Optimistic policy iteration and natural actor-critic: A unifying view and a non-optimality result (extended)

Paul Wagner
Department of Information and Computer Science
Aalto University
FI-00076 Aalto, Finland
paul.wagner@aalto.fi

Abstract

Approximate dynamic programming approaches to the reinforcement learning problem are often categorized into greedy value function methods and value-based policy gradient methods. As our first main result, we show that an important subset of the latter methodology is, in fact, a limiting special case of a general formulation of the former methodology; optimistic policy iteration encompasses not only most of the greedy value function methods but also natural actor-critic methods, and permits one to directly interpolate between them. The resulting continuum adjusts the strength of the Markov assumption in policy improvement and, as such, can be seen as dual in spirit to the continuum in TD(λ)-style algorithms in policy evaluation. As our second main result, we show for a substantial subset of soft-greedy value function approaches that, while having the potential to avoid policy oscillation and policy chattering, this subset can never converge toward an optimal policy, except in a certain pathological case. Consequently, in the context of approximations (either in state estimation or in value function representation), the majority of greedy value function methods seem to be deemed to suffer either from the risk of oscillation/chattering or from the presence of systematic sub-optimality.

1 Introduction

We consider the reinforcement learning problem in which one attempts to find an approximately optimal policy for controlling a stochastic nonlinear dynamical system. We focus on the setting in which the target system is actively sampled during the learning process. Here the sampling policy changes during the learning process in a manner that depends on the main policy being optimized. This learning setting is often called interactive learning [e.g., 25, §3]. Many approaches to the problem are value-based and build on the methodology of simulation-based approximate dynamic programming [25, 4, 9, 21, 8, 23]. The majority of these methods are often categorized into greedy value function methods (critic-only) and value-based policy gradient methods (actor-critic) [e.g., 25, 15].

Within this interactive setting, the policy gradient approach has better convergence guarantees, with the strongest case being for Monte Carlo evaluation with 'compatible' value function approximation. In this case, convergence with probability one (w.p.1) to a local optimum can be established for arbitrary differentiable policy classes under mild assumptions [24, 15, 21]. On the other hand, while the greedy value function approach is often considered to possess practical advantages in terms of convergence speed and representational flexibility, its behavior in the proximity of an optimum is currently not well understood. It is well known that interactively operated approximate hard-greedy value function methods can fail to converge to any single policy and instead become trapped in sustained policy oscillation or policy chattering, which is currently a poorly understood phenomenon [6, 7]. This applies to both non-optimistic and optimistic policy iteration (value iteration being a special case of the latter). In general, the best guarantees for this methodology exist in the form of sub-optimality bounds [6, 7]. The practical value of these bounds, however, is under question (e.g., [2; 7, §6.2.2]), as they can permit very bad solutions. Furthermore, it has been shown that these bounds are tight [7, §6.2.3; 14, §3.2].

A hard-greedy policy is a discontinuous function of its parameters, which has been identified as a key source of problems [20, 11, 19, 24]. In addition to the observation that the class of stochastic policies may often permit much simpler solutions [cf. 22], it is known that continuously stochastic policies can also re-gain convergence: both non-optimistic and optimistic soft-greedy approximate policy iteration using, for example, the Gibbs/Boltzmann policy class, is known to converge with enough softness, 'enough' being problem-specific. This has been shown by Perkins & Precup [20] and Melo et al. [16], respectively, although with no consideration of the quality of the obtained solutions nor with an interpretation of how 'enough' relates to the problem at hand. Unfortunately, the aforementioned sub-optimality bounds are also lost in this case (consider temperature $\tau \to \infty$); while convergence is re-gained, the properties of the obtained solutions are rather unknown.

To summarize, there are considerable shortcomings in the current understanding of the learning dynamics at the very heart of the approximate dynamic programming methodology. We share the belief of Bertsekas [5, 6], expressed in the context of the policy oscillation phenomenon, that a better understanding of these issues "has the potential to alter in fundamental ways our thinking about approximate DP."

In this paper, we provide insight into the convergence behavior and optimality of the generalized optimistic form of the greedy value function methodology by reflecting it against the policy gradient approach. While these two approaches are considered in the literature mostly separately, we are motivated by the belief that it is eventually possible to fully unify them, so as to have the benefits and insights from both in a single framework with no artificial (or historical) boundaries, and that such a unification can eventually resolve the issues outlined above. These issues revolve mainly around the greedy methodology, while at the same time, solid convergence results exist for the policy gradient methodology; connecting these methodologies more firmly might well lead to a fuller understanding of both.

After providing background in Section 2, we take the following steps in this direction. First, we show that natural actor-critic methods from the policy gradient side are, in fact, a limiting special case of optimistic policy iteration (Sec. 3). Second, we show that while having the potential to avoid policy oscillation and chattering, a substantial subset of soft-greedy value function approaches can never converge to an optimal policy, except in a certain pathological case (Sec. 4). We then conclude with a discussion in a broader context and use the results to complete a high-level convergence and optimality property map of the variants of the considered methodology (Sec. 5).

2 Background

A Markov decision process (MDP) is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces. $S_t \in \mathcal{S}$ and $A_t \in \mathcal{A}$ denote random variables at time t. $s, s' \in \mathcal{S}$ and $a, b \in \mathcal{A}$ denote state and action instances. $\mathcal{P}(s, a, s') = \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a)$ defines the transition dynamics and $r(s, a) \in \mathbb{R}$ defines the expected immediate reward function. Non-Markovian aggregate states, i.e., subsets of $\mathcal{S}$, are denoted by y. A policy $\pi(a|s, \theta_k) \in \Pi$ is a stochastic mapping from states to actions, parameterized by $\theta_k \in \Theta$. Improvement is performed with respect to the performance metric $J(\theta) = 1/H \sum_t^H \mathbb{E}[r(S_t, A_t)|\pi(\theta)]$. $\nabla_\theta J(\theta_k) \in \Theta$ denotes a parameter gradient at θ_k. $\nabla_\pi J(\theta_k) \in \Pi$ denotes the corresponding policy gradient in the selected policy space. We define the policy distance $\|\pi_u - \pi_v\|$ as some p-norm of the action probability differences $(\sum_s \sum_a |\pi_u(a|s) - \pi_v(a|s)|^p)^{1/p}$. Action value functions $\bar{Q}(s, a, \hat{w}_k)$ and $Q(s, a, \hat{w}_k)$, parameterized by $\hat{w}_k$, are estimators of the γ-discounted cumulative reward $\sum_t \gamma^t \mathbb{E}[r(S_t, A_t)|S_0 = s, A_0 = a, \pi(\theta_k)]$ for some (s, a) when following some policy $\pi(\theta_k)$. The state value function $V(s, \hat{w}_k)$ is an estimator of such cumulative reward that follows some s. We use ϵ to denote a small positive infinitesimal quantity.

We focus on the Gibbs (Boltzmann) policy class with a linear combination of basis functions ϕ:

$$\pi(a|s,\theta_k) = \frac{e^{\theta_k^\top \phi(s,a)}}{\sum_b e^{\theta_k^\top \phi(s,b)}} \ . \tag{1}$$

We shall use the term 'semi-uniformly stochastic policy' for referring to a policy for which $\pi(a|s) = c_s \vee \pi(a|s) = 0, \ \forall s, a, \ \forall s \ \exists c_s \in [0,1]$. Note that both the uniformly stochastic policy and all deterministic policies are special cases of semi-uniformly stochastic policies.

For the value function, we focus on least-squares linear-in-parameters approximation with the same basis ϕ as in (1). We consider both advantage values [see 24, 21]

$$\bar{Q}_k(s,a,\hat{w}_k) = \hat{w}_k^\top \left(\phi(s,a) - \sum_b \pi(b|s,\theta_k)\phi(s,b) \right) \tag{2}$$

and absolute action values

$$Q_k(s,a,\hat{w}_k) = \hat{w}_k^\top \phi(s,a) \ . \tag{3}$$

Evaluation can be based on either Monte Carlo or temporal difference estimation. We focus on optimistic policy iteration, which contains both non-optimistic policy iteration and value iteration as special cases, and on the policy gradient counterparts of these.

In the general form of optimistic approximate policy iteration (e.g., [7, §6.4]; see also [6, §3.3]), a value function parameter vector w is gradually interpolated toward the most recent evaluation $\hat{w}$:

$$w_{k+1} = w_k + \kappa_k(\hat{w}_k - w_k) \ , \quad \kappa_k \in (0,1] \ . \tag{4}$$

Non-optimistic policy iteration is obtained with $\kappa_k = 1, \ \forall k$ and 'complete' evaluations $\hat{w}_k$ (see below). The corresponding Gibbs soft-greedy policy is obtained by combining (1) and a temperature (softness) parameter τ with

$$\theta_{k+1} = w_{k+1}/\tau_k \ , \quad \tau_k \in (0,\infty) \ . \tag{5}$$

Hard-greedy iteration is obtained in the limit as $\tau \to 0$.

In optimistic policy iteration, policy improvement is based on an incomplete evaluation. We distinguish between two dimensions of completeness, which are evaluation *depth* and evaluation *accuracy*. By evaluation depth, we refer to the look-ahead depth after which truncation with the previous value function estimate occurs. For example, LSPE(0) and LSTD(0) [e.g., 17] implement shallow and deep evaluation, respectively. With shallow evaluation, the current value function parameter vector w_k is required for look-ahead truncation when computing $\hat{w}_{k+1}$. Inaccurate (noisy) evaluation necessitates additional caution in the policy improvement process and is the usual motivation for using (4) with $\kappa < 1$.

It is well known that greedy policy iteration can be non-convergent under approximations [4]. The widely used projected equation approach can manifest convergence behavior that is complex and not well understood, including bounded but potentially severe sustained policy oscillations [6, 7]. Similar consequences arise in the context of partial observability for approximate or incomplete state estimation [e.g., 22, 18]. A detailed description of the oscillation phenomenon can be found in [7, §6.4; 6, §3.5], where it is described in terms of cyclic sequences on the so-called greedy partition of the value function parameter space. It is also noted that the properties of the policy evaluation method (including properties that affect evaluation depth) do not seem to affect the ultimate cycles of policies. It is possible to obtain asymptotic parameter convergence by employing an optimistic policy iteration scheme, but even then the corresponding policy can continue oscillating. This is known as policy chattering. In such cases, the value function parameters spiral toward an attractor on a boundary of the greedy partition, and the obtained value function can fail to meaningfully represent the expected returns of any of the involved policies [6, §3.5]. A novel explanation to the phenomenon in the non-optimistic case was recently proposed in [26, 27], where policy oscillation was re-cast as sustained overshooting over an attractive stochastic policy.

Policy convergence can be established under various restrictions. A common approach is to use θ-independent sampling (non-interactivity), which enables convergence for a well-studied family of approximators (for reviews of this methodology, see [28, §2] and [9]). Interestingly, with the

θ-independent aggregation approach introduced in [6, 5], convergence is obtained without completely sacrificing interactivity. However, the approach is based on a θ-independent projection mapping, which effectively causes representational resources to be allocated in a fixed manner instead of according to the on-policy distribution; one of the natural benefits of interactivity becomes lost, although whether this has practical implications remains to be seen. In the case of Monte Carlo estimation of action values, it is also possible to establish convergence by solely modifying the exploration scheme [e.g., 10]. Importantly to this paper, convergence can be established also with continuously soft-greedy action selection [20, 16], in which case, however, the quality of the obtained solutions is unknown. Finally, the analysis in [12] bears some resemblance to our first main result. The key to seeing the relation is in noting that their main update rule (their Eq. (4.1)) is equivalent to our Eq. (4), except for being defined in the *policy* space. While they establish convergence up to a "breaking point" under relatively restrictive assumptions, we proceed by directly connecting to stronger and more generic results from the policy gradient literature.

In policy gradient reinforcement learning [24, 15, 21, 8], improvement is obtained via stochastic gradient ascent:

$$\theta_{k+1} = \theta_k + \alpha_k G(\theta_k)^{-1} \frac{\partial J(\theta_k)}{\partial \theta} = \theta_k + \alpha_k \eta_k \ , \tag{6}$$

where $\alpha_k \in (0, \infty)$, G is a Riemannian metric tensor that ideally encodes the curvature of the policy parameterization, and η_k is some estimate of the gradient. With value-based policy gradient methods, using (1) together with either (2) or (3) fulfills the 'compatibility condition' [24, 15]. With (2), the value function parameter vector $\hat{w}_k$ becomes the natural gradient estimate for the evaluated policy $\pi(\theta_k)$, leading to natural actor-critic algorithms [13, 21], for which

$$\eta_k = \hat{w}_k \ . \tag{7}$$

For policy gradient learning with a 'compatible' value function and Monte Carlo evaluation, convergence w.p.1 to a local optimum is established under standard assumptions [24, 15]. Temporal difference evaluation can lead to sub-optimal results with a known sub-optimality bound [15, 8].

3 Forgetful natural actor-critic

In this section, we show that an important subset of natural actor-critic algorithms is a limiting special case of optimistic policy iteration. A related connection was recently shown in [26, 27], where a modified form of the natural actor-critic algorithm by Peters & Schaal [21] was shown to correspond to non-optimistic policy iteration. In the following, we generalize and simplify this result: by starting from the more general setting of optimistic policy iteration, we arrive at a unifying view that both encompasses a broader range of greedy methods and permits interpolation between the approaches directly with existing (unmodified) methodology.

We consider the Gibbs policy class from (1) and the linear-in-parameters advantage function from (2), which form a 'compatible' actor-critic setup. We assume deep policy evaluation (cf. Section 2). We begin with the natural actor-critic (NAC) algorithm by Peters & Schaal [21] (cf. (6) and (7)) and generalize it by adding a forgetting term:

$$\theta_{k+1} = \theta_k + \alpha_k \eta_k - \kappa_k \theta_k \ , \tag{8}$$

where $\alpha_k \in (0, \infty)$, $\kappa_k \in (0, 1]$. We refer to this generalized algorithm as the forgetful natural actor-critic algorithm, or NAC(κ). In the following, we show that this algorithm is, within the discussed context, equivalent to the general form of optimistic policy iteration in (4) and (5), with the following translation of the parameterization:

$$\tau_k = \frac{\kappa_k}{\alpha_k} \ , \quad \text{or} \quad \alpha_k = \frac{\kappa_k}{\tau_k} \ . \tag{9}$$

Taking the forgetting factor κ in (8) toward zero leads back toward the original natural actor-critic algorithm, with the implication that the original algorithm is a limiting special case of optimistic policy iteration.

Theorem 1. *For the case of deep policy evaluation (Section 2), the natural actor-critic algorithm for the Gibbs policy class (*(6), (7), (1), (2)*) is a limiting special case of Gibbs soft-greedy optimistic policy iteration (*(4), (5), (1), (2)*).*

Proof. The update rule for Gibbs soft-greedy optimistic policy iteration is given in (4) and (5). By moving the temperature to scale $\hat{w}$ (assume w_0 to be scaled accordingly), we obtain

$$\begin{cases} w'_{k+1} & = w'_k + \kappa_k(\hat{w}_k/\tau_k - w'_k) \\ \theta_{k+1} & = w'_{k+1} \ , \end{cases} \tag{10}$$

again with $\kappa_k \in (0,1], \ \tau_k \in (0,\infty)$. Such a re-formulation effectively re-scales w and is possible only with deep policy evaluation (cf. Section 2), with which the non-scaled w is not needed by the policy evaluation process. We can now remove the redundant second line and rename w' to θ:

$$\theta_{k+1} = \theta_k + \kappa_k(\hat{w}_k/\tau_k - \theta_k) \ . \tag{11}$$

Finally, we open up the last term and encapsulate κ/τ into α:

$$\theta_{k+1} = \theta_k + \kappa_k(\hat{w}_k/\tau_k) - \kappa_k\theta_k \tag{12}$$

$$= \theta_k + \alpha_k\hat{w}_k - \kappa_k\theta_k \ , \tag{13}$$

with $\alpha_k = \kappa_k/\tau_k$. Based on (7), we observe that (13) is equivalent to (8). The original natural actor-critic algorithm is obtained in the limit as $\kappa_k \to 0$, which causes the forgetting term $\kappa_k\theta_k$ to vanish (the effective step size α can still be controlled with τ).

□

This result has some interesting implications. First, it becomes apparent that the implicit effective step size in optimistic policy iteration is, in fact, $\alpha = \kappa/\tau$, i.e., it is inversely related to the temperature τ. If the interpolation factor κ is held fixed, a low temperature, which can lead to policy oscillation, equals a long effective step size. This agrees with the interpretation of policy oscillation as overshooting in [26, 27]. Likewise, a high temperature equals a short effective step size. In [20], convergence is established for a high enough *constant* temperature. This result now becomes translated to showing that convergence is established with a short enough *constant* effective step size,[1] which creates an interesting and more direct connection to convergence results for (batch) steepest descent methods with a constant step size [e.g., 1, 3]. In addition, this connection *might* permit the application of the results in the aforementioned literature to establish, in the considered context, a constant step size convergence result for the natural actor-critic methodology.

Second, we see that the interpolation scheme in optimistic policy iteration, while originally introduced for the sake of countering an inaccurate value function estimate, actually goes in the direction of the policy gradient methodology. Smooth interpolation between policy gradient and greedy value function learning turns out to be possible by simply adjusting the interpolation factor κ while treating the temperature τ as an inverse of the step size (we return to provide an interpretation of the role of κ at a later point). Contrary to the related result in [26], no modifications to existing algorithms are needed. This connection also allows the convergence results from the policy gradient literature to be brought in (see Section 2): convergence w.p.1, under standard assumptions from the referred literature, to an optimal solution is established in the limit for this class of approximate optimistic policy iteration as the interpolation factor κ is taken toward zero and the step size requirements are inversely enforced on the temperature τ.

Third, we observe that in non-optimistic policy iteration ($\kappa = 1$), the forgetting term resets the parameter vector to the origin at the beginning of every iteration, with the implication that solutions that are not within the range of *a single step* from the origin in the direction of the natural gradient cannot be reached in any number of iterations. The choice of the effective step size, which is inversely controlled by the temperature, becomes again decisive: a step size that is too short (the temperature is too high) will cause the algorithm to permanently undershoot the desired optimum, thus trapping it in sustained sub-optimality, while a step size that is too long (the temperature is too low) will cause it to overshoot, which can additionally trap it in sustained oscillation. Unfortunately, even hitting the target exactly with a perfect step size will fail to lead to convergence and optimality at the same time. Our next section examines these issues more closely.

4 Systematic non-optimality of soft-greedy methods

For greedy value function methods, using the hard-greedy policy class trivially prevents convergence to other than deterministic policies. Furthermore, the proximity of an attractive stochastic

[1] Note that the diminishing step size α_t in [20, Fig. 1] concerns policy *evaluation*, not policy *improvement*.

policy can prevent convergence altogether and trap the process in oscillation (cf. Section 2). The Gibbs soft-greedy policy class, on the other hand, *can* represent stochastic policies, fixed points do exist [11, 19], and convergence toward *some* policy is guaranteed with sufficient softness [20, 16]. While convergence toward deterministic optimal decisions is trivially lost as soon as any softness is introduced ($\tau \not\to 0$, and assuming a bounded value function), one might hope that convergence toward stochastic optimal decisions could still occur in some cases. Unfortunately, as we show in the following, this is not the case: in the presence of any softness, this approach can never converge toward any optimal policy (i.e., convergence and optimality become mutually exclusive), except in a certain pathological case.

At this point, we wish to make clear that we are not arguing against the *practical* value of the greedy value function methodology in (interactively) approximated problems; the methodology has some clear merits, and the sub-optimality and oscillations could well be negligible in a given task. Instead, we take the following result, together with existing literature on policy oscillations, as an indication of a fundamental theoretical incompatibility of this methodology to this context: the way by which this methodology deals with stochastic optima seems to be fundamentally flawed, and we believe that a thorough understanding of this flaw will have, in addition to facilitating sound theoretical advances, also immediate practical value by permitting correctly informed trade-off decisions.

Lemma 1. *Consider the Gibbs/Boltzmann distribution* $\mathbb{P}(i|\omega) = e^{\omega^\top \mathbb{I}(i)} / \sum_z e^{\omega^\top \mathbb{I}(z)}$*, where* $\omega \in \mathbb{R}^n$*, and* $\mathbb{I}(i) \in \mathbb{R}^n$ *is the indicator function that picks the* i*th element of its multiplier. This distribution is invariant to, and only to, uniform translations of its argument vector* ω*:*

$$\frac{e^{(\omega+c)^\top \mathbb{I}(i)}}{\sum_z e^{(\omega+c)^\top \mathbb{I}(z)}} = \frac{e^{\omega^\top \mathbb{I}(i)}}{\sum_z e^{\omega^\top \mathbb{I}(z)}} \;, \quad \forall i \;, \tag{14}$$

where $c \in \mathbb{R}^n$*, if and only if*

$$c^\top \mathbb{I}(i) = c^\top \mathbb{I}(z) \;, \quad \forall i, z \;. \tag{15}$$

Furthermore, the Gibbs policy class in (1) *is invariant to, and only to, per-state uniform translations of the term* $\theta^\top \phi(s,a)$*, and if some translation of the parameter vector* θ *operates within this invariance for some* θ_x*, then it does so for all* θ*.*

Proof. Consider some translation c of the argument vector ω of the Gibbs/Boltzmann distribution $\mathbb{P}(i|\omega) = e^{\omega^\top \mathbb{I}(i)} / \sum_z e^{\omega^\top \mathbb{I}(z)}$, with $\omega, c \in \mathbb{R}^n$, and $\mathbb{I}(i) \in \mathbb{R}^n$ as the indicator function that picks the ith element of its multiplier:

$$\frac{e^{(\omega+c)^\top \mathbb{I}(i)}}{\sum_z e^{(\omega+c)^\top \mathbb{I}(z)}} = \frac{e^{\omega^\top \mathbb{I}(i) + c^\top \mathbb{I}(i)}}{\sum_z e^{\omega^\top \mathbb{I}(z) + c^\top \mathbb{I}(z)}} = \frac{e^{\omega^\top \mathbb{I}(i)} e^{c^\top \mathbb{I}(i)}}{\sum_z e^{\omega^\top \mathbb{I}(z)} e^{c^\top \mathbb{I}(z)}} = \frac{e^{\omega^\top \mathbb{I}(i)}}{\sum_z e^{\omega^\top \mathbb{I}(z)}} \;, \quad \forall i \;, \tag{16}$$

if and only if

$$c^\top \mathbb{I}(i) = c^\top \mathbb{I}(z) \;, \quad \forall i, z \;. \tag{17}$$

Importantly to us, whether translating by some c affects the distribution or not, does not depend on the value of ω, as ω does not appear in (17): if a given c is neutral for some ω, then it is neutral for all ω.

We note that the condition (17) only requires $c^\top \mathbb{I}(i)$ to evaluate to the same value for all i, but this condition does not take a stance on what happens inside this term. Thus, we can replace, without modifying the form of the requirement, the indicator function $\mathbb{I}(i)$ with a function $f(\cdot) \in \mathbb{R}^n$ that picks an arbitrary linear combination of elements from its multiplier:

$$c^\top f(i) = c^\top f(z) \;, \quad \forall i, z \;. \tag{18}$$

Finally, we can extend f to accept an additional argument k, with the consequence that (18) must hold simultaneously for all k to permit the last step in (16):

$$\frac{e^{(\omega+c)^\top f(k,i)}}{\sum_z e^{(\omega+c)^\top f(k,z)}} = \frac{e^{\omega^\top f(k,i)}}{\sum_z e^{\omega^\top f(k,z)}} \;, \quad \forall k, i \tag{19}$$

$$\Leftrightarrow \; c^\top f(k,i) = c^\top f(k,z) \;, \quad \forall k, i, z \;. \tag{20}$$

(19) and (20) now apply to (1), stating that (1) is invariant to, and only to, per-state uniform translations of the term $\theta^\top \phi(s,a)$, and that if some translation c of the parameter vector θ operates within this invariance for some θ_x, then it does so for all θ:

$$\begin{aligned} & \pi(a|s,\theta_x + c) = \pi(a|s,\theta_x) \quad \forall s,a \\ \Leftrightarrow\ & \pi(a|s,\theta + c) = \pi(a|s,\theta) \quad \forall s,a,\theta\ . \end{aligned} \tag{21}$$

□

Theorem 2. *Assume an unbiased value function estimator (e.g., Monte Carlo evaluation). Now, for Gibbs soft-greedy policy iteration (*(1), (4) *and* (5)*) using a linear-in-parameters value function approximator (*(2) *or* (3)*), including optimistic and non-optimistic variants (any κ in* (4)*), there cannot exist a fixed point at an optimum, except for the uniformly stochastic policy.*

Proof outline. A fixed point of the update rule (4) must satisfy

$$\hat{w}_k = w_k\ , \tag{22}$$

i.e., at a fixed point, the policy evaluation step $\hat{w}_k := \text{eval}(\pi(w_k/\tau_k))$ for the current parameter vector must yield the same parameter vector as its result:

$$\text{eval}\left(\pi\left(w_k/\tau_k\right)\right) = w_k\ . \tag{23}$$

By applying (22) and (7), we have

$$w_k = \hat{w}_k = \eta_k = G(\theta_k)^{-1} \nabla_\theta J(\theta_k)\ , \tag{24}$$

which shows that the fixed-point policy $\pi(w_k/\tau_k)$ in (23) is defined solely by its own (scaled) performance gradient.

For an optimal policy and an unbiased estimator, this parameter gradient must, by definition, map to the zero policy gradient, i.e., to $\nabla_\pi J(\theta_k) = 0$. Consequently, an optimal policy at a fixed point is defined solely by the zero policy gradient, making the policy equal to $\pi(0)$, which is the uniformly stochastic policy.

Proof. By definition, when the policy defined by some $\theta_x \in \Theta$ is locally optimal, the parameter gradient vector $\eta_{x,0} = G(\theta_x)^{-1}\nabla_\theta J(\theta_x)$ estimated by an unbiased estimator must be such that it maps to the zero policy gradient $\nabla_\pi J(\theta_x) = 0$ at θ_x, i.e., it does not change π at θ_x:

$$\pi(\theta_x + \alpha\eta_{x,0}) = \pi(\theta_x)\ , \quad \forall\alpha\ , \tag{25}$$

with $\pi(\cdot)$ from (1). For this to be possible, $\eta_{x,0}$ must fulfill (20) (in the role of c), which allows us to apply (21) and extend (25) from θ_x to all θ:

$$\begin{aligned} & \pi(\theta_x + \alpha\eta_{x,0}) = \pi(\theta_x)\ , \quad \forall\alpha \\ \Leftrightarrow\ & \pi(\theta + \alpha\eta_{x,0}) = \pi(\theta)\ , \quad \forall\alpha,\theta\ . \end{aligned} \tag{26}$$

Most importantly, (26) implies the following:

$$\begin{aligned} & \pi(\theta_x + \alpha\eta_{x,0}) = \pi(\theta_x)\ , \quad \forall\alpha \\ \Leftrightarrow\ & \pi(\alpha\eta_{x,0}) = \pi(0 + \alpha\eta_{x,0}) = \pi(0)\ , \quad \forall\alpha\ , \end{aligned} \tag{27}$$

where $\pi(0)$ is the uniformly stochastic policy. In words, whenever the evaluated policy is locally optimal, an unbiased gradient estimator must produce such a parameter gradient vector $\eta_{x,0}$ that when added to any policy parameter vector θ, *including the zero parameter vector*, does not change the resulting policy.

A fixed point of the update rule (4) must satisfy

$$\hat{w}_k = w_k\ , \tag{28}$$

regardless of the value of κ. By applying (5), we see that the limiting policy at a fixed point is

$$\pi(\hat{w}_k/\tau_k)\ . \tag{29}$$

For the case of (2), we can apply (7) and re-write this policy as

$$\pi(\alpha_k \eta_k) \;, \tag{30}$$

with $\alpha_k = 1/\tau_k$. Also, the combination of (1) and (2) forms a 'compatible' actor-critic setup, implying that with an unbiased value function estimator, η_k is an unbiased gradient estimate (Section 2). Consequently, for a locally optimal policy, η_k must fulfill (27), with the implication that (30) is the uniformly stochastic policy. Thus, under the given assumptions, whenever the policy is simultaneously locally optimal and a fixed point of (4), it must be the uniformly stochastic policy.

Extension from (2) to (3) follows by noting that the definitions of $\bar{Q}$ and Q differ only by a per-state uniform translation: $Q(s,a) = \bar{Q}(s,a) + V(s)$. As observed in [24], such a translation does not break the compatibility condition, i.e., it does not affect the biasedness of the corresponding gradient. This can be seen also by noting that the translation in $Q(s,a) = \hat{w}^\top \phi(s,a)$ is carried, via (29), to a per-state uniform translation of the term $\theta^\top \phi(s,a)$ in (1), to which the policy class was shown in (19) and (20) to be invariant. Consequently, $\hat{w}_k/\tau_k$ in (29) still defines an unbiased gradient estimate and, for a locally optimal policy, (27) stays in effect.

□

This result still leaves open the possibility of a fixed point residing vanishingly close to an optimum, in which case asymptotic convergence toward such a fixed point would mean asymptotic convergence toward the adjacent optimum. This possibility is ruled out with the following theorem that shows, assuming a smooth gradient field and $\tau \not\to 0$, that the distance between a fixed point and an optimum cannot be vanishingly small, with the exception of an optimum that is semi-uniformly stochastic (assuming unbounded returns) or uniformly stochastic (assuming bounded returns).

Lemma 2. *For the Gibbs/Boltzmann distribution $\mathbb{P}(i|\omega) = e^{(\omega/\tau)^\top \mathbb{I}(i)} / \sum_z e^{(\omega/\tau)^\top \mathbb{I}(z)}$, any finite decrease of the temperature τ has a vanishingly small effect on the distribution if and only if the initial distribution is vanishingly close to being semi-uniformly stochastic (we simplify our notation by noting that decreasing the temperature is equivalent to upscaling the parameter vector):*

$$\left| \frac{e^{\omega^\top \mathbb{I}(i)}}{\sum_z e^{\omega^\top \mathbb{I}(z)}} - \frac{e^{(\alpha\omega)^\top \mathbb{I}(i)}}{\sum_z e^{(\alpha\omega)^\top \mathbb{I}(z)}} \right| < \epsilon \;, \qquad \forall i,\ \forall \alpha > 1,\ \alpha \not\to \infty \tag{31}$$

$$\Leftrightarrow \left| \frac{e^{\omega^\top \mathbb{I}(i)}}{\sum_z e^{\omega^\top \mathbb{I}(z)}} - c \right| < \epsilon \quad \vee \quad \frac{e^{\omega^\top \mathbb{I}(i)}}{\sum_z e^{\omega^\top \mathbb{I}(z)}} < \epsilon \;, \qquad \forall i,\ \exists c \in [0,1] \;. \tag{32}$$

Proof. Informally, this holds because decreasing the temperature makes the distribution more deterministic by amplifying the probability differences present. For this to have no non-vanishing effect on the distribution for any finite decrease of the temperature, it must be that there are no non-vanishing probability differences in non-vanishingly weak dimensions to be amplified, i.e., for all dimensions for which the distribution is not close to zero it must be close to some common constant.

Consider a parameter vector $\bar{\omega}$ that defines some initial Gibbs distribution. Let us choose one of the maximizing dimensions of this distribution and denote it by $i^\star$:

$$\frac{e^{\bar{\omega}^\top \mathbb{I}(i^\star)}}{\sum_z e^{\bar{\omega}^\top \mathbb{I}(z)}} \geq \frac{e^{\bar{\omega}^\top \mathbb{I}(j)}}{\sum_z e^{\bar{\omega}^\top \mathbb{I}(z)}} \;, \quad \forall j \;. \tag{33}$$

There might also exist some dimensions by which the distribution is non-vanishingly smaller than for the maximizing dimension $i^\star$, in case that the distribution is not close to being uniformly stochastic. Let us denote the (possibly empty) set of these dimensions by K:

$$K = \left\{ k \middle| \frac{e^{\bar{\omega}^\top \mathbb{I}(i^\star)}}{\sum_z e^{\bar{\omega}^\top \mathbb{I}(z)}} \gg \frac{e^{\bar{\omega}^\top \mathbb{I}(k)}}{\sum_z e^{\bar{\omega}^\top \mathbb{I}(z)}} \right\} \;, \tag{34}$$

where $\gg$ is a shorthand for a non-vanishing inequality. We then de-normalize (33) and (34) into

$$\begin{cases} e^{\bar{\omega}^\top \mathbb{I}(i^\star)} \geq e^{\bar{\omega}^\top \mathbb{I}(j)} \;, \qquad \forall j \;, & \text{(35a)} \\ e^{\bar{\omega}^\top \mathbb{I}(i^\star)} \gg e^{\bar{\omega}^\top \mathbb{I}(k)} \;, \qquad \forall k \in K \;. & \text{(35b)} \end{cases}$$

We can see that for the near-equality in (31) to hold for the considered $\bar{\omega}$ and for the $i^\star$ chosen, i.e.,

$$\left| \frac{e^{\bar{\omega}^\top \mathbb{I}(i^\star)}}{\sum_z e^{\bar{\omega}^\top \mathbb{I}(z)}} - \frac{e^{(\alpha\bar{\omega})^\top \mathbb{I}(i^\star)}}{\sum_z e^{(\alpha\bar{\omega})^\top \mathbb{I}(z)}} \right| < \epsilon \;, \quad \forall \alpha > 1, \; \alpha \not\to \infty \;, \tag{36}$$

it must be that α makes both the numerator and the denominator of the affected fraction to grow (or shrink for a negative $\bar{\omega}^\top \mathbb{I}(i^\star)$) nearly exactly as fast, so that the effect of α nearly cancels out. Let us use σ_i to denote the scaling factor by which α scales the terms $e^{(\alpha\bar{\omega})^\top \mathbb{I}(i)}$:

$$\sigma_i = \frac{e^{(\alpha\bar{\omega})^\top \mathbb{I}(i)}}{e^{\bar{\omega}^\top \mathbb{I}(i)}} \;. \tag{37}$$

We note that α affects the value of these terms via linear operations and an exponentiation with a positive base, making the effect of α strictly monotone. By combining this monotonicity with (35a), it follows that α can make no term of the affected denominator sum in (36) to grow faster (or shrink slower) than the affected numerator, and, by furthermore combining this monotonicity with (35b), it follows that α will make all terms of the affected denominator sum in (36) that correspond to the dimensions in K to grow non-vanishingly slower (or to shrink non-vanishingly faster for a negative $\bar{\omega}^\top \mathbb{I}(i^\star)$) than the affected numerator:

$$\begin{cases} \sigma_{i^\star} \geq \sigma_j \;, & \forall j & \text{(38a)} \\ \sigma_{i^\star} \gg \sigma_j \;, & \forall j \in K \;. & \text{(38b)} \end{cases}$$

We now see from (38) that for the affected fraction in (36) to stay nearly constant, it must be that *all* of the terms in the denominator sum that correspond to the dimensions in K have a vanishingly small effect on the sum: if any non-vanishingly contributing term is affected by (38b), there can be no other term that would compensate for this change in the sum, due to (38a). In other words, for the near-equality in (31) to hold for the considered $\bar{\omega}$ and for the maximizing $i^\star$ chosen earlier, every term in the denominator sum $\sum_z e^{\bar{\omega}^\top \mathbb{I}(z)}$ of the initial distribution must be either a) vanishingly close to the term for $i^\star$ ($e^{\bar{\omega}^\top \mathbb{I}(i^\star)} \not\gg e^{\bar{\omega}^\top \mathbb{I}(z)}$, so that $z \notin K$), or b) have a vanishingly small effect on the sum ($e^{\bar{\omega}^\top \mathbb{I}(z)} / \sum_{z'} e^{\bar{\omega}^\top \mathbb{I}(z')} < \epsilon$):

$$e^{\bar{\omega}^\top \mathbb{I}(i^\star)} - e^{\bar{\omega}^\top \mathbb{I}(z)} < \epsilon \quad \vee \quad \frac{e^{\bar{\omega}^\top \mathbb{I}(z)}}{\sum_{z'} e^{\bar{\omega}^\top \mathbb{I}(z')}} < \epsilon \;, \qquad \forall z \;. \tag{39}$$

By comparing to the semi-uniformity requirement in (32) while letting $c = e^{\bar{\omega}^\top \mathbb{I}(i^\star)}$, it can be seen that (39) now defines a policy that is vanishingly close to a semi-uniformly stochastic policy. Extension to the Gibbs policy in (1) follows directly by replacing $\mathbb{I}(i)$ with $f(k,i)$ as in Lemma 1.

□

Theorem 3. *Consider the family of methods from Theorem 2. Assume a smooth policy gradient field ($\|\nabla_\pi J(\pi_u) - \nabla_\pi J(\pi_v)\| \to 0$ as $\|\pi_u - \pi_v\| \to 0$) and $\tau \not\to 0$. First, the policy distance between a fixed point policy π^f and an optimal policy $\pi^\star$ cannot be vanishingly small ($\|\pi^f - \pi^\star\| \not< \epsilon$), except if the optimal policy $\pi^\star$ is a semi-uniformly stochastic policy. Second, for bounded returns ($\gamma \not\to 1$ and $r(s,a) \not\to \pm\infty, \forall s, a$), the policy distance between a fixed point policy π^f and an optimal policy $\pi^\star$ cannot be vanishingly small ($\|\pi^f - \pi^\star\| \not< \epsilon$), except if the optimal policy $\pi^\star$ is the uniformly stochastic policy.*

Proof outline. For a policy $\bar{\pi} = \pi(w_k/\tau_k)$ that is vanishingly close to an optimum, an unbiased parameter gradient η_k must, assuming a smooth gradient field, map to a policy gradient that is vanishingly close to zero, i.e., η_k must have a vanishingly small effect on $\bar{\pi}$ with any finite step size:

$$\|\pi(w_k/\tau_k + \alpha\eta_k) - \pi(w_k/\tau_k)\| < \epsilon \;, \quad \forall \alpha > 0, \; \alpha \not\to \infty \;. \tag{40}$$

If $\bar{\pi}$ is also a fixed point, then, by (24), we can substitute both w_k and η_k in (40) with $\hat{w}_k$:

$$\begin{aligned} & \|\pi(\hat{w}_k/\tau_k + \alpha\hat{w}_k) - \pi(\hat{w}_k/\tau_k)\| < \epsilon \;, && \forall \alpha > 0, \; \alpha \not\to \infty \\ \Leftrightarrow\; & \|\pi\left((1/\tau_k + \alpha)\hat{w}_k\right) - \pi((1/\tau_k)\hat{w}_k)\| < \epsilon \;, && \forall \alpha > 0, \; \alpha \not\to \infty \;. \end{aligned} \tag{41}$$

We now see that $\bar{\pi}$ is defined solely by a temperature-scaled version of a vanishingly small policy gradient, and that the condition in (40) is equivalent to stating that any finite decrease of the temperature must not have a non-vanishing effect on $\bar{\pi}$. As only semi-uniformly stochastic policies are invariant to such temperature decreases, it follows that $\bar{\pi}$ must be vanishingly close to such a policy.

Furthermore, if assuming bounded returns, then no dimension of the term $\hat{w}^\top \phi(s,a)$ can approach positive or negative infinity when $\hat{w}$ is estimated using (2) or (3). Consequently, for $\tau \not\to 0$, the uniformly stochastic policy $\pi(0)$ becomes the only semi-uniformly stochastic policy that the Gibbs policy class in (1) can approach, with the implication that $\bar{\pi}$ must be vanishingly close to the uniformly stochastic policy.

Proof. Consider some policy $\bar{\pi} = \pi(w_k/\tau_k)$ from the Gibbs policy class (1) that is vanishingly close to a local optimum $\pi^\star$:

$$\|\pi(w_k/\tau_k) - \pi^\star\| < \epsilon \ . \tag{42}$$

By definition, for a smooth policy gradient field, the parameter gradient vector $\eta_k = G(\bar{\pi})^{-1}\nabla_\theta J(\bar{\pi})$ estimated for this policy by an unbiased value function estimator and approximator (e.g., Monte Carlo evaluation with (2)) must be such that it maps to a vanishingly small policy gradient $\|\nabla_\pi J(\bar{\pi})\| < \epsilon$, i.e., η_k has a vanishingly small effect on $\bar{\pi}$ as long as the step size is finite:

$$\|\pi(w_k/\tau_k + \alpha\eta_k) - \pi(w_k/\tau_k)\| < \epsilon \ , \quad \forall \alpha > 0, \ \alpha \not\to \infty \ . \tag{43}$$

Now assume that $\bar{\pi}$ is a limiting fixed point policy of the update rule (4) when using the policy mapping (5). It follows immediately from (4) that such a fixed point must satisfy

$$\hat{w}_k = w_k \ , \tag{44}$$

regardless of the value of κ. If the policy evaluation step $\hat{w}_k := \mathrm{eval}(\pi(w_k/\tau_k))$ is performed using (2), we can directly use (7), which leads to

$$\hat{w}_k = \eta_k \ . \tag{45}$$

Also, the combination of (1) and (2) forms a 'compatible' actor-critic setup, implying that with an unbiased value function estimator, $\hat{w}_k = \eta_k$ is an unbiased gradient estimate (Section 2). Extension from (2) to (3) follows as in Theorem 2, with the implication that $\hat{w}_k$ is an unbiased gradient estimate also when using (3). This now allows us to take (43) and apply (44) and (45), so as to substitute both w_k and η_k with $\hat{w}_k$:

$$\begin{aligned} &\|\pi(\hat{w}_k/\tau_k + \alpha\hat{w}_k) - \pi(\hat{w}_k/\tau_k)\| < \epsilon \ , \qquad \forall \alpha > 0, \ \alpha \not\to \infty \\ \Leftrightarrow\ &\|\pi\left((1/\tau_k + \alpha)\hat{w}_k\right) - \pi((1/\tau_k)\hat{w}_k)\| < \epsilon \ , \quad \forall \alpha > 0, \ \alpha \not\to \infty \ . \end{aligned} \tag{46}$$

We observe that the sole effect of α in (46) is that it decreases the temperature of the policy, i.e., it makes the policy more deterministic by amplifying the probability differences present in $\pi(\hat{w}_k/\tau_k) = \pi(w_k/\tau_k) = \bar{\pi}$. At the same time, the inequality requirement in (46) states that such amplification must have a vanishingly small effect on the policy *for any finite* α: even huge changes in the policy temperature must have a vanishingly small effect on the policy. As shown by Lemma 2, this is possible only if $\bar{\pi}$ is vanishingly close to a semi-uniformly stochastic policy, so that there are no non-vanishing probability differences in non-vanishingly weak dimensions to be amplified:

$$|\bar{\pi}(a|s) - c_s| < \epsilon \quad \vee \quad \bar{\pi}(a|s) < \epsilon \ , \qquad \forall s, a, \ \forall s \ \exists c_s \in [0,1] \ . \tag{47}$$

The latter case ($\bar{\pi}(a|s) < \epsilon$) is possible for the considered Gibbs policy class only if the term $\theta_k^\top \phi(s,a)$ in (1) can approach negative infinity or some other term $\theta_k^\top \phi(s,z)$ can approach positive infinity (assuming a finite action space). In our case, this translates to the following requirement:

$$(\hat{w}_k/\tau_k)^\top \phi(s,a) \to \pm\infty \ , \quad \exists s, a \ . \tag{48}$$

We recall that with both (2) and (3), $\hat{w}_k^\top \phi(s,a)$ (without the temperature parameter) is a least-squares fit to the estimated value function, possibly with a baseline shift toward centered values when using (2). Consequently, $\hat{w}_k^\top \phi(s,a)$ can approach infinity only if the estimated value function can approach infinity, which, for an unbiased value function estimator, is possible only if the returns

$\sum_t \gamma^t \mathbb{E}[r(S_t, A_t)]$ can approach infinity. This requires that either $\gamma \to 1$ or $r(s,a) \to \pm\infty,\ \exists s, a$. This applies also to $(\hat{w}_k/\tau_k)^\top \phi(s,a)$ (with the temperature parameter) as long as $\tau \not\to 0$.[2] Thus, we have for $\tau \not\to 0$ that (48) and consequently the latter case in (47) can hold only if

$$\gamma \to 1 \quad \vee \quad r(s,a) \to \pm\infty\ ,\ \exists s, a\ . \tag{49}$$

In summary, a fixed point policy $\bar{\pi}$ that is vanishingly close to a local optimum must fulfill (47), i.e., it must be vanishingly close to a semi-uniformly stochastic policy. Furthermore, if the conditions in (49) are not allowed, then $\bar{\pi}$ must be vanishingly close to the uniformly stochastic policy.

□

To interpret the preceding theorems, we observe that the gist of them is that, assuming a well-behaved gradient field, the closer the evaluated policy is to an optimum, the closer the target point of the next greedy update will be to the origin (in policy parameter space). At a fixed point, the policy parameter vector must equal the target point of the next update, causing convergence to or toward a policy that is exactly optimal but not at the origin to be a contradiction (Theorem 2). Convergence to or toward a policy that is vanishingly close to an optimum is also impossible, except if the optimum is (semi-)uniformly stochastic (Theorem 3).

In practical terms, Theorem 2 states that even if the task at hand and the chosen hyperparameters would allow convergence to some policy in a finite number of iterations, the resulting policy can never contain optimal decisions, except for uniformly stochastic ones. Theorem 3 generalizes this result to the case of asymptotic convergence toward some limiting policy: for unbounded returns and any $\tau \not\to 0$, it is impossible to have asymptotic convergence toward any optimal decision in any state, except for semi-uniformly stochastic decisions, and for bounded returns and any $\tau \not\to 0$, it is impossible to have asymptotic convergence toward any non-uniform optimal decision in any state.

If convergence is to occur, then the limiting policy must reside "between" the origin and an optimum, i.e., the result must always undershoot the optimum that the learning process was influenced by. However, we can see in (23) that by decreasing the temperature τ, it is possible to shift this point of convergence further away from the origin and closer to the optimum: in the limit of $\tau \to 0$, (23) can permit the parameter vector $\hat{w}$ to converge toward a point that approaches the origin while, at the same time, allowing the corresponding policy $\pi(\hat{w}/\tau)$ to converge toward a policy that is arbitrarily close to a distant optimum (one can also see that with $\tau \to 0$, the inequality in (41) becomes satisfied for any $\hat{w}_k$, due to $\alpha \not\to \infty$). Unfortunately, as we already know, such manipulation of the distance of the fixed point from an optimum by adjusting τ can ruin convergence altogether in non-Markovian problems. Perkins & Precup [20] report negative convergence results for non-optimistic iteration ($\kappa = 1$) with a too low τ, while for optimistic iteration ($\kappa < 1$), Melo et al. [16] report a lack of positive results. Interestingly, this latter case is exactly what Theorem 1 addressed, showing that there actually *is* a way out and that it is by moving toward natural policy gradient iteration: decreasing the temperature τ toward zero causes the sub-optimality to vanish, while decreasing the interpolation factor κ at the same rate prevents the effective step size from exploding.

Finally, we provide a brief discussion on some questions that may have occurred to the reader by now. First, how does the preceding fit with the well-known soundness of greedy value function methods in the Markovian case? The crucial difference between the Markovian case (fully observable and tabular) and the non-Markovian case (partially observable or non-tabular) follows from the standard result for MDPs that states that in the former, all optima must be deterministic (with the possibility of redundant stochastic optima) [e.g., 25, §A.2]. For the Gibbs policy class, deterministic policies reside at infinity in some direction in the parameter space, with two implications for the Markovian case. First, the distance to an optimum never decreases. Consequently, the value function, being a correction toward an optimum, never vanishes toward a 'neutral' state. Second, only the direction of an optimum is relevant, as the distance can be always assumed to be infinite. This implies that in, and only in Markovian problems, the value function never ceases to retain all necessary information about the current solution, while in non-Markovian problems, relying solely on the value function can lead to losing track of the current solution.

Second, when moving toward an optimum at infinity, how can the value function / natural gradient (encoded by $\hat{w} = \eta$) stay non-zero and continue to properly represent action values while the corre-

[2] With $\tau_k \to 0$, the sum $1/\tau_k + \alpha$ in (46) becomes dominated by the first term (due to $\alpha \not\to \infty$) and the inequality becomes satisfied for any $\hat{w}_k$. We return to this in the following discussion.

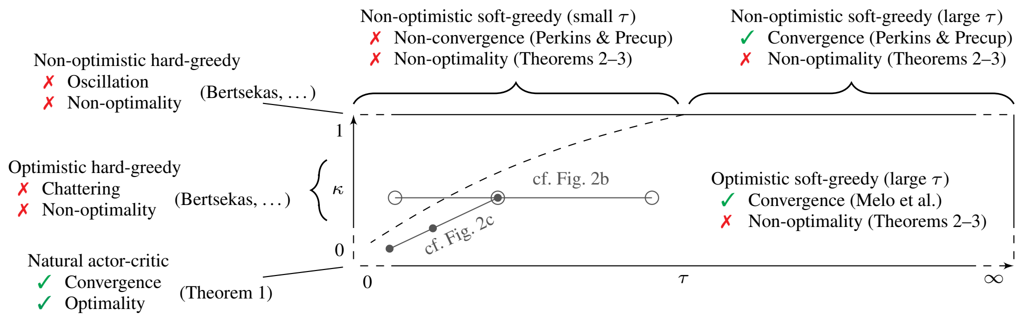
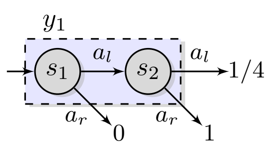
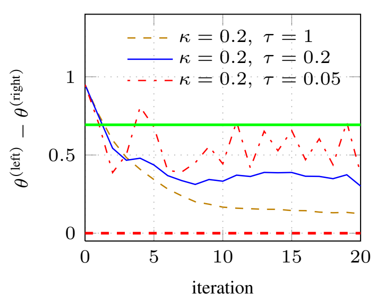


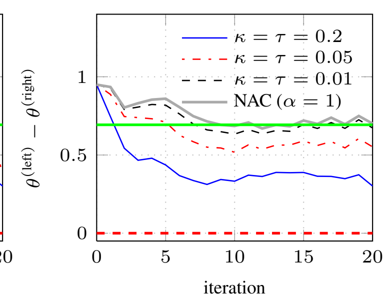
Figure 1: The hyperparameter space of the general form of (approximate) optimistic policy iteration in (4), with known convergence and optimality properties (see text for assumptions).



(a) A non-Markovian problem (adapted from [26]). The incoming arrow indicates the start state. Arrows leading out indicate termination with the shown reward.



(b) Non-optimality or oscillation with $\kappa \not\to 0$. The variants are marked with ○ in Fig. 1 (schematic).



(c) Interpolation toward NAC with $\kappa \to 0$ and $\tau \to 0$. The variants are marked with • in Fig. 1 (schematic).

Figure 2: Empirical illustration of the behavior of optimistic policy iteration ((1), (2), (4) and (5), with tabular ϕ) in the proximity of a stochastic optimum. The problem is shown in Fig. 2a. In Figures 2b and 2c, the optimum at $\theta^{(\text{left})} - \theta^{(\text{right})} = \log(2)$ is denoted by a solid green line. The uniformly stochastic policy is denoted by a dashed red line.

sponding policy gradient $\nabla_\pi J(\theta)$ must approach zero at the same time? We note that the equivalence in (7) is between a value function and a *natural* gradient η. We then recall that the curvature of the Gibbs policy class turns into a plateau at infinity, onto which the policy becomes pushed when moving toward a deterministic optimum. The increasing discrepancy between $\eta = G(\theta)^{-1}\nabla_\theta J(\theta) \not\to 0$ and $\nabla_\pi J(\theta) \to 0$ can be consumed by $G(\theta)^{-1}$ as it captures the curvature of this plateau.

5 Common ground

Figure 1 shows a map of relevant variants of optimistic policy iteration, parameterized as in (4). As is well known, the hard-greedy variants of this methodology (seen on the left edge on the map) can become trapped in non-converging cycles over potentially non-optimal policies (see Section 2 for references and exceptions). For a continuously soft-greedy policy class (toward right on the map), convergence can be established with enough softness [20, 16]. The natural actor-critic algorithm, which is convergent and optimal, is placed to the lower left corner by Theorem 1, while the inevitable non-optimality of soft-greedy variants toward right follows from Theorems 2 and 3. The exact (problem-dependent) place and shape of the line separating non-convergent and convergent soft-greedy variants (dashed line on the map) remains an open problem.

The main value of Theorem 1 is in bringing the greedy value function and policy gradient methodologies closer to each other. In our context, the unifying NAC(κ) formulation in (8) permits interpolation between the methodologies using the κ parameter. As discussed at the end of Section 4, the policy-forgetting term requires a Markovian problem for being justified: a greedy update implicitly stands on a Markov assumption and the κ parameter in (8) can be interpreted as adjusting the strength of this assumption. In this respect, the policy improvement parameter κ in NAC(κ) can be seen (inversely) as a dual in spirit to the policy evaluation parameter λ in TD(λ)-style algorithms. On the policy evaluation side, having $\lambda = 0$ obtains variance reduction by assuming and exploiting Markovianity of the problem, while $\lambda = 1$ obtains unbiased estimates also for non-Markovian problems. On the policy improvement side, with $\kappa = 1$, we have strictly greedy updates that gain in speed as the policy can respond instantly to new opportunities appearing in the value function (for empirical observations of such a speed gain, see [13, 27]), and in representational flexibility due to the lack of continuity constraints between successive policies (for a canonical example, consider fitted Q iteration). This comes at the price of either oscillation or non-optimality if the Markov assumption fails to hold, which is illustrated in Figure 2b for the problem in 2a. With $\kappa \to 0$, we approach natural gradient updates that remain sound also in non-Markovian settings, which is illustrated in Figure 2c. The possibility to interpolate between the approaches might turn out useful in problems with partial Markovianity: a large κ in the NAC(κ) formulation can be used to quickly find the rough direction of the strongest attractors, after which gradually decreasing κ allows a convergent final ascent toward an optimum.

Acknowledgments

This work has been financially supported by the Academy of Finland through project no. 254104, and by the Foundation of Nokia Corporation.

References

[1] Armijo, L. (1966). Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, *16*(1), 1–3.

[2] Baxter, J., & Bartlett, P. L. (2000). Reinforcement learning in POMDP's via direct gradient ascent. In *Proceedings of the Seventeenth International Conference on Machine Learning*, (pp. 41–48).

[3] Bertsekas, D. P. (1997). A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, *7*(4), 913–926.

[4] Bertsekas, D. P. (2005). *Dynamic Programming and Optimal Control*. Athena Scientific.

[5] Bertsekas, D. P. (2010). Pathologies of temporal difference methods in approximate dynamic programming. In *49th IEEE Conference on Decision and Control*, (pp. 3034–3039).

[6] Bertsekas, D. P. (2011). Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, *9*(3), 310–335.

[7] Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

[8] Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., & Lee, M. (2009). Natural actor-critic algorithms. *Automatica*, *45*(11), 2471–2482.

[9] Buşoniu, L., Babuška, R., De Schutter, B., & Ernst, D. (2010). *Reinforcement learning and dynamic programming using function approximators*. CRC Press.

[10] Crook, P. A., & Hayes, G. (2007). Consistent exploration improves convergence of reinforcement learning on POMDPs. In *AAMAS 2007 Workshop on Adaptive and Learning Agents*.

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

[12] Kakade, S., & Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, (pp. 267–274).

[13] Kakade, S. M. (2002). A natural policy gradient. In *Advances in Neural Information Processing Systems*.

[14] Kakade, S. M. (2003). *On the Sample Complexity of Reinforcement Learning*. Ph.D. thesis, University College London.

[15] Konda, V. R., & Tsitsiklis, J. N. (2004). On actor-critic algorithms. *SIAM Journal on Control and Optimization*, *42*(4), 1143–1166.

[16] Melo, F. S., Meyn, S. P., & Ribeiro, M. I. (2008). An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th International Conference on Machine Learning*, (pp. 664–671).

[17] Nedić, A., & Bertsekas, D. P. (2003). Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems: Theory and Applications*, *13*(1–2), 79–110.

[18] Pendrith, M. D., & McGarity, M. J. (1998). An analysis of direct reinforcement learning in non-Markovian domains. In *Proceedings of the Fifteenth International Conference on Machine Learning*.

[19] Perkins, T. J., & Pendrith, M. D. (2002). On the existence of fixed points for Q-learning and sarsa in partially observable domains. In *Proceedings of the Nineteenth International Conference on Machine Learning*, (pp. 490–497).

[20] Perkins, T. J., & Precup, D. (2003). A convergent form of approximate policy iteration. In *Advances in Neural Information Processing Systems*.

[21] Peters, J., & Schaal, S. (2008). Natural actor-critic. *Neurocomputing*, *71*(7-9), 1180–1190.

[22] Singh, S. P., Jaakkola, T., & Jordan, M. I. (1994). Learning without state-estimation in partially observable Markovian decision processes. In *Proceedings of the Eleventh International Conference on Machine Learning*.

[23] Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

[24] Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*.

[25] Szepesvári, C. (2010). *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers.

[26] Wagner, P. (2011). A reinterpretation of the policy oscillation phenomenon in approximate policy iteration. In *Advances in Neural Information Processing Systems 24*, (pp. 2573–2581).

[27] Wagner, P. (to appear). Policy oscillation is overshooting. *Neural Networks*. Author manuscript available at `http://users.ics.aalto.fi/pwagner/`.

[28] Wiering, M., & Van Otterlo, M. (2012). *Reinforcement Learning: State-of-the-Art*. Springer.