^{2p} \le \sum_{i=1}^n x_i^{2p} \le n \max_{1 \le i \le n} x_i^{2p} = n \|x\|_\infty^{2p}. \tag{15}$$

Taking the $2p$-th root of every term in (15) gives $\|x\|_\infty \le \|x\|_{2p} \le n^{1/(2p)} \|x\|_\infty$. This completes the proof. □

B Proof of Theorem 1

Restatement of Theorem 1 (Polynomial contractive certificate).

Let x_t be the unique solution of (3). For convenience, the equation referred to as (3) is repeated here:

$$\frac{d}{dt}x_t = D\big(F(x_t) - x_t\big), \qquad t \geq 0, \qquad x_0 \in \mathbb{R}^n.$$

Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is an ∞-norm contraction: for some $\alpha \in (0, 1)$,

$$\|F(x) - F(y)\|_\infty \leq \alpha\|x - y\|_\infty, \qquad x, y \in \mathbb{R}^n.$$

Let x^* be its unique fixed point, and choose $w_i = 1/d_i$. For any $p \in \mathbb{N}$ satisfying

$$\left\lceil \frac{\ln n}{\ln(\alpha^{-1})} \right\rceil < 2p,$$

we have

$$\begin{aligned}\frac{d}{dt}W_p(x_t - x^*) &\leq -\left(1 - \alpha n^{1/(2p)}\right)\|x_t - x^*\|_{2p}^{2p} \\ &\leq -\frac{2p\left(1 - \alpha n^{1/(2p)}\right)}{w_{\max}}W_p(x_t - x^*),\end{aligned}$$

and hence

$$W_p(x_t - x^*) \leq \exp\left(-\frac{2p\left(1 - \alpha n^{1/(2p)}\right)}{w_{\max}}t\right)W_p(x_0 - x^*), \qquad t \geq 0,$$

where $w_{\max} := \max_{i\in\{1,2,\ldots,n\}} w_i$. Therefore, W_p is a globally smooth polynomial Lyapunov function and x^* is the unique globally exponentially stable equilibrium.

Before estimating the Lyapunov derivative, we make the well-posedness and equilibrium arguments explicit. Since $(\mathbb{R}^n, \|\cdot\|_\infty)$ is complete and F is a contraction, the Banach fixed-point theorem guarantees a unique fixed point x^*. Moreover, every contraction is globally Lipschitz, so the vector field $x \mapsto D(F(x) - x)$ is globally Lipschitz. Hence, Lemma 7 guarantees that (3) has a unique global solution for every initial state. Since $F(x^*) = x^*$, subtracting the constant equilibrium from (3) gives the exact error dynamics

$$\frac{d}{dt}(x_t - x^*) = D\big(F(x_t) - F(x^*) - (x_t - x^*)\big). \tag{16}$$

For the componentwise calculations below, $F_i(x)$ denotes the ith component of $F(x)$, so $F(x) = (F_1(x), \ldots, F_n(x))^\top$. The following three lemmas establish the polynomial comparison, derivative, and integration steps separately.

Lemma 2 (Two-sided comparison for W_p)**.** *Let* $w_{\min} := \min_{1\leq i\leq n} w_i$ *and* $w_{\max} := \max_{1\leq i\leq n} w_i$. *For every* $z \in \mathbb{R}^n$,

$$\frac{w_{\min}}{2p}\|z\|_{2p}^{2p} \leq W_p(z) \leq \frac{w_{\max}}{2p}\|z\|_{2p}^{2p}. \tag{17}$$

Proof. The main-text definition in (2) is repeated here for convenience:

$$\|z\|_{2p,w} := \left(\sum_{i=1}^n w_i z_i^{2p}\right)^{1/(2p)}, \qquad W_p(z) := \frac{1}{2p}\|z\|_{2p,w}^{2p}.$$

Because $2p$ is even, every z_i^{2p} is nonnegative, and the coordinate form is

$$W_p(z) = \frac{1}{2p}\sum_{i=1}^{n} w_i z_i^{2p}. \tag{18}$$

Using $w_{\min} \le w_i \le w_{\max}$ in (18) gives

$$\begin{aligned}\frac{w_{\min}}{2p}\|z\|_{2p}^{2p} = \frac{w_{\min}}{2p}\sum_{i=1}^{n} z_i^{2p} &\le \frac{1}{2p}\sum_{i=1}^{n} w_i z_i^{2p} \\ = W_p(z) &\le \frac{w_{\max}}{2p}\sum_{i=1}^{n} z_i^{2p} = \frac{w_{\max}}{2p}\|z\|_{2p}^{2p},\end{aligned}$$

which proves (17). □

Lemma 3 (Derivative estimate for Theorem 1)**.** *Under the assumptions of Theorem 1,*

$$\begin{aligned}\frac{d}{dt}W_p(x_t - x^*) &\le -\left(1 - \alpha n^{1/(2p)}\right)\|x_t - x^*\|_{2p}^{2p} \\ &\le -\frac{2p\left(1 - \alpha n^{1/(2p)}\right)}{w_{\max}} W_p(x_t - x^*).\end{aligned}$$

Proof. Differentiating the polynomial coordinate form in (18) with respect to z_i gives

$$\frac{\partial W_p(z)}{\partial z_i} = \frac{w_i}{2p}\frac{d}{dz_i} z_i^{2p} = w_i z_i^{2p-1}, \qquad i \in \{1, 2, \dots, n\}, \tag{19}$$

and hence

$$\nabla_z W_p(z) = \begin{bmatrix} w_1 z_1^{2p-1} \\ w_2 z_2^{2p-1} \\ \vdots \\ w_n z_n^{2p-1} \end{bmatrix} = D^{-1}\begin{bmatrix} z_1^{2p-1} \\ z_2^{2p-1} \\ \vdots \\ z_n^{2p-1} \end{bmatrix}.$$

Applying the chain rule using (19) and then applying the error dynamics in (16) coordinatewise yields

$$\begin{aligned}\frac{d}{dt}W_p(x_t - x^*) &= \sum_{i=1}^{n} w_i (x_{t,i} - x_i^*)^{2p-1}\dot{x}_{t,i} \\ &= \sum_{i=1}^{n} w_i d_i (x_{t,i} - x_i^*)^{2p-1} \\ &\quad \times \big(F_i(x_t) - F_i(x^*) - (x_{t,i} - x_i^*)\big).\end{aligned} \tag{20}$$

Using $w_i d_i = 1$ in (20) gives

$$\begin{aligned}\frac{d}{dt}W_p(x_t - x^*) = &\sum_{i=1}^{n} (x_{t,i} - x_i^*)^{2p-1}\big(F_i(x_t) - F_i(x^*)\big) \\ &- \sum_{i=1}^{n} (x_{t,i} - x_i^*)^{2p}.\end{aligned} \tag{21}$$

The exponents $2p/(2p-1)$ and $2p$ are conjugate because

$$\frac{1}{2p/(2p-1)} + \frac{1}{2p} = \frac{2p-1}{2p} + \frac{1}{2p} = 1.$$

Moreover, $(2p-1)\,2p/(2p-1) = 2p$. Therefore, Hölder's inequality and the fact that $2p$ is even give

$$\begin{aligned}
&\left|\sum_{i=1}^{n}(x_{t,i}-x_i^*)^{2p-1}\big(F_i(x_t)-F_i(x^*)\big)\right| \\
&\quad\le \sum_{i=1}^{n}|x_{t,i}-x_i^*|^{2p-1}|F_i(x_t)-F_i(x^*)| \\
&\quad\le \left(\sum_{i=1}^{n}\big(|x_{t,i}-x_i^*|^{2p-1}\big)^{\frac{2p}{2p-1}}\right)^{\frac{2p-1}{2p}} \\
&\qquad\times\left(\sum_{i=1}^{n}\big(F_i(x_t)-F_i(x^*)\big)^{2p}\right)^{\frac{1}{2p}} \\
&\quad= \left(\sum_{i=1}^{n}(x_{t,i}-x_i^*)^{2p}\right)^{\frac{2p-1}{2p}}\left(\sum_{i=1}^{n}\big(F_i(x_t)-F_i(x^*)\big)^{2p}\right)^{\frac{1}{2p}} \\
&\quad= \|x_t-x^*\|_{2p}^{2p-1}\|F(x_t)-F(x^*)\|_{2p}. \qquad (22)
\end{aligned}$$

The finite-dimensional norm comparison in Lemma 1 and the contraction assumption further give

$$\begin{aligned}
\|F(x_t)-F(x^*)\|_{2p} &\le n^{1/(2p)}\|F(x_t)-F(x^*)\|_\infty \\
&\le \alpha n^{1/(2p)}\|x_t-x^*\|_\infty \\
&\le \alpha n^{1/(2p)}\|x_t-x^*\|_{2p}. \qquad (23)
\end{aligned}$$

Using (22) in (21), and then using (23), yields the complete chain

$$\begin{aligned}
\frac{d}{dt}W_p(x_t-x^*) &\le \left|\sum_{i=1}^{n}(x_{t,i}-x_i^*)^{2p-1}\big(F_i(x_t)-F_i(x^*)\big)\right| - \sum_{i=1}^{n}(x_{t,i}-x_i^*)^{2p} \\
&\le \|x_t-x^*\|_{2p}^{2p-1}\|F(x_t)-F(x^*)\|_{2p} - \|x_t-x^*\|_{2p}^{2p} \\
&\le \|x_t-x^*\|_{2p}^{2p-1}\left(\alpha n^{1/(2p)}\|x_t-x^*\|_{2p}\right) - \|x_t-x^*\|_{2p}^{2p} \\
&= \alpha n^{1/(2p)}\|x_t-x^*\|_{2p}^{2p} - \|x_t-x^*\|_{2p}^{2p} \\
&= -\left(1-\alpha n^{1/(2p)}\right)\|x_t-x^*\|_{2p}^{2p}. \qquad (24)
\end{aligned}$$

This proves the first inequality in (4), which is repeated here:

$$\frac{d}{dt}W_p(x_t-x^*) \le -\left(1-\alpha n^{1/(2p)}\right)\|x_t-x^*\|_{2p}^{2p}.$$

The upper estimate in Lemma 2, applied with $z = x_t - x^*$, gives

$$W_p(x_t-x^*) \le \frac{w_{\max}}{2p}\|x_t-x^*\|_{2p}^{2p}.$$

Equivalently,

$$\|x_t-x^*\|_{2p}^{2p} \ge \frac{2p}{w_{\max}}W_p(x_t-x^*). \qquad (25)$$

Moreover, the stated condition on p implies

$$2p > \frac{\ln n}{\ln(\alpha^{-1})}, \qquad \frac{\ln n}{2p} < \ln(\alpha^{-1}), \qquad \alpha n^{1/(2p)} = \exp\left(\ln\alpha + \frac{\ln n}{2p}\right) < 1. \qquad (26)$$

Thus, $1-\alpha n^{1/(2p)} > 0$. Because this coefficient is positive, using (25) in the negative term of (24) gives

$$\begin{aligned}\frac{d}{dt}W_p(x_t-x^*) &\le -\left(1-\alpha n^{1/(2p)}\right)\|x_t-x^*\|_{2p}^{2p}\\ &\le -\left(1-\alpha n^{1/(2p)}\right)\frac{2p}{w_{\max}}W_p(x_t-x^*)\\ &= -\frac{2p\left(1-\alpha n^{1/(2p)}\right)}{w_{\max}}W_p(x_t-x^*). \end{aligned} \tag{27}$$

This proves the second inequality in (4), which is repeated here:

$$\frac{d}{dt}W_p(x_t-x^*) \le -\frac{2p\left(1-\alpha n^{1/(2p)}\right)}{w_{\max}}W_p(x_t-x^*).$$

□

Lemma 4. *Under the assumptions of Theorem 1, let* $w_{\min} := \min_i w_i$. *Then*

$$W_p(x_t-x^*) \le \exp\left(-\frac{2p(1-\alpha n^{1/(2p)})}{w_{\max}}t\right)W_p(x_0-x^*), \qquad t\ge 0,$$

and

$$\|x_t-x^*\|_{2p} \le \left(\frac{w_{\max}}{w_{\min}}\right)^{1/(2p)}\exp\left(-\frac{1-\alpha n^{1/(2p)}}{w_{\max}}t\right)\|x_0-x^*\|_{2p}, \qquad t\ge 0.$$

Proof. By (26),

$$\frac{2p\left(1-\alpha n^{1/(2p)}\right)}{w_{\max}} > 0. \tag{28}$$

Since $W_p(x_t-x^*)$ is nonnegative and continuously differentiable, (27) can be written as

$$\frac{d}{dt}W_p(x_t-x^*) \le -\frac{2p\left(1-\alpha n^{1/(2p)}\right)}{w_{\max}}W_p(x_t-x^*). \tag{29}$$

Applying Grönwall's inequality to (29) with zero additive term gives

$$W_p(x_t-x^*) \le \exp\left(-\frac{2p\left(1-\alpha n^{1/(2p)}\right)}{w_{\max}}t\right)W_p(x_0-x^*), \qquad t\ge 0. \tag{30}$$

This is the main-text estimate in (5). Applying the lower estimate in Lemma 2 at time t, the bound in (30), and the upper estimate in Lemma 2 at time 0 gives

$$\begin{aligned}\frac{w_{\min}}{2p}\|x_t-x^*\|_{2p}^{2p} &\le W_p(x_t-x^*)\\ &\le \exp\left(-\frac{2p(1-\alpha n^{1/(2p)})}{w_{\max}}t\right)W_p(x_0-x^*)\\ &\le \exp\left(-\frac{2p(1-\alpha n^{1/(2p)})}{w_{\max}}t\right)\frac{w_{\max}}{2p}\|x_0-x^*\|_{2p}^{2p}. \end{aligned} \tag{31}$$

Multiplying (31) by $2p/w_{\min}$ and taking the positive $2p$th root yields

$$\|x_t-x^*\|_{2p} \le \left(\frac{w_{\max}}{w_{\min}}\right)^{1/(2p)}\exp\left(-\frac{1-\alpha n^{1/(2p)}}{w_{\max}}t\right)\|x_0-x^*\|_{2p}. \tag{32}$$

Because D is nonsingular, every equilibrium of (3) satisfies $F(x)=x$. The contraction property gives the unique fixed point x^*, and (32) proves global exponential stability. □

Proof of Theorem 1. The two-sided polynomial comparison is proved in Lemma 2. The derivative inequalities stated in the theorem follow from Lemma 3, and the exponential estimates follow from Lemma 4. Since D is nonsingular, every equilibrium of the ODE is a fixed point of F, and the contraction property makes x^* the unique equilibrium. This proves the theorem. □

C Proof of Theorem 2

Restatement of Theorem 2 (Polynomial perturbation certificate).

Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is globally Lipschitz continuous, and let x_t be the unique global solution of (3). For convenience, the equation referred to as (3) is repeated here:

$$\frac{d}{dt}x_t = D\big(F(x_t) - x_t\big), \qquad t \geq 0, \qquad x_0 \in \mathbb{R}^n.$$

Suppose that $F' : \mathbb{R}^n \to \mathbb{R}^n$ is an ∞-norm contraction with factor $\alpha \in (0, 1)$ and fixed point x^*:

$$\|F'(x) - F'(y)\|_\infty \leq \alpha\|x - y\|_\infty, \qquad x, y \in \mathbb{R}^n.$$

Suppose also that F satisfies

$$\|F(x) - F'(x)\|_\infty \leq \eta, \qquad x \in \mathbb{R}^n,$$

for some $\eta > 0$. Choose $w_i = 1/d_i$, and let $w_{\min} := \min_i w_i$ and $w_{\max} := \max_i w_i$. For any $p \in \mathbb{N}$ satisfying

$$\left\lceil \frac{\ln n}{\ln(\alpha^{-1})} \right\rceil < 2p,$$

we have

$$\frac{d}{dt}W_p(x_t - x^*) \leq -\left(1 - \alpha n^{1/(2p)}\right)\|x_t - x^*\|_{2p}^{2p} + n^{1/(2p)}\eta\|x_t - x^*\|_{2p}^{2p-1}.$$

Moreover,

$$\|x_t - x^*\|_{2p,w} \leq \|x_0 - x^*\|_{2p,w}\exp\left(\frac{\alpha n^{1/(2p)} - 1}{w_{\max}}t\right) + \frac{n^{1/(2p)}\eta w_{\max}}{\left(1 - \alpha n^{1/(2p)}\right) w_{\min}^{(2p-1)/(2p)}}, \qquad t \geq 0.$$

In particular, $W_p(x_t - x^*)$ decreases whenever $\|x_t - x^*\|_{2p} > n^{1/(2p)}\eta/(1 - \alpha n^{1/(2p)})$.

Lemma 5 (Derivative estimate for Theorem 2)**.** *Under the assumptions of Theorem 2,*

$$\frac{d}{dt}W_p(x_t - x^*) \leq -\left(1 - \alpha n^{1/(2p)}\right)\|x_t - x^*\|_{2p}^{2p} + n^{1/(2p)}\eta\|x_t - x^*\|_{2p}^{2p-1}.$$

Proof. For the componentwise calculations, $F_i(x)$ and $F_i'(x)$ denote the ith components of $F(x)$ and $F'(x)$, respectively. The coordinate derivative in (19), the chain rule, and the ODE repeated in the theorem restatement give

$$\begin{aligned}\frac{d}{dt}W_p(x_t - x^*) &= \sum_{i=1}^n w_i(x_{t,i} - x_i^*)^{2p-1}\dot{x}_{t,i} \\ &= \sum_{i=1}^n (x_{t,i} - x_i^*)^{2p-1}\big(F_i(x_t) - x_{t,i}\big),\end{aligned} \tag{33}$$

where $w_i d_i = 1$ was used in the second equality. Since $F'(x^*) = x^*$,

$$F_i(x_t) - x_{t,i} = F_i'(x_t) - F_i'(x^*) + F_i(x_t) - F_i'(x_t) - (x_{t,i} - x_i^*). \tag{34}$$

Substituting (34) into (33) yields

$$\frac{d}{dt}W_p(x_t - x^*) = \sum_{i=1}^{n}(x_{t,i} - x_i^*)^{2p-1}\big(F_i'(x_t) - F_i'(x^*)\big) + \sum_{i=1}^{n}(x_{t,i} - x_i^*)^{2p-1}\big(F_i(x_t) - F_i'(x_t)\big) - \sum_{i=1}^{n}(x_{t,i} - x_i^*)^{2p}. \tag{35}$$

Applying Hölder's inequality to the first sum in (35) gives

$$\left|\sum_{i=1}^{n}(x_{t,i} - x_i^*)^{2p-1}\big(F_i'(x_t) - F_i'(x^*)\big)\right| \le \|x_t - x^*\|_{2p}^{2p-1}\|F'(x_t) - F'(x^*)\|_{2p}. \tag{36}$$

The same conjugate exponents applied to the second sum in (35) give

$$\left|\sum_{i=1}^{n}(x_{t,i} - x_i^*)^{2p-1}\big(F_i(x_t) - F_i'(x_t)\big)\right| \le \|x_t - x^*\|_{2p}^{2p-1}\|F(x_t) - F'(x_t)\|_{2p}. \tag{37}$$

Using (36) and (37) in (35) yields

$$\frac{d}{dt}W_p(x_t - x^*) \le \|x_t - x^*\|_{2p}^{2p-1}\|F'(x_t) - F'(x^*)\|_{2p} + \|x_t - x^*\|_{2p}^{2p-1}\|F(x_t) - F'(x_t)\|_{2p} - \|x_t - x^*\|_{2p}^{2p}. \tag{38}$$

The contraction assumption and Lemma 1 imply

$$\begin{aligned}\|F'(x_t) - F'(x^*)\|_{2p} &\le n^{1/(2p)}\|F'(x_t) - F'(x^*)\|_{\infty}\\ &\le \alpha n^{1/(2p)}\|x_t - x^*\|_{\infty}\\ &\le \alpha n^{1/(2p)}\|x_t - x^*\|_{2p},\end{aligned} \tag{39}$$

whereas the uniform perturbation bound gives

$$\begin{aligned}\|F(x_t) - F'(x_t)\|_{2p} &\le n^{1/(2p)}\|F(x_t) - F'(x_t)\|_{\infty}\\ &\le n^{1/(2p)}\eta.\end{aligned} \tag{40}$$

Substituting (39) and (40) into (38) yields

$$\frac{d}{dt}W_p(x_t - x^*) \le -\left(1 - \alpha n^{1/(2p)}\right)\|x_t - x^*\|_{2p}^{2p} + n^{1/(2p)}\eta\|x_t - x^*\|_{2p}^{2p-1}. \tag{41}$$

Equation (41) proves the main-text estimate in (6), which is repeated here:

$$\frac{d}{dt}W_p(x_t - x^*) \le -\left(1 - \alpha n^{1/(2p)}\right)\|x_t - x^*\|_{2p}^{2p} + n^{1/(2p)}\eta\|x_t - x^*\|_{2p}^{2p-1}.$$

□

******

Lemma 6 (Scalar comparison for Theorem 2)**.** *Under the assumptions of Theorem 2,*

$$\|x_t - x^*\|_{2p,w} \le \|x_0 - x^*\|_{2p,w}\exp\left(-\frac{1 - \alpha n^{1/(2p)}}{w_{\max}}t\right) + \frac{n^{1/(2p)}\eta w_{\max}}{\left(1 - \alpha n^{1/(2p)}\right)w_{\min}^{(2p-1)/(2p)}}, \qquad t \ge 0.$$

Moreover, $W_p(x_t - x^*)$ is strictly decreasing whenever $\|x_t - x^*\|_{2p} > n^{1/(2p)}\eta/(1 - \alpha n^{1/(2p)})$.

Proof. As shown in (26), the condition on p ensures that $1-\alpha n^{1/(2p)} > 0$. For $x_t \neq x^*$, the identity

$$W_p(x_t - x^*) = \frac{1}{2p}\|x_t - x^*\|_{2p,w}^{2p} \tag{42}$$

gives, by differentiation,

$$\frac{d}{dt}W_p(x_t - x^*) = \|x_t - x^*\|_{2p,w}^{2p-1}\frac{d}{dt}\|x_t - x^*\|_{2p,w}. \tag{43}$$

The weighted and unweighted norms satisfy

$$\begin{aligned}\|x_t - x^*\|_{2p}^{2p} &\geq \frac{1}{w_{\max}}\|x_t - x^*\|_{2p,w}^{2p},\\ \|x_t - x^*\|_{2p}^{2p-1} &\leq \frac{1}{w_{\min}^{(2p-1)/(2p)}}\|x_t - x^*\|_{2p,w}^{2p-1}.\end{aligned} \tag{44}$$

Using (44) in (41) and dividing according to (43) gives

$$\frac{d}{dt}\|x_t - x^*\|_{2p,w} \leq -\frac{1-\alpha n^{1/(2p)}}{w_{\max}}\|x_t - x^*\|_{2p,w} + \frac{n^{1/(2p)}\eta}{w_{\min}^{(2p-1)/(2p)}}. \tag{45}$$

The norm trajectory is nonnegative and absolutely continuous on every finite time interval. At almost every time at which it is zero, its derivative is also zero; therefore, (45) holds almost everywhere. These properties justify applying Grönwall's inequality, including when the trajectory passes through x^*:

$$\begin{aligned}\|x_t - x^*\|_{2p,w} &\leq \exp\left(-\frac{1-\alpha n^{1/(2p)}}{w_{\max}}t\right)\|x_0 - x^*\|_{2p,w}\\ &\quad + \frac{n^{1/(2p)}\eta}{w_{\min}^{(2p-1)/(2p)}}\int_0^t \exp\left(-\frac{1-\alpha n^{1/(2p)}}{w_{\max}}(t-\tau)\right)d\tau\\ &= \exp\left(-\frac{1-\alpha n^{1/(2p)}}{w_{\max}}t\right)\|x_0 - x^*\|_{2p,w}\\ &\quad + \frac{n^{1/(2p)}\eta w_{\max}}{\left(1-\alpha n^{1/(2p)}\right)w_{\min}^{(2p-1)/(2p)}}\left[1-\exp\left(-\frac{1-\alpha n^{1/(2p)}}{w_{\max}}t\right)\right].\end{aligned} \tag{46}$$

Since the expression in square brackets in (46) lies in $[0,1]$, it follows that

$$\begin{aligned}\|x_t - x^*\|_{2p,w} &\leq \exp\left(-\frac{1-\alpha n^{1/(2p)}}{w_{\max}}t\right)\|x_0 - x^*\|_{2p,w}\\ &\quad + \frac{n^{1/(2p)}\eta w_{\max}}{\left(1-\alpha n^{1/(2p)}\right)w_{\min}^{(2p-1)/(2p)}}, \qquad t \geq 0.\end{aligned} \tag{47}$$

This is the main-text estimate in (7). Finally, factoring (41) gives

$$\frac{d}{dt}W_p(x_t - x^*) \leq \|x_t - x^*\|_{2p}^{2p-1}\left[-\left(1-\alpha n^{1/(2p)}\right)\|x_t - x^*\|_{2p} + n^{1/(2p)}\eta\right]. \tag{48}$$

The sign of the bracket in (48) proves the stated strict-decrease condition. □

Proof of Theorem 2. The derivative statement follows from Lemma 5. The weighted-norm estimate and the strict-decrease condition follow from Lemma 6. This proves the theorem. □

D Proof of Theorem 3

Restatement of Theorem 3.

Consider (8) with $H \in \{H_{\max}, H^{\lambda}_{\mathrm{lse}}, H^{\lambda}_{\mathrm{mm}}\}$, and let Q_t be its solution. The referenced ODE is repeated here:

$$\frac{d}{dt}Q_t = DR + \gamma DPH(Q_t) - DQ_t, \qquad t \geq 0, \qquad Q_0 \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}.$$

Let Q_e denote the corresponding fixed point in (10). The fixed-point relations in (10) are repeated here:

$$Q^*_{\max} = F_{\max}(Q^*_{\max}), \qquad Q^{\lambda}_{\mathrm{lse}} = F^{\lambda}_{\mathrm{lse}}(Q^{\lambda}_{\mathrm{lse}}), \qquad Q^{\lambda}_{\mathrm{mm}} = F^{\lambda}_{\mathrm{mm}}(Q^{\lambda}_{\mathrm{mm}}).$$

Set $n = |\mathcal{S}\times\mathcal{A}|$, choose $w_i = 1/d_i$, and let $w_{\max} := \max_i w_i$. For any $p \in \mathbb{N}$ satisfying

$$\left\lceil \frac{\ln n}{\ln(\gamma^{-1})} \right\rceil < 2p,$$

we have

$$W_p(Q_t - Q_e) \leq \exp\left(-\frac{2p(1-\gamma n^{1/(2p)})}{w_{\max}}t\right) W_p(Q_0 - Q_e), \qquad t \geq 0.$$

Therefore, Q_e is the unique globally exponentially stable equilibrium of the corresponding ODE.

Proof of Theorem 3. Fix $H \in \{H_{\max}, H^{\lambda}_{\mathrm{lse}}, H^{\lambda}_{\mathrm{mm}}\}$ and let F be the corresponding Bellman operator in (9). By Lemma 11, F is a γ-contraction in the ∞-norm. Hence, the Banach fixed-point theorem guarantees the unique fixed point Q_e specified in the theorem.

Moreover, the condition on p implies

$$\gamma n^{1/(2p)} < 1.$$

Therefore, applying Theorem 1 with $x_t = Q_t$, $x^* = Q_e$, and $\alpha = \gamma$ gives

$$W_p(Q_t - Q_e) \leq \exp\left(-\frac{2p\left(1-\gamma n^{1/(2p)}\right)}{w_{\max}}t\right) W_p(Q_0 - Q_e), \qquad t \geq 0. \tag{49}$$

Since D is nonsingular, the equilibria of (9) coincide with the fixed points of F. Thus, Q_e is the unique globally exponentially stable equilibrium. □

E Proof of Theorem 4

Restatement of Theorem 4.

Consider (8) with $H = H^{\lambda}_{\mathrm{bz}}$, and let Q_t be its solution. The referenced ODE is repeated here:

$$\frac{d}{dt}Q_t = DR + \gamma DPH^{\lambda}_{\mathrm{bz}}(Q_t) - DQ_t, \qquad t \geq 0, \qquad Q_0 \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}.$$

Set $n = |\mathcal{S}\times\mathcal{A}|$, choose $w_i = 1/d_i$, and let $w_{\min} := \min_i w_i$ and $w_{\max} := \max_i w_i$. For any $p \in \mathbb{N}$ satisfying

$$\left\lceil \frac{\ln n}{\ln(\gamma^{-1})} \right\rceil < 2p,$$

we have

$$\begin{aligned}\frac{d}{dt}W_p(Q_t - Q^*_{\max}) \leq & -\left(1-\gamma n^{1/(2p)}\right)\|Q_t - Q^*_{\max}\|^{2p}_{2p} \\ & + n^{1/(2p)}\gamma\lambda\ln(|\mathcal{A}|)\|Q_t - Q^*_{\max}\|^{2p-1}_{2p},\end{aligned}$$

and

$$\|Q_t - Q^*_{\max}\|_{2p,w} \le \|Q_0 - Q^*_{\max}\|_{2p,w} \exp\left(\frac{\gamma n^{1/(2p)} - 1}{w_{\max}} t\right) + \frac{n^{1/(2p)} w_{\max} \gamma\lambda \ln(|\mathcal{A}|)}{\left(1 - \gamma n^{1/(2p)}\right) w_{\min}^{(2p-1)/(2p)}}, \qquad t \ge 0.$$

Proof of Theorem 4. If $|\mathcal{A}| = 1$, then $H^\lambda_{\mathrm{bz}} = H_{\max}$ and $F^\lambda_{\mathrm{bz}} = F_{\max}$. The derivative estimate follows from Theorem 1 with $\alpha = \gamma$. Taking the positive $2p$th root of (5) and using (42) gives the stated weighted-norm bound with zero additive term. Thus, it remains to consider $|\mathcal{A}| \ge 2$, for which $\gamma\lambda \ln(|\mathcal{A}|) > 0$.

By Lemma 15, F^λ_{bz} is globally Lipschitz continuous. By Lemma 14, the Boltzmann and max Bellman operators satisfy

$$\|F^\lambda_{\mathrm{bz}}(Q) - F_{\max}(Q)\|_\infty \le \gamma\lambda \ln(|\mathcal{A}|), \qquad Q \in \mathbb{R}^n. \tag{50}$$

Moreover, Lemma 11 gives

$$\|F_{\max}(Q) - F_{\max}(Q')\|_\infty \le \gamma \|Q - Q'\|_\infty, \qquad Q, Q' \in \mathbb{R}^n, \tag{51}$$

and $F_{\max}(Q^*_{\max}) = Q^*_{\max}$. Therefore, Theorem 2 applies with

$$\begin{aligned} &F' = F_{\max}, && F = F^\lambda_{\mathrm{bz}}, && x_t = Q_t, && x^* = Q^*_{\max}, \\ &\alpha = \gamma, && \eta = \gamma\lambda \ln(|\mathcal{A}|). \end{aligned} \tag{52}$$

The general derivative estimate in (6) is repeated here:

$$\frac{d}{dt} W_p(x_t - x^*) \le -\left(1 - \alpha n^{1/(2p)}\right) \|x_t - x^*\|_{2p}^{2p} + n^{1/(2p)} \eta \|x_t - x^*\|_{2p}^{2p-1}.$$

Substituting (52) into this repeated estimate gives

$$\frac{d}{dt} W_p(Q_t - Q^*_{\max}) \le -\left(1 - \gamma n^{1/(2p)}\right) \|Q_t - Q^*_{\max}\|_{2p}^{2p} + n^{1/(2p)} \gamma\lambda \ln(|\mathcal{A}|) \|Q_t - Q^*_{\max}\|_{2p}^{2p-1}. \tag{53}$$

This proves the main-text estimate in (12). To obtain the scalar differential inequality explicitly, first use (2) at any time such that $Q_t \ne Q^*_{\max}$:

$$\begin{aligned} \frac{d}{dt} W_p(Q_t - Q^*_{\max}) &= \frac{1}{2p} \frac{d}{dt} \|Q_t - Q^*_{\max}\|_{2p,w}^{2p} \\ &= \|Q_t - Q^*_{\max}\|_{2p,w}^{2p-1} \frac{d}{dt} \|Q_t - Q^*_{\max}\|_{2p,w}. \end{aligned} \tag{54}$$

The norm comparisons in Lemma 1 give

$$\begin{aligned} \|Q_t - Q^*_{\max}\|_{2p}^{2p} &\ge \frac{1}{w_{\max}} \|Q_t - Q^*_{\max}\|_{2p,w}^{2p}, \\ \|Q_t - Q^*_{\max}\|_{2p}^{2p-1} &\le \frac{1}{w_{\min}^{(2p-1)/(2p)}} \|Q_t - Q^*_{\max}\|_{2p,w}^{2p-1}. \end{aligned} \tag{55}$$

Because $1-\gamma n^{1/(2p)} > 0$, the lower bound in (55) applies to the negative term of (53), whereas the upper bound applies to its nonnegative term. Dividing by the positive factor in (54) therefore yields

$$\begin{aligned}
\frac{d}{dt}\|Q_t - Q^*_{\max}\|_{2p,w} &= \frac{\frac{d}{dt}W_p(Q_t - Q^*_{\max})}{\|Q_t - Q^*_{\max}\|_{2p,w}^{2p-1}} \\
&\le -\left(1-\gamma n^{1/(2p)}\right)\frac{\|Q_t - Q^*_{\max}\|_{2p}^{2p}}{\|Q_t - Q^*_{\max}\|_{2p,w}^{2p-1}} \\
&\quad + n^{1/(2p)}\gamma\lambda\ln(|\mathcal{A}|)\frac{\|Q_t - Q^*_{\max}\|_{2p}^{2p-1}}{\|Q_t - Q^*_{\max}\|_{2p,w}^{2p-1}} \\
&\le -\frac{1-\gamma n^{1/(2p)}}{w_{\max}}\frac{\|Q_t - Q^*_{\max}\|_{2p,w}^{2p}}{\|Q_t - Q^*_{\max}\|_{2p,w}^{2p-1}} \\
&\quad + \frac{n^{1/(2p)}\gamma\lambda\ln(|\mathcal{A}|)}{w_{\min}^{(2p-1)/(2p)}}\frac{\|Q_t - Q^*_{\max}\|_{2p,w}^{2p-1}}{\|Q_t - Q^*_{\max}\|_{2p,w}^{2p-1}} \\
&= -\frac{1-\gamma n^{1/(2p)}}{w_{\max}}\|Q_t - Q^*_{\max}\|_{2p,w} + \frac{n^{1/(2p)}\gamma\lambda\ln(|\mathcal{A}|)}{w_{\min}^{(2p-1)/(2p)}}. \qquad (56)
\end{aligned}$$

The norm trajectory is nonnegative and absolutely continuous on every finite time interval. Its derivative is zero at almost every time at which the norm is zero, so (56) holds almost everywhere. This justifies the application of Grönwall's inequality below even when $Q_t = Q^*_{\max}$ at some time. Set

$$a_p := \frac{1-\gamma n^{1/(2p)}}{w_{\max}} > 0. \qquad (57)$$

Applying Grönwall's inequality to (56) gives

$$\begin{aligned}
\|Q_t - Q^*_{\max}\|_{2p,w} \le e^{-a_p t}\|Q_0 - Q^*_{\max}\|_{2p,w} \\
+ \frac{n^{1/(2p)}\gamma\lambda\ln(|\mathcal{A}|)}{a_p w_{\min}^{(2p-1)/(2p)}}, \qquad t \ge 0. \qquad (58)
\end{aligned}$$

Substituting (57) into (58) yields

$$\begin{aligned}
\|Q_t - Q^*_{\max}\|_{2p,w} \le \exp\left(-\frac{1-\gamma n^{1/(2p)}}{w_{\max}}t\right)\|Q_0 - Q^*_{\max}\|_{2p,w} \\
+ \frac{n^{1/(2p)} w_{\max}\gamma\lambda\ln(|\mathcal{A}|)}{\left(1-\gamma n^{1/(2p)}\right) w_{\min}^{(2p-1)/(2p)}}, \qquad t \ge 0. \qquad (59)
\end{aligned}$$

This is the main-text estimate in (13). In addition, (53) shows that $W_p(Q_t - Q^*_{\max})$ is strictly decreasing whenever

$$\|Q_t - Q^*_{\max}\|_{2p} > \frac{n^{1/(2p)}\gamma\lambda\ln(|\mathcal{A}|)}{1-\gamma n^{1/(2p)}}.$$

□

F Basics of nonlinear system theory

We use several standard notions from nonlinear-system theory in the ODE analysis. Consider the system

$$\frac{d}{dt}x_t = f(x_t), \quad t \ge 0, \quad x_0 \in \mathbb{R}^n, \qquad (60)$$

where $x_t \in \mathbb{R}^n$ is the state at time $t \ge 0$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear mapping. For simplicity, we assume that the solution to (60) exists and is unique. This holds if f is globally Lipschitz continuous.

Lemma 7 ((Khalil, 2002, Theorem 3.2)). *Consider the nonlinear system (60) and assume that f is globally Lipschitz continuous, i.e.,*

$$\|f(x) - f(y)\| \le L\|x - y\|, \ \forall x, y \in \mathbb{R}^n, \tag{61}$$

for some real number $L > 0$ and norm $\|\cdot\|$. Then, the system has a unique solution x_t for every $t \ge 0$ and every $x_0 \in \mathbb{R}^n$.

A central concept in nonlinear-system analysis is an equilibrium point. A point $x = x^e$ in the state space is said to be an equilibrium point of (60) if it has the property that whenever the state of the system starts at x^e, it will remain at x^e (Khalil, 2002). For (60), the equilibrium points are the real roots of the equation $f(x) = 0$. The equilibrium point x^e is said to be globally asymptotically stable if it is Lyapunov stable and $x_t \to x^e$ as $t \to \infty$ for every initial state $x_0 \in \mathbb{R}^n$. Here, Lyapunov stability means that, for every $\varepsilon > 0$, there exists $\delta > 0$ such that $\|x_0 - x^e\| < \delta$ implies $\|x_t - x^e\| < \varepsilon$ for all $t \ge 0$. The equilibrium point is said to be globally exponentially stable if there exist a norm $\|\cdot\|$ and constants $C \ge 1$ and $r > 0$, independent of the initial state x_0, such that

$$\|x_t - x^e\| \le Ce^{-rt}\|x_0 - x^e\|, \qquad t \ge 0, \qquad x_0 \in \mathbb{R}^n.$$

G SUPPORTING LEMMAS

This section collects the supporting lemmas used in the subsequent analysis. First, the classical differential form of Grönwall's inequality is stated as follows.

Lemma 8 ((Gronwall, 1919)). *Let $\psi_t : [0, \infty) \to [0, \infty)$ be an absolutely continuous function satisfying the differential inequality*

$$\frac{d}{dt}\psi_t \le -a\psi_t + b, \quad \text{for almost every } t \ge 0$$

where $a > 0, b \ge 0$. Then,

$$\psi_t \le \psi_0 e^{-at} + \frac{b}{a}, \quad \forall t \ge 0.$$

The inequality in Lemma 8 is used repeatedly to establish stability of the ODEs considered below. The following relations are essential for analyzing the smooth Q-learning algorithms. Although proofs of these relations appear in Pan et al. (2020) and related studies, we include them here for completeness and convenience. In particular, the proof of the bound for the Boltzmann softmax operator is simpler than the proof in Pan et al. (2020).

Lemma 9. *For any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and any $\lambda > 0$, we have*

1. $h_{\max}(x) \le h^{\lambda}_{\text{lse}}(x) \le h_{\max}(x) + \lambda \ln n$;
2. $h_{\max}(x) - \lambda \ln n \le h^{\lambda}_{\text{mm}}(x) \le h_{\max}(x)$;
3. $h_{\max}(x) - \lambda \ln n \le h^{\lambda}_{\text{bz}}(x) \le h_{\max}(x)$;
4. $h^{\lambda}_{\text{bz}}(x) \le h^{\lambda}_{\text{lse}}(x) \le h^{\lambda}_{\text{bz}}(x) + \lambda \ln n$.

Proof. Since $x_i \le h_{\max}(x)$ for every i and equality holds for at least one coordinate,

$$e^{h_{\max}(x)/\lambda} \le \sum_{i=1}^{n} e^{x_i/\lambda} \le ne^{h_{\max}(x)/\lambda}. \tag{62}$$

Taking logarithms in (62) and multiplying by λ gives

$$h_{\max}(x) \le \lambda \ln\left(\sum_{i=1}^{n} e^{x_i/\lambda}\right) \le h_{\max}(x) + \lambda \ln n, \tag{63}$$

which proves the first statement. Dividing (62) by n before taking logarithms gives

$$h_{\max}(x) - \lambda \ln n \le \lambda \ln\left(\frac{1}{n}\sum_{i=1}^{n} e^{x_i/\lambda}\right) \le h_{\max}(x), \tag{64}$$

which proves the second statement.

For the Boltzmann operator, define

$$Z := \sum_{j=1}^{n} e^{x_j/\lambda}, \qquad \pi_i := \frac{e^{x_i/\lambda}}{Z}, \qquad i \in \{1, \dots, n\}. \tag{65}$$

Then $\pi_i \ge 0$, $\sum_i \pi_i = 1$, and

$$h^{\lambda}_{\mathrm{bz}}(x) = \sum_{i=1}^{n} \pi_i x_i \le h_{\max}(x). \tag{66}$$

For the lower bound, we use the convexity of $h^{\lambda}_{\mathrm{lse}}$ and its gradient formula (Gao & Pavel, 2017). In terms of the weights in (65),

$$\nabla h^{\lambda}_{\mathrm{lse}}(x)^{\top} x = \sum_{i=1}^{n} \pi_i x_i = h^{\lambda}_{\mathrm{bz}}(x). \tag{67}$$

The first-order inequality for a differentiable convex function gives

$$h^{\lambda}_{\mathrm{lse}}(y) \ge h^{\lambda}_{\mathrm{lse}}(x) + \nabla h^{\lambda}_{\mathrm{lse}}(x)^{\top}(y - x), \qquad x, y \in \mathbb{R}^n. \tag{68}$$

Setting $y = 0$ and using (67) yields

$$\lambda \ln n = h^{\lambda}_{\mathrm{lse}}(0) \ge h^{\lambda}_{\mathrm{lse}}(x) - h^{\lambda}_{\mathrm{bz}}(x). \tag{69}$$

Together with the lower bound in (63), this implies

$$h_{\max}(x) \le h^{\lambda}_{\mathrm{lse}}(x) \le h^{\lambda}_{\mathrm{bz}}(x) + \lambda \ln n, \tag{70}$$

so $h^{\lambda}_{\mathrm{bz}}(x) \ge h_{\max}(x) - \lambda \ln n$. Together with (66), this proves the third statement. Finally, the fourth statement follows from (69) and the inequalities $h^{\lambda}_{\mathrm{bz}}(x) \le h_{\max}(x) \le h^{\lambda}_{\mathrm{lse}}(x)$ already proved above. □

Furthermore, the following lemma will be used in the analysis.

Lemma 10. *For every $\lambda > 0$,*

$$\frac{h_{\max}(cx)}{c}, \qquad \frac{h^{\lambda}_{\mathrm{lse}}(cx)}{c}, \qquad \frac{h^{\lambda}_{\mathrm{mm}}(cx)}{c}, \qquad \frac{h^{\lambda}_{\mathrm{bz}}(cx)}{c}$$

all converge to $h_{\max}(x)$ as $c \to \infty$, uniformly on every compact subset of $\mathbb{R}^n$.

Proof. The max operator is positively homogeneous, so

$$\frac{h_{\max}(cx)}{c} = h_{\max}(x), \qquad c > 0. \tag{71}$$

Applying Lemma 9 to cx and dividing each inequality by c gives

$$0 \le \frac{h^\lambda_{\mathrm{lse}}(cx)}{c} - h_{\max}(x) \le \frac{\lambda \ln n}{c}, \tag{72}$$

$$0 \le h_{\max}(x) - \frac{h^\lambda_{\mathrm{mm}}(cx)}{c} \le \frac{\lambda \ln n}{c}, \tag{73}$$

$$0 \le h_{\max}(x) - \frac{h^\lambda_{\mathrm{bz}}(cx)}{c} \le \frac{\lambda \ln n}{c}. \tag{74}$$

Consequently,

$$\sup_{x \in \mathbb{R}^n} \left| \frac{h^\lambda_{\mathrm{lse}}(cx)}{c} - h_{\max}(x) \right| \le \frac{\lambda \ln n}{c}, \tag{75}$$

$$\sup_{x \in \mathbb{R}^n} \left| \frac{h^\lambda_{\mathrm{mm}}(cx)}{c} - h_{\max}(x) \right| \le \frac{\lambda \ln n}{c}, \tag{76}$$

$$\sup_{x \in \mathbb{R}^n} \left| \frac{h^\lambda_{\mathrm{bz}}(cx)}{c} - h_{\max}(x) \right| \le \frac{\lambda \ln n}{c}. \tag{77}$$

The right-hand sides tend to zero as $c \to \infty$ and do not depend on x. Together with (71), this proves uniform convergence on $\mathbb{R}^n$, and therefore on every compact subset. □

The following lemma collects the order and translation properties used in the contraction argument.

Lemma (Monotonicity and translation equivariance)**.** *Each mapping in* $\{h_{\max}, h^\lambda_{\mathrm{lse}}, h^\lambda_{\mathrm{mm}}\}$ *is monotone in the sense of Definition 1. Moreover, for every* $z \in \mathbb{R}^{|\mathcal{A}|}$ *and* $c \in \mathbb{R}$,

$$\begin{aligned} h_{\max}(z + c\mathbf{1}) &= h_{\max}(z) + c, \\ h^\lambda_{\mathrm{lse}}(z + c\mathbf{1}) &= \lambda \ln \left(\sum_i e^{(z_i + c)/\lambda} \right) = c + h^\lambda_{\mathrm{lse}}(z), \\ h^\lambda_{\mathrm{mm}}(z + c\mathbf{1}) &= \lambda \ln \left(\frac{1}{|\mathcal{A}|} \sum_i e^{(z_i + c)/\lambda} \right) = c + h^\lambda_{\mathrm{mm}}(z). \end{aligned} \tag{78}$$

Proof. If $u \le v$ componentwise, then $h_{\max}(u) \le h_{\max}(v)$. The same implication holds for h^λ_{lse} and h^λ_{mm} because the exponential, summation, and logarithm operations preserve the componentwise order. The three translation identities follow by factoring $e^{c/\lambda}$ from the corresponding exponential sums; the max identity is immediate. □

The following result states that the Bellman operators induced by the max, LSE, and mellowmax operators are contractions.

Lemma 11 (Bellman contraction for the max, LSE, and mellowmax operators)**.** *For every* $F \in \{F_{\max}, F^\lambda_{\mathrm{lse}}, F^\lambda_{\mathrm{mm}}\}$,

$$\|F(x) - F(y)\|_\infty \le \gamma \|x - y\|_\infty, \qquad x, y \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}.$$

Proof. Let $h \in \{h_{\max}, h^\lambda_{\mathrm{lse}}, h^\lambda_{\mathrm{mm}}\}$ and let $u, v \in \mathbb{R}^{|\mathcal{A}|}$. Then

$$v - \|u - v\|_\infty \mathbf{1} \le u \le v + \|u - v\|_\infty \mathbf{1} \tag{79}$$

componentwise. Applying the preceding monotonicity and translation-equivariance lemma to (79) gives

$$\begin{aligned} h(v) - \|u - v\|_\infty &= h(v - \|u - v\|_\infty \mathbf{1}) \le h(u) \\ &\le h(v + \|u - v\|_\infty \mathbf{1}) = h(v) + \|u - v\|_\infty. \end{aligned} \tag{80}$$

Therefore,

$$|h(u) - h(v)| \le \|u - v\|_\infty. \tag{81}$$

Applying (81) state by state yields

$$\|H(x) - H(y)\|_\infty \le \|x - y\|_\infty. \tag{82}$$

Since every row of P is a probability vector, $\|Pz\|_\infty \le \|z\|_\infty$. Hence

$$\begin{aligned} \|F(x) - F(y)\|_\infty &= \gamma \|P(H(x) - H(y))\|_\infty \\ &\le \gamma \|H(x) - H(y)\|_\infty \le \gamma \|x - y\|_\infty. \end{aligned} \tag{83}$$

This proves the claim. □

H AUXILIARY STABILITY RESULTS FOR PRECONDITIONED OPERATOR DYNAMICS

Consider the system in (3):

$$\frac{d}{dt} x_t = DF(x_t) - Dx_t, \quad \forall t \ge 0, \quad x_0 \in \mathbb{R}^n,$$

where $t \ge 0$ is the continuous time, $x_t \in \mathbb{R}^n$ is the state at time t, $F : \mathbb{R}^n \to \mathbb{R}^n$ is a mapping that will be specified later, and $D \in \mathbb{R}^{n \times n}$ is a positive definite diagonal matrix with strictly positive diagonal elements $d_i > 0, i \in \{1, 2, \dots, n\}$. The following result is an important intermediate step for our stability analysis of this ODE.

Lemma 12. *Consider the system in (3). Let $x_t \in \mathbb{R}^n$, $t \ge 0$, be its unique solution, and suppose that x^* is the unique fixed point of F, i.e., $x^* = F(x^*)$. Moreover, suppose that $p \in \mathbb{N}$ and set $w_i = 1/d_i$ for all $i \in \{1, 2, \dots, n\}$. Then, at every $t \ge 0$ such that $x_t \ne x^*$, we have*

$$\frac{d}{dt} \|x_t - x^*\|_{2p,w} \le - \frac{\|x_t - x^*\|_{2p}^{2p}}{\|x_t - x^*\|_{2p,w}^{2p-1}} + \frac{\|x_t - x^*\|_{2p}^{2p-1}}{\|x_t - x^*\|_{2p,w}^{2p-1}} \|F(x_t) - F(x^*)\|_{2p}.$$

Proof. Fix any $t \ge 0$ such that $x_t \ne x^*$. Using

$$\frac{\partial}{\partial x_i} \left(\sum_{j=1}^n w_j {x_j}^{2p} \right)^{1/(2p)} = \frac{w_i {x_i}^{2p-1}}{\|x\|_{2p,w}^{2p-1}},$$

the chain rule gives

$$
\begin{aligned}
&\frac{d}{dt}\|x_t - x^*\|_{2p,w} \\
=&\frac{d}{dt}\left(\sum_{j=1}^n w_j (x_{t,j} - x_j^*)^{2p}\right)^{1/(2p)} \\
=&\sum_{i=1}^n \frac{\partial}{\partial x_i}\left(\sum_{j=1}^n w_j (x_j - x_j^*)^{2p}\right)^{1/(2p)}\Bigg|_{x_i = x_{t,i}} \frac{dx_{t,i}}{dt} \\
=&\frac{1}{\|x_t - x^*\|_{2p,w}^{2p-1}} \begin{bmatrix} w_1(x_{t,1} - x_1^*)^{2p-1} \\ w_2(x_{t,2} - x_2^*)^{2p-1} \\ \vdots \\ w_n(x_{t,n} - x_n^*)^{2p-1} \end{bmatrix}^\top (DF(x_t) - Dx_t + Dx^* - DF(x^*)) \\
=&T_1 + T_2,
\end{aligned}
$$

where T_1 and T_2 are defined below and $x^* = F(x^*)$ is used in the last line. For T_1, we obtain

$$
\begin{aligned}
T_1 =&\frac{1}{\|x_t - x^*\|_{2p,w}^{2p-1}} \begin{bmatrix} w_1(x_{t,1} - x_1^*)^{2p-1} \\ w_2(x_{t,2} - x_2^*)^{2p-1} \\ \vdots \\ w_n(x_{t,n} - x_n^*)^{2p-1} \end{bmatrix}^\top (Dx^* - Dx_t) \\
=&-\frac{\|x_t - x^*\|_{2p}^{2p}}{\|x_t - x^*\|_{2p,w}^{2p-1}},
\end{aligned}
$$

where we use $w_i = \frac{1}{d_i}$ for all $i \in \{1, 2, \ldots, n\}$ in the second line. Moreover, T_2 can be bounded as

$$
\begin{aligned}
T_2 =&\frac{1}{\|x_t - x^*\|_{2p,w}^{2p-1}} \begin{bmatrix} (x_{t,1} - x_1^*)^{2p-1} \\ (x_{t,2} - x_2^*)^{2p-1} \\ \vdots \\ (x_{t,n} - x_n^*)^{2p-1} \end{bmatrix}^\top (F(x_t) - F(x^*)) \\
\le&\frac{1}{\|x_t - x^*\|_{2p,w}^{2p-1}} \left\| \begin{bmatrix} (x_{t,1} - x_1^*)^{2p-1} \\ (x_{t,2} - x_2^*)^{2p-1} \\ \vdots \\ (x_{t,n} - x_n^*)^{2p-1} \end{bmatrix} \right\|_q \|F(x_t) - F(x^*)\|_{2p} \\
=&\frac{\|x_t - x^*\|_{2p}^{2p-1}}{\|x_t - x^*\|_{2p,w}^{2p-1}} \|F(x_t) - F(x^*)\|_{2p},
\end{aligned}
$$

where Hölder's inequality is used in the first inequality, and $1/(2p) + 1/q = 1$ is used in the remaining parts. □

The preceding derivative estimate yields the following stability result.

Theorem 5. *Consider the system in (3), and let $x_t \in \mathbb{R}^n$, $t \ge 0$, be its unique solution. Suppose that $p \in \mathbb{N}$ and the mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is a contraction with respect to $\|\cdot\|_{2p}$, i.e., $\|F(x) - F(y)\|_{2p} \le \alpha \|x - y\|_{2p}, \forall x, y \in \mathbb{R}^n$ for some $\alpha \in (0, 1)$ so that it admits the unique fixed point $F(x^*) = x^*$. Then, we have*

$$
\frac{d}{dt}\|x_t - x^*\|_{2p,w} \le \frac{(\alpha - 1)}{w_{\max}} \|x_t - x^*\|_{2p,w}, \quad \forall t \ge 0, \quad x_0 \in \mathbb{R}^n \tag{84}
$$

and

$$\|x_t - x^*\|_{2p,w} \leq \|x_0 - x^*\|_{2p,w} \exp\left(\frac{(\alpha-1)t}{w_{\max}}\right), \quad \forall t \geq 0, \quad x_0 \in \mathbb{R}^n,$$

where $w_{\min} := \min_{i\in\{1,2,\ldots,n\}} w_i$, $w_{\max} := \max_{i\in\{1,2,\ldots,n\}} w_i$, *and* $w_i = \frac{1}{d_i}, \forall i \in \{1,2,\ldots,n\}$. *Therefore,* x^* *is the unique globally exponentially stable equilibrium point.*

Proof. Using Lemma 12 and the hypothesis

$$\|F(x) - F(y)\|_{2p} \leq \alpha \|x - y\|_{2p}, \forall x, y \in \mathbb{R}^n$$

we have

$$\begin{aligned}\frac{d}{dt}\|x_t - x^*\|_{2p,w} \leq & (\alpha-1)\frac{\|x_t - x^*\|_{2p}^{2p}}{\|x_t - x^*\|_{2p,w}^{2p-1}} \leq (\alpha-1)\frac{1}{w_{\max}}\frac{\|x_t - x^*\|_{2p,w}^{2p}}{\|x_t - x^*\|_{2p,w}^{2p-1}} \\ = & \frac{(\alpha-1)}{w_{\max}}\|x_t - x^*\|_{2p,w},\end{aligned}$$

where Lemma 1 is used in the second line. Next, using Grönwall's inequality in Lemma 8 yields the desired conclusion. □

The result in Theorem 5 holds for every $p \in \mathbb{N}$. Moreover, (84) shows that $V(x) := \|x - x^*\|_{2p,w}$ is a Lyapunov function (Khalil, 2002). Note that when $p = \infty$, $\|x - x^*\|_{2p}$ becomes non-differentiable in x. In order to bypass this issue, we can use the approximation property in Lemma 1 instead of directly dealing with ∞-norm. This approach provides simpler stability analysis compared to the approach in Borkar & Soumyanath (1997).

Lemma 13. *Suppose that the mapping* $F' : \mathbb{R}^n \to \mathbb{R}^n$ *is a contraction mapping, i.e.*

$$\|F'(x) - F'(y)\|_{2p} \leq \alpha \|x - y\|_{2p}, \quad \forall x, y \in \mathbb{R}^n$$

for some $\alpha \in (0,1)$, *where* $p \in \mathbb{N}$. *Moreover, suppose that the mapping* $F : \mathbb{R}^n \to \mathbb{R}^n$ *satisfies* $\|F(x) - F'(x)\|_{2p} \leq \eta, \forall x \in \mathbb{R}^n$, *where* $\eta > 0$ *is some real number. Then, we have* $\|F(x) - F'(y)\|_{2p} \leq \alpha\|x - y\|_{2p} + \eta, \forall x, y \in \mathbb{R}^n$.

Proof. The triangle inequality gives

$$\begin{aligned}\|F(x) - F'(y)\|_{2p} \leq & \|F(x) - F'(x)\|_{2p} + \|F'(x) - F'(y)\|_{2p} \\ \leq & \eta + \alpha\|x - y\|_{2p},\end{aligned}$$

where the last line follows from the hypotheses on F and F'. This completes the proof. □

Next, we study properties related to the convergence of the ODE (3) when F is not a contraction. To this end, we first need the following result, which serves as an intermediate step toward establishing the stability of the ODE (3).

Theorem 6. *Suppose that* $F : \mathbb{R}^n \to \mathbb{R}^n$ *is globally Lipschitz continuous, and let* $x_t \in \mathbb{R}^n$, $t \geq 0$, *be the unique global solution of* (3). *Suppose that* $p \in \mathbb{N}$, *the mapping* $F' : \mathbb{R}^n \to \mathbb{R}^n$ *is a contraction mapping with respect to the* $2p$*-norm, i.e.,*

$$\|F'(x) - F'(y)\|_{2p} \leq \alpha \|x - y\|_{2p}, \quad \forall x, y \in \mathbb{R}^n$$

for some $\alpha \in (0,1)$, *and the corresponding fixed point is* x^*. *Moreover, suppose that* F *satisfies* $\|F(x) - F'(x)\|_{2p} \leq \eta, \forall x \in \mathbb{R}^n$, *where* $\eta > 0$ *is some real number. Then, for every initial*

state $x_0 \in \mathbb{R}^n$, the map $t \mapsto \|x_t - x^*\|_{2p,w}$ is absolutely continuous on every finite time interval and satisfies

$$\frac{d}{dt}\|x_t - x^*\|_{2p,w} \le \frac{\alpha - 1}{w_{\max}}\|x_t - x^*\|_{2p,w} + \frac{\eta}{w_{\min}^{(2p-1)/(2p)}}$$

for almost every $t \ge 0$. Moreover,

$$\|x_t - x^*\|_{2p,w} \le \|x_0 - x^*\|_{2p,w} \exp\left(\frac{\alpha - 1}{w_{\max}} t\right) + \frac{\eta w_{\max}}{(1-\alpha) w_{\min}^{(2p-1)/(2p)}}, \qquad t \ge 0,$$

where $w_{\min} := \min_{i\in\{1,2,\dots,n\}} w_i$, $w_{\max} := \max_{i\in\{1,2,\dots,n\}} w_i$, $w_i = 1/d_i$ for all $i \in \{1, 2, \dots, n\}$, and $d_i > 0$ are the diagonal elements of D in (3). Therefore,

$$x_t \to \mathcal{H} := \left\{ x \in \mathbb{R}^n : \|x - x^*\|_{2p,w} \le \frac{\eta w_{\max}}{(1-\alpha) w_{\min}^{(2p-1)/(2p)}} \right\}$$

as $t \to \infty$.

Proof. Fix a time $t \ge 0$ such that $x_t \neq x^*$. Proceeding as in Lemma 12, we obtain

$$\begin{aligned}
&\frac{d}{dt}\|x_t - x^*\|_{2p,w} \\
=&\frac{1}{\|x_t - x^*\|_{2p,w}^{2p-1}} \begin{bmatrix} w_1(x_{t,1} - x_1^*)^{2p-1} \\ w_2(x_{t,2} - x_2^*)^{2p-1} \\ \vdots \\ w_n(x_{t,n} - x_n^*)^{2p-1} \end{bmatrix}^\top (DF(x_t) - Dx_t + Dx^* - DF'(x^*)) \\
=&T_1 + T_2
\end{aligned}$$

where T_1 and T_2 are defined below. For T_1, we obtain

$$\begin{aligned}
T_1 =&\frac{1}{\|x_t - x^*\|_{2p,w}^{2p-1}} \begin{bmatrix} w_1(x_{t,1} - x_1^*)^{2p-1} \\ w_2(x_{t,2} - x_2^*)^{2p-1} \\ \vdots \\ w_n(x_{t,n} - x_n^*)^{2p-1} \end{bmatrix}^\top (Dx^* - Dx_t) \\
=&-\frac{\|x_t - x^*\|_{2p}^{2p}}{\|x_t - x^*\|_{2p,w}^{2p-1}},
\end{aligned}$$

where we use $w_i = \frac{1}{d_i}$ for all $i \in \{1, 2, \dots, n\}$ in the second line. Moreover, T_2 can be bounded as

$$\begin{aligned}
T_2 =&\frac{1}{\|x_t - x^*\|_{2p,w}^{2p-1}} \begin{bmatrix} (x_{t,1} - x_1^*)^{2p-1} \\ (x_{t,2} - x_2^*)^{2p-1} \\ \vdots \\ (x_{t,n} - x_n^*)^{2p-1} \end{bmatrix}^\top (F(x_t) - F'(x^*)) \\
\le&\frac{1}{\|x_t - x^*\|_{2p,w}^{2p-1}} \left\| \begin{bmatrix} (x_{t,1} - x_1^*)^{2p-1} \\ (x_{t,2} - x_2^*)^{2p-1} \\ \vdots \\ (x_{t,n} - x_n^*)^{2p-1} \end{bmatrix} \right\|_q \|F(x_t) - F'(x^*)\|_{2p} \\
=&\frac{\|x_t - x^*\|_{2p}^{2p-1}}{\|x_t - x^*\|_{2p,w}^{2p-1}} \|F(x_t) - F'(x^*)\|_{2p},
\end{aligned}$$

where Hölder's inequality is used in the first inequality, and $1/(2p) + 1/q = 1$ is used in the remaining parts. Next, using Lemma 13, one has

$$\begin{aligned} T_2 &\leq \frac{\|x_t - x^*\|_{2p}^{2p-1}}{\|x_t - x^*\|_{2p,w}^{2p-1}} \|F(x_t) - F'(x^*)\|_{2p} \\ &\leq \alpha \frac{\|x_t - x^*\|_{2p}^{2p}}{\|x_t - x^*\|_{2p,w}^{2p-1}} + \eta \frac{\|x_t - x^*\|_{2p}^{2p-1}}{\|x_t - x^*\|_{2p,w}^{2p-1}} \\ &\leq \alpha \frac{\|x_t - x^*\|_{2p}^{2p}}{\|x_t - x^*\|_{2p,w}^{2p-1}} + \frac{\eta}{w_{\min}^{(2p-1)/(2p)}}. \end{aligned}$$

Here, the second inequality uses Lemma 13:

$$\begin{aligned} \|F(x_t) - F'(x^*)\|_{2p} =& \|F(x_t) - F'(x_t) + F'(x_t) - F'(x^*)\|_{2p} \\ \leq& \|F(x_t) - F'(x_t)\|_{2p} + \|F'(x_t) - F'(x^*)\|_{2p} \\ \leq& \eta + \alpha \|x_t - x^*\|_{2p}. \end{aligned}$$

Combining the bounds on T_1 and T_2 gives

$$\begin{aligned} \frac{d}{dt}\|x_t - x^*\|_{2p,w} &\leq (\alpha - 1) \frac{\|x_t - x^*\|_{2p}^{2p}}{\|x_t - x^*\|_{2p,w}^{2p-1}} + \frac{\eta}{w_{\min}^{(2p-1)/(2p)}} \\ &\leq (\alpha - 1) \frac{\|x_t - x^*\|_{2p,w}^{2p}}{w_{\max}\|x_t - x^*\|_{2p,w}^{2p-1}} + \frac{\eta}{w_{\min}^{(2p-1)/(2p)}} \\ &= \frac{\alpha - 1}{w_{\max}} \|x_t - x^*\|_{2p,w} + \frac{\eta}{w_{\min}^{(2p-1)/(2p)}}. \end{aligned}$$

The map $t \mapsto \|x_t - x^*\|_{2p,w}$ is nonnegative and absolutely continuous on every finite time interval, since x_t is continuously differentiable and the norm is Lipschitz continuous. At any interior time at which this map is differentiable and equals zero, it has a local minimum, so its derivative is zero. The additive term in the preceding bound is nonnegative; hence the same inequality holds at these times. Since an absolutely continuous function is differentiable almost everywhere, we obtain

$$\frac{d}{dt}\|x_t - x^*\|_{2p,w} \leq \frac{\alpha - 1}{w_{\max}} \|x_t - x^*\|_{2p,w} + \frac{\eta}{w_{\min}^{(2p-1)/(2p)}}$$

for almost every $t \geq 0$. Applying Grönwall's inequality in Lemma 8 on each finite time interval gives

$$\|x_t - x^*\|_{2p,w} \leq \|x_0 - x^*\|_{2p,w} \exp\left(\frac{\alpha - 1}{w_{\max}} t\right) + \frac{\eta w_{\max}}{(1 - \alpha) w_{\min}^{(2p-1)/(2p)}}, \qquad t \geq 0.$$

The final claim follows from this bound. This completes the proof. □

Theorem 6 uses the $2p$-norm as a Lyapunov function for a finite $p \in \mathbb{N}$. This argument can also be extended to the case $p \to \infty$. Theorem 8 shows that without the contraction mapping property, the ODE in (3) may still approach an error set around the fixed point x^* of F'.

The following result is required to prove the stability of the ODE (3) with the Boltzmann softmax operator and Theorem 6.

Lemma 14. *For the Boltzmann softmax operator, the mapping $F_{\mathrm{bz}}^{\lambda}$ satisfies*

$$\left\|F_{\mathrm{bz}}^{\lambda}(x) - F_{\max}(x)\right\|_{\infty} \leq \gamma\lambda \ln(|\mathcal{A}|), \quad \forall x \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}.$$

Proof. To begin with, Lemma 9 gives $H_{\max}(x) - \lambda \ln(|\mathcal{A}|)\mathbf{1} \le H^{\lambda}_{\mathrm{bz}}(x) \le H_{\max}(x)$, where $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}|}$ has all entries equal to one. Equivalently, $0 \le H_{\max}(x) - H^{\lambda}_{\mathrm{bz}}(x) \le \lambda \ln(|\mathcal{A}|)\mathbf{1}$. Therefore,

$$\left\| H^{\lambda}_{\mathrm{bz}}(x) - H_{\max}(x) \right\|_{\infty} \le \lambda \ln(|\mathcal{A}|). \tag{85}$$

Then, we have

$$\begin{aligned} \left\| F_{\max}(x) - F^{\lambda}_{\mathrm{bz}}(x) \right\|_{\infty} \le & \left\| \gamma P H_{\max}(x) - \gamma P H^{\lambda}_{\mathrm{bz}}(x) \right\|_{\infty} \\ \le & \gamma \left\| H_{\max}(x) - H^{\lambda}_{\mathrm{bz}}(x) \right\|_{\infty} \le \gamma\lambda \ln(|\mathcal{A}|), \end{aligned}$$

where the last inequality is due to (85). This concludes the proof. □

The next lemma shows that the Boltzmann operator is globally Lipschitz. The precise constant is not important for the stability argument; we only need global Lipschitz continuity.

Lemma 15. *For the Boltzmann softmax operator, the mapping $H^{\lambda}_{\mathrm{bz}}$ satisfies*

$$\left\| H^{\lambda}_{\mathrm{bz}}(x) - H^{\lambda}_{\mathrm{bz}}(y) \right\|_{\infty} \le L_{\mathrm{bz}} \|x - y\|_{\infty}, \quad \forall x, y \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|},$$

where one may take $L_{\mathrm{bz}} := 1 + 2|\mathcal{A}|/e$.

Proof. It suffices to prove the claim for the scalar operator $h^{\lambda}_{\mathrm{bz}} : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}$. Define

$$Z(x) := \sum_j e^{x_j/\lambda}, \qquad \pi_j(x) := \frac{e^{x_j/\lambda}}{Z(x)}, \qquad h^{\lambda}_{\mathrm{bz}}(x) = \sum_j \pi_j(x) x_j.$$

These functions are smooth on all of $\mathbb{R}^{|\mathcal{A}|}$. Differentiating the quotient directly gives

$$\frac{\partial \pi_j(x)}{\partial x_i} = \frac{1}{\lambda} \pi_j(x) \big(\mathbf{1}_{\{i=j\}} - \pi_i(x) \big),$$

and hence

$$\begin{aligned} \frac{\partial h^{\lambda}_{\mathrm{bz}}(x)}{\partial x_i} &= \pi_i(x) + \sum_j x_j \frac{\partial \pi_j(x)}{\partial x_i} \\ &= \pi_i(x) + \frac{1}{\lambda} \pi_i(x) \left(x_i - \sum_j \pi_j(x) x_j \right) \\ &= \pi_i(x) \left(1 + \frac{x_i - h^{\lambda}_{\mathrm{bz}}(x)}{\lambda} \right). \end{aligned}$$

This identity holds at every x, including points with multiple maximizing coordinates, because no derivative of the max function has been used.

To bound the gradient, use $x_i \le h_{\max}(x)$ and $h_{\mathrm{bz}}^{\lambda}(x) \le h_{\max}(x)$ to obtain

$$\begin{aligned}
\sum_{i=1}^{|\mathcal{A}|} \left| \frac{\partial h_{\mathrm{bz}}^{\lambda}(x)}{\partial x_i} \right| &= \sum_{i=1}^{|\mathcal{A}|} \pi_i(x) \left| 1 + \frac{x_i - h_{\mathrm{bz}}^{\lambda}(x)}{\lambda} \right| \\
&\le 1 + \frac{1}{\lambda} \sum_{i=1}^{|\mathcal{A}|} \pi_i(x) \left| x_i - h_{\mathrm{bz}}^{\lambda}(x) \right| \\
&\le 1 + \frac{1}{\lambda} \sum_{i=1}^{|\mathcal{A}|} \pi_i(x) \left[h_{\max}(x) - x_i + h_{\max}(x) - h_{\mathrm{bz}}^{\lambda}(x) \right] \\
&= 1 + \frac{2}{\lambda} \sum_{i=1}^{|\mathcal{A}|} \pi_i(x) \left(h_{\max}(x) - x_i \right) \\
&= 1 + 2 \frac{\sum_{i=1}^{|\mathcal{A}|} \frac{h_{\max}(x) - x_i}{\lambda} \exp\left(-\frac{h_{\max}(x) - x_i}{\lambda} \right)}{\sum_{j=1}^{|\mathcal{A}|} \exp\left(-\frac{h_{\max}(x) - x_j}{\lambda} \right)} \\
&\le 1 + 2 \sum_{i=1}^{|\mathcal{A}|} \frac{h_{\max}(x) - x_i}{\lambda} \exp\left(-\frac{h_{\max}(x) - x_i}{\lambda} \right) \\
&\le 1 + \frac{2|\mathcal{A}|}{e} = L_{\mathrm{bz}}.
\end{aligned} \tag{86}$$

The first equality after the inequalities uses $\sum_i \pi_i(x) = 1$ and $h_{\max}(x) - h_{\mathrm{bz}}^{\lambda}(x) = \sum_i \pi_i(x)(h_{\max}(x) - x_i)$. The denominator in (86) is at least one because a maximizing coordinate makes one exponential term equal to one. The last inequality uses $ue^{-u} \le e^{-1}$ for $u \ge 0$, applied to each nonnegative ratio $(h_{\max}(x) - x_i)/\lambda$.

For any $u, v \in \mathbb{R}^{|\mathcal{A}|}$, the integral form of the mean-value theorem along the line segment from v to u gives

$$\begin{aligned}
h_{\mathrm{bz}}^{\lambda}(u) - h_{\mathrm{bz}}^{\lambda}(v) &= \int_0^1 \frac{d}{dt} h_{\mathrm{bz}}^{\lambda}\big(v + t(u - v)\big)\, dt \\
&= \int_0^1 \nabla h_{\mathrm{bz}}^{\lambda}\big(v + t(u - v)\big)^{\top} (u - v)\, dt.
\end{aligned} \tag{87}$$

Using (86) in (87) yields

$$\begin{aligned}
\left| h_{\mathrm{bz}}^{\lambda}(u) - h_{\mathrm{bz}}^{\lambda}(v) \right| &\le \int_0^1 \sum_{i=1}^{|\mathcal{A}|} \left| \frac{\partial h_{\mathrm{bz}}^{\lambda}}{\partial x_i}\big(v + t(u - v)\big) \right| |u_i - v_i|\, dt \\
&\le \|u - v\|_{\infty} \int_0^1 \left\| \nabla h_{\mathrm{bz}}^{\lambda}\big(v + t(u - v)\big) \right\|_1 dt \\
&\le L_{\mathrm{bz}} \|u - v\|_{\infty}.
\end{aligned} \tag{88}$$

Finally, for $x, y \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$, apply (88) to the action-value vectors at each state:

$$\begin{aligned}
\left\| H_{\mathrm{bz}}^{\lambda}(x) - H_{\mathrm{bz}}^{\lambda}(y) \right\|_{\infty} &= \max_{s \in \mathcal{S}} \left| h_{\mathrm{bz}}^{\lambda}\big(x(s, \cdot)\big) - h_{\mathrm{bz}}^{\lambda}\big(y(s, \cdot)\big) \right| \\
&\le L_{\mathrm{bz}} \max_{s \in \mathcal{S}} \|x(s, \cdot) - y(s, \cdot)\|_{\infty} \\
&= L_{\mathrm{bz}} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |x(s, a) - y(s, a)| \\
&= L_{\mathrm{bz}} \|x - y\|_{\infty}.
\end{aligned}$$

This proves the claim. □

Lemma 16. *Consider Algorithm 1 with the Boltzmann softmax operator and step-sizes satisfying $\alpha_k > 0$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Then,*

$$\sup_{k \geq 0} \|Q_k\|_\infty < \infty$$

for every sample path.

Proof. The step-size condition implies $\alpha_k \to 0$. Hence, there exists an integer $K \geq 0$ such that $0 < \alpha_k \leq 1$ for every $k \geq K$. Set

$$B := \max \left\{ \max_{0 \leq j \leq K} \|Q_j\|_\infty, \frac{\max_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} |r(s,a,s')|}{1-\gamma} \right\}.$$

The number B is finite because only finitely many iterates precede K and the state and action spaces are finite. Moreover, $h_{\mathrm{bz}}^{\lambda}(q)$ is a convex combination of the coordinates of q, and therefore $|h_{\mathrm{bz}}^{\lambda}(q)| \leq \|q\|_\infty$. Suppose that $\|Q_k\|_\infty \leq B$ for some $k \geq K$. For the updated coordinate, the update rule and $0 < \alpha_k \leq 1$ give

$$\begin{aligned}
|Q_{k+1}(s_k,a_k)| &= \left|(1-\alpha_k)Q_k(s_k,a_k) + \alpha_k \left[r(s_k,a_k,s_k') + \gamma h_{\mathrm{bz}}^{\lambda}\big(Q_k(s_k',\cdot)\big)\right]\right| \\
&\leq (1-\alpha_k)|Q_k(s_k,a_k)| + \alpha_k \left|r(s_k,a_k,s_k') + \gamma h_{\mathrm{bz}}^{\lambda}\big(Q_k(s_k',\cdot)\big)\right| \\
&\leq (1-\alpha_k)|Q_k(s_k,a_k)| + \alpha_k |r(s_k,a_k,s_k')| + \alpha_k\gamma \left|h_{\mathrm{bz}}^{\lambda}\big(Q_k(s_k',\cdot)\big)\right| \\
&\leq (1-\alpha_k)\|Q_k\|_\infty + \alpha_k |r(s_k,a_k,s_k')| + \alpha_k\gamma \|Q_k(s_k',\cdot)\|_\infty \\
&\leq (1-\alpha_k)\|Q_k\|_\infty + \alpha_k \max_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} |r(s,a,s')| + \alpha_k\gamma \|Q_k\|_\infty \\
&\leq (1-\alpha_k)B + \alpha_k \left[\max_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} |r(s,a,s')| + \gamma B\right] \\
&\leq (1-\alpha_k)B + \alpha_k \left[(1-\gamma)B + \gamma B\right] \\
&= (1-\alpha_k)B + \alpha_k B = B.
\end{aligned}$$

The last inequality follows from the definition of B, which gives $\max_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} |r(s,a,s')| \leq (1-\gamma)B$. For every unselected pair $(s,a) \neq (s_k,a_k)$, the coordinate remains unchanged:

$$|Q_{k+1}(s,a)| = |Q_k(s,a)| \leq B.$$

Thus $\|Q_{k+1}\|_\infty \leq B$. Since $\|Q_K\|_\infty \leq B$, induction proves the bound for every $k \geq K$; the finitely many earlier iterates are already included in the definition of B. □

I WEIGHTED-NORM COUNTERPARTS OF THE POLYNOMIAL RESULTS

The main text formulates the central stability results through the smooth polynomial Lyapunov function W_p. This section collects the corresponding estimates stated directly in the weighted $2p$-norm. It also gives the limiting ∞-norm estimates and the positively invariant sets used in the stochastic Boltzmann analysis.

Theorem 7 (Weighted-norm contractive estimate)**.** *Let x_t be the unique solution of (3). Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is an ∞-norm contraction with factor $\alpha \in (0,1)$ and unique fixed point x^*. Choose $w_i = 1/d_i$ and let $w_{\max} := \max_i w_i$. For any $p \in \mathbb{N}$ satisfying*

$$\left\lceil \frac{\ln n}{\ln(\alpha^{-1})} \right\rceil < 2p,$$

we have

$$\frac{d}{dt}\|x_t - x^*\|_{2p,w} \leq \frac{\alpha n^{1/(2p)} - 1}{w_{\max}} \|x_t - x^*\|_{2p,w}, \tag{89}$$

and

$$\|x_t - x^*\|_{2p,w} \le \|x_0 - x^*\|_{2p,w} \exp\left(\frac{\alpha n^{1/(2p)} - 1}{w_{\max}} t\right), \qquad t \ge 0. \tag{90}$$

Therefore, x^* is the unique globally exponentially stable equilibrium.

Proof. At every $t \ge 0$ such that $x_t \ne x^*$, differentiating the weighted $2p$-norm directly and using $w_i d_i = 1$ gives

$$\begin{aligned}
\frac{d}{dt}\|x_t - x^*\|_{2p,w} &= \frac{\sum_{i=1}^{n} (x_{t,i} - x_i^*)^{2p-1} \big(F_i(x_t) - F_i(x^*) - (x_{t,i} - x_i^*)\big)}{\|x_t - x^*\|_{2p,w}^{2p-1}} \\
&\le \frac{\|x_t - x^*\|_{2p}^{2p-1} \|F(x_t) - F(x^*)\|_{2p} - \|x_t - x^*\|_{2p}^{2p}}{\|x_t - x^*\|_{2p,w}^{2p-1}} \\
&\le -\left(1 - \alpha n^{1/(2p)}\right) \frac{\|x_t - x^*\|_{2p}^{2p}}{\|x_t - x^*\|_{2p,w}^{2p-1}} \\
&\le \frac{\alpha n^{1/(2p)} - 1}{w_{\max}} \|x_t - x^*\|_{2p,w}.
\end{aligned}$$

Here, the first inequality follows from Hölder's inequality, the second from the ∞-norm contraction and Lemma 1, and the third from $\|x_t - x^*\|_{2p,w}^{2p} \le w_{\max} \|x_t - x^*\|_{2p}^{2p}$. If $x_t = x^*$ at some time, the trajectory remains at x^* thereafter, so the same estimate holds almost everywhere. This proves (89). Applying Lemma 8 with zero additive term gives (90). □

Theorem 7 is the direct weighted-norm counterpart of Theorem 1. The threshold on p ensures that $\alpha n^{1/(2p)} < 1$. Weighted $2p$-norm Lyapunov functions have been used for mappings that are contractive in the same norm (Borkar & Soumyanath, 1997; Borkar, 2008). Here, the finite-norm estimate is obtained from an ∞-norm contraction through the norm comparisons in Lemma 1.

Theorem 8 (Weighted-norm perturbation estimate)**.** *Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is globally Lipschitz continuous, and let x_t be the unique global solution of (3). Suppose that $F' : \mathbb{R}^n \to \mathbb{R}^n$ is an ∞-norm contraction with factor $\alpha \in (0,1)$ and fixed point x^*, and suppose that*

$$\|F(x) - F'(x)\|_\infty \le \eta, \qquad x \in \mathbb{R}^n,$$

for some $\eta > 0$. Choose $w_i = 1/d_i$, let $w_{\min} := \min_i w_i$ and $w_{\max} := \max_i w_i$, and take any $p \in \mathbb{N}$ satisfying

$$\left\lceil \frac{\ln n}{\ln(\alpha^{-1})} \right\rceil < 2p.$$

Then

$$\frac{d}{dt}\|x_t - x^*\|_{2p,w} \le \frac{\alpha n^{1/(2p)} - 1}{w_{\max}} \|x_t - x^*\|_{2p,w} + \frac{n^{1/(2p)} \eta}{w_{\min}^{(2p-1)/(2p)}}, \tag{91}$$

for almost every $t \ge 0$, and

$$\|x_t - x^*\|_{2p,w} \le \|x_0 - x^*\|_{2p,w} \exp\left(\frac{\alpha n^{1/(2p)} - 1}{w_{\max}} t\right) \tag{92}$$

$$+ \frac{n^{1/(2p)} \eta w_{\max}}{\left(1 - \alpha n^{1/(2p)}\right) w_{\min}^{(2p-1)/(2p)}}, \qquad t \ge 0. \tag{93}$$

Moreover,

$$\|x_t - x^*\|_\infty \le \|x_0 - x^*\|_\infty \exp\left(\frac{\alpha - 1}{w_{\max}} t\right) + \frac{\eta w_{\max}}{(1-\alpha) w_{\min}}, \qquad t \ge 0. \tag{94}$$

Consequently,

$$\mathcal{H} := \left\{ x \in \mathbb{R}^n : \|x - x^*\|_\infty \le \frac{\eta w_{\max}}{(1-\alpha) w_{\min}} \right\} \tag{95}$$

is positively invariant and satisfies $\operatorname{dist}(x_t, \mathcal{H}) \to 0$ *as* $t \to \infty$.

Proof. The calculation leading to (45) gives (91) whenever $x_t \neq x^*$. The norm trajectory is nonnegative and absolutely continuous on every finite time interval. At almost every time at which it equals zero, its derivative is zero; hence (91) holds almost everywhere. Applying Lemma 8 on each finite time interval gives (93) for every $t \ge 0$. Applying Lemma 1 to (93) gives

$$\|x_t - x^*\|_\infty \le \left(\frac{n w_{\max}}{w_{\min}}\right)^{1/(2p)} \|x_0 - x^*\|_\infty \exp\left(\frac{\alpha n^{1/(2p)} - 1}{w_{\max}} t\right) + \frac{n^{1/(2p)} \eta w_{\max}}{\left(1 - \alpha n^{1/(2p)}\right) w_{\min}}.$$

Taking $p \to \infty$ proves (94).

It remains to verify positive invariance. At any boundary point of $\mathcal{H}$, each coordinate satisfying $|x_{t,i} - x_i^*| = \|x_t - x^*\|_\infty$ obeys

$$\operatorname{sgn}(x_{t,i} - x_i^*) \frac{d}{dt}(x_{t,i} - x_i^*) \le d_i \left(\eta - (1-\alpha)\|x_t - x^*\|_\infty\right) \le 0,$$

because $w_{\max} \ge w_{\min}$. Thus, the vector field does not point outside $\mathcal{H}$. Finally, (94) implies $\operatorname{dist}(x_t, \mathcal{H}) \to 0$. □

Theorem 8 states the norm estimate corresponding to the polynomial perturbation result in Theorem 2. It makes the limiting ∞-norm neighborhood explicit and provides the invariant-set statement used below.

Theorem 9 (Weighted-norm estimate for contractive Q-learning ODEs)**.** *Consider* (8) *with* $H \in \{H_{\max}, H_{\mathrm{lse}}^\lambda, H_{\mathrm{mm}}^\lambda\}$*, and let* Q_t *be its solution. Let* Q_e *denote the corresponding fixed point in* (10)*, set* $n = |\mathcal{S} \times \mathcal{A}|$*, choose* $w_i = 1/d_i$*, and let* $w_{\max} := \max_i w_i$*. For any* $p \in \mathbb{N}$ *satisfying*

$$\left\lceil \frac{\ln n}{\ln(\gamma^{-1})} \right\rceil < 2p,$$

we have

$$\|Q_t - Q_e\|_{2p,w} \le \|Q_0 - Q_e\|_{2p,w} \exp\left(\frac{\gamma n^{1/(2p)} - 1}{w_{\max}} t\right), \qquad t \ge 0. \tag{96}$$

Proof. The Bellman operators $F_{\max}$, F_{lse}^λ, and F_{mm}^λ are γ-contractions by Lemma 11. Applying Theorem 7 with $x_t = Q_t$, $x^* = Q_e$, and $\alpha = \gamma$ proves the result. □

Theorem 9 is the weighted-norm counterpart of Theorem 3. It gives exponential convergence for the max, LSE, and mellowmax ODEs in the weighted finite norm.

Theorem 10 (Weighted-norm estimate for the Boltzmann ODE)**.** *Consider* (8) *with* $H = H^{\lambda}_{\mathrm{bz}}$*, and let* Q_t *be its solution. Set* $n = |\mathcal{S} \times \mathcal{A}|$*, choose* $w_i = 1/d_i$*, and let* $w_{\min} := \min_i w_i$ *and* $w_{\max} := \max_i w_i$*. For any* $p \in \mathbb{N}$ *satisfying*

$$\left\lceil \frac{\ln n}{\ln(\gamma^{-1})} \right\rceil < 2p,$$

we have

$$\|Q_t - Q^*_{\max}\|_{2p,w} \le \|Q_0 - Q^*_{\max}\|_{2p,w} \exp\left(\frac{\gamma n^{1/(2p)} - 1}{w_{\max}} t\right) \tag{97}$$

$$+ \frac{n^{1/(2p)} w_{\max}}{\left(1 - \gamma n^{1/(2p)}\right) w_{\min}^{(2p-1)/(2p)}} \gamma\lambda \ln(|\mathcal{A}|), \qquad t \ge 0, \tag{98}$$

and

$$\|Q_t - Q^*_{\max}\|_{\infty} \le \|Q_0 - Q^*_{\max}\|_{\infty} \exp\left(\frac{\gamma - 1}{w_{\max}} t\right) + \frac{\gamma\lambda \ln(|\mathcal{A}|) w_{\max}}{(1-\gamma) w_{\min}}, \qquad t \ge 0. \tag{99}$$

Consequently,

$$\mathcal{H} := \left\{ Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : \|Q - Q^*_{\max}\|_{\infty} \le \frac{\gamma\lambda \ln(|\mathcal{A}|) w_{\max}}{(1-\gamma) w_{\min}} \right\} \tag{100}$$

is positively invariant and satisfies $\operatorname{dist}(Q_t, \mathcal{H}) \to 0$ *as* $t \to \infty$.

Proof. If $|\mathcal{A}| = 1$, then $F^{\lambda}_{\mathrm{bz}} = F_{\max}$ and $\mathcal{H} = \{Q^*_{\max}\}$. The weighted-norm estimate follows from Theorem 9; applying Lemma 1 and taking $p \to \infty$ gives the stated ∞-norm estimate with zero additive term. Since $Q^*_{\max}$ is an equilibrium, $\mathcal{H}$ is positively invariant, and the error estimate gives $\operatorname{dist}(Q_t, \mathcal{H}) \to 0$.

For $|\mathcal{A}| \ge 2$, we have $\gamma\lambda \ln(|\mathcal{A}|) > 0$. By Lemma 15, $F^{\lambda}_{\mathrm{bz}}$ is globally Lipschitz continuous. By Lemma 14,

$$\|F^{\lambda}_{\mathrm{bz}}(Q) - F_{\max}(Q)\|_{\infty} \le \gamma\lambda \ln(|\mathcal{A}|).$$

Applying Theorem 8 with $F' = F_{\max}$, $F = F^{\lambda}_{\mathrm{bz}}$, $\alpha = \gamma$, $x_t = Q_t$, $x^* = Q^*_{\max}$, and $\eta = \gamma\lambda \ln(|\mathcal{A}|)$ proves the result. □

Theorem 10 is the weighted-norm counterpart of Theorem 4. It provides the exact ∞-norm invariant set used in the stochastic set-convergence result. The following example illustrates convergence toward this guaranteed neighborhood.

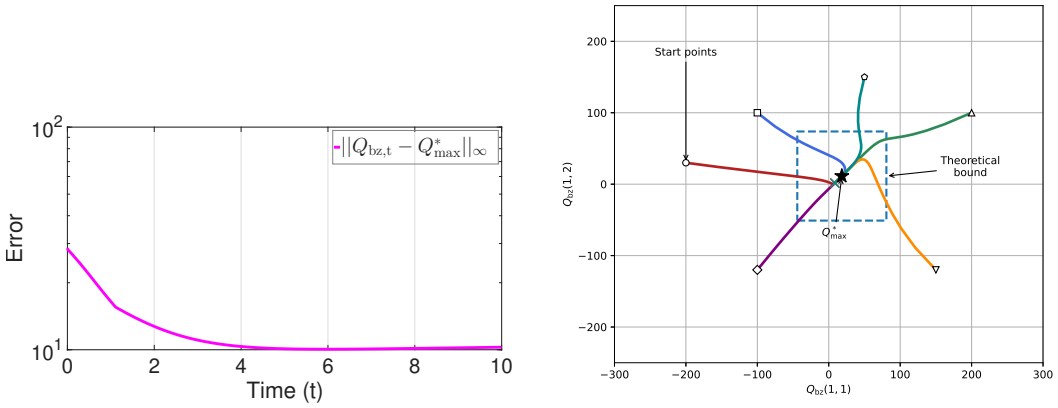
J ODE-BASED STOCHASTIC APPROXIMATION UNDER I.I.D. SAMPLING

Because of its generality, the ODE method is widely used to analyze convergence of RL algorithms (Bhatnagar et al., 2013; Kushner & Yin, 2003). It studies a stochastic recursion through the stability of its limiting ODE, using the fact that diminishing-step-size iterates track the ODE asymptotically. A standard version of this approach is based on the Borkar–Meyn theorem (Borkar & Meyn, 2000). We summarize the assumptions needed below.

$$\theta_{k+1} = \theta_k + \alpha_k(f(\theta_k) + \varepsilon_{k+1}) \tag{101}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear mapping and the integer $k \ge 0$ is the iteration step. Basic technical assumptions are given below.

Assumption 2.



(a) Error convergence curve

(b) Phase plane trajectories

Figure 1: Empirical stability analysis of the ODE with the Boltzmann operator. (a) Evolution of $\|Q_t - Q^*_{\max}\|_\infty$. (b) Solutions from different initial conditions.

1. *The mapping $f : \mathbb{R}^n \to \mathbb{R}^n$ is globally Lipschitz continuous.*
2. *There exists a unique globally asymptotically stable equilibrium $\theta^e \in \mathbb{R}^n$ for the ODE $\dot{x}_t = f(x_t)$.*
3. *There exists a function $f_\infty : \mathbb{R}^n \to \mathbb{R}^n$ such that, for every compact set $K \subset \mathbb{R}^n$,*
$$\lim_{c\to\infty} \sup_{x\in K} \left\| \frac{f(cx)}{c} - f_\infty(x) \right\|_2 = 0.$$
4. *The origin in $\mathbb{R}^n$ is the unique globally asymptotically stable equilibrium for the ODE $\dot{x}_t = f_\infty(x_t)$.*
5. *The sequence $\{\varepsilon_k, \mathcal{G}_k, k \geq 1\}$ with $\mathcal{G}_k = \sigma(\theta_i, \varepsilon_i, i \leq k)$ is a martingale difference sequence. In addition, there exists a constant $C_0 < \infty$ such that for any initial $\theta_0 \in \mathbb{R}^n$, we have $\mathbb{E}[\|\varepsilon_{k+1}\|_2^2 | \mathcal{G}_k] \leq C_0(1 + \|\theta_k\|_2^2), \forall k \geq 0$.*
6. *The step-sizes satisfy*
$$\alpha_k > 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{102}$$

Lemma 17 ((Borkar & Meyn, 2000, Borkar and Meyn theorem)). *Under Assumption 2, for any initial $\theta_0 \in \mathbb{R}^n$, we have $\sup_{k\geq 0} \|\theta_k\|_2 < \infty$ with probability one. In addition, $\theta_k \to \theta^e$ as $k \to \infty$ with probability one.*

The ODE approach in Lemma 17 has been widely used to prove convergence of RL algorithms, such as the synchronous Q-learning algorithm (Borkar & Meyn, 2000), synchronous temporal-difference (TD) learning (Borkar & Meyn, 2000), asynchronous Q-learning (Lee & He, 2020), gradient TD-learning algorithms (Sutton et al., 2008; 2009; Ghiassian et al., 2020; Lee et al., 2023), Q-learning with linear function approximation (Melo et al., 2008), and related algorithms (Bhatnagar et al., 2013). A related route uses the Robbins–Monro stochastic-approximation framework (Robbins & Monro, 1951; Bhatnagar et al., 2013; Borkar, 2008). The corresponding technical assumptions are given below.

Assumption 3.

1. *The mapping $f : \mathbb{R}^n \to \mathbb{R}^n$ is globally Lipschitz continuous.*
2. *The ODE $\dot{x}_t = f(x_t)$ has a closed bounded positively invariant set $\mathcal{H}$ that uniformly attracts bounded sets. Specifically, for every bounded set $B \subset \mathbb{R}^n$,*
$$\lim_{t\to\infty} \sup_{x_0 \in B} \operatorname{dist}(x_t, \mathcal{H}) = 0,$$
where x_t is the solution initialized at x_0.
3. *The iterates defined by (101) remain bounded with probability one, i.e.,*
$$\sup_{k\geq 0} \|\theta_k\|_2 < \infty$$
with probability one.
4. *The sequence $\{\varepsilon_k, \mathcal{G}_k, k \geq 1\}$ with $\mathcal{G}_k = \sigma(\theta_i, \varepsilon_i, i \leq k)$ is a martingale difference sequence. In addition, there exists a constant $C_0 < \infty$ such that for any initial $\theta_0 \in \mathbb{R}^n$, we have $\mathbb{E}[\|\varepsilon_{k+1}\|_2^2 | \mathcal{G}_k] \leq C_0(1 + \|\theta_k\|_2^2), \forall k \geq 0$ with probability one.*
5. *The step-sizes satisfy $\alpha_k > 0, \sum_{k=0}^{\infty} \alpha_k = \infty, \sum_{k=0}^{\infty} \alpha_k^2 < \infty$.*

Lemma 18 (ODE method for set convergence (Borkar, 2008, Chapter 2))**.** *Under Assumption 3, for any initial $\theta_0 \in \mathbb{R}^n$, we have*
$$\operatorname{dist}(\theta_k, \mathcal{H}) \to 0$$
with probability one.

The set-convergence form of the ODE method follows the same stochastic-approximation principle as the classical Robbins–Monro theorem (Robbins & Monro, 1951), but uses attraction of a closed invariant set rather than convergence to a single equilibrium. The main difference is that Lemma 18 assumes bounded iterates in Assumption 3, whereas Lemma 17 establishes boundedness without imposing it as a separate assumption. This advantage comes at the cost of stricter conditions on the underlying ODE model in Borkar & Meyn (2000), including a unique globally stable equilibrium and the existence and stability of the auxiliary ODE $\dot{x}_t = f_\infty(x_t)$.

K Detailed convergence analysis for Q-learning variants

This section provides the detailed convergence analysis summarized in Section 6. We prove convergence of Algorithm 1. The update in Algorithm 1 can be expressed as
$$Q_{k+1} = Q_k + \alpha_k(f(Q_k) + \varepsilon_{k+1}) \tag{103}$$
with the stochastic error
$$\varepsilon_{k+1} = (e_{a_k} \otimes e_{s_k})(r(s_k, a_k, s_k{}') + \gamma h(Q_k(s_k{}', \cdot)) - Q_k(s_k, a_k)) - f(Q_k)$$
and $f(Q_k) := DR + \gamma DPH(Q_k) - DQ_k$, where $P \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}$ is the state-action pair to state transition probability matrix defined in (1), $Q_k \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, $R \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is an enumeration of the expected reward defined in (1), $D \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ is a diagonal matrix whose diagonal elements are an enumeration of (14), and the order of the enumeration is defined by $d(s, a) = p(s)\beta(a|s) = (e_a \otimes e_s)^\top D(e_a \otimes e_s)$. Under Assumption 1, D is nonsingular and has strictly positive diagonal elements. Equation (103) is the mean-field decomposition of the asynchronous single-coordinate recursion in Algorithm 1. Since $\sum_{s,a} d(s, a) = 1$, setting $D = I$ does not correspond to a sampling distribution for this recursion when $|\mathcal{S} \times \mathcal{A}| > 1$. The synchronous full-update recursion is defined separately by updating all coordinates at each iteration; its mean field is $R + \gamma PH(Q) - Q$, which corresponds to $D = I$. The ODE stability results apply to both mean fields, while the calculations below are written for Algorithm 1. We first establish convergence of Algorithm 1 for the max, LSE, and mellowmax operators. Since the corresponding Bellman operator is a contraction, the proof follows by applying the Borkar–Meyn theorem (Borkar & Meyn, 2000).

Theorem 11. *Under the i.i.d. sampling model specified for Algorithm 1, assume that the step-sizes satisfy $\alpha_k > 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. For the LSE, mellowmax, and max operators, the iterates converge with probability one to the corresponding fixed point in (10).*

Proof. It suffices to verify Assumption 2 and apply the Borkar–Meyn theorem. Consider the system $\frac{dQ_t}{dt} = f(Q_t)$, where $f(Q) = DF(Q) - DQ$ and $F \in \{F_{\max}, F_{\text{lse}}^{\lambda}, F_{\text{mm}}^{\lambda}\}$. For the first statement, f is Lipschitz continuous because

$$\begin{aligned}\|f(x) - f(y)\|_\infty \leq& \|DF(x) - DF(y)\|_\infty + \|D(x-y)\|_\infty \\ \leq& \|D\|_\infty \|F(x) - F(y)\|_\infty + \|D\|_\infty \|x - y\|_\infty \\ \leq& \gamma \|D\|_\infty \|x - y\|_\infty + \|D\|_\infty \|x - y\|_\infty \\ =& (\gamma + 1)\|D\|_\infty \|x - y\|_\infty\end{aligned}$$

where Lemma 11 is used in the third line. For the second statement, we note that by Lemma 11, F is a contraction mapping with respect to $\|\cdot\|_\infty$, and by Theorem 1, the fixed point is the unique globally asymptotically stable equilibrium point. For the third statement, define $f_\infty(Q) := \gamma DPH_{\max}(Q) - DQ$. By Lemma 10, the convergence $H(cQ)/c \to H_{\max}(Q)$ is uniform on compact sets; in fact, for the LSE and mellowmax operators,

$$\sup_{Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}} \left\| \frac{H(cQ)}{c} - H_{\max}(Q) \right\|_\infty \leq \frac{\lambda \ln(|\mathcal{A}|)}{c},$$

while the difference is zero for the max operator. Therefore, for every compact set K,

$$\begin{aligned}&\sup_{Q \in K} \left\| \frac{f(cQ)}{c} - f_\infty(Q) \right\|_\infty \\ &\quad \leq \frac{\|DR\|_\infty}{c} + \gamma \|D\|_\infty \sup_{Q \in K} \left\| \frac{H(cQ)}{c} - H_{\max}(Q) \right\|_\infty \longrightarrow 0.\end{aligned}$$

Since all norms on the finite-dimensional space are equivalent, the same convergence holds in the 2-norm required by Assumption 2. For the fourth statement, consider

$$\frac{dQ_t}{dt} = f_\infty(Q_t),$$

where $f_\infty(Q) = D\bar{F}(Q) - DQ$ and $\bar{F}(Q) = \gamma PH_{\max}(Q)$. The same argument as in Lemma 11 shows that $\bar{F}$ is a contraction with respect to $\|\cdot\|_\infty$. Moreover, the fixed point is the origin, which is the unique globally asymptotically stable equilibrium point by Theorem 1.

For the fifth statement, define the information σ-algebra $\mathcal{G}_k := \sigma(Q_0, Q_i, \varepsilon_i, 1 \leq i \leq k)$ and the process $(M_k)_{k=0}^{\infty}$ by $M_k := \sum_{i=1}^{k} \varepsilon_i$. Under the i.i.d. sampling model and the definition of $f(Q_k)$ as the conditional mean update, $\mathbb{E}[\varepsilon_{k+1} \mid \mathcal{G}_k] = 0$. Hence,

$$\begin{aligned}\mathbb{E}[M_{k+1}|\mathcal{G}_k] =& \mathbb{E}\left[\sum_{i=1}^{k+1} \varepsilon_i \middle| \mathcal{G}_k\right] = \mathbb{E}[\varepsilon_{k+1}|\mathcal{G}_k] + \mathbb{E}\left[\sum_{i=1}^{k} \varepsilon_i \middle| \mathcal{G}_k\right] \\ =& \mathbb{E}\left[\sum_{i=1}^{k} \varepsilon_i \middle| \mathcal{G}_k\right] = M_k,\end{aligned}$$

where the final equality also uses the $\mathcal{G}_k$-measurability of M_k. Therefore, $(M_k)_{k=0}^{\infty}$ is a martingale, and $\varepsilon_{k+1} = M_{k+1} - M_k$ is a martingale difference. Moreover, for the second-moment bound in

item 5 of Assumption 2, we have

$$
\begin{aligned}
&\mathbb{E}[\|\varepsilon_{k+1}\|_2^2 \mid \mathcal{G}_k] \\
=\ &\mathbb{E}[\|(e_{a_k} \otimes e_{s_k})(r(s_k, a_k, s_k') + \gamma h(Q_k(s_k', \cdot)) - Q_k(s_k, a_k)) - f(Q_k)\|_2^2 \mid \mathcal{G}_k] \\
=\ &\mathbb{E}[\|(e_{a_k} \otimes e_{s_k})(r(s_k, a_k, s_k') + \gamma h(Q_k(s_k', \cdot)) - Q_k(s_k, a_k))\|_2^2 \mid \mathcal{G}_k] \\
&\quad - \mathbb{E}[\|f(Q_k)\|_2^2 \mid \mathcal{G}_k] \\
\le\ &\mathbb{E}[\|(e_{a_k} \otimes e_{s_k})(r(s_k, a_k, s_k') + \gamma h(Q_k(s_k', \cdot)) - Q_k(s_k, a_k))\|_2^2 \mid \mathcal{G}_k] \\
=\ &\mathbb{E}[(r(s_k, a_k, s_k') + \gamma h(Q_k(s_k', \cdot)) - Q_k(s_k, a_k))^2 \mid \mathcal{G}_k] \\
\le\ &3\mathbb{E}[r(s_k, a_k, s_k')^2 | \mathcal{G}_k] + 3\gamma^2 \mathbb{E}[h(Q_k(s_k', \cdot))^2 | \mathcal{G}_k] \\
&\quad + 3\mathbb{E}[Q_k(s_k, a_k)^2 | \mathcal{G}_k] \\
\le\ &3R_{\max}^2 + 3\gamma^2 \mathbb{E}\left[2\|Q_k\|_\infty^2 + C_h \mid \mathcal{G}_k\right] \\
&\quad + 3\mathbb{E}[\|Q_k\|_\infty^2 | \mathcal{G}_k] \\
\le\ &3R_{\max}^2 + 3\gamma^2 C_h \\
&\quad + (6\gamma^2 + 3)\mathbb{E}[\|Q_k\|_2^2 | \mathcal{G}_k]
\end{aligned}
$$

where $R_{\max} := \max_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} |r(s, a, s')|$. The second inequality uses $\|a+b+c\|_2^2 \le 3\|a\|_2^2 + 3\|b\|_2^2 + 3\|c\|_2^2$ for any $a, b, c \in \mathbb{R}^n$. We also use $|h(q)| \le \|q\|_\infty$ for $h \in \{h_{\max}, h_{\mathrm{mm}}^\lambda\}$, because the mellowmax value lies between the minimum and maximum coordinates of q, and $|h_{\mathrm{lse}}^\lambda(q)|^2 \le 2\|q\|_\infty^2 + 2(\lambda \ln(|\mathcal{A}|))^2$. Thus, $C_h = 0$ for $h \in \{h_{\max}, h_{\mathrm{mm}}^\lambda\}$ and $C_h = 2(\lambda \ln(|\mathcal{A}|))^2$ for $h = h_{\mathrm{lse}}^\lambda$. The final inequality follows from $\|\cdot\|_\infty \le \|\cdot\|_2$. This completes the proof. □

Next, we consider the Boltzmann softmax operator in Algorithm 1. Since the Boltzmann softmax operator is, in general, not non-expansive with respect to the ∞-norm (Asadi & Littman, 2017), the induced mapping F may fail to be a contraction in the ∞-norm. Consequently, the contraction-based approach used for the max, LSE, and mellowmax operators does not apply directly. Instead, we show that Algorithm 1 with the Boltzmann softmax operator converges to a bounded neighborhood of $Q_{\max}^*$.

Theorem 12. *Under the i.i.d. sampling model specified for Algorithm 1, assume that the step-sizes satisfy $\alpha_k > 0$, $\sum_{k=0}^\infty \alpha_k = \infty$, and $\sum_{k=0}^\infty \alpha_k^2 < \infty$. Set $w_i = 1/d_i$, $w_{\min} := \min_i w_i$, and $w_{\max} := \max_i w_i$. For the Boltzmann softmax operator, the corresponding iterates satisfy $\mathrm{dist}(Q_k, \mathcal{H}) \to 0$ with probability one, where*

$$
\mathcal{H} := \left\{ Q \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|} : \|Q - Q_{\max}^*\|_\infty \le \frac{\gamma \lambda \ln(|\mathcal{A}|) w_{\max}}{(1-\gamma) w_{\min}} \right\}.
$$

Proof. It suffices to verify Assumption 3 and apply Lemma 18. Consider the system $\frac{dQ_t}{dt} = f(Q_t)$, where $f(Q) = DF_{\mathrm{bz}}^\lambda(Q) - DQ$. For the first condition, f is Lipschitz continuous because

$$
\begin{aligned}
\|f(x) - f(y)\|_\infty \le&\left\|DF_{\mathrm{bz}}^\lambda(x) - DF_{\mathrm{bz}}^\lambda(y)\right\|_\infty + \|D(x-y)\|_\infty \\
\le&\|D\|_\infty \left\|F_{\mathrm{bz}}^\lambda(x) - F_{\mathrm{bz}}^\lambda(y)\right\|_\infty + \|D\|_\infty \|x-y\|_\infty \\
\le&(\gamma L_{\mathrm{bz}} + 1)\|D\|_\infty \|x-y\|_\infty,
\end{aligned}
$$

where Lemma 15 is used in the third line. By its definition, $\mathcal{H}$ is closed and bounded. By Theorem 10, it is positively invariant, and the explicit exponential estimate is uniform over every bounded set of initial conditions. Therefore, the second condition in Assumption 3 is satisfied. By Lemma 16, the iterates are bounded for the step-sizes considered here, so the third condition is also satisfied. Finally, the martingale-difference condition and the second-moment bound are proved as in the proof of Theorem 11, using $|h_{\mathrm{bz}}^\lambda(q)| \le \|q\|_\infty$. Applying Lemma 18 gives $\mathrm{dist}(Q_k, \mathcal{H}) \to 0$ with probability one. This completes the proof. □

From Theorem 12, we see that as λ decreases to zero, the limiting error set around the true solution $Q^*_{\max}$ becomes smaller. This is consistent with $h^{\lambda}_{\mathrm{bz}}(x) \to h_{\max}(x)$ as $\lambda \downarrow 0$.

L EXAMPLES

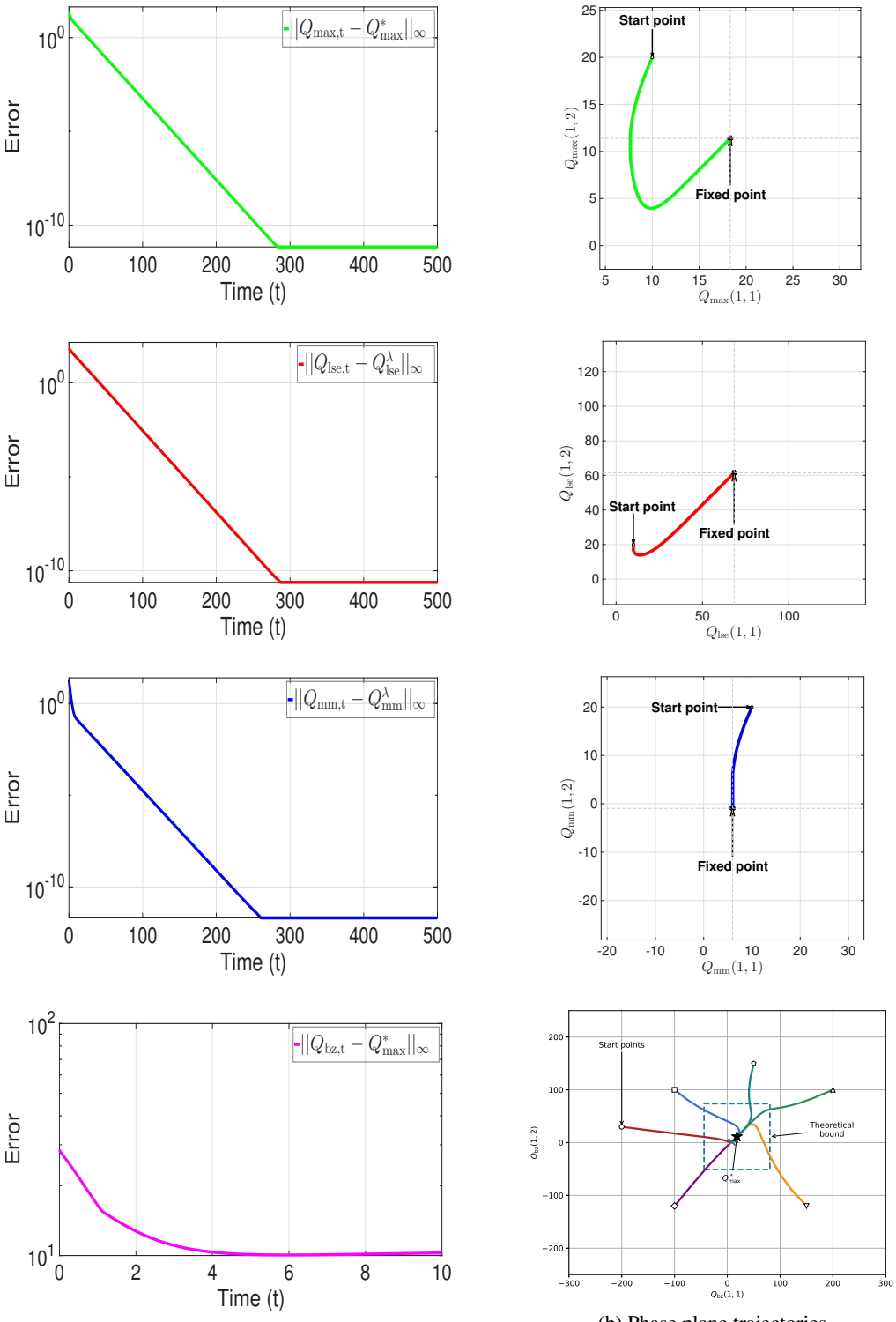
In this section, we present simulation results corresponding to Figure 2 and Figure 3. We consider a simple MDP with $|\mathcal{S}| = 4$, $|\mathcal{A}| = 2$, $\gamma = 0.9$, $R = [0, 5, 0, -5, 0, 5, 0, -5]^\top$, and

$$P_1 = \begin{bmatrix} 0.1 & 0.8 & 0.0 & 0.1 \\ 0.1 & 0.1 & 0.8 & 0.0 \\ 0.0 & 0.1 & 0.1 & 0.8 \\ 0.8 & 0.0 & 0.1 & 0.1 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 0.1 & 0.1 & 0.0 & 0.8 \\ 0.8 & 0.1 & 0.1 & 0.0 \\ 0.0 & 0.8 & 0.1 & 0.1 \\ 0.1 & 0.0 & 0.8 & 0.1 \end{bmatrix},$$

where P_1 and P_2 represent the state transition probability matrices under actions $a = 1$ and $a = 2$, respectively. The temperature parameter for the LSE, mellowmax, and Boltzmann operators is set to $\lambda = 10$. For the continuous-time simulations, we use the synchronous mean field with $D = I$.

The continuous-time ODE results are shown in Figure 2, whose first three rows correspond to the standard max, LSE, and mellowmax operators. In each error plot, $\|Q_t - Q_e\|_\infty$ decays approximately linearly on a logarithmic scale, consistent with the exponential convergence predicted by Theorem 3. The phase portraits in the right column use a common initial point and illustrate convergence toward the respective fixed points. We note that the trajectory plots are obtained by projecting the dynamics onto a two-dimensional plane for visualization, since the true Q_t is eight-dimensional. In the fourth row, we focus on the Boltzmann operator. Trajectories from different initial points are shown to converge toward a ball $\mathcal{H}$ around $Q^*_{\max}$, which is obtained theoretically in Theorem 10. The dashed square shown in the figure is the two-dimensional coordinate-plane cross-section of the actual set $\mathcal{H}$. Regardless of the initial points, the trajectories consistently remain around the theoretical error ball $\mathcal{H}$ defined in Theorem 10.

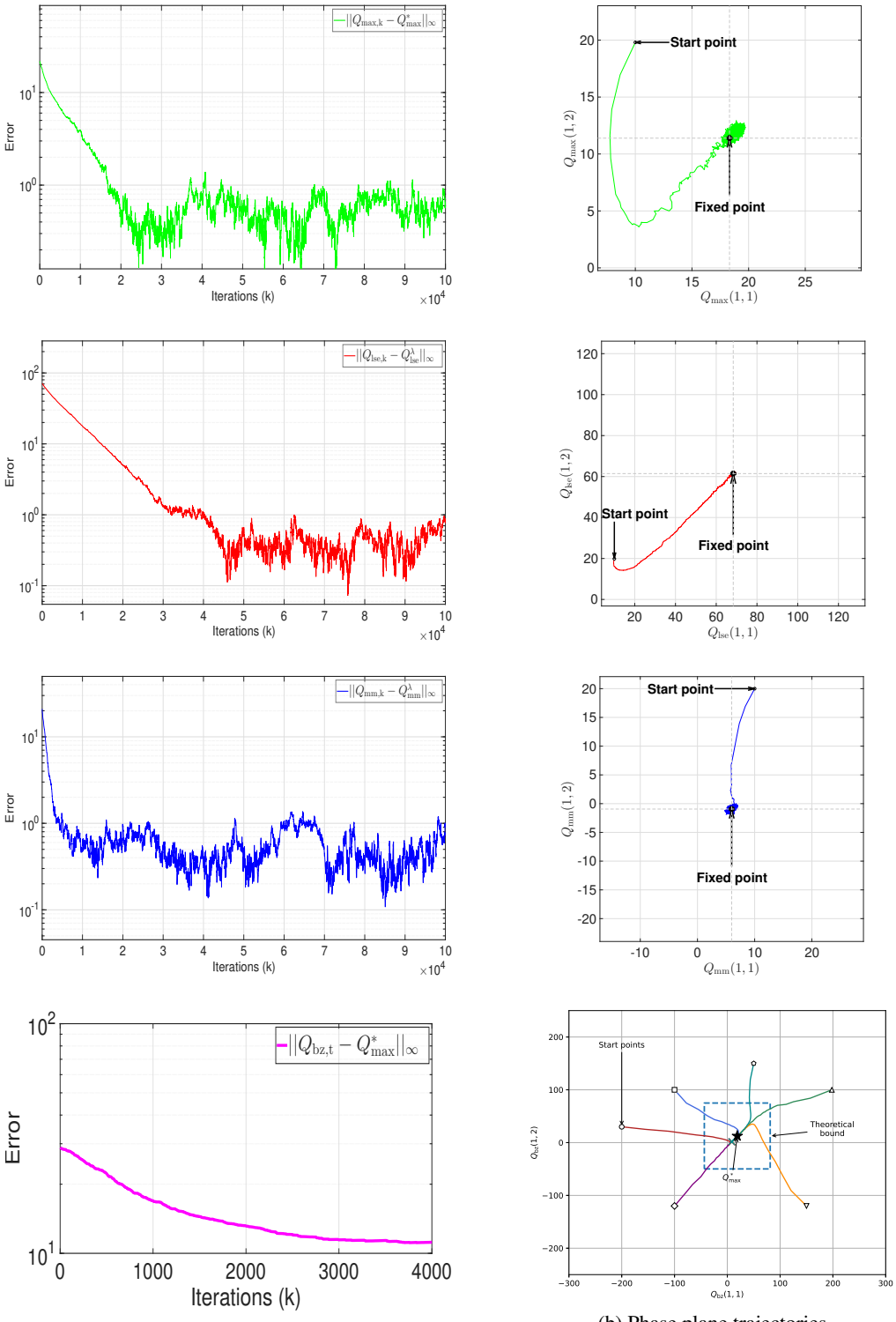
Figure 3 presents the results for the corresponding RL versions with a constant step-size $\alpha = 0.01$ over $100,000$ iterations. The simulation follows a single Markov trajectory, whereas the convergence theorems assume i.i.d. generative-model samples. In addition, the constant step-size does not satisfy the diminishing step-size condition in (102). Therefore, these simulations are illustrative and are not direct numerical instances of Theorems 11 and 12. Over the simulated horizon, the first three rows (max, LSE, and mellowmax) show trajectories approaching neighborhoods of their respective fixed points. We note that the trajectory plots are obtained by projecting the dynamics onto a two-dimensional plane for visualization, since the true Q_k is eight-dimensional. In the fourth row, the Boltzmann operator exhibits an initial decay followed by a persistent residual. This behavior is qualitatively consistent with convergence toward a neighborhood of $Q^*_{\max}$, but the constant-step experiment does not establish the probability-one set-convergence conclusion of Theorem 12. The dashed square shown in the figure is the two-dimensional coordinate-plane cross-section of the actual set $\mathcal{H}$. These observations provide qualitative illustrations of the ODE-based stability mechanisms in discrete-time stochastic implementations.



(a) Error convergence curves

(b) Phase plane trajectories

Figure 2: Empirical stability analysis under different operators. Panel (a) shows the error curves, and panel (b) shows the phase-plane trajectories toward their respective fixed points.



(a) Error convergence curves

(b) Phase plane trajectories

Figure 3: Empirical stability analysis under different operators in RL settings. Panel (a) shows the error curves, and panel (b) shows the phase-plane trajectories toward their respective fixed points.