The Projected Bellman Equation in Reinforcement Learning

Sean Meyn

***Abstract*—Q-learning has become an important part of the reinforcement learning toolkit since its introduction in the dissertation of Chris Watkins in the 1980s. In the original tabular formulation, the goal is to compute exactly a solution to the discounted-cost optimality equation, and thereby obtain the optimal policy for a Markov Decision Process. The goal today is more modest: obtain an approximate solution within a prescribed function class.**

The standard algorithms are based on the same architecture as formulated in the 1980s, with the goal of finding a value function approximation that solves the so-called projected Bellman equation. While reinforcement learning has been an active research area for over four decades, there is little theory providing conditions for convergence of these Q-learning algorithms, or even existence of a solution to this equation.

The purpose of this paper is to show that a solution to the projected Bellman equation does exist, provided the function class is linear and the input used for training is a form of ε-greedy policy with sufficiently small ε. Moreover, under these conditions it is shown that the Q-learning algorithm is stable, in terms of bounded parameter estimates. Convergence remains one of many open topics for research.

I. Introduction

Much of reinforcement learning concerns optimal control of state space models, typically cast in a Markov Decision Process (MDP) setting. Following standard notation from the control systems literature, the state process is denoted $\boldsymbol{X} = \{X_k : k \geq 0\}$, the input process $\boldsymbol{U} = \{U_k : k \geq 0\}$, and $c(X_k, U_k)$ denotes the one-stage cost at time k.

This paper concerns Q-learning algorithms, motivated by the same objective as in the first formulation of Watkins [54], [53]: the infinite-horizon optimal control problem, with state-action value function

$$Q^\star(x,u) = \min \sum_{k=0}^{\infty} \gamma^k \mathsf{E}[c(X_k, U_k) \mid X_0 = x\,,\ U_0 = u] \quad (1)$$

where $\gamma \in (0,1)$ is the discount factor. The minimum in (1) is over all history dependent input sequences. This is the *Q-function* of Q-learning.

Under standard assumptions an optimal input is obtained by state feedback, $U_k^* = \phi^\star(X_k^*)$ for each k, where an optimal policy $\phi^\star \colon \mathsf{X} \to \mathsf{U}$ is obtained via $\phi^\star(x) \in \arg\min_u Q^\star(x,u)$ for each x [8]. To avoid technicalities (in particular to avoid discussion of measurability), it is assumed in this paper that the state process evolves on a finite state space X, and the input takes values a finite set U.

The Q-function solves the Bellman equation $Q^\star = \mathcal{T}Q^\star$, in which the *Bellman operator* $\mathcal{T}$ acts on functions $H \colon \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ via

$$\mathcal{T}H\,(x,u) = c(x,u) + \mathsf{E}[\underline{H}(X_{n+1}) \mid X_n = x\,,\ U_n = u]$$

where throughout the paper $\underline{H}(x) := \min_u H(x,u)$, $x \in \mathsf{X}$. It is helpful to express the Bellman equation in sample path form: for any adapted input, and $k \geq 0$,

$$Q^\star(X_k\,, U_k) = c(X_k\,, U_k) + \gamma \mathsf{E}[\underline{Q^\star}(X_{k+1}) \mid \mathcal{F}_k] \quad (2)$$

in which $\mathcal{F}_k = \sigma\{X_i\,, U_i : i \leq k\}$ is the *history up to time* k.

The objective of Q-learning is to obtain an approximate solution among a parameterized class $\{Q^\theta : \theta \in \mathbb{R}^d\}$. Given an approximation we obtain a policy defined in analogy with the optimal policy,

$$\phi^\theta(x) \in \arg\min_u Q^\theta(x,u)\,, \quad x \in \mathsf{X}\,, \quad (3)$$

with some fixed rule in place in case of ties.

The most common criterion for success is the solution to the *projected Bellman equation*: find $\theta^* \in \mathbb{R}^d$ such that

$$0 = \mathsf{E}\big[\{c(X_n,U_n) + \gamma \underline{Q}^{\theta^*}(X_{n+1}) - Q^{\theta^*}(X_n,U_n)\}\zeta_n^{\theta^*}\big] \quad (4)$$

in which the expectation is in steady-state, and $\zeta_n^\theta = \nabla_\theta Q^\theta(X_n,U_n)$ for each n and $\theta \in \mathbb{R}^d$. Alternatives are discussed in Section IV-D.

The theoretical results in this paper are obtained in the special case of linear function approximation:

$$Q^\theta = \theta^\intercal \psi \qquad \textit{giving} \quad \zeta_n^\theta = \psi(X_n,U_n)\,, \quad (5)$$

with ψ a vector of basis functions. In this case the projected Bellman equation may be expressed in the Hilbert space notation of [50],

$$Q^{\theta^*} = \Pi \mathcal{T} Q^{\theta^*}$$

in which Π denotes the projection onto the d-dimensional subspace $L_2(\pi_{\theta^*})$; the definition of the probability mass function (pmf) π_{θ^*} may be found below (39b).

Much of the present article focuses on a generalization of the original algorithm of Watkins: For initialization $\theta_0 \in \mathbb{R}^d$, define the sequence of estimates recursively:

$$\theta_{n+1} = \theta_n + \alpha_{n+1}\mathcal{D}_{n+1}\zeta_n \quad (6a)$$

$$\mathcal{D}_{n+1} = c(X_n,U_n) + \gamma \underline{Q}^{\theta_n}(X_{n+1}) - Q^{\theta_n}(X_n,U_n)\,, \quad (6b)$$

in which $\{\alpha_n\}$ is a non-negative step-size sequence, $\{\zeta_n := \zeta_n^{\theta_n}\}$ are known as the *eligibility vectors* (entirely analogous to

Financial support from ARO award W911NF2010055 and NSF award CCF 2306023 is gratefully acknowledged. Many thanks to Caio Lauand at UF for comments on a draft manuscript.

S. Meyn is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA (email: meyn@ece.ufl.edu).

the eligibility vectors used in the TD(0) algorithm [46], [50]), and $\{\mathcal{D}_{n+1}\}$ is known as the temporal difference sequence.

The recursion (6a) reduces to Watkins' algorithm when using a tabular basis [54], [53] (see Section III-A for definitions). See [45], [47], [37] for a range of interpretations of the algorithm.

Soon after Q-learning was introduced, it was recognized that the algorithm can be cast within the framework of stochastic approximation (SA) [49], [24]. To explain the contributions and approach to analysis in this paper it is necessary to first explain why (6a) can also be cast as an SA recursion, subject to mild assumptions on the input used for training.

Some history The central open issue motivating the research surveyed in this paper is this: *it is not known if the projected Bellman equation* (4) *has a solution outside of very special cases*.

Success stories surveyed in [47] include the special case of binning [24], which is a generalization of the tabular setting, and the criterion in [35] and its improvement in [28], for which the assumptions are not easily verified in practice. The progress report in [47, Section 3.3.2] states that the *only known convergence result is due to Melo et al.* [35]. See [45, Section 11.2] for further discussion, and [22] for recent insight.

This open problem was a topic of discussion throughout the Simons program on reinforcement learning held in 2020, especially during the bootcamp lectures [48].

Thms. IV.1 and IV.5 resolve this open problem for Q-learning with optimistic training. Following several preliminaries, the proof of Thm. IV.1 is similar to the proof of convergence of TD(λ) learning from the dissertation of Van Roy [50], [51], and the assumptions are related to the assumptions in this prior work, even though the setting is very different.

An approximate projected Bellman equation is considered in [17], in which the minimum defining $\underline{Q}^{\theta_n}$ is replaced with a soft-min. Under assumptions similar to those imposed here they establish the existence of a solution [17, Theorem 5.1]. This result is similar to Prop. IV.2 (ii) of the present paper.

The recent paper [29] considers Q-learning with linear function approximation and oblivious training (meaning that the input used for training does not depend directly on parameter estimates). With sufficiently large regularization they obtain a unique equilibrium for the algorithm that approximates the solution to the projected Bellman equation.

Also recent is the work of [15], which is cast in a similar setting: Q-learning with linear function approximation and oblivious training. It is argued that the use of a target network combined with a carefully constructed projection of parameters improves performance, and their error bounds are consistent with their claims. While the paper is a significant step forward, they leave open the question of existence of a solution to the projected Bellman equation. With vanishing step-size, if convergence is established with or without a target network, the limit must be a solution to the projected Bellman equation (see [37, Proposition 5.10] for proof in the case of deterministic optimal control—the arguments in the stochastic setting are identical).

The lack of theory motivated Baird's gradient descent approach [4] as well as GQ learning [31], in which the root finding problem is replaced with the minimization of a loss function. See [3] for recent theory and Section IV-D for further discussion.

Zap stochastic approximation was introduced to ensure convergence, and also provide acceleration [20]. While originally proposed for Q-learning with linear function approximation, it was later shown to be convergent even with nonlinear function approximation [14], and the general technique applies to any application in which stochastic approximation is used. The Zap-Zero algorithm introduced in [37] and improved recently in [38] is designed to avoid matrix inversion.

Much recent research has focused on *linear* MDPs, notably [55], [25], in which the system dynamics are partially known: for a known "feature map" $\phi\colon \mathsf{X}\times\mathsf{U}\to\mathbb{R}^d$ and an unknown sequence of probability measures $\{\mu_i : 1\le i\le d\}$ on X, a linear MDP is assumed to have a controlled transition matrix of the form $P_u(x,x') = \sum \phi_i(x,u)\mu_i(x')$. There is now a relatively complete theory for this special case, in which the algorithm is designed based on knowledge of the feature map.

The reader is encouraged to see [5], [30], [33], [32] for new approaches to Q-learning based on convex programming approaches to MDPs. It is hoped that the analytical techniques presented in this paper may be adapted to these new algorithms.

Overview Section II surveys relevant recent results from stochastic approximation theory, and Section III provides a review of Q-learning, for which the vast majority of theory is restricted to linear function approximation and oblivious training.

Consideration of optimistic policies is postponed to Section IV, which contains the main contributions of the paper: if a smooth approximation of the ε-greedy policy is used for training, then under mild conditions the parameter estimates are bounded, and there exists a solution to the projected Bellman equation (see Thm. IV.1).

II. Background and Assumptions

This section is devoted to three topics: assumptions surrounding the MDP model, a brief summary of results from the theory of stochastic approximation, followed by assumptions surrounding the Q-learning algorithms to be considered.

Notation: $Q^\star$: Discounted-cost value function, (1).

- f_{n+1}, c_n and $\psi_{(n)}$, (38).
- ϕ^θ: Q^θ-greedy policy, (3). $\widetilde{\phi}^\theta$ randomized policy, (31).
- $\mathcal{C}^\Theta$: region of policy continuity, (29).
- $\underline{H}(x) := \min_u H(x,u)$, appearing in DCOE (2).
- Q-learning notation from (6b): step-size α_n, eligibility vector ζ_n, temporal difference $\mathcal{D}_{n+1}$.
- $\theta^*\in\mathbb{R}^d$: solves projected Bellman equation (4).
- θ_n parameter estimate, θ_n^{PR} PR-average, (12).
- Errors: $\tilde\theta_n = \theta_n - \theta^*$, $\tilde\theta_n^{\mathsf{PR}} = \theta_n^{\mathsf{PR}} - \theta^*$.
- P_u: controlled transition matrix, (8).
- $\bar f\colon \mathbb{R}^d\to\mathbb{R}^d$, vector field for mean-flow, (10b).
- $\bar f_\infty$, vector field for ODE@∞, (22).

- ϑ_t: solution to mean-flow, (11).
- $A = \partial_\theta \bar{f}$ and $A^* := A(\theta^*)$, (19).
- Σ_Θ asymptotic covariance, (24a).
- $\Sigma_\Theta^{\mathsf{PR}} = G\Sigma_\Delta^* G^\intercal$ with $G = -(A^*)^{-1}$ and Σ_Δ^*, (24b).
- $U_k = (1 - B_k)\mathcal{U}_k + B_k \mathcal{W}_k$ training policy, (30).

A. Markov Decision Process

While the search for an optimal policy may be restricted to static state feedback under the assumptions imposed below, in reinforcement learning it is standard practice to introduce randomization in policies as a way of introducing exploration during training. We restrict to randomized policies of the form,

$$U_k = \phi(X_k, \theta_k, I_k)\,, \qquad k \ge 0\,, \tag{7}$$

in which $\boldsymbol{I} = \{I_1, I_2, \dots\}$ is an i.i.d. sequence. Under the assumption that X and U are finite, we can assume without loss of generality that $\boldsymbol{I}$ evolves on a finite set.

The input-state dynamics are assumed to be defined by a controlled Markov chain, with controlled transition matrix P. For any randomized stationary policy, the following holds for each $x, x' \in \mathsf{X}$, $u \in \mathsf{U}$, and $k \ge 0$:

$$\mathsf{P}\{X_{k+1} = x' \mid X_k = x\,,\ U_k = u\} = P_u(x, x') \tag{8}$$

The dynamic programming equation $Q^\star = \mathcal{T}Q^\star$ (equivalently (2)) may be expressed, for $x \in \mathsf{X}$, $u \in \mathsf{U}$, by

$$Q^\star(x,u) = c(x,u) + \gamma \sum_{x' \in \mathsf{X}} P_u(x,x')\ \underline{Q}^\star(x') \tag{9}$$

B. What is stochastic approximation?

A fuller answer may be found in any of the standard monographs, such as [12] (see also [37] for a crash course).

The goal of SA is to solve the root finding problem $\bar{f}(\theta^*) = 0$, where the function is defined in terms of an expectation, $\bar{f}(\theta) = \mathsf{E}[f(\theta, \Phi)]$ for $\theta \in \mathbb{R}^d$ and with Φ a random vector. The general SA algorithm is expressed in two forms:

$$\theta_{n+1} = \theta_n + \alpha_{n+1} f(\theta_n\,, \Phi_{n+1}) \tag{10a}$$

$$= \theta_n + \alpha[\bar{f}(\theta_n) + \Delta_{n+1}]\,, \quad n \ge 0\,, \tag{10b}$$

where $\Delta_{n+1} := f(\theta_n, \Phi_{n+1}) - \bar{f}(\theta_n)$. It is assumed that the sequence $\{\Phi_n\}$ converges in distribution to Φ.

The algorithm is motivated by ordinary differential equation (ODE) theory, and this theory plays a large part in establishing convergence of (10a) along with convergence rates. These results are obtained by comparing solutions (10a) to solutions of the *mean flow*,

$$\tfrac{d}{dt}\vartheta_t = \bar{f}(\vartheta_t). \tag{11}$$

In particular, θ^* is a stationary point of this ODE.

Averaging A large step-size $\{\alpha_{n+1}\}$ in (10a) is desirable for quick transient response, but this typically leads to high variance. There is no conflict if the "noisy" parameter estimates are averaged. The averaging technique of Polyak and Ruppert defines

$$\theta_n^{\mathsf{PR}} = \frac{1}{n}\sum_{k=1}^{n} \theta_k\,, \qquad n \ge 1. \tag{12}$$

Thm. II.1 illustrates the value of this approach.

Basic SA assumptions The following are imposed in this section, and in some others that follow.

It is assumed that the step-size sequence $\{\alpha_n : n \ge 1\}$ is deterministic, satisfies $0 < \alpha_n \le 1$,

$$\sum_{n=1}^{\infty} \alpha_n = \infty \quad \textit{and} \quad \sum_{n=1}^{\infty} \alpha_n^2 < \infty \tag{13}$$

These conditions hold for $\alpha_n = gn^\rho$ with $\frac{1}{2} < \rho \le 1$ and $g > 0$; see [27] for theory justifying larger step-sizes obtained using $0 < \rho \le \frac{1}{2}$. We sometimes require two time-scale algorithms in which there is a second step-size sequence $\{\beta_n : n \ge 1\}$ that is relatively large:

$$\lim_{n\to\infty} \frac{\alpha_n}{\beta_n} = 0 \tag{14}$$

Parameter dependent noise A construction of the process $\boldsymbol{\Phi}$ appearing in the SA recursion reveals one complication. To make the construction entirely explicit, consider a state space realization of the MDP, $X_{k+1} = F(X_k, U_k, D_k)$ in which $\boldsymbol{D}$ evolves on a finite set, and F is a function taking values in the finite set X. In view of (7) we are in the setting of parameter-dependent noise: instead of one Markov chain, one considers a family $\{\boldsymbol{\Phi}^\theta : \theta \in \mathbb{R}^d\}$.

In the present setting, the Markov chain $\boldsymbol{\Phi}^\theta$ is defined by freezing the parameter in (7) to define

$$U_k^\theta = \phi(X_k^\theta, \theta, I_k) \quad X_{k+1}^\theta = F(X_k^\theta, U_k^\theta, D_k)\,, \qquad k \ge 0\,.$$

Note that $\boldsymbol{X}^\theta$ is itself a Markov chain on X, whose transition matrix is denoted P_θ. A simple choice is then $\Phi_k^\theta = (X_k, I_k, D_k)$. Letting $\xi = (x; \iota; \delta)$, $\xi' = (x'; \iota'; \delta')$ denote two state values, the transition matrix is denoted $\mathcal{P}_\theta$ and has the simple realization,

$$\mathcal{P}_\theta(\xi, \xi') = P_\theta(x, x')\mu_I(\iota')\mu_D(\delta') \tag{15}$$

where μ_I, μ_D are the respective pmfs for I_k, D_k (independent of k). It is assumed that $\mathcal{P}_\theta$ admits a unique invariant pmf ϖ_θ for each θ, from which we obtain an expression for the vector field in the mean flow,

$$\bar{f}(\theta) = \mathsf{E}[f(\theta, \Phi^\theta)]\,, \qquad \Phi^\theta \sim \varpi_\theta \tag{16}$$

That is, Φ^θ is distributed according to ϖ_θ in the expectation.

Theory for convergence of SA with parameter-dependent noise began in the seminal paper [36], and the theory is nearly as mature as in the classical setting with exogenous noise [26]. The following assumptions for the general SA recursion (10) are much stronger than those imposed in this prior work:

Assumptions for convergence:

SA1 The function f is globally Lipschitz continuous in its first variable, with subgradients satisfying $\sup_\theta |\partial_{\theta_i} f_j(\theta, \xi)| < \infty$ for each i, j and ξ.

SA2 For each θ, the time-homogeneous Markov chain $\boldsymbol{\Phi}^\theta$ evolves on a finite set. Moreover,

(i) A uniform minorization condition holds: for some $N \ge 1$, a constant $\delta_\Phi > 0$, and a state $\xi^\bullet$,

$$\sum_{k=1}^{N} \mathcal{P}_\theta^k(\xi, \xi^\bullet) \ge \delta_\Phi\,, \quad \textit{for each } \theta, \xi. \tag{17}$$

It is assumed moreover that $\mathbf{\Phi}^\theta$ is aperiodic.

Consequently, there is a unique invariant pmf ϖ_θ.

(ii) The transition matrix $\mathcal{P}_\theta$ is continuously differentiable in θ, with with vanishing gradient for large θ: for each $1 \le i \le d$,

$$\sup_{\xi,\xi',\theta} \big|\partial_{\theta_i} \mathcal{P}_\theta(\xi,\xi')\big| \|\theta\| < \infty \tag{18}$$

SA3 The mean flow (11) is globally asymptotically stable, with unique equilibrium θ^*.

Assumptions for convergence rates: The following is used to obtain useful bounds on the rate of convergence, which requires the existence of a linearization (at least in a neighborhood of θ^*). Denote

$$A(\theta) = \partial_\theta \bar{f}\,(\theta) \tag{19}$$

SA4 The derivative (19) is a bounded and continuous function of θ, and $A^* := A(\theta^*)$ is a Hurwitz matrix (its eigenvalues lie in the strict left hand plane).

Assumptions (SA1)–(SA3) imply that $\bar{f}$ is globally Lipschitz continuous. Hence the only part of (SA4) that goes beyond the previous assumptions is the Hurwitz condition.

These assumptions combined with theory in [36], [26] imply convergence of $\{\theta_n\}$ to θ^* almost surely from each initial condition, provided one more property is established:

The parameter sequence $\{\theta_n : n \ge 0\}$ is bounded with probability one from each initial condition. (20)

Lyapunov techniques provide a means of establishing (20):

(v4) For a globally Lipschitz continuous and C^1 function $V\colon \mathbb{R}^d \to [1,\infty)$, and a constant $\delta_v > 0$,

$$\frac{d}{dt} V(\vartheta_t) \le -\delta_v V(\vartheta_t)\,, \qquad \textit{when } \|\vartheta_t\| \ge \delta_v^{-1}. \tag{21}$$

The designation "v4" comes from an analogous bound appearing in stability theory of Markov chains [39].

An alternative that is often easily verified for RL algorithms is the so-called Borkar-Meyn theorem of [13], [12].

ODE@∞ The time-homogeneous ODE $\frac{d}{dt} x = \bar{f}_\infty(x)$ with vector field,

$$\bar{f}_\infty(\theta) := \lim_{r\to\infty} r^{-1} \bar{f}(r\theta). \tag{22}$$

We always have $\bar{f}_\infty(0) = 0$, which means that the origin is an equilibrium for the ODE@∞. It is also radially homogeneous, $\bar{f}_\infty(r\theta) = r\bar{f}_\infty(\theta)$ for any $\theta \in \mathbb{R}^d$ and $r > 0$. Based on these properties it is known that local asymptotic stability of the origin implies global exponential asymptotic stability [13].

Stability of the ODE@∞ is equivalent to (v4) whenever the limit (22) exits for each θ.

It is shown in [13] that (20) holds provided the ODE@∞ is locally asymptotically stable, and $\{\Delta_n\}$ appearing in (10b) is a martingale difference sequence. This statistical assumption does *not* hold in many applications of reinforcement learning. Extensions of [13] are given in [9], [41], [10].

The article [10] and its followup [27] require minimal assumptions on the Markov chain (there is no need for a finite state space). While these papers consider exogenous noise, the proof extends to the setting of this paper. See Appendix A for explanation.

Theorem II.1. *Suppose that (SA1) and (SA2) hold for the SA recursion* (10a)*, and in addition that the origin is locally asymptotically stable for the ODE@∞, or that (v4) holds. Then,*

(i) *The bound* (20) *holds in a strong sense: there is a fixed constant B_Θ such that for each initial condition (θ_0, Φ_0),*

$$\limsup_{n\to\infty} \|\theta_n\| \le B_\Theta \quad a.s.. \tag{23}$$

(ii) *If in addition (SA3) holds then $\lim_{n\to\infty} \theta_n = \theta^*$ almost surely from each initial condition.* ■

Based on theory in [10] we can expect to also establish mean-square convergence rates—part (iii) that follows is stated as a conjecture at this stage. A functional Central Limit Theorem is easily obtained under the assumptions of Thm. II.1 and also (SA4) [7], [12], implying that the limits in (24) will hold under an L_p bound on $\{\tilde{\theta}_n/\sqrt{\alpha_n} : n \ge 1\}$ for some $p > 2$, where the tilde denotes error: $\tilde{\theta}_n = \theta_n - \theta^*$, $\tilde{\theta}_n^{\mathsf{PR}} = \theta_n^{\mathsf{PR}} - \theta^*$. An L_4 bound is established in [10] for exogenous noise.

(iii) Suppose that (SA1)–(SA4) hold, and that $\alpha_n = gn^\rho$, $n \ge 1$, with $\frac{1}{2} < \rho < 1$ and $g > 0$. We then have convergence in mean square, and the following limits exist and are finite:

$$\lim_{n\to\infty} \frac{1}{\alpha_n} \mathsf{E}[\tilde{\theta}_n \tilde{\theta}_n^\intercal] = \Sigma_\Theta \tag{24a}$$

$$\lim_{n\to\infty} n\mathsf{E}[\tilde{\theta}_n^{\mathsf{PR}} \{\tilde{\theta}_n^{\mathsf{PR}}\}^\intercal] = \Sigma_\Theta^{\mathsf{PR}} \tag{24b}$$

The covariance matrix $\Sigma_\Theta^{\mathsf{PR}}$ is minimal in a matricial sense, made precise in [43], [40]. It has the explicit form $\Sigma_\Theta^{\mathsf{PR}} = G\Sigma_\Delta^* G^\intercal$ in which $G = -(A^*)^{-1}$, and

$$\Sigma_\Delta^* = \sum_{k=-\infty}^{\infty} \mathsf{E}[\Delta_k^* \{\Delta_k^*\}^\intercal] \tag{25}$$

where $\{\Delta_k^* := f(\theta^*, \Phi_k^{\theta^*}) : k \in \mathbb{Z}\}$, with $\mathbf{\Phi}^{\theta^*}$ a stationary version of the Markov chain on the two-sided time interval. An alternative representation for Σ_Δ^* is contained in Appendix A.

In practice we rarely make use of these formulae: the covariance matrix $\Sigma_\Theta^{\mathsf{PR}}$ can be estimated using the batch means method, which requires performing many relatively short runs with distinct initial conditions [2].

A criterion for stationary points The existence of a suitable Lyapunov function implies the existence of a stationary point.

Proposition II.2 (Lyapunov Criterion for Existence of a Stationary Point)**.** *For an ODE* (11) *with globally Lipschitz continuous vector field, suppose there is a function $V\colon \mathbb{R}^d \to \mathbb{R}_+$ with locally Lipschitz continuous gradient, satisfying for some $b^{\text{II.2}}$,*

$$\nabla V(\theta)^\intercal \bar{f}(\theta) \le -1\,, \quad \textit{whenever } \|\theta\| \ge b^{\text{II.2}}.$$

Suppose moreover that V is convex and coercive. Then there exists a solution to $\bar{f}(\theta^*) = 0$.

Proof. Let $L_\delta(\theta) = \theta + \delta\bar{f}(\theta)$ for $\theta \in \mathbb{R}^d$, with $\delta > 0$ to be chosen. For $\delta > 0$ sufficiently small we construct a convex and compact set S_δ for which $L_\delta(\theta) \in S_\delta$ for each $\theta \in S_\delta$.

It follows from Brouwer's fixed-point theorem that there is a solution to $L_\delta(\theta^*) = \theta^*$. This is equivalent to the desired conclusion $\bar{f}(\theta^*) = 0$.

Denote $b_\delta = \sup\{V(L_\delta(\theta)) : \|\theta\| \le b^{\text{II.2}}\}$, and $S_\delta = \{\theta : V(\theta) \le b_\delta\}$; a convex and compact set subject to the assumptions on V.

We next show that S_δ is invariant under L_δ if δ is small. We consider two cases, based on whether or not θ lies in the set $S = \{\theta : \|\theta\| \le b^{\text{II.2}}\}$

1. If $\theta \in S_\delta \cap S$, then $L_\delta(\theta) \in S_\delta$ by construction of S_δ.

2. If $\theta \in S_\delta \setminus S$ then we apply convexity combined with the drift condition: denoting $\theta^+ = L_\delta(\theta)$,

$$V(\theta) \ge V(\theta^+) + \nabla V(\theta^+)^\intercal(\theta - \theta^+) = V(\theta^+) - \delta \nabla V(\theta^+)^\intercal \bar{f}(\theta)$$

Since the gradient is locally Lipschitz continuous and $\bar{f}$ is globally Lipschitz continuous, there is b_v satisfying

$$V(\theta) \ge V(\theta^+) - \delta \nabla V(\theta)^\intercal \bar{f}(\theta) - b_v \delta^2\,, \qquad \theta \in S_\delta \setminus S$$

The value of b_v can be chosen independent of $\delta \in (0, 1]$.

Under the assumed drift condition this gives $V(\theta^+) \le V(\theta) - \delta + b_v\delta^2$. Choosing $\delta = 1/b_v$ gives $V(\theta^+) \le V(\theta) \le b_\delta$, in which the second inequality holds because $\theta \in S_\delta \setminus S$. Hence $L_\delta(\theta) = \theta^+ \in S_\delta$ as desired. ■

When stability of the mean flow cannot be established, stability can typically be assured using a matrix gain algorithm.

Zap stochastic approximation. This is a two time-scale algorithm introduced in [20]. For initialization $\theta_0 \in \mathbb{R}^d$, and $\widehat{A}_0 \in \mathbb{R}^{d\times d}$, obtain the sequence of estimates $\{\theta_n : n \ge 0\}$ recursively:

$$\theta_{n+1} = \theta_n - \alpha_{n+1}\widehat{A}_{n+1}^{-1} f(\theta_n\,, \Phi_{n+1}) \tag{26a}$$

$$\widehat{A}_{n+1} = \widehat{A}_n + \beta_{n+1}[A_{n+1} - \widehat{A}_n]\,, \tag{26b}$$

with $A_{n+1} := \partial_\theta f_{n+1}(\theta_n)$.

The two gain sequences $\{\alpha_n\}$ and $\{\beta_n\}$ satisfy (14). This ensures that the ODE approximation for the parameter estimates is the Newton-Raphson flow $\frac{d}{dt}\bar{f}(\vartheta) = -\bar{f}(\vartheta)$. Hence stability is assured if $\|\bar{f}\|$ is coercive [14].

The recursion (26b) requires modification for the applications considered here, in which the transition law for the Markov chain depends on the parameter estimate. See discussion surrounding eq. (51) in Section IV-D.

C. Compatible assumptions for Q-learning

The basic Q-learning algorithm (6a) is an instance of stochastic approximation, for which we can apply general theory subject to assumptions on the input used for training (recall (7)). Two settings are considered:

Oblivious training This means that (7) simplifies to

$$U_k = \phi(X_k, I_k)\,, \qquad k \ge 0\,, \tag{27}$$

in which it is always assumed that $\{I_k\}$ is i.i.d..

It follows that the pair process $\{(X_k, U_k) : k \ge 0\}$ is a time homogeneous Markov chain. It is assumed to be unichain (i.e., the invariant pmf π is unique). In the expression $f_{n+1}(\theta_n) = f(\theta_n\,, \Phi_{n+1})$ we take $\{\Phi_k = (X_k; X_{k+1}; U_k) : k \ge 0\}$, which is also a time homogeneous Markov chain, for which its invariant pmf is also unique and easily expressed in terms of π and the controlled transition matrix.

If the function class is linear $\{Q^\theta = \theta^\intercal\psi : \theta \in \mathbb{R}^d\}$, then the autocorrelation matrix is assumed full rank

$$R_0 = \mathsf{E}_\pi[\psi(X_n, U_n)\psi(X_n, U_n)^\intercal] \tag{28}$$

where the expectation is taken in steady-state

Optimistic training In this non-oblivious approach the input sequence depends on the parameter sequence, and is designed to approximate the Q^θ-greedy policy ϕ^θ defined in (3).

There are only a finite number of deterministic stationary policies, so ϕ^θ is necessarily discontinuous in θ. The region on which continuity holds is denoted

$$\mathcal{C}^\Theta = \left\{ \begin{matrix} \theta \in \mathbb{R}^d : \text{there is } \varepsilon > 0 \text{ s.t. } \phi^\theta(x) = \phi^{\theta'}(x) \\ \text{for all } x \text{ when } \|\theta - \theta'\| \le \varepsilon \end{matrix} \right\} \tag{29}$$

The training policy is taken of the form,

$$U_k = (1 - B_k)\mathcal{U}_k + B_k\mathcal{W}_k \tag{30}$$

in which $\{B_k\}$ is an i.i.d. Bernoulli sequence with $\mathsf{P}\{B_k = 1\} = \varepsilon$, and $\{\mathcal{W}_k\}$ is an i.i.d. sequence taking values in U and independent of $\{B_k\}$. The U-valued random variable $\mathcal{U}_k$ depends on the parameter θ_k, and is independent of $(B_k; \mathcal{W}_k)$ for each k.

The sequences $\{U_k, \mathcal{U}_k : k \ge 0\}$ are defined by randomized stationary policies $\{\widetilde{\phi}^\theta\,, \widetilde{\phi}_0^\theta : \theta \in \mathbb{R}^d\}$. Both $\widetilde{\phi}^\theta(\cdot \mid x)$ and $\widetilde{\phi}_0^\theta(\cdot \mid x)$ are pmfs on U for each x and θ. Based on the assumptions imposed after (30), we have

$$\begin{aligned} \mathsf{P}\{U_k = u \,|\mathcal{F}_k^-; X_k = x\} &= \widetilde{\phi}^{\theta_k}(u \mid x) \\ &= (1 - \varepsilon)\widetilde{\phi}_0^{\theta_k}(u \mid x) + \varepsilon\nu_{\mathcal{W}}(u) \end{aligned} \tag{31}$$

with $\nu_{\mathcal{W}}$ the common pmf for $\{\mathcal{W}_k\}$, and $\mathcal{F}_k^- = \sigma\{X_i, U_i : i < k; B_i, \mathcal{W}_i : i \le k\}$ (a partial history of observations up to iteration k).

Special cases are described in the following.

1. ε-greedy. The choice $\mathcal{U}_k = \phi^{\theta_k}(X_k)$, so that

$$\widetilde{\phi}_0^\theta(u \mid x) = \mathbb{1}\{u = \phi^\theta(x)\} \tag{32}$$

The mean flow has many attractive properties (see Prop. A.6 in the Appendix). However, because $\{\phi^\theta : \theta \in \mathbb{R}^d\}$ is a piecewise constant function of θ, it follows that the vector field $\bar{f}$ is not continuous in θ as required in Thm. II.1.

2. Gibbs approximation For fixed constant $\kappa > 0$ define

$$\widetilde{\phi}_0^\theta(u \mid x) = \frac{1}{\mathcal{Z}_\kappa^\theta(x)} \exp\bigl(-\kappa Q^\theta(x, u)\bigr) \tag{33}$$

in which $\mathcal{Z}_\kappa^\theta(x)$ is normalization. This is indeed an approximation of (32): for $\theta \in \mathcal{C}^\Theta$,

$$\lim_{r\to\infty} \frac{1}{\mathcal{Z}_\kappa^{r\theta}(x)} \exp\bigl(-\kappa Q^{r\theta}(x, u)\bigr) = \mathbb{1}\{u = \phi^\theta(x)\} \tag{34}$$

The limit (34) has two important implications. First is that the vector field $\bar{f}_\infty$ for the ODE@∞ is unchanged whether we consider (32) or its smooth approximation (33). Second is that discontinuity of $\bar{f}_\infty$ implies that $\bar{f}$ is not globally Lipschitz continuous, which violates an assumption of Thm. II.1.

3. Tamed Gibbs approximation This is a modification of (33) in which κ depends on θ:

$$\widetilde{\phi}_0^\theta(u \mid x) = \frac{1}{\mathcal{Z}_\kappa^\theta(x)} \exp\bigl(-\kappa_\theta Q^\theta(x,u)\bigr) \tag{35}$$

For analysis the following structure is helpful: choose a large constant $\kappa_0 > 0$, and assume that

$$\kappa_\theta \begin{cases} = \frac{1}{\|\theta\|}\kappa_0 & \|\theta\| \ge 1 \\ \ge \frac{1}{2}\kappa_0 & \textit{else} \end{cases} \tag{36}$$

This will be called the (ε, κ_0)-*tamed Gibbs policy* when it is necessary to make the policy parameters explicit.

The equality in (36) ensures the following holds for all x, u:

$$\widetilde{\phi}^{r\theta}(u \mid x) = \widetilde{\phi}^{\theta}(u \mid x) \text{ for all } r \ge 1 \text{ and } \|\theta\| \ge 1. \tag{37}$$

The Q-learning algorithm (6) can be cast as stochastic approximation when the input is defined using any of the training policies described above, in which we take $\Phi_{n+1} = (X_n, X_{n+1}, U_n)$ since these three variables appear in (6).

It is assumed in Thm. II.1 that $\mathbf{\Phi}$ is *exogenous*—its transition matrix does not depend on the parameter sequence. Fortunately, there is now well developed theory that allows for parameter-dependent dynamics for $\mathbf{\Phi}$ in the SA recursion (10a)—see the recent paper [56] for history and recent results. In particular, theory of convergence and asymptotic statistics is now mature.

The question is then, *how can we apply SA theory to make statements about convergence and convergence rates?*

III. Trouble with Tabular

We begin with a useful representation for the mean flow vector field $\bar{f}$ for Q-learning with linear function approximation (5), for arbitrary basis. The projected Bellman equation (4) is the root finding problem, $\bar{f}(\theta^*) = 0$.

To avoid long equations we adopt the shorthand notation,

$$\begin{aligned} &f_{n+1}(\theta_n) = f(\theta_n, \Phi_{n+1}) \\ &c_n = c(X_n, U_n)\,, \quad \psi_{(n)} = \psi(X_n, U_n)\,. \end{aligned} \tag{38}$$

The eligibility vector in (6a) is then $\zeta_n = \psi_{(n)}$.

If the parameter θ is frozen, so that $U_k \sim \widetilde{\phi}^\theta(\cdot \mid X_k)$ for each k, then the controlled state process $\boldsymbol{X}^\theta$ is a time homogeneous Markov chain with transition matrix,

$$P_\theta(x,x') := \sum_u \widetilde{\phi}^\theta(u \mid x) P_u(x,x')\,, \qquad x, x' \in \mathsf{X}\,. \tag{39a}$$

The pair process $\{(X_k, U_k) : k \ge 0\}$ is also Markovian, with transition matrix denoted

$$T_\theta(z,z') := P_u(x,x')\widetilde{\phi}^\theta(u' \mid x')\,, \tag{39b}$$

for each $z = (x,u)\,,\ z' = (x',u') \in \mathsf{X} \times \mathsf{U}$. It is assumed that T_θ has a unique invariant pmf π_θ.

Of course, the parameter θ is never frozen in any algorithm. The transition matrices P_θ and T_θ are introduced for analysis.

Q-learning in the form (6a) is an instance of stochastic approximation, with mean flow vector field,

$$\bar{f}(\theta) = \mathsf{E}_{\pi_\theta}[\psi_{(n)} \mathcal{B}(X_n, U_n; \theta)]\,, \tag{40a}$$

$$\mathcal{B}(x,u;\theta) = c(x,u) - Q^\theta(x,u) + \gamma \sum_{x'} P_u(x,x') \underline{Q}^\theta(x') \tag{40b}$$

An alternative formula is valuable for analysis.

Lemma III.1. *The vector field* (40a) *may be expressed*

$$\begin{aligned} \bar{f}(\theta) &= A(\theta)\theta - b(\theta) \\ &\textit{with} \quad A(\theta) = -\mathsf{E}_{\pi_\theta}\bigl[\psi_{(n)}\{\psi_{(n)} - \gamma\psi^\theta_{(n+1)}\}^\intercal\bigr] \\ &\qquad\quad b(\theta) = -\mathsf{E}_{\pi_\theta}[\psi_{(n)} c_n] \end{aligned} \tag{41}$$

and $\psi^\theta_{(n+1)} = \psi(X_{n+1}, u)$ *with* $u = \phi^\theta(X_{n+1})$.

The vector field $\bar{f}$ *is globally Lipschitz continuous when using the training policy* (31) *with tamed Gibbs policy* (35)*, for any value of* $\varepsilon \in [0,1]$ *and* $\kappa > 0$. ■

The representation (41) follows directly from (40b). Lipschitz continuity follows Lemma A.1 combined with Prop. A.2 (each postponed to the Appendix).

Note that $\varepsilon = 1$ corresponds to an oblivious policy, so the lemma provides a large collection of policies for which $\bar{f}$ fits the standard SA theory. The tamed Gibbs policy is the only choice among the optimistic training rules for which $\bar{f}$ satisfies the smoothness conditions required in Thm. II.1.

In the remainder of this section we restrict to oblivious training. The main results of this paper in Section IV concern optimistic training.

A. Tabular Q-learning, the good and the bad

In the tabular setting we take $d = |\mathsf{X}| \times |\mathsf{U}|$ in (5), and

$$\psi_i(x,u) = \mathbb{1}\{(x,u) = (x^i, u^i)\}\,, \qquad x \in \mathsf{X}\,,\ u \in \mathsf{U} \tag{42}$$

where $\{(x^i, u^i) : 1 \le i \le d\}$ is any ordering of state-action pairs. Hence $Q^{\theta_n}(x^i, u^i) = \theta_n(i)$ for each n, i.

It is typical to use a diagonal matrix gain,

$$\theta_{n+1} = \theta_n + \alpha_{n+1} G_n \mathcal{D}_{n+1} \zeta_n \tag{43}$$

in which $G_n^{-1}(i,i)$ indicates the number of times the pair (x^i, u^i) is visited up to time n (set to unity when this is zero).

The mean flow (11) associated with (43) is

$$\tfrac{d}{dt}\vartheta_t = A(\vartheta_t)\vartheta_t - b \tag{44}$$

with b the d-dimensional vector with entries $b_i = -c(x^i, u^i)$, and the matrix-valued function A is piecewise constant.

The good news: The statistical properties of the algorithm are attractive because $\{\Delta_{n+1}\}$ appearing in (10b) is a martingale difference sequence in the tabular setting.

The best news is stability: we have $A(\theta) = -[I - \gamma M(\theta)]$, in which $M_{i,j}(\theta) = P_{u^i}(x^i, x^j)\mathbb{1}\{u^j = \phi^\theta(x^j)\}$. The induced operator norm of $M(\theta)$ in ℓ_∞ is no greater than one, meaning $\max_i |\sum_j M_{i,j}(\theta) v_j| \le \|v\|_\infty := \max_i |v_i|$ for any vector v

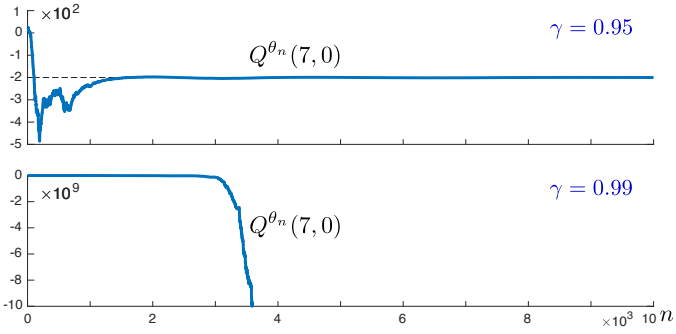

Fig. 1. Evolution of the Q-function approximations for two values of discount factor, and using an ε-greedy policy with common value of $\varepsilon = 0.5$.

and any θ. It follows that the ℓ_∞ norm serves as a Lyapunov function: Letting $\tilde{\vartheta}_t = \vartheta_t - \theta^*$ and $V(\theta) = \|\theta\|_\infty$,

$$\tfrac{d}{dt} V(\vartheta_t) \le -(1-\gamma) V(\vartheta_t)$$

This is how convergence is established for tabular Q-learning.

The bad: The matrix $A(q)$ has an eigenvalue at $-(1-\gamma)$ for each θ, which is a reason for slow convergence when the discount factor is close to unity. One consequence is that the asymptotic covariance Σ_Θ appearing in (24a) is not finite if $\gamma > 1/2$ and the step-size sequence is $\alpha_n = 1/n$ (see [20] and the sample complexity analysis that followed in [52]).

B. Change your goals

A reader with experience in SA would counter that $\alpha_n = 1/n$ is a poor choice of step-size. Use instead $\alpha_n = 1/n^\rho$, with $\rho \in (\frac{1}{2}, 1)$, and then average using (12) to obtain $\{\theta_n^{\text{PR}}\}$. It is found that averaging fails for this example for large discount factors, even though it is known that these estimates achieve the optimal asymptotic covariance [37], [19], [18].

The observed numerical instability is a consequence of the eigenvalue at $-(1-\gamma)$ for $A^* := A(\theta^*)$ (recall (19)). The eigenvalue can be moved through a change in objective. For example, construct an algorithm that estimates the *relative Q-function*,

$$H^\star(x,u) = Q^\star(x,u) - \delta \langle \nu, Q^\star \rangle \tag{45}$$

where ν is a fixed pmf on $\mathsf{X} \times \mathsf{U}$ and $\delta > 0$. Subtracting a constant doesn't change the minimizer over u, and has enormous benefits.

The function $H^\star$ satisfies a DP equation which motivates *relative Q-learning*. It is shown in [21] that the eigenvalues of A^* remain bounded away from the imaginary axis uniformly for all $0 \le \gamma \le 1$, resulting in much faster convergence. See [37] for generalizations.

IV. Stability with Optimism

The theory surveyed in the preceding section imposed oblivious training. In the case of Watkins' Q-learning this assumption was imposed in part for historical reasons, though we will see that the analysis is somewhat more complex when we consider parameter dependent policies. The technical challenges for Zap Q-learning are far more interesting because

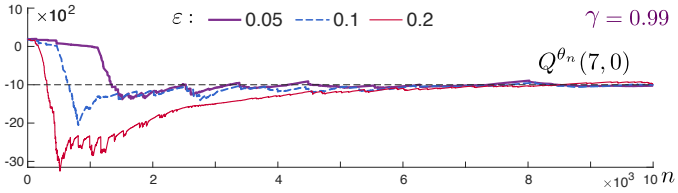

Fig. 2. Evolution of the Q-function approximations when using an ε-greedy policy. Convergence holds when $\varepsilon > 0$ is sufficiently small.

the definition of the linearization $A(\theta)$ is not obvious. See the conclusions for further discussion.

We begin with a motivating example.

A. Baird's star example

We refer the reader to the source [4]—the final page contains a full description of the model considered in the experiments surveyed here. See [45] for a fuller discussion.

There are seven states $\mathsf{X} = \{1, \dots, 7\}$ and two actions $\mathsf{U} = \{0, 1\}$, in which $X_{k+1} = 7$ with probability one whenever $U_k = 0$. In [4] it is assumed the cost is identically zero. We take $c(x,u) = 0$ if $x \le 6$ and $c(7,u) = -10$ (independent of u). The Q-function is linearly parameterized with dimension $d = 14$. With a well-motivated oblivious policy it was shown that the parameter estimates diverge when the discount factor is sufficiently large.

Fig. 1 shows trajectories from the Q-learning algorithm (6a) with an ε-greedy policy using $\varepsilon = 0.5$. The ideal behavior is that $Q^{\theta_n}(x,u) \to Q^\star(x,u) = -10/(1-\gamma)$ as $n \to \infty$ when $(x,u) = (7,0)$. The figure shows convergence when $\gamma = 0.95$, but the parameters are divergent with discount factor $\gamma = 0.99$.

With the larger discount factor we obtain stability when using a smaller value of $\varepsilon > 0$. Fig. 2 shows typical results for three small values. The dashed line indicates $Q^\star(7,0)$.

The step-size sequence was taken to be $\alpha_n = \min(\bar{\alpha}, g/n^\rho)$ using $g = 1/(1-\gamma)$, $\rho = 0.85$, and $\bar{\alpha} = 0.1$ in each run. The Matlab code is available on arXiv—see the Appendix of [38].

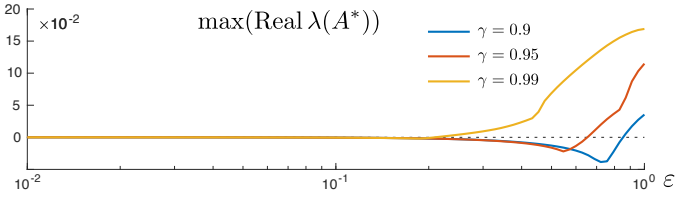

Fig. 3. The maximum eigenvalue of A^* as a function of ε. The matrix is Hurwitz for sufficiently small $\varepsilon > 0$, but some eigenvalues approach zero with vanishing ε.

See Section III-B for an explanation for slow convergence with a large discount factor, and [21] for explanation of the choice $g = 1/(1-\gamma)$ based on consideration of the linearization matrix A^*. Fig. 3 shows a plot of the maximum real part of A^* as a function of $\varepsilon > 0$, estimated via Monte-Carlo. For larger values of $\varepsilon > 0$ we see that A^* is not Hurwitz for the three choices of discount factor. There is also trouble for very small $\varepsilon > 0$: The discussion following Thm. II.1 *suggests* that the asymptotic covariance will be very large

when $\max(\text{Real}\,\lambda(A^*))$ is close to zero, but the covariance Σ^*_Δ must also be considered to make any conclusions.

Applications to change detection Similar experiments were conducted in [16] for application to quickest change detection. Much like in Baird's example it was found that ε-greedy training was successful, but only for extremely small values of $\varepsilon > 0$. Zap Q-learning was far more reliable over a large range of $\varepsilon \in (0, 1)$: testing revealed that the final parameter $\theta^\bullet$ defining the policy was nearly optimal. However, the matrix $A(\theta^\bullet)$ was not Hurwitz, which suggests that the standard algorithm (6) is unable to recover $\theta^\bullet$.

These findings illustrate that significant understanding of RL theory is essential for practical success in many applications.

B. Sufficient optimism

The main result of this paper shows how exploration using a policy of the form (30) encourages stability of the Q-learning algorithm (6) with linear function approximation (5). Analysis of (6) requires the family of autocorrelation matrices,

$$R^{\Theta}(\theta) = \mathsf{E}_{\pi_\theta}\big[\psi(X_n, \phi^\theta(X_n))\psi(X_n, \phi^\theta(X_n))^\intercal\big] \tag{46a}$$

$$R^{\mathcal{W}}(\theta) = \mathsf{E}_{\pi_\theta}\big[\psi(X_n, \mathcal{W}_n)\psi(X_n, \mathcal{W}_n)^\intercal\big] \tag{46b}$$

$$R(\theta) = \mathsf{E}_{\pi_\theta}[\psi_{(n)}\psi_{(n)}^\intercal] = (1-\varepsilon)R^{\Theta}(\theta) + \varepsilon R^{\mathcal{W}}(\theta) \tag{46c}$$

The expectations are in steady-state, with stationary pmf π_θ induced by the randomized stationary policy with fixed parameter.

A special case is considered in the assumptions, in which we take $\varepsilon = 1$, and the randomized policy is then denoted $\widetilde{\phi}^{\mathcal{W}}$, giving $\widetilde{\phi}^{\mathcal{W}}(\,\cdot \mid x) = \nu_{\mathcal{W}}(u)$ for all x, u. The (assumed unique) invariant pmf is denoted $\pi_{\mathcal{W}}$, and the autocorrelation matrix

$$R^{\mathcal{W}} = \mathsf{E}_{\pi_{\mathcal{W}}}\big[\psi(X_n, \mathcal{W}_n)\psi(X_n, \mathcal{W}_n)^\intercal\big] \quad \textit{using } U_k = \mathcal{W}_k \textit{ for all } k. \tag{46d}$$

We have $R^{\mathcal{W}} > 0$ in Baird's star example whenever the distribution of $\mathcal{W}_k$ is not degenerate.

The following assumptions are required in the main results of this section:

The randomized policy $\widetilde{\phi}^{\mathcal{W}}$ gives rise to an aperiodic and uni-chain Markov chain, with unique invariant pmf $\pi_{\mathcal{W}}$, and the autocorrelation matrix $R^{\mathcal{W}}$ defined in (46d) is positive definite. (47a)

The inverse temperature κ_θ is twice continuously differentiable (C^2) in θ, and the first and second derivatives of κ_θ are continuous and bounded. (47b)

We also require small $\varepsilon > 0$ in specification of the policies. Denote

$$\varepsilon_\gamma := \frac{(1-\gamma)^2}{(1-\gamma)^2 + \gamma^2} \tag{48}$$

Theorem IV.1. *Consider the Q-learning algorithm* (6a) *with linear function approximation, and training policy* (30) *defined using the tamed Gibbs policy* (35)*. Suppose moreover that* (47) *holds. Then, for any $\varepsilon \in (0, \varepsilon_\gamma)$ there is $\kappa_{\varepsilon,\gamma} > 0$ for which the following hold using the (ε, κ_0)-tamed Gibbs policy, using $\kappa_0 \ge \kappa_{\varepsilon,\gamma}$:*

(i) *The parameter estimates $\{\theta_n\}$ are bounded: there is a fixed constant B_Θ, independent of $\kappa_0 \ge \kappa_{\varepsilon,\gamma}$, such that* (23) *holds with probability one from each initial condition.*

(ii) *There exists at least one solution to the projected Bellman equation* (4)*.*

See Section IV-C for an extension of (ii) to the ε-greedy policy.

To see why (i) is plausible, consider an algorithm approximating (6a), in which the minimum defining $\underline{Q}^{\theta_n}(X_{n+1})$ is replaced by substitution of the input used for training:

$$\begin{aligned} \theta_{n+1} &= \theta_n + \alpha_{n+1}\widetilde{\mathcal{D}}_{n+1}\zeta_n\,. \\ \widetilde{\mathcal{D}}_{n+1} &= c_n - Q^{\theta_n}(X_n, U_n) + \gamma Q^{\theta_n}(X_{n+1}, U^-_{n+1}) \end{aligned} \tag{49}$$

in which U^-_{n+1} is obtained by sampling from $\widetilde{\phi}^\theta(\,\cdot \mid x)$ using $x = X_{n+1}$ and $\theta = \theta_n$. The recursion (49) is a variant of the SARSA algorithm [42], [45].

Stability of the ODE@∞ is then relatively easy, from which we obtain the following:

Proposition IV.2. *Consider the recursion* (49) *with linear function approximation, and training policy* (30) *defined as the (ε, κ_0)-tamed Gibbs policy* (35) *with $\varepsilon \in (0, 1)$ and $\kappa_0 > 0$. Suppose moreover that* (47) *holds. Then, we obtain the conclusions of Thm. IV.1:*

(i) *The parameter estimates $\{\theta_n\}$ are bounded with probability one from each initial condition.*

(ii) *There exists at least one solution θ^* to $\bar{f}(\theta^*) = 0$, with $\bar{f}$ the mean flow for* (49)*.*

The proof of Thm. IV.1 is postponed to the Appendix—we proceed here with the proof of Prop. IV.2.

To begin, suppose that U^-_{n+1} is replaced by U_{n+1} in (49). This does not lead to a practical algorithm, since $\widetilde{\mathcal{D}}_{n+1}$ would then depend on θ_{n+1}, but it may be regarded as an approximation since $\theta_{n+1} \approx \theta_n$. The approximation leads to a recursion similar to the TD(0) learning algorithm:

$$\theta_{n+1} = \theta_n + \alpha_{n+1}\big[\psi_{(n)}c_n - \psi_{(n)}\{\psi_{(n)} - \gamma\psi_{(n+1)}\}^\intercal\theta_n\big]$$

This motivates consideration of the family of autocorrelation matrices $R_k(\theta) = \mathsf{E}_{\pi_\theta}[\psi_{(n+k)}\psi_{(n)}^\intercal]$ for $n, k \ge 0$, so that $R_0(\theta) = R(\theta)$ in the notation (46c).

The vector field for the mean flow associated with (49) is Lipschitz continuous and has an attractive form in terms of the vector and matrix valued functions,

$$b(\theta) = -\mathsf{E}_{\pi_\theta}[\psi_{(n)}c_n]\,, \quad A(\theta) = -R_0(\theta) + \gamma R_{-1}(\theta)$$

Lemma IV.3. *Under the assumptions of Prop. IV.2 the following hold for* (49)*:*

(i) *The vector field for the mean flow is $\bar{f}(\theta) = A(\theta)\theta - b(\theta)$.*

(ii) *The limit defining $\bar{f}_\infty$ in* (22) *exists and may be expressed $\bar{f}_\infty(\theta) = A_\infty(\theta)\theta$ where $A_\infty(\theta) = A(\theta/\|\theta\|)$ for $\theta \ne 0$.*

Proof. Identification of $\bar{f}$ follows immediately from (49) since θ is held fixed in the definition of the mean flow. The representation of the ODE@∞ follows from structure of the

policy highlighted in (37), which implies the following for all $r \geq 1$ and $\theta \in \mathbb{R}^d$ satisfying $\|\theta\| \geq 1$:

$$\pi_{r\theta} = \pi_\theta\,, \quad A(r\theta) = A(\theta)\,, \quad \textit{and} \quad b(r\theta) = b(\theta)$$

■

Lemma IV.4. *Suppose that* (47a) *holds. Then, for the recursion* (49) *there exists $\delta_\psi > 0$, independent of θ such that*

$$\begin{aligned} R_0(\theta) &\geq \delta_\psi I \qquad &\textit{for all } \theta \in \mathbb{R}^d \\ \theta^\intercal A(\theta)\theta &\leq -(1-\gamma)\delta_\psi \qquad &\textit{for all } \theta \in \mathbb{R}^d,\ \|\theta\| \geq 1. \end{aligned}$$

Proof. The proof of the lower bound on $R_0(\theta)$ is identical to the proof of Lemma A.3 in the Appendix. From Lemma IV.3 (i) we have for $\theta \in \mathbb{R}^d$ satisfying $\|\theta\| \geq 1$,

$$\begin{aligned} \theta^\intercal A(\theta)\theta &= -\theta^\intercal R_0(\theta)\theta + \gamma\theta^\intercal R_{-1}(\theta)\theta \\ &\leq -(1-\gamma)\theta^\intercal R_0(\theta)\theta \leq -(1-\gamma)\delta_\psi\|\theta\|^2 \end{aligned}$$

■

Proof of Prop. IV.2. Let $V_1(\theta) = \frac{1}{2}\|\theta\|^2$ and apply Lemmas IV.3 and IV.4 to obtain, whenever $\|\vartheta_t\| \geq 1$,

$$\begin{aligned} \tfrac{d}{dt}V_1(\vartheta_t) = \vartheta_t^\intercal \bar{f}(\vartheta_t) &= \vartheta_t^\intercal\{A(\vartheta_t)\vartheta_t - b(\vartheta_t)\} \\ &\leq -\delta_1\|\vartheta_t\|^2 + \|\vartheta_t\|\|b(\vartheta_t)\| \end{aligned}$$

with $\delta_1 = (1-\gamma)\delta_\psi$. This gives, with $\bar{b} = \sup_\theta \|b(\theta)\| < \infty$,

$$\tfrac{d}{dt}V_1(\vartheta_t) \leq -\tfrac{1}{2}\delta_1\|\vartheta_t\|^2\,, \qquad \|\vartheta_t\| \geq \max(1, 2\bar{b})$$

We then obtain (v4) using $V(\theta) = \sqrt{V_1(\theta)} = \|\theta\|$ for $\|\theta\| \geq \max(1, 2\bar{b})$ (modified in a neighborhood of the origin to impose the C^1 condition):

$$\tfrac{d}{dt}V(\vartheta_t) \leq -\delta_v V(\vartheta_t)\,, \qquad \|\vartheta_t\| \geq \max(1, 2\bar{b})\,,$$

with $\delta_v = \delta_1/4$. Part (i) then follows from Thm. II.1 (i) and part (ii) from Prop. II.2. ■

C. Implications to the ε-greedy policy

A full analysis of Q-learning using the ε-greedy policy for training is beyond the scope of this paper due to discontinuity of the vector field. We find here that Thm. IV.1 admits a partial extension.

We consider here the mean flow (40a), and also the algorithm with matrix gain, whose mean flow vector field is

$$\bar{f}^{\mathsf{zap}}(\theta) = -[A(\theta)]^{-1}\bar{f}(\theta) = -\theta + [A(\theta)]^{-1}b(\theta)\,, \quad \theta \in \mathcal{C}^\Theta$$

This defines the dynamics expected when using Zap Q-learning based on (26).

The set $\mathcal{C}^\Theta$ in (29) may be expressed as the disjoint union,

$$\mathcal{C}^\Theta = \bigcup_i \mathcal{C}_i^\Theta$$

in which each $\mathcal{C}_i^\Theta$ is an open convex polyhedron, with $\phi^\theta = \phi^{\theta'}$ for all $\theta, \theta' \in \mathcal{C}_i^\Theta$. Consequently, both $\bar{f}$ and $\bar{f}^{\mathsf{zap}}$ are constant on each set $\mathcal{C}_i^\Theta$.

For each $\theta \in \mathbb{R}^d$, denote by Φ^θ the set of all randomized Q^θ-greedy policies: if $\widetilde{\phi} \in \Phi^\theta$ then

$$\sum_u \widetilde{\phi}(u \mid x)Q^\theta(x,u) = \underline{Q}^\theta(x)\,, \qquad x \in \mathsf{X}\,.$$

If $\theta \in \mathcal{C}^\Theta$ then $\Phi^\theta = \{\phi^\theta\}$ is a singleton.

Theorem IV.5. *Suppose that* (47a) *holds. Then, the following hold for the mean flows associated with the Q-learning algorithm with ε-greedy training, provided $0 < \varepsilon < \varepsilon_\gamma$:*

(i) *There exists $\theta^* \in \mathbb{R}^d$ and $\widetilde{\phi}^* \in \Phi^{\theta^*}$ such that $\bar{f}(\theta^*) = 0$, with $\bar{f}$ defined in* (40a) *in which the expectation is taken in steady-state using π_{θ^*} obtained from the randomized policy,*

$$\widetilde{\phi}^{\theta^*}(u \mid x) = (1-\varepsilon)\widetilde{\phi}^*(u \mid x) + \varepsilon\nu_{\mathcal{W}}(u) \tag{50}$$

(ii) *If $\theta^* \in \mathcal{C}^\Theta$ then θ^* is locally asymptotically stable for the mean flow with vector field $\bar{f}$.*

(iii) *If $\theta^* \in \mathcal{C}_i^\Theta$ for some i, then θ^* is locally asymptotically stable for the mean flow with vector field $\bar{f}^{\mathsf{zap}}$, with domain of attraction including all of $\mathcal{C}_i^\Theta$.*

Proof. The proof of (i) is contained in Appendix E.

If $\bar{f}(\theta^*) = 0$ with $\theta^* \in \mathcal{C}^\Theta$, it then follows from the definition of the vector field that $\theta^* = [A(\theta^*)]^{-1}b(\theta^*)$. Consequently, for θ in a neighborhood of θ^* contained in $\mathcal{C}^\Theta$,

$$\bar{f}(\theta) = A(\theta^*)\big(\theta - \theta^*\big)$$

See Prop. A.6 for a proof that $A(\theta^*)$ is Hurwitz, so that θ^* is locally asymptotically stable as claimed in (ii).

We have under the assumptions of (iii),

$$\bar{f}^{\mathsf{zap}}(\theta) = -\theta + \theta^*\,, \quad \theta \in \mathcal{C}_i^\Theta$$

If $\vartheta_0 \in \mathcal{C}_i^\Theta$ it follows that the solution to $\frac{d}{dt}\vartheta_t = \bar{f}^{\mathsf{zap}}(\vartheta_t)$ is given by $\vartheta_t = \theta^* + [\vartheta_t - \theta^*]e^{-t}$. Convexity of $\mathcal{C}_i^\Theta$ ensures that $\vartheta_t \in \mathcal{C}_i^\Theta$ for all t, which completes the proof of (iii). ■

D. Extensions from the basic algorithm

Theory for the basic Q-learning algorithm will have implications to other algorithms:

Stochastic Gradient Descent (SGD) The mean flow for the GQ learning algorithm of [31] can be expressed $\frac{d}{dt}\vartheta_t = -\nabla_\theta L\,(\vartheta_t)$, with $L(\theta) = \bar{f}(\theta)^\intercal Z\bar{f}(\theta)$ and Z positive definite. The theory in the present paper provides sufficient conditions under which the non-singularity condition (L3) of [31] holds, and most important is the new finding that $\min_\theta L(\theta) = 0$.

A numerical challenge with GQ learning or gradient descent [4], [3] is that the condition number of the linearization is squared. In particular, for GQ-learning the linearization is expressed $\frac{d}{dt}\tilde{\vartheta}_t \approx -[\nabla_\theta^2 L\,(\theta^*)]\tilde{\vartheta}_t$, and $\lambda_{\mathsf{min}}(\nabla_\theta^2 L\,(\theta^*)) = O(|1-\gamma|^2)$. Hence some of the bad news reviewed in Section III-A is exacerbated using these SGD methods.

Relative Q-learning The mean flow vector field for this algorithm is a modification of (41): $\bar{f}(\theta) = [A(\theta) - Z]\theta - b(\theta)$, in which Z is a rank one matrix chosen by the user (the specific form follows from (45)). [37, Proposition 9.23] (adapted from [21]) tells us that that the maximum eigenvalue of A^* remains bounded away from 0 for relative Q-learning in the tabular setting. It is conjectured that Thm. IV.1 can be extended to these algorithms, with ε_γ sufficiently small, but independent of the discount factor $\gamma \in (0,1)$. This may require a fresh look at the choice of Z.

Regularization Similar to relative Q-learning, in the regularized Q-learning algorithm of [29] the mean flow becomes $\bar{f}(\theta) = [A(\theta) - Z]\theta - b(\theta)$. It is possible that the conditions on Z for convergence may be relaxed based on the theory in this paper.

Double Q-learning Stability has been established in the tabular setting [23]. The algorithm with linear function approximation has a $2d$-dimensional mean flow whose state $\vartheta_t = [\vartheta_t^A; \vartheta_t^B]$ satisfies the mean-flow equations

$$\tfrac{d}{dt}\vartheta_t = \begin{bmatrix} -R(\vartheta_t) & \gamma M^B(\vartheta_t) \\ \gamma M^A(\vartheta_t) & -R(\vartheta_t) \end{bmatrix} \vartheta_t - \begin{bmatrix} b(\vartheta_t) \\ b(\vartheta_t) \end{bmatrix}$$

with b defined in (41), where $\theta = [\theta^A; \theta^B] \in \mathbb{R}^{2d}$ in the definition of π_θ. The matrix R is unchanged from (46c), and

$$M^q(\theta) = \mathsf{E}_{\pi_\theta}\left[\psi_{(n)}\psi_{(n+1)}^{\theta^q}\}^\intercal\right], \qquad q = A, B\,.$$

It is not clear how to establish stability of the mean flow using the techniques of this paper: the intuition following Prop. IV.2 is valid only if (following a transient) each of the policies $\phi^{\vartheta_t^A}$, $\phi^{\vartheta_t^B}$ approximates the ε-greedy policy for each of the two Q-function approximations.

Zap Q-learning Success requires that $\partial_\theta \bar{f}\,(\theta)$ be non-singular for "most" θ. Based on theory surrounding the Actor-Critic method, we have for a policy of the form (31),

$$\partial_\theta \bar{f}\,(\theta) = A_0(\theta) + Z(\theta)\,, \tag{51}$$

$A_0(\theta) := \mathsf{E}_{\pi_\theta}[\partial_\theta f_{n+1}\,(\theta)]$ and $Z(\theta) := \mathsf{E}_{\pi_\theta}[\hat{f}_n\,(\theta)\Lambda_n(\theta)^\intercal]$.

The random vectors in these definitions are:

- f_{n+1} is given in (38).
- $\Lambda_n(\theta) = \nabla_\theta \log \widetilde{\phi}^\theta(u \mid x)$, evaluated at $u = U_n$ and $x = X_n$, the *score function* associated with the randomized policy.
- $\hat{f}_n$ solves a certain Poisson equation. If the transition matrix T_θ is aperiodic, then for a stationary realization of $\{X_n, U_n : n \geq 0\}$ we have

$$\mathsf{E}_{\pi_\theta}[\hat{f}_n(\theta)\Lambda_n(\theta)^\intercal] = \sum_{k=0}^{\infty} \mathsf{E}_{\pi_\theta}[[f_{n-k}(\theta) - \bar{f}(\theta)]\Lambda_n(\theta)^\intercal]$$

Based on this representation we can obtain unbiased estimates of $\partial_\theta \bar{f}\,(\theta_n)$ by adopting concepts from actor-critic algorithms. See [37, Ch. 10] for a survey in the style of this paper.

Analysis of the resulting algorithms will be considered in future research. The most likely path to success is to establish that the matrix norm of $Z(\theta)$ is uniformly small, since approximations in this paper imply that A_0 has desirable properties.

V. Conclusions

In the vast majority of application-oriented papers on reinforcement learning the policy used for training is not oblivious. Motivation for ε-greedy policies and their variants comes largely from the belief that optimism will lead to faster training. This paper makes clear that optimism is invaluable to ensure algorithmic stability.

There are of course a myriad of open questions.

- Can we obtain sharp bounds establishing that optimism leads to better sample complexity bounds (or perhaps better bounds on the asymptotic covariance)? The results in this paper combined with [15] might provide a first step. Ideally we want bounds sharp enough to inform the choice of ε, and for this it is likely best to begin with examination of asymptotic statistics.
- There is a large literature on SA with discontinuous dynamics, such as [6], [11]. It may be possible to extend Thm. IV.1 to ε-greedy policies (going beyond Thm. IV.5).
- Extension to average cost optimal control is straightforward through consideration of [1], as well as extension to relative Q-learning [21], [37].
- We should consider other paradigms for algorithm design. The recent approaches [34], [5], [30] are based on the linear programming formulation of optimal control, and are likely to lead to algorithms that respect desired performance bounds.

Appendix

This Appendix contains proofs of the main technical results concerning Q-learning subject to the linear function class assumption (5) and optimistic training. We begin with some general SA theory.

A. Stochastic approximation theory

Convergence theory begins with a decomposition of the "disturbance" appearing in (10b):

$$\Delta_{n+1} = \mathcal{W}_{n+2} - \mathcal{T}_{n+2} + \mathcal{T}_{n+1} - \alpha_{n+1}\Upsilon_{n+2}\,. \tag{52}$$

Under the assumptions of [10], the dominating term in analysis is $\{\mathcal{W}_{n+2}\}$, which is a martingale difference sequence. The sequence $\{-\mathcal{T}_{n+2} + \mathcal{T}_{n+1}\}$ is telescoping so is insignificant in an ODE approximation. The sequence $\{\alpha_{n+1}\Upsilon_{n+2}\}$ is small relative to the parameter sequence.

Lemma A.1 that follows provides bounds on the terms on the right hand side of (52) that are identical to those used in [10] to obtain the conclusions of Thm. II.1 and finer results. Consequently, the proof of Thm. II.1 is obtained by following identical steps in this prior work.

Moreover, if the limits in (24) hold, then the covariance matrix (25) admits the alternate representation

$$\Sigma_\Delta^* = \mathsf{E}[\mathcal{W}_k^*\{\mathcal{W}_k^*\}^\intercal]$$

in which $\{\mathcal{W}_k^* : k \in \mathbb{Z}\}$ is a stationary version of the martingale difference sequence obtained from a stationary realization of $\mathbf{\Phi}^{\theta^*}$.

The terms in (52) admit representations in terms of a family of solutions to Poisson's equation, following [36]. For each θ, one version of the fundamental matrix associated with $\mathcal{P}_\theta$ is the matrix inverse $\mathcal{Z}_\theta = [I - \widetilde{\mathcal{P}}_\theta]^{-1}$, where $\widetilde{\mathcal{P}}_\theta(\xi, \xi') := \mathcal{P}_\theta(\xi, \xi') - \varpi_\theta(\xi')$ for each ξ, ξ'. Under (SA2) this may be expressed as the sum, $\mathcal{Z}_\theta = \sum_{n=0}^{\infty} \widetilde{\mathcal{P}}_\theta^n$. Writing $\hat{f}(\theta, \xi) = \sum_{\xi'} \mathcal{Z}_\theta(\xi, \xi')f(\theta, \xi)$ and $\Delta(\theta, \xi) := f(\theta, \xi) - \bar{f}(\theta)$, the desired Poisson equation is solved:

$$\sum_{\xi'} \mathcal{P}_\theta(\xi, \xi')\hat{f}(\theta, \xi') = \hat{f}(\theta, \xi) - \Delta(\theta, \xi)$$

Lemma A.1. *Under the assumptions of Thm. II.1,*

$$\sup_{\theta}\big\|\partial_{\theta_i}\bar{f}\,(\theta)\big\| < \infty \quad \sup_{\xi,\theta}\big\|\partial_{\theta_i}\bar{f}\,(\theta,\xi)\big\| < \infty \tag{53}$$

Moreover, (52) *holds with*

$$\begin{aligned}\mathcal{W}_{n+2} &:= \hat{f}(\theta_n,\Phi_{n+2}) - \mathsf{E}[\hat{f}(\theta_n,\Phi_{n+2}) \mid \mathcal{F}_{n+1}],\\ \mathcal{T}_{n+1} &:= \hat{f}(\theta_n,\Phi_{n+1}),\\ \Upsilon_{n+2} &:= -\frac{1}{\alpha_{n+1}}\big[\hat{f}(\theta_{n+1},\Phi_{n+2}) - \hat{f}(\theta_n,\Phi_{n+2})\big]\end{aligned} \tag{54}$$

Proof. The decomposition with terms in (54) is obtained exactly as in [10, Section 2] following [36], so it remains to establish the bounds on $\bar{f}$ and $\hat{f}$ in (53).

Assumption (SA2) is a uniform Doeblin condition that implies $\|\mathcal{Z}_\theta\|$ is uniformly bounded in θ [39]. Write the invariance equation in operator theoretic notation $\varpi_\theta \mathcal{P}_\theta = \varpi_\theta$. On differentiating both sides we obtain the sensitivity formula of [44]: $\partial_{\theta_i}\varpi_\theta = \varpi_\theta[\partial_{\theta_i}\mathcal{P}_\theta]\mathcal{Z}_\theta$. Consequently, the invariant pmf ϖ_θ enjoys the same smoothness properties as $\mathcal{P}_\theta$, giving $\sup_{\xi,\theta,i}|\partial_{\theta_i}\varpi_\theta(\xi)|\|\theta\| < \infty$. This and (SA1) imply the bound on $\partial_{\theta_i}\bar{f}\,(\theta)$ in (53).

The bound for the solution to Poisson's equation follows from the identity $\partial_{\theta_i}\mathcal{Z}_\theta = \mathcal{Z}_\theta\big[\partial_{\theta_i}\widetilde{\mathcal{P}}_\theta\big]\mathcal{Z}_\theta$. This combined with (SA1), (SA2) completes the proof of (53). ∎

In the following we explain why the SA assumptions hold for Q-learning under the training policies of interest.

B. Validating SA assumptions for Q-learning

The proof of the following is postponed to the end of this subsection:

Proposition A.2. *Under the assumptions of Thm. IV.1 the Q-learning algorithm satisfies Assumptions (SA1) and (SA2) with parameter-dependent noise* $\Phi_k = (X_k, I_k, D_k)$.

Consider the oblivious policy defined by $U_k \equiv \mathcal{W}_k$ in the definition of $R^{\mathcal{W}}$ in (46d). The transition matrix for the joint process $\{(X_k, U_k) : k \ge 0\}$ can be obtained from (39b):

$$T_{\mathcal{W}}(z,z') = P_u(x,x')\nu_{\mathcal{W}}(u')\,,$$

for any $z = (x,u)$, $z' = (x',u') \in \mathsf{X}\times\mathsf{U}$. The invariance equation $\pi_{\mathcal{W}}(z') = \sum_z \pi_{\mathcal{W}}(z)T_{\mathcal{W}}(z,z')$ implies that the invariant pmf is product form:

$$\pi_{\mathcal{W}}(z') = \mu_{\mathcal{W}}(x')\nu_{\mathcal{W}}(u')\,, \qquad z' = (x',u') \in \mathsf{X}\times\mathsf{U}\,.$$

in which $\mu_{\mathcal{W}}(x') = \sum_u \pi_{\mathcal{W}}(x',u)$ is the steady-state marginal distribution of $\boldsymbol{X}$ under this policy.

Similar notation is adopted for each of the invariant pmfs,

$$\mu_\theta(x) = \sum_u \pi_\theta(x,u)\,, \qquad x \in \mathsf{X}\,,\ \theta \in \mathbb{R}^d\,.$$

These are the invariant pmfs for the transition matrices $\{P_\theta\}$. Recall that these transition matrices and $\{T_\theta\}$ are defined in (39).

Let X_0 denote the support of $\mu_{\mathcal{W}}$ and U_0 the support of $\nu_{\mathcal{W}}$.

Lemma A.3. *Suppose that* (47a) *holds, so that* $\pi_{\mathcal{W}}$ *is the unique invariant pmf. Consider any one of the three choices of* $\{\mathcal{U}_k\}$ *used in* (30) *with* $\varepsilon < 1$ *and any choice of* κ *in the case of* (33) *or* $\{\kappa_\theta\}$ *in the case of* (35)*. The following conclusions then hold:*

(i) T_θ *is aperiodic, and for some* $N \ge 1$, $\delta_T > 0$,

$$\sum_{k=1}^{N} T_\theta^k(z,z') \ge \delta_T\,, \quad z \in \mathsf{X}\times\mathsf{U},\ z' \in \mathsf{X}_0\times\mathsf{U}_0,\ \theta\in\mathbb{R}^d.$$

(ii) *There is* $\delta_\bullet > 0$ *such that* $\pi_\theta(z) \ge \delta_\bullet \pi_{\mathcal{W}}(z)$ *for all* z, θ.

(iii) $\pi_\theta(x,u) \ge \varepsilon\mu_\theta(x)\nu_{\mathcal{W}}(u)$ *for all* x,u, θ.

(iv) $R^{\mathcal{W}}(\theta) \ge \delta_\bullet R^{\mathcal{W}}$ *for all* θ.

The constants $\delta_T, \delta_\bullet$ *may depend on the policy parameters, but not* θ.

Proof. In view of (30) we have the bound $T_\theta^k(z,z') \ge \varepsilon^k T_{\mathcal{W}}^k(z,z')$ for any k. Hence aperiodicity of T_θ follows from the assumed aperiodicity of $T_{\mathcal{W}}$.

The lower bound in (i) holds for the oblivious policy under (47a): there is $N \ge 1$ and $\delta_{\mathcal{W}} > 0$ such that

$$\sum_{k=1}^{N} T_{\mathcal{W}}^k(z,z') \ge \delta_{\mathcal{W}}\,, \quad \textit{for } z \in \mathsf{X}\times\mathsf{U},\ z' \in \mathsf{X}_0\times\mathsf{U}_0.$$

This implies the desired lower bound in (i) with $\delta_T = \varepsilon^N\delta_{\mathcal{W}}$.

Part (ii) follows from the bounds above and invariance:

$$\pi_\theta(z') = \sum_z \pi_\theta(z)\Big(\frac{1}{N}\sum_{k=1}^{N} T_\theta^k(z,z')\Big) \ge \frac{1}{N}\delta_T$$

Part (iii) also follows from invariance in the following one-step form: we have from (39b),

$$\begin{aligned}\pi_\theta(z') &= \sum_z \pi_\theta(z)T_\theta(z,z')\\ &= \sum_{x,u} \pi_\theta(x,u)P_u(x,x')\widetilde{\phi}^\theta(u' \mid x')\\ &\ge \varepsilon\sum_{x,u} \pi_\theta(x,u)P_u(x,x')\nu_{\mathcal{W}}(u') = \varepsilon\mu_\theta(x')\nu_{\mathcal{W}}(u')\end{aligned}$$

The inequality follows from the bound $\widetilde{\phi}^\theta(u' \mid x') \ge \varepsilon\nu_{\mathcal{W}}(u')$.

For part (iv), we begin with the definitions (46), giving

$$R^{\mathcal{W}}(\theta) = \sum_{x,u} \mu_\theta(x)\nu_{\mathcal{W}}(u)\psi(x,u)\psi(x,u)^{\intercal}$$

Applying (ii) gives $\mu_\theta(x) \ge \delta_\bullet\mu_{\mathcal{W}}(x)$ for all x, and hence the desired bound:

$$R^{\mathcal{W}}(\theta) \ge \delta_\bullet \sum_{x,u} \mu_{\mathcal{W}}(x)\nu_{\mathcal{W}}(u)\psi(x,u)\psi(x,u)^{\intercal} = \delta_\bullet R^{\mathcal{W}}$$

∎

Proof of Prop. A.2. Assumption (SA1) is follows directly from the form of the recursion (6a), which gives $f(\theta,\xi) = \psi(x,u)[c(x,u) + \gamma\min_{u'}\{\psi(x',u')^{\intercal}\theta\}]$, in which x,u,x' is a function of $\xi = (x;\iota;\delta)$.

Part (i) of (SA2) follows directly from Lemma A.3. It remains to establish (ii).

For this we apply (15) and (39a), which imply it is enough to show that a similar bound holds for the policy:

$$\sup_{u,x,\theta}\big|\partial_{\theta_i}\widetilde{\phi}^\theta(u \mid x)\big|\|\theta\| < \infty$$

This follows from the construction, giving $\widetilde{\phi}^{\theta} = \widetilde{\phi}^{r\theta}$ for each $r \geq 1$ and parameter satisfying $\|\theta\| \geq 1$. ■

C. Mean flow for the ε-greedy policy

In this subsection the input is chosen to be the ε-greedy policy (30). The motivation is in part the fact that establishing stability of the ODE@∞ in this case is far easier than the tamed Gibbs approximation.

The transition matrix (39b) becomes

$$T_\theta(z,z') = P_u(x,x')\big\{(1-\varepsilon)\mathbb{1}\{u' = \phi^\theta(x')\} + \varepsilon\nu^{\mathcal{W}}(u')\big\} \quad (55)$$

for $z = (x,u)$, $z' = (x',u') \in \mathsf{X} \times \mathsf{U}$. The family $\{T_\theta : \theta \in \mathbb{R}^d\}$ is finite because there are only a finite number of deterministic stationary policies; it takes on a constant value on each connected component of $\mathcal{C}^{\Theta}$ (recall (29)).

Compact representations of f and $\bar{f}$ are obtained with additional notation. For $n \geq 0$ denote

$$\begin{aligned} \psi^{\Theta}_{(n)} &= \psi(X_n, \phi^{\theta_n}(X_n)) & \psi^{\mathcal{W}}_{(n)} &= \psi(X_n, \mathcal{W}_n) \\ c^{\Theta}_n &= c(X_n, \phi^{\theta_n}(X_n)) & c^{\mathcal{W}}_n &= c(X_n, \mathcal{W}_n) \end{aligned} \quad (56)$$

We have under the ε-greedy policy (30, 32),

$$\begin{aligned} f_{n+1}(\theta_n) = (1-B_n)\big(c^{\Theta}_n + \big[\gamma\psi^{\Theta}_{(n+1)} - \psi^{\Theta}_{(n)}\big]^{\intercal}\theta_n\big)\psi^{\Theta}_{(n)} \\ + B_n\big(c^{\mathcal{W}}_n + \big[\gamma\psi^{\Theta}_{(n+1)} - \psi^{\mathcal{W}}_{n}\big]^{\intercal}\theta_n\big)\psi^{\mathcal{W}}_{(n)} \end{aligned} \quad (57)$$

Lemma A.4. $\mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi^{\Theta}_{(n+1)}\}^{\intercal}\big] = R_{-1}(\theta) + \varepsilon D(\theta)$, *in which*

$$D(\theta) = \mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi^{\Theta}_{(n+1)} - \psi^{\mathcal{W}}_{(n+1)}\}^{\intercal}\big] \quad (58)$$

Proof. Starting with the definition

$$R_{-1}(\theta) = \mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi_{(n+1)}\}^{\intercal}\big]$$

we have under the ε-greedy policy, $R_{-1}(\theta) =$

$$\begin{aligned} &(1-\varepsilon)\mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi^{\Theta}_{(n+1)}\}^{\intercal}\big] + \varepsilon\mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi^{\mathcal{W}}_{(n+1)}\}^{\intercal}\big] \\ &= \mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi^{\Theta}_{(n+1)}\}^{\intercal}\big] + \varepsilon\mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi^{\mathcal{W}}_{(n+1)} - \psi^{\Theta}_{(n+1)}\}^{\intercal}\big] \end{aligned}$$

■

Lemma A.5. *The vector fields for the mean flow and the ODE@∞ for the ε-greedy policy are*

$$\bar{f}(\theta) = A(\theta)\theta - b(\theta) \qquad \bar{f}_\infty(\theta) = A(\theta)\theta \quad (59a)$$

$$\textit{in which} \quad A(\theta) = -\big[R_0(\theta) - \gamma R_{-1}(\theta)\big] + \varepsilon\gamma D(\theta) \quad (59b)$$

$$b(\theta) = (1-\varepsilon)b^{\Theta}(\theta) + \varepsilon b^{\mathcal{W}}(\theta) \quad (59c)$$

$b^{\Theta}(\theta) = -\mathsf{E}_{\pi_\theta}[\psi^{\Theta}_{(n)}c^{\Theta}_n]$ *and* $b^{\mathcal{W}}(\theta) = -\mathsf{E}_{\pi_\theta}[\psi^{\mathcal{W}}_{(n)}c(X_n,\mathcal{W}_n)]$.

Proof. The representation (57) is equivalently expressed $f_{n+1}(\theta_n) = A_{n+1}\theta_n - b_{n+1}$, in which

$$\begin{aligned} A_{n+1} &= \psi_{(n)}\big[\gamma\psi^{\Theta}_{(n+1)} - \psi_{(n)}\big]^{\intercal} \\ b_{n+1} &= (1-B_n)\psi^{\Theta}_{(n)}c(X_n,\phi^{\theta}(X_n)) + B_n\psi^{\mathcal{W}}_{(n)}c(X_n,\mathcal{W}_n) \end{aligned}$$

The expression for $b(\theta)$ in the expression $\bar{f}(\theta) = \mathsf{E}_{\pi_\theta}[f_{n+1}(\theta)] = A(\theta)\theta - b(\theta)$ is immediate.

We have $A(\theta) = -R_0(\theta) + \gamma\mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi^{\Theta}_{(n+1)}\}^{\intercal}\big]$, so that (59b) follows from Lemma A.4.

The expression for $\bar{f}_\infty$ follows from the fact that A and b are invariant under positive scaling of their arguments: $A(r\theta) = A(\theta)$ and $b(r\theta) = b(\theta)$ for any θ and $r > 0$. ■

The mean flow (11) is a differential inclusion because the vector field $\bar{f}$ is not continuous.

The form of the expression for $A(\theta)$ in (59b) is intended to evoke the similar formula (IV.3) obtained for (49).

The following conclusions are based on arguments similar to what is used to obtain stability of on-policy TD-learning [50]. Recall the definition (48): $\varepsilon_\gamma := (1-\gamma)^2/[(1-\gamma)^2+\gamma^2]$.

Proposition A.6. *If $\varepsilon < \varepsilon_\gamma$, then there is $\beta_\varepsilon > 0$ such that $v^{\intercal}A(\theta)v \leq -\beta_\varepsilon\|v\|^2$ for each $v, \theta \in \mathbb{R}^d$.*

Proof. Applying Lemma A.5 gives for any v, θ,

$$v^{\intercal}A(\theta)v \leq -(1-\gamma)v^{\intercal}R_0(\theta)v + \varepsilon\gamma v^{\intercal}D(\theta)v \quad (60)$$

The inequality follows from the bound $v^{\intercal}R_k(\theta)v \leq v^{\intercal}R_0(\theta)v$, valid for any k.

We are left to bound the term involving D. Write $v^{\intercal}D(\theta)v = d^{\Theta}_v(\theta) - d^{\mathcal{W}}_v(\theta)$ with

$$\begin{aligned} d^{\Theta}_v(\theta) &= \mathsf{E}_{\pi_\theta}\big[(v^{\intercal}\psi_{(n)})(v^{\intercal}\psi^{\Theta}_{(n+1)})\big] \\ d^{\mathcal{W}}_v(\theta) &= \mathsf{E}_{\pi_\theta}\big[(v^{\intercal}\psi_{(n)})(v^{\intercal}\psi^{\mathcal{W}}_{(n+1)})\big] \end{aligned}$$

Using the bound $xy \leq \frac{1}{2}[x^2+y^2]$ for $x, y \in \mathbb{R}$, we obtain for any $\delta_{\mathcal{W}}, \delta_{\Theta} > 0$,

$$\begin{aligned} \big|d^{\Theta}_v(\theta)\big| &\leq \tfrac{1}{2}\delta_{\Theta}^{-1}v^{\intercal}R_0(\theta)v + \tfrac{1}{2}\delta_{\Theta}v^{\intercal}R^{\Theta}_0(\theta)v \\ \big|d^{\mathcal{W}}_v(\theta)\big| &\leq \tfrac{1}{2}\delta_{\mathcal{W}}^{-1}v^{\intercal}R_0(\theta)v + \tfrac{1}{2}\delta_{\mathcal{W}}v^{\intercal}R^{\mathcal{W}}_0(\theta)v \end{aligned}$$

Recall from (46c) that $R_0(\theta) = (1-\varepsilon)R^{\Theta}_0(\theta) + \varepsilon R^{\mathcal{W}}_0(\theta)$. Set $\delta_{\mathcal{W}} = \varepsilon\eta$, $\delta_{\Theta} = (1-\varepsilon)\eta$, with $\eta > 0$ to be chosen. Then,

$$\begin{aligned} v^{\intercal}D(\theta)v &\leq \tfrac{1}{2}\Big[\big(\delta_{\mathcal{W}}^{-1} + \delta_{\Theta}^{-1}\big)v^{\intercal}R_0(\theta)v \\ &\qquad + \delta_{\Theta}v^{\intercal}R^{\Theta}_0(\theta)v + \delta_{\mathcal{W}}v^{\intercal}R^{\mathcal{W}}_0(\theta)v\Big] \\ &= \tfrac{1}{2}\Big[\Big(\frac{1}{\varepsilon} + \frac{1}{1-\varepsilon}\Big)\frac{1}{\eta} + \eta\Big]v^{\intercal}R_0(\theta)v \end{aligned}$$

The value $\eta^*_\varepsilon = \sqrt{\varepsilon^{-1} + (1-\varepsilon)^{-1}}$ minimizes the right hand side, and on substitution, $v^{\intercal}D(\theta)v \leq \eta^*_\varepsilon\, v^{\intercal}R_0(\theta)v$.

Substitution into (60) gives the final bound,

$$v^{\intercal}A(\theta)v \leq \big[-(1-\gamma) + \varepsilon\gamma\eta^*_\varepsilon\big]v^{\intercal}R_0(\theta)v$$

The coefficient is negative for positive ε if and only if $\varepsilon < \varepsilon_\gamma$. We obtain the desired bound with

$$\beta_\varepsilon = \big[(1-\gamma) - \varepsilon\gamma\eta^*_\varepsilon\big]\min_\theta \lambda_{\min}(R_0(\theta))$$

Lemma A.3 implies that the minimum is strictly positive. ■

The extension of Prop. A.6 to the tamed Gibbs policy requires approximations summarized in the next subsection.

D. Entropy and Gibbs bounds

Consider a single Gibbs pmf on U with energy $E\colon \mathsf{U} \to \mathbb{R}$ and inverse temperature $\kappa > 0$:

$$p_\kappa(u) = \frac{1}{\mathcal{Z}_\kappa}\exp(-\kappa E(u))\,, \qquad u \in \mathsf{U}\,,$$

The normalizing factor $\mathcal{Z}_\kappa$ is known as the *partition function*. The entropy of p_κ is denoted

$$H_\kappa = -\sum_u p_\kappa(u)\log(p_\kappa(u)) = \sum_u p_\kappa(u)\big[\kappa E(u) + \log(\mathcal{Z}_\kappa)\big]$$

Bounds on entropy imply bounds on the quality of the softmin approximation. Denote $\underline{E} := \min_u E(u)$.

Lemma A.7. *For any $\kappa > 0$,*

$$\underline{E} \le \sum_u p_\kappa(u) E(u) \le \underline{E} + \frac{1}{\kappa} \log(|\mathsf{U}|)$$

Proof. The uniform distribution maximizes entropy, giving

$$\sum_u p_\kappa(u) \big[\kappa E(u) + \log(\mathcal{Z}_\kappa)\big] \le \log(|\mathsf{U}|)$$

The bound $\log(\mathcal{Z}_\kappa) = \log \sum_u \exp(-\kappa E(u)) \ge -\kappa \underline{E}$ completes the proof. ■

An implication of the lemma to the policy (35): for any initial distribution for (X_0, U_0),

$$\begin{aligned} \underline{Q}^\theta(X_{k+1}) &\le \mathsf{E}\big[Q^\theta(X_{k+1}, \mathcal{U}_{k+1}) \mid X_0^{k+1}, U_0^k\big] \\ &\le \underline{Q}^\theta(X_{k+1}) + \frac{1}{\kappa_\theta} \log(|\mathsf{U}|)\,, \quad k \ge 0\,. \end{aligned} \tag{61}$$

E. Proof of Thms. IV.1 and IV.5

The proof of Thm. IV.1 closely follows the proof of Prop. A.6. We begin a companion to Lemma A.4:

Lemma A.8. *We have for the (ε, κ_0)-tamed Gibbs policy,*

$$\mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi^{\ominus}_{(n+1)}\}^\intercal\big] = R_{-1}(\theta) + \varepsilon D(\theta) + (1-\varepsilon) E(\theta) \tag{62a}$$

in which

$$D(\theta) = \mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi^{\ominus}_{(n+1)} - \psi^{\mathcal{W}}_{(n+1)}\}^\intercal\big] \tag{62b}$$

$$E(\theta) = \mathsf{E}_{\pi_\theta}\big[\psi_{(n)}\{\psi^{\ominus}_{(n+1)} - \psi^{\mathcal{U}}_{(n+1)}\}^\intercal\big] \tag{62c}$$

with $\psi^{\mathcal{U}}_{(n+1)} = \psi(X_{n+1}, \mathcal{U}_{n+1})$.

We have a partial extension of Prop. A.6:

Lemma A.9. *The following holds for the (ε, κ_0)-tamed Gibbs policy, subject to* (47) *and $\varepsilon < \varepsilon_\gamma$: there is $\beta_\varepsilon > 0$ such that $\theta^\intercal A(\theta)\theta \le -\beta_\varepsilon \|\theta\|^2$ for all $\kappa_0 > 0$ sufficiently large, and all $\|\theta\| \ge 1$.*

Proof. Applying Lemma A.8 to (41), and following the same steps as in the proof of Prop. A.6 we obtain

$$\begin{aligned} \theta^\intercal A(\theta)\theta &\le -\beta^0_\varepsilon\, \theta^\intercal R_0(\theta)\theta + \gamma(1-\varepsilon)\theta^\intercal E(\theta)\theta \\ \textit{with} \quad \beta^0_\varepsilon &= \big[(1-\gamma) - \varepsilon\gamma\eta^*_\varepsilon\big] > 0 \\ \eta^*_\varepsilon &= \sqrt{\varepsilon^{-1} + (1-\varepsilon)^{-1}} \end{aligned}$$

From the definition (62c) we have

$$\theta^\intercal E(\theta)\theta = \mathsf{E}_{\pi_\theta}\big[Q^\theta(X_n, U_n)\{\underline{Q}^\theta(X_{n+1}) - Q^\theta(X_{n+1}, \mathcal{U}_{n+1})\}\big]$$

Applying (61) and the expression for κ_θ in (36), we obtain for $\|\theta\| \ge 1$,

$$\begin{aligned} \big|\theta^\intercal E(\theta)\theta\big| &\le \frac{1}{\kappa_0}\|\theta\| \log(|\mathsf{U}|) \mathsf{E}_{\pi_\theta}\big[\big|Q^\theta(X_n, U_n)\big|\big] \\ &\le \frac{1}{\kappa_0}\|\theta\|^2 \log(|\mathsf{U}|)\sqrt{\lambda_{\max}} \end{aligned}$$

with $\lambda_{\max}$ the maximum over all θ of the maximum eigenvalue of $R_0(\theta)$. Combining these bounds completes the proof. ■

Proof of Thm. IV.1. Precisely as in the proof of Prop. IV.2 we obtain a solution to (v4) using $V(\theta) = \|\theta\|$ (recall (21)), which implies (23) exactly as in the case when $\mathbf{\Phi}$ is exogenous.

The existence of θ^* follows from Prop. II.2, exactly as in the proof in Prop. IV.2 ■

Proof of Thm. IV.5. Let θ^{κ_0} denote the solution to the projected Bellman equation for the (ε, κ_0)-tamed Gibbs policy, in which $\varepsilon < \varepsilon_\gamma$ is fixed.

Observe that in Lemma A.9 we obtain a uniform bound over all large κ_0. An examination of the proof of Prop. II.2 shows that there is a constant b_ε such that $\|\theta^{\kappa_0}\| \le b_\varepsilon$ for all sufficiently large κ_0.

Hence we can find a subsequence $\kappa_0^n \to \infty$ as $n \to \infty$, for which the following limits exist:

$$\theta^* = \lim_{n\to\infty} \theta^{\kappa_0^n}\,, \quad \pi^* = \lim_{n\to\infty} \pi_n\,,$$

in which π_n is the invariant pmf obtained from the policy using $\theta^{\kappa_0^n}$.

The invariant pmfs have the form

$$\pi_n(x,u) = \mu_n(x)\widetilde{\phi}^n(u \mid x)$$

with $\widetilde{\phi}^n$ defined in (35) using κ_0^n, and μ_n the first marginal of π_n. It follows that the limiting invariant pmf has the same structure, $\pi^*(x,u) = \mu^*(x)\widetilde{\phi}^{\theta^*}(u \mid x)$. Since $\kappa_0^n \uparrow\uparrow \infty$, convergence implies that $\widetilde{\phi}^{\theta^*}$ is of the form (50) with $\widetilde{\phi}^* \in \Phi^{\theta^*}$.

Letting $\bar{f}_n$ denote the vector field obtained using $\theta^{\kappa_0^n}$ we must have convergence for each θ:

$$\bar{f}(\theta) = \lim_{n\to\infty} \bar{f}_n(\theta) = \mathsf{E}_{\pi^*}[\psi_{(n)}\mathcal{B}(X_n, U_n; \theta)]\,,$$

in which U_n is defined using the randomized ε-greedy policy $\widetilde{\phi}^{\theta^*}$, and $\mathcal{B}$ defined in (40b) is a continuous function of θ. Since $\bar{f}_n(\theta_n) = 0$ for each n, we conclude that $\bar{f}(\theta^*) = 0$ as desired. ■

REFERENCES

[1] J. Abounadi, D. Bertsekas, and V. S. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

[2] S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, New York, 2007.

[3] K. E. Avrachenkov, V. S. Borkar, H. P. Dolhare, and K. Patil. Full gradient DQN reinforcement learning: A provably convergent scheme. In *Modern Trends in Controlled Stochastic Processes:*, pages 192–220. Springer, 2021.

[4] L. Baird. Residual algorithms: Reinforcement learning with function approximation. In A. Prieditis and S. Russell, editors, *Proc. Machine Learning*, pages 30–37. Morgan Kaufmann, San Francisco (CA), 1995.

[5] J. Bas Serrano, S. Curi, A. Krause, and G. Neu. Logistic Q-learning. In A. Banerjee and K. Fukumizu, editors, *Proc. of The Intl. Conference on Artificial Intelligence and Statistics*, volume 130, pages 3610–3618, 13–15 Apr 2021.

[6] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions, Part II: applications. *Mathematics of Operations Research*, 31(4):673–695, 2006.

[7] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, Berlin Heidelberg, 2012.

[8] D. P. Bertsekas. *Dynamic programming and optimal control. Vol. II*. Athena Scientific, Belmont, MA, fourth edition, 2012.

[9] S. Bhatnagar. The Borkar–Meyn Theorem for asynchronous stochastic approximations. *Systems & control letters*, 60(7):472–478, 2011.

[10] V. Borkar, S. Chen, A. Devraj, I. Kontoyiannis, and S. Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *arXiv e-prints:2110.14427*, pages 1–50, 2021.

[11] V. S. Borkar. Stochastic approximation with 'controlled Markov' noise. *Systems & control letters*, 55(2):139–145, 2006.

[12] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency, Delhi, India, 2nd edition, 2021.

[13] V. S. Borkar and S. P. Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2):447–469, 2000.

[14] S. Chen, A. M. Devraj, F. Lu, A. Bušić, and S. Meyn. Zap Q-Learning with nonlinear function approximation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Proc. Conference on Neural Information Processing Systems (NeurIPS), and arXiv e-prints 1910.05405*, volume 33, pages 16879–16890, 2020.

[15] Z. Chen, J.-P. Clarke, and S. T. Maguluri. Target network and truncation overcome the deadly triad in Q-learning. *SIAM Journal on Mathematics of Data Science*, 5(4):1078–1101, 2023.

[16] A. Cooper and S. Meyn. Reinforcement learning design for quickest change detection. *Submitted for publication, and arXiv preprint arXiv:2403.14109*, 2024.

[17] D. De Farias and B. Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization Theory and Applications*, 105(3):589–608, 2000.

[18] A. M. Devraj. *Reinforcement Learning Design with Optimal Learning Rate*. PhD thesis, University of Florida, 2019.

[19] A. M. Devraj, A. Bušić, and S. Meyn. Fundamental design principles for reinforcement learning algorithms. In K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, editors, *Handbook on Reinforcement Learning and Control*, Studies in Systems, Decision and Control series (SSDC, volume 325). Springer, 2021.

[20] A. M. Devraj and S. P. Meyn. Zap Q-learning. In *Proc. of the Intl. Conference on Neural Information Processing Systems*, pages 2232–2241, 2017.

[21] A. M. Devraj and S. P. Meyn. Q-learning with uniformly bounded variance. *IEEE Trans. on Automatic Control*, 67(11):5948–5963, 2022.

[22] A. Gopalan and G. Thoppe. Approximate Q-learning and SARSA(0) under the ϵ-greedy policy: a differential inclusion analysis. *arXiv preprint arXiv:2205.13617*, 2022.

[23] H. V. Hasselt. Double Q-learning. In *Proc. Advances in Neural Information Processing Systems*, pages 2613–2621, 2010.

[24] T. Jaakola, M. Jordan, and S. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6:1185–1201, 1994.

[25] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

[26] P. Karmakar and S. Bhatnagar. Stochastic approximation with iterate-dependent Markov noise under verifiable conditions in compact state space with the stability of iterates not ensured. *IEEE Transactions on Automatic Control*, 66(12):5941–5954, 2021.

[27] C. K. Lauand and S. Meyn. Revisiting step-size assumptions in stochastic approximation. *arXiv 2405.17834*, 2024.

[28] D. Lee and N. He. A unified switching system perspective and ODE analysis of Q-learning algorithms. *arXiv*, page arXiv:1912.02270, 2019.

[29] H.-D. Lim, D. W. Kim, and D. Lee. Regularized Q-learning. *preprint arXiv:2202.05404*, 2022.

[30] F. Lu, P. G. Mehta, S. P. Meyn, and G. Neu. Convex analytic theory for convex Q-learning. In *IEEE Conference on Decision and Control*, pages 4065–4071, Dec 2022.

[31] H. R. Maei, C. Szepesvári, S. Bhatnagar, and R. S. Sutton. Toward off-policy learning control with function approximation. In *Proc. ICML*, pages 719–726, USA, 2010. Omnipress.

[32] A. Martinelli, M. Gargiani, M. Draskovic, and J. Lygeros. Data-driven optimal control of affine systems: A linear programming perspective. *IEEE Control Systems Letters*, 6:3092–3097, 2022.

[33] A. Martinelli, M. Gargiani, and J. Lygeros. Data-driven optimal control with a relaxed linear program. *Automatica*, 136:110052, 2022.

[34] P. G. Mehta and S. P. Meyn. Q-learning and Pontryagin's minimum principle. In *Proc. of the Conf. on Dec. and Control*, pages 3598–3605, Dec. 2009.

[35] F. S. Melo, S. P. Meyn, and M. I. Ribeiro. An analysis of reinforcement learning with function approximation. In *Proc. ICML*, pages 664–671, New York, NY, 2008.

[36] M. Metivier and P. Priouret. Theoremes de convergence presque sure pour une classe d'algorithmes stochastiques a pas decroissants. *Prob. Theory Related Fields*, 74:403–428, 1987.

[37] S. Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, Cambridge, 2022.

[38] S. Meyn. Stability of Q-learning through design and optimism. *arXiv 2307.02632*, 2023.

[39] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. Published in the Cambridge Mathematical Library. 1993 edition online.

[40] B. T. Polyak. A new method of stochastic approximation type. *Avtomatika i telemekhanika (in Russian). translated in Automat. Remote Control, 51 (1991)*, pages 98–107, 1990.

[41] A. Ramaswamy and S. Bhatnagar. A generalization of the Borkar-Meyn Theorem for stochastic recursive inclusions. *Mathematics of Operations Research*, 42(3):648–661, 2017.

[42] G. A. Rummery and M. Niranjan. On-line Q-learning using connectionist systems. Technical report 166, Cambridge Univ., Dept. Eng., Cambridge, U.K. CUED/F-INENG/, 1994.

[43] D. Ruppert. Efficient estimators from a slowly convergent Robbins-Monro processes. Technical Report Tech. Rept. No. 781, Cornell University, School of Operations Research and Industrial Engineering, Ithaca, NY, 1988.

[44] P. J. Schweitzer. Perturbation theory and finite Markov chains. *J. Appl. Prob.*, 5:401–403, 1968.

[45] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

[46] R. S. Sutton. Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3(1):9–44, 1988.

[47] C. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

[48] C. Szepesvari, E. Brunskill, S. Bubeck, A. Malek, S. Meyn, A. Tewari, and M. Wang. Theory of Reinforcement Learning Boot Camp. Aug 31 to Sep 4, 2020. https://simons.berkeley.edu/workshops/rl-2020-bc.

[49] J. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

[50] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Control*, 42(5):674–690, 1997.

[51] B. Van Roy. *Learning and Value Function Approximation in Complex Decision Processes*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1998. AAI0599623.

[52] M. J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *CoRR*, abs/1905.06265, 2019.

[53] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, Cambridge, UK, 1989.

[54] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

[55] L. Yang and M. Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756, 2020.

[56] F. Zarin Faizal and V. Borkar. Functional Central Limit Theorem for two timescale stochastic approximation. *arXiv e-prints*, page arXiv:2306.05723, June 2023.

Sean Meyn was raised by the beach in Santa Monica, California. Following his BA in mathematics at UCLA, he moved on to pursue a PhD with Peter Caines at McGill University. After about 20 years as a professor of ECE at the University of Illinois, in 2012 he moved to beautiful Gainesville. He is now Professor and Robert C. Pittman Eminent Scholar Chair in the Department of Electrical and Computer Engineering at the University of Florida, director of the Laboratory for Cognition and Control, and Inria International Chair at Inria, France. He is an IEEE CSS distinguished lecturer. His interests span many aspects of stochastic control, stochastic processes, information theory, and optimization. For the past decade, his applied research has focused on engineering, markets, and policy in energy systems.