Towards Blackwell Optimality: Bellman Optimality Is All You Can Get

Victor Boone[†*] victor.boone@irit.fr

Adrienne Tuynman[‡*] adrienne.tuynman@inria.fr

[*]Equal contribution

[†]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

[†]IRIT, Université de Toulouse, CNRS, Toulouse INP, Toulouse, France

[‡]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189-CRIStAL, F-59000 Lille, France

Abstract

Although average gain optimality is a commonly adopted performance measure in Markov Decision Processes (MDPs), it is often too asymptotic. Further incorporating measures of immediate losses leads to the hierarchy of bias optimalities, all the way up to Blackwell optimality. In this paper, we investigate the problem of identifying policies of such optimality orders. To that end, for each order, we construct a learning algorithm with vanishing probability of error. Furthermore, we characterize the class of MDPs for which identification algorithms can stop in finite time. That class corresponds to the MDPs with a unique Bellman optimal policy, and does not depend on the optimality order considered. Lastly, we provide a tractable stopping rule that when coupled to our learning algorithm triggers in finite time whenever it is possible to do so.

1 INTRODUCTION

What a simple task it is to find the optimal policy of a Markov decision process for which only a single reward — or transition, is unknown. And yet, even if the uniqueness of the optimal policy is guaranteed beforehand, even if there is a single parameter to learn, even then, that learning task is surprisingly more delicate than it seems, especially when the optimality at stake is of higher order. Because when one is trying to learn higher order optimalities, standard learning approaches and intuitions fail.

Let us consider the simple learning problem depicted in Figure 1.

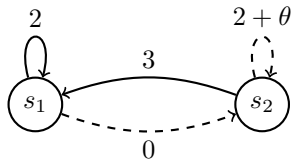

Figure 1: A class of deterministic transition Markov decision processes parameters by $\theta \in (-\varepsilon, \varepsilon)$. Choices of actions are represented by arrows that represent transitions. Labels are mean rewards. Rewards are reduced Gaussians, i.e., of the form $\mathrm{N}(-, 1)$.

In Figure 1, we consider a class of deterministic Markov decision processes $\mathcal{M} \equiv \{M_\theta : \theta \in (-\varepsilon, \varepsilon)\}$ where the only thing to learn is the variable θ in a neighborhood of 0. In all cases, the full-line policy has an asymptotic reward of 2, and when starting in state s_2, an immediate reward of 3. Meanwhile, the dashed-line policy has an asymptotic reward of $2 + \theta$ and an immediate reward of 0 when starting in s_1. Clearly, the full-line policy is the best when $\theta < 0$, while the dashed-lined one is better when $\theta > 0$. In case of the tiebreak $\theta = 0$, we see that the full-line policy is strictly better than the other because of its immediate reward from s_2 — in formal terms, it is said bias-optimal. At $\theta = 0$, we do not have both policies optimal from a higher order viewpoint. As such, the set of optimal policies is discontinuous at $\theta = 0$. Because no statistical test can ever determine that $\theta = 0$, it is impossible to determine which policy is optimal, despite there always only being one.

So, is bias-optimality impossible to learn?

Despite its simplicity, Figure 1 provides a glimpse at

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

the answer: if "$\theta = 0$" is likely to hold under the current statistical data, the full-line policy is a better guess for bias-optimality than the dashed-line one. The idea is that if $\theta \neq 0$, the learning agent should be able at some point to reject the possibility that "$\theta = 0$". So, claiming that the full-line policy is bias-optimal — when "$\theta = 0$" is likely enough — is the better temporary guess, although it can never be quantifiably certain.

In this paper, we investigate the problem of learning high order optimalities in Markov decision processes and generalize the conclusions drawn from the example of Figure 1. In opposition to best arm identification in bandits, the learnability of the optimal policy is not determined by its uniqueness. Instead, there exists one order of optimality that decides the learnability of all others: Bellman optimality. Indeed, the learnability of optimal policies, whatever the order, is equivalent to the uniqueness of the Bellman optimal policy.

2 PRELIMINARIES

Let us start by giving a few common definitions and notations for our setting. In general, given a finite set $\mathcal{X}$, we write $\mathcal{P}(\mathcal{X})$ the set of probability distributions on $\mathcal{X}$, and $2^{\mathcal{X}}$ the power set of $\mathcal{X}$. For $p \in \mathcal{P}(\mathcal{X})$ and $u \in \mathbb{R}^{\mathcal{X}}$, we denote $pu := \sum_{x\in\mathcal{X}} u_x p_x$ the expectation of $u(X)$ for $X \sim p$. We further write $\mathrm{sp}(u) := \max(u) - \min(u)$ the **span semi-norm** of $u \in \mathbb{R}^{\mathcal{X}}$. Lastly, for $u \in \mathbb{R}^{\mathcal{X}}_+$, we let

$$\mathrm{dmin}(u) := \min\{u(x) : x \in \mathcal{X} \text{ and } u(x) > 0\}$$

be the **definite minimum** of u, with the convention $\mathrm{dmin}(u) = 0$ for $u = 0$.

2.1 MDPs and high order optimalities

A Markov decision process is a tuple $M \equiv (\mathcal{Z}, \nu, p)$, where $\mathcal{Z} \equiv \prod_{s\in\mathcal{S}}\{s\} \times \mathcal{A}(s)$ is the pair space, constructed as the dependent product of the state space $\mathcal{S}$ with the action space $\mathcal{A}$; the function $\nu : \mathcal{Z} \to \mathcal{P}([0,1])$ is the reward function, that maps state-action pairs to probability distributions over $[0,1]$; the function $p : \mathcal{Z} \to \mathcal{P}(\mathcal{S})$ is the transition kernel, that maps pairs to probability distributions over states. The mean reward of the pair $z \in \mathcal{Z}$ is the expected reward upon playing z and is given by $r(z) := \int x \, \mathrm{d}\nu(z)(x)$. We finally define the distance between two models M and M' by

$$d_\infty(M, M') := \max\{\|r - r'\|_\infty, \|p - p'\|_\infty\}$$

2.1.1 Interacting with a MDP

At time $t \geq 1$, the controller observes the current state S_t, picks an action $A_t \in \mathcal{A}(S_t)$ and observes a reward $R_t \sim \nu(S_t, A_t)$ and the next state $S_{t+1} \sim p(S_t, A_t)$, both sampled independently from past observations. The played pair is denoted $Z_t \equiv (S_t, A_t)$. The resulting interaction is summarized into a single vector $H_t := (S_1, A_1, R_1, S_2, \ldots, S_t)$, called the history of observations. That vector lives in the history space, given by $\mathcal{H} := \bigcup_{t=1}^{\infty} \mathcal{H}_t$ where $\mathcal{H}_t := (\mathcal{Z} \times [0,1])^{t-1} \times \mathcal{S}$. Formally, the choice of action A_t is $\sigma(H_t)$-measurable and can be described as a kernel $\Lambda : \mathcal{H} \to \mathcal{A}$ from possible histories to actions, that we refer to as a **sampling rule**. Finally, the choice of a Markov decision process M, of a sampling rule Λ and of an initial state $s_0 \in \mathcal{S}$ properly defines a probability space on the history space $\mathcal{H}$, of which we write $\mathbb{P}^{M,\Lambda}_{s_0}(-)$ and $\mathbb{E}^{M,\Lambda}_{s_0}[-]$ the probability and expectation. Often, the Markov decision process is controlled by sampling rules that are independent of time, called **policies**, written $\pi \in \Pi := \mathcal{S} \to \mathcal{A}$.

A few assumptions must be added to the Markov decision process to guarantee that learning objectives can be met. In some models, there are multiple groups of disconnected states: some states are not reachable from others. If high rewards are in those unreachable states, how is the agent supposed to learn them? Learning agents need to explore everywhere, and should not be punished for ending up trapped in low-reward states. To prevent this problem, we will exclusively consider models in which good states are always accessible.

Assumption 1. *The model M is **communicating**. That is, for every pair of states s, s', there exists a deterministic stationary policy π such that*

$$\exists n \geq 1, \quad \mathbb{P}^{M,\pi}_s(S_n = s') > 0.$$

Finally, in some cases, we will look closely at what we call **unichain policies**. Those are policies π such that states are either reachable from every other state (and thus are visited infinitely many times, whatever the starting state), or are not reachable from the first set of states (and are thus visited finitely many times, whatever the starting state).

2.1.2 Gain, High Order Biases & Optimalities

The following definitions are deduced from Puterman (1994), and detailed in Appendix A. Given a policy $\pi \in \Pi$, its **gain** and **bias** functions are:

$$g^\pi(s, M) := \lim_{T\to+\infty} \frac{1}{T} \mathbb{E}^{M,\pi}_s \left[\sum_{t=1}^{T} R_t\right]$$

$$b^\pi(s, M) := \operatorname*{Cesaro\text{-}lim}_{T\to+\infty} \mathbb{E}^{M,\pi}_s \left[\sum_{t=1}^{T} (R_t - g^\pi(S_t, M))\right],$$

where Cesaro-lim stands for the Cesàro limit, given by $\text{Cesaro-lim}_{T\to+\infty} u_T := \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} u_t$ when the

limit exists — Note that the limit and the Cesàro limit of a convergent sequence coincide.

The gain is the asymptotic average amount of reward that the policy collects, while the bias is the amount of immediate reward gathered in addition to the gain. These quantities will also be written as $y_{-1} \equiv g$ and $y_0 \equiv b$. By convention, we let $y_{-2} = 0$. Beyond the gain and the bias, the n-th order bias y_n^π of π (for $n \geq 1$) is the bias of π for which the reward function $r(s,a)$ is changed to the previous bias value $-y_{n-1}^\pi(s)$:

$$y_n^\pi(s;M) := -\operatorname*{Cesaro-lim}_{T\to+\infty} \mathbb{E}_s^{M,\pi}\left[\sum_{t=1}^{T} y_{n-1}^\pi(S_t;M)\right].$$

The **optimal gain** and **bias** functions are defined in a nested fashion. For $n=-2$, we set $\Pi_{-2}^*(M) = \Pi$ as the set of all policies. For $n \geq -1$, we define the optimal bias $y_n^*(M)$ of order n as well the set of optimal policies $\Pi_n^*(M)$ of order n, also called **n-optimal policies**, in a co-inductive way:

$$\begin{cases} y_n^*(s,M) := \sup\limits_{\pi\in\Pi_{n-1}^*(M)} y_n^\pi(s,M) \\ \Pi_n^*(M) := \{\pi \in \Pi : \forall k \leq n, y_k^\pi(M) = y_k^*(M)\}. \end{cases}$$

Note that $\Pi_n^*(M) \subseteq \Pi_{n-1}^*(M)$. The set of **Blackwell optimal policies** is the set of policies that are optimal at every order, i.e., $\Pi_\infty^*(M) := \bigcap_{n=-1}^{+\infty} \Pi_n^*(M)$. It is known that Blackwell optimal exist unconditionally, i.e., that $\Pi_\infty^*(M) \neq \emptyset$, see (Puterman, 1994, Theorem 10.1.4) for example.

Lastly the n-th order **gaps** relative to a policy $\pi \in \Pi$ are defined as

$$\Delta_n^\pi(s,a) := y_n(s) + y_{n-1}(s) - p(s,a)y_n - r_n(s,a)$$

for (s,a) a state action pair, where $r_n := \mathbb{1}_{n=0} r$ is the n-th order reward function. They lead to what we call high order Bellman optimal policies.

Definition 1. *A policy is n-**Bellman optimal** if, for all $m \leq n$, we have for every state action pair $z \in \mathcal{Z}$,*

$$[\forall \ell < m, \Delta_\ell^\pi(z) = 0] \implies \Delta_m^\pi(z) \geq 0$$

As a matter of fact, Definition 1 states that the policy satisfies the first $n+2$ nested Bellman optimality equations (Puterman, 1994, §10.2). In other words, a policy is n-Bellman optimal if, to every order up to n, it beats the other actions that were optimal until then. When we say that policy π is **Bellman optimal** without specifying the order, we mean that it is 0-Bellman optimal, and we write $\pi \in \Pi^{\mathrm{B}}(M)$. This notion of Bellman optimality coincides with the one motivated by Gast et al. (2023). As we will see further down, those Bellman optimal policies have special connections to optimal policies of order n.

2.2 Fixed-Confidence Identification and δ-Probably Correct Algorithms

In identification, the underlying Markov decision process is unknown and the objective is to find an optimal policy as quickly as possible. As such, the learning algorithm explores the hidden Markov decision process by playing actions. Doing so, it gathers information from which a stream of potential optimal policies is produced. Formally, an identification algorithm consists in a pair of

(1) A **sampling rule** $\Lambda : \mathcal{H} \to \mathcal{A}$ that decides how the Markov decision process is navigated;
(2) A **recommendation rule** $\hat{\pi} : \mathcal{H} \to \Pi$ that emits a stream of potential optimal policies.

The recommended policy at time $t \geq 1$, denoted $\hat{\pi}_t$, is the policy that the algorithm believes to be the best guess as an element of $\Pi^*(M)$.

From there, there are two main ways to appreciate the learning performance of a learning agent.

The more straight-forward approach is to minimize the probability of error $\mathbb{P}^{M,\Lambda}(\hat{\pi}_t \notin \Pi^*(M))$ as t grows. At the very least, one expects that this probability vanishes asymptotically, leading to the definition of consistency from Kaufmann et al. (2016), see Definition 2 below.

Definition 2 (Consistency)**.** *An identification algorithm $(\Lambda, \hat{\pi})$ is said **consistent** with respect to a class of models $\mathcal{M}$ and an optimality map $\Pi^* : \mathcal{M} \to 2^\Pi$ if*

$$\forall M \in \mathcal{M}, \quad \mathbb{P}^{M,\Lambda}(\hat{\pi}_t \notin \Pi^*(M)) = \mathrm{o}(1).$$

The natural performance criterion for consistent algorithms is the behaviour of the error probability $\mathbb{P}^{M,\Lambda}(\hat{\pi}_t \notin \Pi^*(M))$ when $t \to \infty$.

Beyond consistency alone, we are further interested in *optimality certificates*. That is, in addition to the mere recommendation $\hat{\pi}_t$, one requires the algorithm to estimate how likely $\hat{\pi}_t$ is to be optimal. In the literature, this learning task is typically formulated the other way around: one picks a confidence threshold $\delta > 0$ and looks for a stopping time τ_δ at which the recommendation $\hat{\pi}_{\tau_\delta}$ is correct with probability greater than $1-\delta$. Such algorithms are said δ-probably correct (δ-PC, Boone and Gaujal (2023)), see Definition 3.

Definition 3 (Probable correctedness)**.** *An identification algorithm $(\Lambda, \hat{\pi})$ together with a stopping time τ_δ is said δ-**probably correct** (δ-PC) with respect to a class of models $\mathcal{M}$ and an optimality map $\Pi^* : \mathcal{M} \to 2^\Pi$ if*

$$\forall M \in \mathcal{M}, \quad \mathbb{P}^{M,\Lambda}(\hat{\pi}_{\tau_\delta} \notin \Pi^*(M)) \leq \delta.$$

The natural performance criterion for δ-PC algorithms is their **sample complexity** $\mathbb{E}^{M,\Lambda}[\tau_\delta]$ (Kaufmann et al.

(2016); Mannor and Tsitsiklis (2004)), that measures the expected time before the algorithm stops.

In this paper, we will be interested in consistent algorithms enriched with a δ-PC property.

2.3 Related Works

The problem of identifying optimal policies can be likened to best arm identification (Audibert and Bubeck (2010); Mannor and Tsitsiklis (2004)) in multi-armed bandits, that are *de facto* state-less RL problems. The time complexity of best arm identification is thoroughly described by Kaufmann et al. (2016), where the complexity is shown to stem from the proximity of the mean rewards. With classical multi-armed bandits, barring specific settings such as risk-sensitive, an arm is optimal or it is not; hence, all optimality orders are absolutely equivalent, therefore occluding much of the nuances of the RL setting.

A worth-mentioning pool of works that sits closely to this work is perhaps the problem of the identification of gain optimal policies in average reward RL in the (ε, δ)-PAC setting (Jin and Sidford (2021); Wang et al. (2024); Zurek and Chen (2024)), that has seen significant advances in the recent years. In this setting, the goal is to find an ε-gain optimal policy with probability $1-\delta$ with as few samples of the underlying instance as possible. Despite the impossibility to estimate it (Tuynman et al. (2024)), the bias of the optimal policy appears in the sample complexity bounds (Lee et al. (2025); Zurek and Chen (2025)), highlighting the dependencies between the various optimality orders and the difficulties to determine them.

Beyond identification problems, other lines of work are focusing on the computation of Blackwell optimal policies. Notably, Grand-Clément and Petrik (2023); Mukherjee and Kalyanakrishnan (2025) have worked off the historical definition of Blackwell (1962), stating that a Blackwell optimal policy must be discount optimal for all discount factors beyond a threshold $\gamma_0 < 1$, providing new complexity guarantees regarding their computation. The value of the threshold is discontinuous in M however, thus complicating the approach when further incorporating noise and uncertainty.

At the intersection of identification and Blackwell optimal policies, it has been shown recently that, in the space of Markov decision processes with deterministic transitions, the learnability of Blackwell optimal policies depends on a few criteria on the gain and bias optimal policies (Boone and Gaujal (2023)). More specifically, the uniqueness of the bias optimal policy, and the unichain character of all gain optimal policies, were shown to be necessary conditions for learnability.

2.4 Contributions

We start our paper by showing that Blackwell optimal policies can be computed even in the presence of noise. More precisely, in Section 3, we show that consistent algorithms exist. For that, we provide `HOPE`, standing for `H`igher `O`rder `P`olicy iteration `E`psilonized, that we show to be consistent for Π^*_∞.

We stress that a natural set of MDPs to focus on are those for which consistent algorithms can be stopped in finite time. Such instances are said *non-degenerate*. Although non-degeneracy obviously depends on the order of optimality that we consider, we characterize non-degenerate MDPs at every order. To that end, we first prove in Section 4 that non-degenerate MDPs necessarily have a unique Bellman-optimal policy, generalizing the results of Boone and Gaujal (2023) to stochastic transition MDPs. Lastly, in Section 5, we provide a stopping rule for `HOPE`, therefore proving that MDPs with a unique Bellman optimal policy are non-degenerate. As a result, a MDP is non-degenerate if and only if it has a unique Bellman optimal policy, whatever the order of optimality that we consider.

3 CONSISTENT ALGORITHMS AND MODEL DEGENERACY

As shown in Figure 1, simply returning the best policy of the empirical policy is not necessarily a good idea, as the probability of error is not necessarily asymptotically vanishing. This raises the following question: for some order n, is there an algorithm that is consistent for the optimality mapping Π^*_n?

3.1 Existence of Consistent Algorithms

We show that consistent algorithms exist indeed for every order. This means that there exist an algorithm $(\Lambda, \hat{\pi})$ with vanishing probability of error for any order $n \geq -1$ and model M, i.e., $\mathbb{P}^{M,\Lambda}(\hat{\pi}_t \notin \Pi^*(M)) \to 0$.

For that, we build off Policy Iteration from Puterman (1994). The Policy Iteration algorithm maintains a policy that is improved gradually using a subroutine called Policy Improvement. That subroutine guarantees that the aggregate bias vector $(y^\pi_m)_{m\geq -1}$ is improved everytime the policy is updated. At a proof level, the algorithm progressively discovers the optimal biases $(y^*_m)_{m\geq -1}$. Once the algorithm is sure that $y^\pi_m = y^*_m$ for $m \leq k$, it will restrict the Policy Improvement subroutine to select state-action pairs such that $\Delta^\pi_m(s,a) = 0$ for $m \leq k-1$. At that point, the policy is improved by looking for state-action pairs for which $\Delta^\pi_k(s,a) < 0$ or $\Delta^\pi_{k+1}(s,a) < 0$. The issue of replicating this algorithm

in a learning setting is the constraint

$$\Delta_m^\pi(s,a) = 0, \quad \forall m \leq k-1 \tag{1}$$

because the estimation of the biases (y_m^π), therefore of the gaps Δ_m^π, is noisy.

Dealing with noise in Policy Iteration. We thus construct Higher Order Policy Iteration ($\texttt{HOPI}(n,\varepsilon)$), Algorithm 2 (described in detail in Appendix C), that, for each order, finds the set of policies that are *nearly* Bellman optimal for that order. It does so by relaxing the constraint (1) to

$$\Delta_m^\pi(s,a) \leq \varepsilon, \quad \forall m \leq k-1. \tag{2}$$

In the example of Figure 1 with $\theta = 0$, even if θ is a bit overestimated, the full-line policy will remain near gain-optimal, and will thus be considered for bias optimality instead of being disqualified.

Running HOPI with a small enough ε will return exactly the correct policies, even in the presence of noise — provided that this noise is small enough. Indeed, a small noise on the MDP translates to a small noise in the biases. We will show in Lemma B.11 that there exists a function $\psi : \mathbb{R}_+ \to [0,\infty]$ with

$$\sup_{\pi\in\Pi} \left| r_n + p y_n^\pi - r'_n - p' y'^\pi_n \right| \leq \psi(d_\infty(M,M'))$$

such that $\psi(\varepsilon) = \mathrm{O}(\varepsilon)$ when $\varepsilon \to 0$.

In Corollary C.8, we use that result to show that there exists a certain function f such that, when $d_\infty(M,M') \leq f(\varepsilon, M)$, $\texttt{HOPI}(n,0,M)$ gives the exact same result as $\texttt{HOPI}(n,\varepsilon,M')$, because the sequence of policies of $\texttt{HOPI}(n,\varepsilon,M')$ is a valid sequence of policies for $\texttt{HOPI}(n,0,M)$.

This is the basis for our Higher Order Policy iteration Epsilonized (HOPE) algorithm (Algorithm 1).

Algorithm 1: Higher Order Policy iteration Epsilonized

Require: Desired order of optimality $n \geq -1$.
  $t \leftarrow 0$
  loop
    Construct empirical MDP $\hat{M}_t$
    $\varepsilon \leftarrow t^{-1/4}$
    $\prod_s \mathcal{A}_n(s) \leftarrow \texttt{HOPI}_{n,\varepsilon}(\hat{M}_t)$
    Recommend any policy in $\prod_s \mathcal{A}_n(s)$
    Sample uniformly $a_t \in \mathcal{A}(s_t)$, observe r_t and s_{t+1}
    $t \leftarrow t+1$
  end loop

We will show in Appendix C that the choice of $\varepsilon_t := t^{-1/4}$ is sufficient to get vanishing probability of error.

Theorem 1. *For all $n \geq -1$, HOPE(n) (Algorithm 1) is consistent for Π_n^*.*

3.2 MDPs Learnable in Finite Time

Since consistent algorithms exist, we have shown that against all odds, high order optimal policies can be found even in the presence of noise. However, even if the recommendation is correct *at some point*, there might be no way of knowing when this happens. In Figure 1, for instance, since the optimal policy of M_θ changes depending on whether $\theta \geq 0$ or $\theta < 0$, it is not possible to reject $\theta < 0$ in M_0. The problem consists in giving a finite stopping time, that is, being able to at some time t give a certificate of optimality of the recommended policy. We define non-degenerate models as those in which, for some consistent algorithm, there exists a stopping time that is finite on the model.

Definition 4 (Non-degenerate models)**.** *A model $M \in \mathcal{M}$ is said non-degenerate with respect to an optimality map Π^* if there exists a consistent algorithm $(\Lambda, \hat{\pi})$ and a collection of stopping rules $(\tau_\delta)_{\delta>0}$ such that, regardless of the confidence parameter $\delta > 0$, we have*

(1) *$(\Lambda, \hat{\pi}, \tau_\delta)$ is δ-PC on $\mathcal{M}$ with respect to Π^*;*
(2) *τ_δ is almost-surely finite in M, i.e., we have $\mathbb{P}^{M,\Lambda}(\tau_\delta < \infty) = 1$.*

The question that we raise now is the following: which models are degenerate according to this definition? On which models is the construction δ-PC algorithms hopeless? This is the central question of our paper, and here is our final answer:

Theorem 2 (Characterization of non-degeneracy)**.** *Let $n \geq -1$. A Markov decision process $M \in \mathcal{M}$ is non-degenerate w.r.t. $\Pi_n^*(M)$ if, and only if M has a unique Bellman optimal policy.*

It is well-known that once the Bellman optimal policy π^* is unique, then it is not only the unique bias optimal policy, but also the unique Blackwell optimal policy, i.e., that $\Pi_0^*(M) = \Pi_1^*(M) = \ldots = \Pi_\infty^*(M) = \{\pi^*\}$, see (Puterman, 1994, Corollary 10.1.7).

Combined with this observation, Theorem 2 has a spectacular consequence: non-degeneracy does not depend on the order of optimality (Corollary 3). In more simple terms, be it the identification of gain optimal policies, of bias optimal policies, or of Blackwell optimal policies, consistent algorithms can only stop on the same exact class of Markov decision processes: exactly those that have a unique Bellman optimal policy.

Corollary 3 (Degeneracy collapse)**.** *The set of degenerate models with regards to Π_{-1}^*, Π_0^*, Π_1^*, …, Π_∞^* are all the same.*

Theorem 2 is a strict generalization of Theorem 1 from Boone and Gaujal (2023), in which the uniqueness of

the bias optimal policy is shown necessary for the learnability of Blackwell-optimal policies in deterministic policies. Our result extends this to stochastic models, focuses on consistent algorithms, and gives an exact and simpler characterization of the unlearnable models.

The following sections will be dedicated to proving and explaining Theorem 2.

4 DEGENERATE MODELS AND STOPPING IN FINITE TIME

The main result (Theorem 2) claims that for a consistent algorithm to stop in finite time, the Bellman optimal policy of the underlying instance must be unique. This section unravels the reason behind this result with Corollary 8: if multiple Bellman optimal policies coexist, then a consistent algorithm will indefinitely struggle to determine which one is supposed to be the best.

4.1 Non-Degeneracy and Locally Optimal Policies

We begin by drawing a general observation on non-degenerate instances. Non-degenerate instances admit policies that remain optimal under perturbations of the instance itself. This observation does not depend on the optimality order, or even on the optimality notion considered, see Proposition 4 below.

Proposition 4. *Let $\Pi^* : \mathcal{M} \to 2^\Pi$ be a measurable optimality map. If $M \in \mathcal{M}$ is non-degenerate, then there is a policy $\pi \in \Pi^*(M)$ that remains optimal in a neighbourhood of M, i.e.,*

$$\exists \varepsilon > 0, \quad \bigcap \{\Pi^*(M') : d_\infty(M, M') < \varepsilon\} \neq \emptyset.$$

Proof Sketch. If M is non-degenerate, there exists a consistent algorithm that identifies an optimal policy $\hat{\pi}$ of M in finite time. Because the algorithm only takes decisions from observed histories, it is very likely to recommend $\hat{\pi}$ in finite time for every instance M' that produces histories similar enough to M. In other words, the recommendation of the algorithm is continuous, so that if M and M' are close, the algorithm will recommend $\hat{\pi}$ for both M and M'. Accordingly, by consistency, $\hat{\pi}$ is an optimal policy of M' as well. □

Proposition 4 drives the remaining of the proof, and we will work with a converse version of it: if, for an optimality map $\Pi^* : \mathcal{M} \to 2^\Pi$, every $\pi \in \Pi^*(M)$ can be made into the unique optimal policy of an instance M' arbitrarily close to M, then M cannot be non-degenerate unless it has a unique optimal policy.

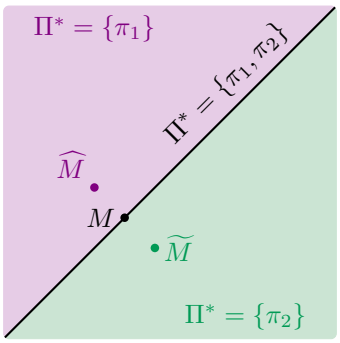

Figure 2: Illustration of Proposition 4 — as long as we can put $\widehat{M}$ and $\widetilde{M}$ arbitrarily close to M but having no common optimal policy, we can never learn an optimal policy for M

4.2 Making a Bellman Optimal Policy into the Unique Gain Optimal Policy

Following Proposition 4, to show that non-degenerate instances have a unique Bellman optimal policy, we prove that a Bellman optimal policy can be isolated as the unique gain optimal policy through an ergodification technique, that we call *shattering*.

Proposition 5 (Shattering Bellman optimality)**.** *Let M be a communicating instance. For every unichain Bellman optimal policy $\pi \in \Pi^{\mathrm{B}}(M)$ and precision $\varepsilon > 0$, there exists $M' \in \mathcal{M}$ with $d_\infty(M, M') < \varepsilon$ for which π is the unique gain optimal policy, i.e., $\Pi^*_{-1}(M') = \{\pi\}$.*

Note that Proposition 5 only claims that *unichain* Bellman optimal policies may be isolated in such fashion — that detail will be taken care of later on.

Proposition 5 is derived as a two-step process. First, we make $\pi \in \Pi^{\mathrm{B}}(M)$ the *unique* Bellman optimal policy of $M' \approx M$ by degrading the mean reward at every state-action pair that π does not play (Lemma 6). Then, we argue that if an instance has a unique Bellman optimal policy, then it can be made into the unique gain optimal policy under a well-chosen (infinitesimal) ergodic transform.

Lemma 6. *Let $M \equiv (\mathcal{Z}, \nu, p)$ be a communicating instance. For every unichain Bellman optimal policy $\pi \in \Pi^{\mathrm{B}}(M)$ and $\varepsilon > 0$, there exists $M' \equiv (\mathcal{Z}, r', p)$ with $d_\infty(M, M') < \varepsilon$ for which*

(1) *π is the unique Bellman optimal policy;*
(2) *$g^*(M') = g^*(M)$;*
(3) *$b^*(M') = b^*(M)$.*

Proof Sketch of Lemma 6. We consider the penalized Markov decision process $M'_\varepsilon \equiv (\mathcal{Z}, r'_\varepsilon, p)$ with rewards

$$r'_\varepsilon(s, a) := r_\varepsilon(s, a) - \varepsilon \mathbb{1}(a \neq \pi(s))$$

and check that the family (M'_ε) works as intended. □

Consider Figure 3. The red policy is unichain and Bellman optimal, but is not the only one. We can thus penalise every non-red action by substracting ε to their rewards: this can be an arbitrarily small modification. The red policy is the only Bellman-optimal policy. However, it is not the only gain-optimal policy.

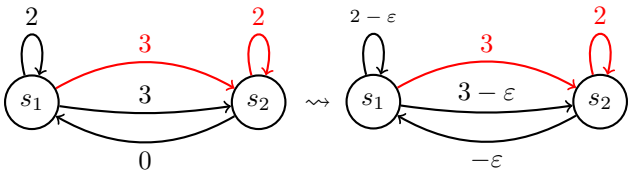

Figure 3: An illustration of Lemma 6, making a (non-unique) bias optimal policy (in red) into the unique Bellman-optimal policy.

It is to be noted that the model M' given by Lemma 6 is not $[0,1]$-reward in general. In general, this is not an issue because the latter property can be forced *a posteriori* via a linear reward transform.

In turn, we provide the construction for Proposition 5.

Proof Sketch of Proposition 5. Given a precision $\varepsilon > 0$ and a unichain Bellman optimal policy $\pi \in \Pi^{\mathrm{B}}(M)$, let M' be the instance isolating π as the unique Bellman optimal policy provided by Lemma 6. Then, we consider the ergodic transform $M'' \equiv (\mathcal{Z}, r'', p'')$ with reward and kernel functions given by

$$r''(s,a) := r'(s,a) + (p''(s,a) - p'(s,a))b^{\pi}(M')$$
$$p''(s'|s,a) := (1-\varepsilon)p'(s'|s,a) + \frac{\varepsilon}{|\mathcal{A}(s)|}.$$

We check that $d_\infty(M, M'') \leq C\varepsilon$ for some $C \geq 0$, and that for ε small enough, we have $\Pi^*_{-1}(M'') = \{\pi\}$. To force the rewards to be $[0,1]$, we further apply the linear reward transform $\varphi(x) := C\varepsilon + (1-2C\varepsilon)x$. □

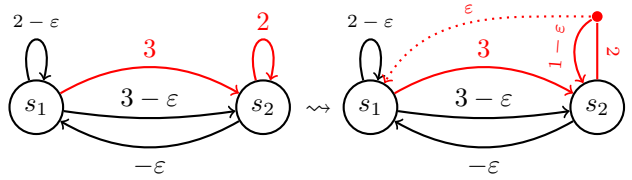

Figure 4: An example of the shattering operation, making the unique Bellman-optimal policy (in red) into the unique gain optimal policy.

Let us illustrate this idea with our example. In Figure 4, we make it so taking the looping action in s_2 has a small probability of going back to s_1. That way, s_1 becomes a recurrent state instead of a transitive one. The action taken in s_1 matters, as the reward incurred will impact asymptotic gain. That way, the red policy has become the unique gain-optimal policy.

4.3 Non-Degeneracy and Unique Bellman Optimal Policies

We can almost directly plug Propositions 4 and 5, but we have one issue to circumvent: Proposition 5 can only shatter *unichain* Bellman optimal policies. Therefore, we need to prove that if the Bellman optimal policy isn't unique, then there are multiple unichain Bellman optimal policies. Thankfully, this is indeed true.

Lemma 7. *If a communicating Markov decision process has multiple Bellman optimal policies, then it has multiple unichain Bellman optimal policies.*

The proof of Lemma 7 relies on the intimate structure of Bellman optimal policies; we defer it to Appendix.

Corollary 8. *Given $n \geq -1$, every Π^*_n-non-degenerate instance has a unique Bellman optimal policy.*

Proof. Let $M \in \mathcal{M}$ be a non-degenerate communicating instance. Denote $\Pi^{\mathrm{B}}_u(M)$ its set of unichain Bellman optimal policies. By Proposition 5, for every unichain Bellman optimal policy $\pi \in \Pi^{\mathrm{B}}_u(M)$ and precision $\varepsilon > 0$, there exists $M^\pi_\varepsilon \in \mathcal{M}$ such that $\Pi^*_{-1}(M^\pi_\varepsilon) = \{\pi\}$ and $d_\infty(M, M^\pi_\varepsilon) < \varepsilon$. The instance M is non-degenerate so by Proposition 4, for $\varepsilon > 0$ small enough, we have

$$\emptyset \neq \bigcap \{\Pi^*_n(M') : d_\infty(M, M') < \varepsilon\}$$
$$\subseteq \bigcap_{\pi \in \Pi^{\mathrm{B}}_u(M)} \Pi^*_{-1}(M^\pi_\varepsilon) = \bigcap_{\pi \in \Pi^{\mathrm{B}}_u(M)} \{\pi\}.$$

So $\Pi^{\mathrm{B}}_u(M)$ is necessarily a singleton and thus by Lemma 7, $\Pi^{\mathrm{B}}(M)$ is necessarily a singleton as well. □

5 MDPS WITH A UNIQUE BELLMAN OPTIMAL POLICY

At this point of the paper, you should be convinced that non-degenerate instances have a unique Bellman optimal policy. In this section, we show that the converse is equally true, therefore proving Theorem 2.

The heart of the rationale is that if the Bellman optimal policy of M is unique, then this policy remains the unique optimal policy of instances in a neighbourhood of M. Showing that this neighbourhood exists — and measuring it, is a hard result. Most of the proof details of the results down below are deferred to Appendix.

5.1 Deciding the Uniqueness of Bellman Optimal Policies

To begin with, we provide an important structural property of Bellman optimal policies that are unique: they have to be unichain (Lemma 9). This observation is not only very important to the analysis, it also has a huge computational interest, because deciding the uniqueness of the Bellman optimal policy becomes polynomial in time-complexity.

Lemma 9. *Let M be a communicating instance. Then*

(1) *If M has a unique Bellman optimal policy, then this policy is unichain.*
(2) *Conversely, if there exists a unichain policy $\pi \in \Pi$ such that $g^\pi(s) + b^\pi(s) > r(s,a) + p(s,a)b^\pi$ for all $a \neq \pi(s)$ and $s \in \mathcal{S}$, then π is the unique Bellman optimal policy of M.*

Lemma 9 provides a simple algorithm to determine whether a communicating Markov decision process has a unique Bellman optimal policy or not. First, compute a Bellman optimal policy π^* using Policy Iteration. If that policy isn't unichain, reject; otherwise, verify that every gap is positive. These tests can be realized in time $\mathrm{O}(|\mathcal{Z}|)$ once π^* is known, which is overall negligble in front of the computation of π^*, also polynomial.

5.2 A Stopping Rule with δ-Probable Correctness

Once the Bellman optimal policy is proven to be unichain, we can rely on a range of results that bound the variations of its gain and bias functions with respect to $d_\infty(M, M')$, that can be found in Appendix. Without the unichain property, such bounds are impossible to obtain without the assumption that M and M' have transition kernels of similar supports, and this assumption is not reasonable in identification problems.

Now, given a communicating instance $M \equiv (\mathcal{Z}, \nu, p)$ with unique Bellman optimal policy π^*, we introduce the threshold:

$$\beta(M) := \min\left\{ \frac{\mathrm{dmin}(\Delta^*)}{(1+4\alpha)(2+\mathrm{sp}(b^*))}, \frac{1}{\alpha} \right\} \qquad (3)$$

where we set $\alpha := \min_{s_\infty \in \mathcal{S}_\infty} \alpha(s_\infty)$ with $\alpha(s_\infty) := \max_{s \in \mathcal{S}} \mathbb{E}_s^{\pi^*}[\inf\{t \geq 1 : S_t = s_\infty\}]$ is the worst hitting time to s_∞ under the optimal policy π^*, and $\mathcal{S}_\infty$ is the set of recurrent states of π^*.

The threshold $\beta(M)$ provides a simple measure of the radius of the neighbourhood in which all instances have the same Bellman optimal policy as M. This is formalized in Lemma 10 thereafter.

Lemma 10. *Let $M \equiv (\mathcal{Z}, \nu, p)$ be a communicating instance with unique Bellman optimal policy π^*. For every $M' \equiv (\mathcal{Z}, \nu', p')$ such that $\mathrm{supp}(p') \supseteq \mathrm{supp}(p)$, if*

$$d_\infty(M, M') < \beta(M)$$

then π^* is the unique Bellman optimal policy of M'.

From a complexity perspective, the quantity $\beta(M)$ is computed in time $\mathrm{O}(|\mathcal{Z}| + |\mathcal{S}|^4)$. Indeed, the recurrent class $\mathcal{S}_\infty$ of π^* is determined in time $\mathrm{O}(|\mathcal{S}|^2)$. Given a recurrent state $s_\infty \in \mathcal{S}_\infty$, the gain and bias functions g^*, b^* are obtained by matrix inversions in $\mathrm{O}(|\mathcal{S}|^3)$ and Δ^* is deduced in $\mathrm{O}(|\mathcal{Z}|)$. Finally, the family of hitting times to s_∞ (under π^*) is the solution of a linear system that can be solved in time $\mathrm{O}(|\mathcal{S}|^3)$. Doing it for every $s_\infty \in \mathcal{S}_\infty$ yields α, for a total time cost $\mathrm{O}(|\mathcal{S}|^4)$.

Since the threshold $\beta(M)$ is easy to compute, Lemma 10 provides a simple stopping rule for `HOPE`, our consistent algorithm from Section 3. First, we design a high probability ball $\{M' : d_\infty(\hat{M}_t, M') < \xi_\delta(t)\}$ that contains M with probability $1-\delta$ uniformly over time, where the radius $\xi_\delta(t)$ is set to

$$\xi_\delta(t) := \sqrt{\frac{|\mathcal{S}| \log\left(\frac{2|\mathcal{Z}|(1+t)}{\delta}\right)}{\min_{z \in \mathcal{Z}} N_t(z)}}. \qquad (4)$$

Then, we stop `HOPE` as soon as $\xi_\delta(t) \leq \beta(\hat{M}_t)$, i.e., when all the instances within the confidence ball share the same unique Bellman optimal policy.

Proposition 11. *Upon running with the stopping rule*

$$\tau_\delta := \inf\left\{ t \geq 1 : \xi_\delta(t) \leq \beta(\hat{M}_t) \text{ and } \Pi^{\mathrm{B}}(\hat{M}_t) = \{\hat{\pi}_t\} \right\}$$

where $\xi_\delta(t)$ and $\beta(t)$ are respectively given by Equations (3) and (4), `HOPE` is δ-PC for Π_n^*, whatever $n \geq 1$. Moreover, for every instance M with unique Bellman optimal policy, we have $\mathbb{P}^{M,\texttt{HOPE}}(\tau_\delta < \infty) = 1$.

With this, we have all the ingredients necessary to finish our proof.

5.3 Instances with Unique Bellman Optimal Policies are Non-Degenerate

At last, the non-degeneracy of instances with unique Bellman optimal policies is a direct consequence of Proposition 11, concluding the proof of Theorem 2.

Corollary 12. *If a communicating Markov decision process has a unique Bellman optimal policy, then it is Π_n^*-non-degenerate for all $n \geq -1$.*

Proof. By Theorem 1, `HOPE` is consistent for Π_n^*. By Proposition 11, once enriched with the stopping rule τ_δ of Proposition 11, it is δ-PC. Lastly, by Proposition 11 again, we have $\mathbb{P}^{M,\texttt{HOPE}}(\tau_\delta < \infty) = 1$ for every instance M with unique Bellman optimal policy. □

6 CONCLUSION

As the main take-away from this paper, Bellman optimality is a fundamental limit to identifying n-optimal policies. This collapsing result bodes badly for learning in average reward: can we never try to learn policies optimal in both the long and the short term, then? There is hope in the PAC setting, in which one investigates *near* optimal policies.

Another obvious research path, now that we have isolated the models in which optimality is learnable, would be to investigate the necessary and achievable sample complexities.

ACKNOWLEDGMENTS

The authors acknowledge the funding of the French National Research Agency under the project FATE (ANR22-CE23-0016-01), the PEPR IA FOUNDRY project (ANR-23-PEIA-0003), the ANR LabEx CIMI (grant ANR-11-LABX-0040) within the French State Programme Investissements d'Avenir, and the AI Interdisciplinary Institute ANITI, funded by the France 2030 program under the Grant agreement n°ANR-23-IACL-0002. The second author is member of the Inria team Scool.

References

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p.

Blackwell, D. (1962). Discrete dynamic programming. *The Annals of Mathematical Statistics*, pages 719–726. Publisher: JSTOR.

Boone, V. (2024). Optimal regrets in markov decision processes. PhD thesis, see victor-boone.github.io.

Boone, V. (2025). Asymptotically optimal regret in communicating Markov Decision Processes.

Boone, V. and Gaujal, B. (2023). Identification of Blackwell Optimal Policies for Deterministic MDPs. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 7392–7424. PMLR.

Boone, V. and Gaujal, B. (2025). Logarithmic Regret of Exploration in Average Reward Markov Decision Processes. _eprint: 2502.06480.

Bourel, H., Maillard, O., and Talebi, M. S. (2020). Tightening Exploration in Upper Confidence Reinforcement Learning. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1056–1066. PMLR.

Gast, N., Gaujal, B., and Khun, K. (2023). Testing indexability and computing Whittle and Gittins index in subcubic time. *Mathematical Methods of Operations Research*, 97(3):391–436. Publisher: Springer.

Grand-Clément, J. and Petrik, M. (2023). Reducing Blackwell and Average Optimality to Discounted MDPs via the Blackwell Discount Factor. In *NeurIPS 2023*.

Jin, Y. and Sidford, A. (2021). Towards tight bounds on the sample complexity of average-reward MDPs. In *International Conference on Machine Learning*, pages 5055–5064. PMLR.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42.

Lee, J., Bravo, M., and Cominetti, R. (2025). Near-optimal sample complexity for MDPs via anchoring. *arXiv preprint arXiv:2502.04477*.

Mannor, S. and Tsitsiklis, J. N. (2004). The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648.

Mukherjee, D. and Kalyanakrishnan, S. (2025). Efficient Computation of Blackwell Optimal Policies using Rational Functions.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

Tuynman, A., Degenne, R., and Kaufmann, E. (2024). Finding good policies in average-reward Markov Decision Processes without prior knowledge. In *Advances in Neural Information Processing Systems*, volume 37, pages 109948–109979.

Wang, S., Blanchet, J., and Glynn, P. (2024). Optimal Sample Complexity for Average Reward Markov Decision Processes.

Zurek, M. and Chen, Y. (2024). Span-based optimal sample complexity for weakly communicating and general average reward MDPs. *Advances in Neural Information Processing Systems*, 37:33455–33504.

Zurek, M. and Chen, Y. (2025). Span-agnostic optimal sample complexity and oracle inequalities for average-reward RL. In Haghtalab, N. and Moitra, A., editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 6156–6209. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A A REFRESHER ON HIGH ORDER OPTIMALITIES

High order optimalities (Puterman, 1994, Chapter 9) can be introduced in several ways. In this paper, they are introduced via nested Bellman equations. The book of (Puterman, 1994), that serves as the reference for standard Markov decision process material of this paper, introduces them via *sensitive discounted optimalities* instead. Although the links between the two approaches are well known and documented, this is a bit of a niche subject. For better accessibility, we sketch the main line of this theory in the paragraphs below.

In all the subsequent section, we fix a Markov decision process M.

The dependency on M is dropped from now on.

Given a policy π and $\gamma \in [0,1)$ a discount factor, the **γ-discounted value** of that policy is

$$V_\gamma^\pi(s) := \mathbb{E}_s^\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t\right]. \tag{A.1}$$

The optimal discounted value is $V_\gamma^*(s) := \sup_\pi V_\gamma^\pi(s)$. From standard theory, it is known that for $\gamma < 1$, discounted optimal policies exist, i.e., policies $\pi \in \Pi_\gamma^*$ such that that $V_\gamma^\pi = V_\gamma^*$. Ever since (Blackwell, 1962), it is also known that in a neighbourhood of 1, the set of discounted optimal policies Π_γ^* do not depend on the discount anymore; such policies are said **Blackwell optimal**, and this is the historical definition of Blackwell's optimality. As claimed by in the main body, this optimality can be also obtained as the ultimate refinement of high order optimalities. This starts with the following remark. Given a policy π, its discounted value function admits a Laurent series expansion at $\gamma = 1$ as a rational function of γ, see (Puterman, 1994, Theorem 8.2.3):

$$V_\gamma^\pi = \frac{y_{-1}^\pi}{1-\gamma} + \frac{y_0^\pi}{\gamma} + \sum_{k=1}^{\infty} \frac{(1-\gamma)^k}{\gamma^{k+1}} y_k^\pi \tag{A.2}$$

with $y_k^\pi \in \mathbb{R}^{\mathcal{S}}$ for all $k \geq -1$. The coefficients (y_k^π) coincide among Blackwell optimal policies, and denoted (y_k^*).

- (Puterman, 1994, Corollary 8.2.4) For every policy $\pi \in \Pi$, $g^\pi = y_{-1}^\pi$ and $b^\pi = y_0^\pi$;
- (Puterman, 1994, Definition 10.1.1) A policy $\pi^* \in \Pi$ is **n-discount optimal** if, for every $\pi \in \Pi$, we have $\liminf_{\gamma \to 1} (1-\gamma)^n (V_\gamma^{\pi^*} - V_\gamma^\pi) \geq 0$. (Puterman, 1994, Theorem 10.2.3) A policy $\pi^* \in \Pi$ is n-discount optimal if, and only if $(y_k^{\pi^*})_{k=-1}^n$ is the maximal of $\{(y_k^\pi)_{k=-1}^n : \pi \in \Pi\}$ under the lexicographic order on $\mathbb{R}^{n+2}$.
- (Puterman, 1994, Theorem 10.1.5) Blackwell optimal policies correspond to policies that are n-discount optimal for all $n \geq -1$.
- In particular, $g^* = y_{-1}^*$ and $b^* = y_0^*$;
- (Puterman, 1994, Theorem 10.2.1) Let π^* be a Blackwell optimal policy. Then $(y_n^{\pi^*})_{n \geq -1}$ satisfies the nested optimality equations from the main text. Therefore, so does $(y_n^*)_{n \geq -1}$.
- (Puterman, 1994, Theorem 8.2.8) Since $(y_n^*)_{n \geq -1} = (y_n^{\pi^*})_{n \geq -1}$ where π^* is any Blackwell optimal policy, we have

$$y_n^*(s) := -\operatorname*{Cesaro-lim}_{T \to \infty} \mathbb{E}_s^{\pi^*}\left[\sum_{t=1}^{T} y_{n-1}(S_t)\right] \tag{A.3}$$

  for all $n \geq 1$. In particular, the coefficients $(y_n^*)_{n \geq -1}$ of the Laurent series expansion of $V_\gamma^{\pi^*}$ coincides with the similarly denoted $(y_n^*)_{n \geq -1}$ from the main text.

B PRELIMINARY TECHNICAL RESULTS

In this Appendix, we provide a variety of preliminary results.

We start with some concentration results in Appendix B.1, quantifying the speed at which the empirical MDP $\hat{M}_t$ converges towards the real MDP in terms of the distance d_∞. We then give in Appendix B.2 a numerical result that we'll use in the following. In Appendix B.3, we study the number of visits in each state-action pair under uniform exploration. Finally, for two MDPs with "small" distance, we study in Appendix B.4 how closely their bias vectors are, for unichain policies and in general.

B.1 A Few Concentration Results from Probability Theory

Lemma B.1 (Maximal Azuma)**.** *Let (X_m) be a martingale difference sequence, with almost surely X_m bounded between -0.5 and 0.5 for all $m \in \mathbb{N}$. Denote $\hat{\mu}_n := \frac{1}{n}\sum_{m=1}^{n} X_m$ the empirical mean of the first n variables. For all $x \geq 0$, we have*

$$\mathbb{P}(\exists m \geq n, \hat{\mu}_m \geq x) \leq \exp(-2nx^2).$$

Lemma B.2 (Time-uniform Weissman, Boone (2024))**.** *Let q be a distribution over $\{1, \dots, d\}$. Let (X_m) be a sequence of i.i.d. random variables of distribution q, and $\hat{q}_n$ be the empirical mean over n samples, $\frac{1}{n}\sum_{m=1}^{n} X_m$. Then, for all $\delta > 0$,*

$$\mathbb{P}\left(\exists n \geq 1, \|\hat{q}_n - q\|_1 \geq \sqrt{\frac{d\log\big((2\sqrt{1+n})/\delta\big)}{n}}\right) \leq \delta.$$

Lemma B.3 (Time-uniform Azuma, Lemma 5 of Bourel et al. (2020))**.** *Let (Y_m) be a sequence of i.i.d. random variables bounded between 0 and 1, with $\hat{\mu}_n$ the empirical mean estimate after n samples and μ the real mean. Then, for all $\delta > 0$,*

$$\mathbb{P}\left(\exists n \geq 1, |\hat{\mu}_n - \mu| \geq \sqrt{\left(1+\frac{1}{n}\right)\frac{\log\big(\sqrt{1+n}/\delta\big)}{2n}} \leq \delta\right)$$

Corollary B.4 (Time-uniform MDP concentration)**.** *Let M be a Markov decision process with pair and state spaces $\mathcal{Z}$ and $\mathcal{S}$ respectively. Let $\hat{M}_t$ be its empirical estimation, as given in (...). For all $\delta > 0$, we have*

$$\mathbb{P}\left(\exists t \geq 1, d_\infty(M, \hat{M}_t) \geq \sqrt{\frac{|\mathcal{S}|\log\big((4|\mathcal{Z}|\sqrt{1+\min N_t})/\delta\big)}{\min N_t}}\right) \leq \delta$$

Proof. Write $M \equiv (\mathcal{Z}, p, r)$. Let $\xi(n,\delta) := \sqrt{|\mathcal{S}|\log((4|\mathcal{Z}|\sqrt{1+n})/\delta)/n}$ the error as given by Corollary B.4. We have

$$\begin{aligned}
\mathbb{P}\Big(\exists t \in \mathbb{N}, d_\infty(M, \hat{M}_t) \geq \xi(\min(N_t), \delta)\Big) &\overset{\circ 1}{\leq} \sum_{(s,a)\in\mathcal{Z}} \begin{bmatrix} \mathbb{P}(\exists t \in \mathbb{N}, \|\hat{p}_t(s,a) - p(s,a)\|_1 \geq \xi(\min(N_t), \delta)) \\ + \\ \mathbb{P}(\exists t \in \mathbb{N}, |\hat{r}_t(s,a) - r(s,a)| \geq \xi(\min(N_t), \delta)) \end{bmatrix} \\
&\overset{\circ 2}{\leq} \sum_{(s,a)\in\mathcal{Z}} \begin{bmatrix} \mathbb{P}(\exists t \in \mathbb{N}, \|\hat{p}_t(s,a) - p(s,a)\|_1 \geq \xi(N_t(s,a), \delta)) \\ + \\ \mathbb{P}(\exists t \in \mathbb{N}, |\hat{r}_t(s,a) - r(s,a)| \geq \xi(N_t(s,a), \delta)) \end{bmatrix} \\
&\overset{\circ 3}{\leq} |\mathcal{Z}| \cdot \frac{\delta}{2|\mathcal{Z}|} + |\mathcal{Z}| \cdot \frac{\delta}{2|\mathcal{Z}|} = \delta
\end{aligned}$$

where ($\circ$1) unfolds the definition of $d_\infty(-)$ and uses a union-bound over the set of pairs; ($\circ$2) uses that, for $x > 0$ and $y > 1$, $x \mapsto \frac{1}{x} \cdot \log y\sqrt{1+x}$ is decreasing, so that $\min N_t$ can be changed to $N_t(s,a)$; and ($\circ$3) bounds the deviations of transition kernels with Lemma B.2 and the deviations of rewards with Lemma B.3). □

B.2 A Numerical Result

We will also be using the following folklore bound.

Lemma B.5 (Folklore)**.** *Let p, q be two probability distributions over $\{1, ..., d\}$ and let $u \in \mathbb{R}^d$. Then $|(q-p)u| \leq \frac{1}{2}\mathrm{sp}(u)\|q-p\|_1$.*

Proof. Let $e := (1, \dots, 1)$ be the unitary vector. Because $\sum_{i=1}^{d} p_i = \sum_{i=1} q_i$, we have $(q-p)u = (q-p)(u+\lambda e)$ for all $\lambda \in \mathbb{R}$. We find

$$|(q-p)u| = \inf_{\lambda\in\mathbb{R}} |(q-p)(u+\lambda e)| \overset{(\dagger)}{\leq} \inf_{\lambda\in\mathbb{R}} \|q-p\|_1 \|u+\lambda e\|_\infty \overset{(\ddagger)}{=} \frac{1}{2}\mathrm{sp}(u)\|q-p\|_1$$

where ($\dagger$) follows from Hölder's inequality; and ($\ddagger$) follows by setting $\lambda := -\frac{1}{2}(\min(u) + \max(u))$. □

B.3 Visit Counts under Uniform Exploration

Algorithm `HOPE` uses uniform exploration. We thus need to determine how often each pair is visited, to use in conjunction with Corollary B.4.

Lemma B.6. *Let $M \equiv (\mathcal{Z}, p, r)$ a communicating Markov decision process and let π be the uniformly random policy, given by $\pi(a|s) := |\mathcal{A}(s)|^{-1}$. There exist constants $\beta, \gamma, \lambda > 0$ such that*

$$\mathbb{P}^{M,\pi}(\exists t \geq T, \exists z \in \mathcal{Z}, N_t(z) < \beta t) \leq \exp(-\gamma T + \lambda)$$

Proof. Fix a pair $z_0 \in \mathcal{Z}$. Consider the deterministic reward function $f(z) := \mathbb{1}(z = z_0)$, and let $M_{z_0} := (\mathcal{Z}, p, f)$ the copy of M with reward function f. In M_{z_0}, we get a reward only when we sample z_0. Let $g^\pi_{z_0}$, $b^\pi_{z_0}$ and $\Delta^\pi_{z_0}$ the gain, 0-bias and 0-gap functions of π in M_{z_0}. Because M_{z_0} is communicating and π is uniform, every pair z is recurrent under π, so in particular z_0 is recurrent. Therefore, $g_{z_0}(z) > 0$. Set $\gamma_{z_0} := \max\{\mathrm{sp}(b_{z_0}), \|\Delta^*_{z_0}\|_\infty\}$. We have

$$\begin{aligned}
N^M_t(z_0) = N^{M_{z_0}}_t(z_0) &= \sum_{i=1}^t (Z_i = z_0) \\
&\overset{\circ 1}{=} \sum_{i=1}^t \left(g_{z_0} + b_{z_0}(S_i) - p(Z_i) b_{z_0} - \Delta_{z_0}(Z_i)\right) \\
&\overset{\circ 2}{\geq} t g_{z_0} + b_{z_0}(S_1) - p(Z_t) b_{z_0} + \sum_{i=1}^t \left(e_{S_{i+1}} - p(Z_i)\right) b_{z_0} - \sum_{i=1}^t \left(e_{A_i} - \pi(S_{i-1})\right) \Delta_{z_0}(S_i) \\
&\geq t g_{z_0} - \gamma_{z_0} + \sum_{i=1}^t \left((e_{S_{i+1}} - p(Z_i)) b_{z_0} + (e_{A_i} - \pi(S_{i-1})) \Delta_{z_0}(S_i)\right)
\end{aligned}$$

where ($\circ$1) comes from the definition of Δ_{z_0} ; ($\circ$2) uses that $\pi(s) \cdot \Delta_{z_0}(s) = 0$ for any state s, where $\Delta_{z_0}(s) := (\Delta_{z_0}(s,a))_{a \in \mathcal{A}_s}$. In the last inequation, we recognize in the right hand side term a sum of martingale differences, each bounded by $2\gamma_{z_0}$. Therefore, by Lemma B.1 applied to

$$(\nu_i)_i := \left(\frac{(e_{S_{i+1}} - p(Z_i)) b_{z_0} + (e_{A_i} - \pi(S_{i-1})) \Delta_{z_0}(S_i)}{4\gamma_{z_0}}\right)_i$$

we get for any $x \geq 0$

$$\begin{aligned}
\exp\left(-\frac{T x^2}{8\gamma^2_{z_0}}\right) &\geq \mathbb{P}\left(\exists t \geq T, \frac{1}{t}\sum_{i=1}^t \nu_i \leq -\frac{x}{4\gamma_{z_0}}\right) \\
&\geq \mathbb{P}\left(\exists t \geq T, \sum_{i=1}^t (e_{S_{i+1}} - p(Z_i)) b_{z_0} + (e_{A_i} - \pi(S_{i-1})) \Delta_{z_0}(S_i) \leq -xt\right) \\
&\geq \mathbb{P}\left(\exists t \geq T, N_t(z_0) \leq t g_{z_0} - \gamma_{z_0} - xt\right)
\end{aligned}$$

and we get the result by picking $x := \frac{1}{3}\min_z g_z > 0$, with $\beta := x$, $\gamma := \frac{(\min_z g_z)^2}{72(\max_z \gamma_z)^2}$ and $\lambda := \frac{3\gamma \max_z \gamma_z}{\min_z g_z} + \ln|\mathcal{Z}|$. □

B.4 Distances Between Two MDPs

For two MDPs, it is expected that, as they get close, their behaviour should also become similar: gap functions, bias functions... In the following, we describe the variation of those quantities with regards to the distance between the models. We start with the special case of unichain policies.

B.4.1 Unichain Policies

The goal of this paragraph is to provide a bound on the variations of the gap function $\Delta^\pi(s,a) := g^\pi(s) + b^\pi(s) - p(s,a)b^\pi - r(s,a)$ of a *unichain* policy π under changes of the reward r and kernel function p (Lemma B.10). Recall that the distance between model M and M' in ℓ_∞-norm is given by

$$d_\infty(M, M') := \max_{z \in \mathcal{Z}}\{|r'(z) - r(z)| + \|p'(z) - p(z)\|_1\}.$$

Because the gap function is a combination of the gain and the bias function, the result is obtained by combining deviational bounds on these to quantities (Lemma B.7 and Lemma B.9). The technique is adapted and generalized from (Boone, 2025, Appendix D).

Lemma B.7 (Variations of the gain, unichain case, Boone and Gaujal (2025))**.** *Let $M \equiv (\mathcal{Z}, p, r)$ be a Markov decision process and fix $\pi \in \Pi$ a unichain policy of M. For all $M' \equiv (\mathcal{Z}, p', r')$, we have*

$$\|g^\pi(M') - g^\pi(M)\|_\infty \leq (1 + \mathrm{sp}(b^\pi(M)) d_\infty(M, M').$$

Lemma B.8 (Variations of reaching times, unichain case)**.** *Let $M \equiv (\mathcal{Z}, p, r)$ be a Markov decision process and fix $\pi \in \Pi$ a unichain policy of M. Let $s_\infty \in \mathcal{S}$ be a recurrent state of π and denote $\tau_\infty := \inf\{t \geq 1 : S_t = s_\infty\}$ the reaching time to s_∞. For every model $M' \equiv (\mathcal{Z}, p', r')$ such that $\mathrm{supp}(p') \supseteq \mathrm{supp}(p)$, for all $s \in \mathcal{S}$, we have*

$$\left|\mathbb{E}_s^{M',\pi}[\tau_\infty] - \mathbb{E}_s^{M,\pi}[\tau_\infty]\right| \leq \frac{1}{2}\mathbb{E}_s^{M',\pi}[\tau_\infty] \max_{s' \in \mathcal{S}} \left\{\mathbb{E}_{s'}^{M,\pi}[\tau_\infty]\right\} d_\infty(M, M').$$

In particular, if $d_\infty(M, M') \leq \left(\max_{s' \in \mathcal{S}} \mathbb{E}_{s'}^{M,\pi}[\tau_\infty]\right)^{-1}$, then $\mathbb{E}_s^{M',\pi}[\tau_\infty] \leq 2\mathbb{E}_s^{M,\pi}[\tau_\infty]$ for all $s \in \mathcal{S}$.

Proof. Let $M_0 \equiv (\mathcal{Z}, p_0, r_0)$ be the model with reward function $r_0(s, a) := \mathbb{1}(s \neq s_\infty)$ and kernel

$$p_0(s'|s, a) := \begin{cases} p(s'|s, a) & \text{if } s \neq s_\infty; \\ 0 & \text{if } s = s_\infty \text{ and } s' \neq s_\infty; \\ 1 & \text{if } s \neq s_\infty \text{ and } s' = s_\infty. \end{cases}$$

Define M_0' similarly. We see that s_∞ is eventually absorbing under π in M_0; This is also true in M_0' since $\mathrm{supp}(p') \supseteq \mathrm{supp}(p)$. So $g^\pi(M_0) = g^\pi(M_0') = 0$ and $b^\pi(s_\infty; M_0) = b^\pi(s_\infty; M_0') = 0$. Now, for $s \in \mathcal{S}$, we have

$$\begin{aligned}
\mathbb{E}_s^{M',\pi}[\tau_\infty] &= \mathbb{E}_s^{M_0',\pi}\left[\sum_{t=1}^{\tau_\infty - 1} r_0(s \neq s_\infty)\right] \\
&\overset{\circ 1}{=} \mathbb{E}_s^{M_0',\pi}\left[\sum_{t=1}^{\tau_\infty - 1} (g^\pi(M_0) + (e_{S_t} - p(Z_t)) b^\pi(M_0))\right] \\
&= b^\pi(s; M_0) - b^\pi(s_\infty; M_0) + \mathbb{E}_s^{M_0',\pi}\left[\sum_{t=1}^{\tau_\infty - 1} (p'(Z_t) - p(Z_t)) b^\pi(M_0)\right]
\end{aligned}$$

where ($\circ$1) follows from the Poisson equation of π in M_0. Similarly, by unfolding the Poisson equation of π in M_0'', we find that $\mathbb{E}_s^{M'',\pi}[\tau_\infty] = b^\pi(s; M_0'')$ for every $M'' \in \{M, M'\}$. Applying Lemma B.5, we conclude that

$$\begin{aligned}
\left|\mathbb{E}_s^{M',\pi}[\tau_\infty] - \mathbb{E}_s^{M,\pi}[\tau_\infty]\right| &\leq \frac{1}{2}\mathrm{sp}(b^\pi(M_0))\mathbb{E}_s^{M',\pi}[\tau_\infty]\|p' - p\|_\infty \\
&\leq \frac{1}{2}\mathbb{E}_s^{M',\pi}[\tau_\infty] \max_{s' \in \mathcal{S}} \left\{\mathbb{E}_{s'}^{M,\pi}[\tau_\infty]\right\} d_\infty(M, M').
\end{aligned}$$

This concludes the proof. □

Lemma B.9 (Variations of the bias, unichain case)**.** *Let $M \equiv (\mathcal{Z}, p, r)$ be a Markov decision process and fix $\pi \in \Pi$ a unichain policy of M. Let $s_\infty \in \mathcal{S}$ be a recurrent state of π and denote $\tau_\infty := \inf\{t \geq 1 : S_t = s_\infty\}$ the reaching time to s_∞. For every model $M' \equiv (\mathcal{Z}, p', r')$ such that $\mathrm{supp}(p') \supseteq \mathrm{supp}(p)$ and for all $s \in \mathcal{S}$, we have*

$$|b^\pi(s; M') - b^\pi(s; M) + b^\pi(s_\infty; M) - b^\pi(s_\infty; M')| \leq 4\mathbb{E}_s^{M,\pi}[\tau_\infty](1 + \mathrm{sp}(b^\pi(M)))d_\infty(M, M')$$

provided that $d_\infty(M, M') \leq \left(\max_{s' \in \mathcal{S}} \mathbb{E}_{s'}^{M,\pi}[\tau_\infty]\right)^{-1}$.

Proof. Note that since $\mathrm{supp}(p') \supseteq \mathrm{supp}(p)$ and that π is unichain in M, it is unichain in M' as well. We further have $\mathbb{E}_s^{M',\pi}[\tau_\infty] < \infty$ regardless of the initial state $s \in \mathcal{S}$. Now, fix $s \in \mathcal{S}$. We have

$$b^\pi(s; M') - b^\pi(s_\infty; M')$$

$$
\begin{aligned}
&:= \mathbb{E}_s^{M',\pi}\left[\sum_{t=0}^{\tau_\infty-1}(r'(Z_t)-g^\pi(M'))\right]\\
&= \mathbb{E}_s^{M',\pi}\left[\sum_{t=0}^{\tau_\infty-1}(r(Z_t)-g^\pi(M))\right]+\mathbb{E}_s^{M',\pi}\left[\sum_{t=0}^{\tau_\infty-1}(r'(Z_t)-r(Z_t))\right]+\mathbb{E}_s^{M',\pi}\left[\sum_{t=0}^{\tau_\infty-1}(g^\pi(M)-g^\pi(M'))\right]\\
&\overset{\circ 1}{=} \mathbb{E}_s^{M',\pi}\left[\sum_{t=0}^{\tau_\infty-1}(p(Z_t)b^\pi(M)-b^\pi(S_t;M))\right]+\mathrm{d}r+\mathrm{d}g\\
&\overset{\circ 2}{=} b^\pi(s;M)-b^\pi(s_\infty;M)+\mathbb{E}_s^{M',\pi}\left[\sum_{t=0}^{\tau_\infty-1}(p(Z_t)-p'(Z_t))b^\pi(M)\right]+\mathrm{d}r+\mathrm{d}g\\
&\overset{\circ 3}{=} b^\pi(s;M)-b^\pi(s_\infty;M)+\mathrm{d}r+\mathrm{d}g+\mathrm{d}p
\end{aligned}
$$

where ($\circ$1) follows from the Poisson equation of π in M and introduces $\mathrm{d}r$ and $\mathrm{d}g$ for the two noise terms; ($\circ$2) is obtained by recognizing a telescopic sum; and ($\circ$3) introduces $\mathrm{d}p$ as the last noise term.

To bound each term, notice that since $d_\infty(M,M') \le 1/\max_{s'\in\mathcal{S}}\mathbb{E}_{s'}^{M,\pi}[\tau_\infty]$, we have $\mathbb{E}_s^{M',\pi}[\tau_\infty] \le 2\mathbb{E}_s^{M,\pi}[\tau_\infty]$ (Lemma B.8). Combined with the bound on gain deviations of Lemma B.7 and Lemma B.5, we obtain the desired result. □

Lemma B.10 (Variations of the gap function, unichain case)**.** *Let $M \equiv (\mathcal{Z},p,r)$ be a Markov decision process and fix $\pi\in\Pi$ a unichain policy of M. Let $s_\infty\in\mathcal{S}$ be a recurrent state of π and denote $\tau_\infty := \inf\{t\ge 1 : S_t = s_\infty\}$ the reaching time to s_∞. For every model $M'\equiv(\mathcal{Z},p',r')$ such that $\mathrm{supp}(p')\supseteq\mathrm{supp}(p)$, we have*

$$\|\Delta^\pi(M')-\Delta^\pi(M')\|_\infty \le \left(1+4\max_{s'\in\mathcal{S}}\mathbb{E}_{s'}^{M,\pi}[\tau_\infty]\right)(2+\mathrm{sp}(b^\pi(M)))d_\infty(M,M').$$

provided that $d_\infty(M,M') \le \left(\max_{s'\in\mathcal{S}}\mathbb{E}_{s'}^{M,\pi}[\tau_\infty]\right)^{-1}$.

Proof. Fix $(s,a)\in\mathcal{Z}$. We have

$$
\begin{aligned}
&|\Delta^\pi(s,a;M')-\Delta^\pi(s,a;M)|\\
&= |g^\pi(s;M')-g^\pi(s;M)| + |r'(s,a)-r(s,a)| + |(e_s-p'(s,a))b^\pi(M')-(e_s-p(s,a))b^\pi(M)|\\
&\overset{\circ 1}{\le} \left(2+\frac{1}{2}\mathrm{sp}(b^\pi(M))\right)d_\infty(M,M') + |(e_s-p'(s,a))(b^\pi(M')-b^\pi(M))| + |(p'(s,a)-p(s,a))b^\pi(M)|\\
&\overset{\circ 2}{\le} (2+\mathrm{sp}(b^\pi(M)))d_\infty(M,M') + \inf_{\lambda\in\mathbb{R}}|(e_s-p'(s,a))(b^\pi(M')-b^\pi(M)+\lambda e)|\\
&\overset{\circ 3}{\le} \left(1+4\max_{s'\in\mathcal{S}}\mathbb{E}_{s'}^{M,\pi}[\tau_\infty]\right)(2+\mathrm{sp}(b^\pi(M)))d_\infty(M,M')
\end{aligned}
$$

where ($\circ$1) follows by Lemma B.7; ($\circ$2) uses Lemma B.5 to bound the left-most term; and ($\circ$3) chooses $\lambda = b^\pi(s_\infty;M')-b^\pi(s_\infty;M)$ and invokes Lemma B.9. □

B.4.2 General Case

For the general case, we import a result from Boone (2025), see Lemma B.12 below, that bounds the variations of the bias function between two chains in the general setting. In the general setting, this bound requires that $\mathrm{supp}(p_1)=\mathrm{supp}(p_2)$, i.e., that the transition kernel of the two models have the same support. It further requires that $\|p_2-p_1\|$ is small relatively to the **generalized diameter** of p_1, that we recall below.

Definition B.1 (Generalized diameter of a kernel)**.** *Let p be a transition kernel over some finite space $\mathcal{S}$. Let $\mathcal{S}_1,\dots,\mathcal{S}_m\subseteq\mathcal{S}$ be the recurrent components of the Markov chain induced by p over $\mathcal{S}$. The **generalized diameter of** p is*

$$D(p) := \max_{s_0\in\mathcal{S}}\max_{(s_i)\in\prod_i(\mathcal{S}_i)}\mathbb{E}_s^p[\inf\{t\ge 1 : S_t\in\{s_1,\dots,s_m\}\}]. \tag{B.1}$$

That is, the generalized diameter is the worst hitting time to a covering of the recurrent components. Note that $D(p) < \infty$ as soon as $\mathcal{S}$ is finite. Because, given a MDP, a policy $\pi \in \Pi$ induces a kernel p_π, we will write $D(\pi) \equiv D(p_\pi)$ in abuse of notation. We finally define the **worst diameter** of a MDP as the largest generalized diameter over its policies, see below.

Definition B.2. *Let M be a Markov decision process with finitely many states. Its **worst diameter** is*

$$D_*(M) := \max_{\pi \in \Pi} D(\pi). \tag{B.2}$$

The worst diameter is always finite. Then, we prove the following.

Lemma B.11. *Let $M \equiv (\mathcal{Z}, p, r)$ a communicating Markov decision process. For all $n \geq 1$, there exists a constant $\alpha_n(M) \in \mathbb{R}$ such that, for all $\hat{M}$ with $d^*_\infty(M, \hat{M}) \leq 1/D_*(M)$, we have*

$$\max_{\pi \in \Pi} \max_{z \in \mathcal{Z}} |r_n(z) + p(z) y^\pi_n - \hat{r}_n(z) - \hat{p}(z) \hat{y}^\pi_n| \leq \alpha_n(M) d^*_\infty(M, \hat{M}) \tag{B.3}$$

Lemma B.12 (Variations of the bias function of a MRP, (Boone, 2025, Lemma D.14))**.** *Let $(M_i)_{i=1,2} \equiv (p_i, r_i)_{i=1,2}$ be a pair of MRPs with common state space $\mathcal{S}$. Denote $b_i \in \mathbb{R}^{\mathcal{S}}$ the bias function of (p_i, r_i). If $\|p_2 - p_1\|^*_\infty \leq 1/D(p_1)$, then*

$$\|b_2 - b_1\|_\infty \leq (12 + (16 + |\mathcal{S}|) D(p_1)) D(p_1) d_\infty(M_1, M_2)$$

where $d_\infty(M_1, M_2) := \max\{\|r_2 - r_1\|_\infty, \|p_2 - p_1\|_\infty\}$.

Proof of Lemma B.11. Because Π is a finite set, it is enough to establish Equation (B.3) for a single policy $\pi \in \Pi$. To begin with, note that by triangular inequality, we have

$$\begin{aligned}
\max_{z \in \mathcal{Z}} |r_n(z) + p(z) y^\pi_n - \hat{r}_n(z) - \hat{p}(z) \hat{y}^\pi_n| &\leq \|r_n - \hat{r}_n\|_\infty + \max_{z \in \mathcal{Z}} |(p(z) - \hat{p}(z)) y^\pi_n| + \max_{z \in \mathcal{Z}} |\hat{p}(z)(y^\pi_n - \hat{y}^\pi_n)| \\
&\leq \left(1 + \frac{1}{2} \mathrm{sp}(y^\pi_n)\right) d_\infty(M, \hat{M}) + \|y^\pi_n - \hat{y}^\pi_n\|_\infty
\end{aligned}$$

so that the result is about proving that there exists $\alpha'_n(M) < \infty$ such that $\|y^\pi_n - \hat{y}^\pi_n\| \leq \alpha'_n(M) d_\infty(M, \hat{M})$. Because the bias y^π_{n+1} of order $n+1$ is the bias of the Markov reward process $(-y^\pi_n, p^\pi)$, the proof naturally goes by induction on $n \geq 0$. For $n = 0$, since $D_*(M) \geq D(p^\pi)$ by definition of $D_*(M)$, the result follows readily from Lemma B.12, that bounds the variations of the 0-th order bias, by setting $\alpha'_0(M) := (12 + (16 + |\mathcal{S}|) D(p_1)) D(p_1)$. Supposing $n \geq 1$, the result follows by a straight-forward computation:

$$\begin{aligned}
\|y^\pi_n - \hat{y}^\pi_n\|_\infty &\overset{\circ 1}{\leq} (12 + (16 + |\mathcal{S}|) D(p_1)) D(p_1) \cdot d_\infty\big((-y^\pi_{n-1}, p^\pi), (-\hat{y}^\pi_{n-1}, \hat{p}^\pi)\big) \\
&\overset{\circ 2}{\leq} (12 + (16 + |\mathcal{S}|) D(p_1)) D(p_1) \cdot \alpha'_{n-1}(M) d_\infty(M, \hat{M}) \\
&\equiv (12 + (16 + |\mathcal{S}|) D(p_1))^{n+1} D(p_1)^{n+1} d_\infty(M, \hat{M})
\end{aligned}$$

where ($\circ$1) follows by Lemma B.12; and ($\circ$2) follows by induction. Note that on the way, we show that $\alpha'_n(M) = (12 + (16 + |\mathcal{S}|) D(p_1))^{n+1} D(p_1)^{n+1}$. This concludes the proof. □

C CONSISTENCY AND PROOFS OF SECTION 3

C.1 The `HOPI` Algorithm

We start off by presenting `HOPI`, our algorithm capable of computing optimal policies of the desired order. The pseudo-code is provided in Algorithm 2.

The algorithm is essentially an "epsilonized" version of the high order Policy Iteration algorithm (Puterman, 1994, §10), that is, a version of Policy Iteration that takes noise into account during each of its iterations. Because we will have to adapt the proof of Policy Iteration to a noisy setting, we provide a high level description of how the algorithm works. The key idea of the algorithm is to compute the set of state-action pairs $\mathcal{Z}_m$ that are optimal at order m, and to progressively increase m to further refine the optimality order of the output policy until the

Algorithm 2: Higher Order Policy Iteration

Input: Order n, slack $\varepsilon \geq 0$
Initialize $k \leftarrow 1$ and some arbitrary policy π_k
▷ *Initialize a Bellman optimal policy of order* 0
loop
  Compute g^{π_k} and b^{π_k}
  if g^{π_k} is not constant **then**
    Select π_{k+1} that is constant and such that $g^{\pi_{k+1}}(s) = \max_{s' \in \mathcal{S}} g^{\pi_k}(s')$
    $k \leftarrow k+1$
  else if there is $s \in \mathcal{S}$ with $\pi_k(s) \notin \operatorname{soft}_\varepsilon \operatorname{argmax}_a \{r(s,a) + p(s,a) b^{\pi_k}\}$ **then**
    $\pi_{k+1}(s) \leftarrow \operatorname{soft}_\varepsilon \operatorname{argmax}_a \{r(s,a) + p(s,a) b^{\pi_k}\}$ (break ties arbitrarily)
    $\pi_{k+1}(s') \leftarrow \pi_k(s')$ for $s' \neq s$
    $k \leftarrow k+1$
  else
    $\mathcal{A}_{-2} \leftarrow \mathcal{A}$, $k_{-2} \leftarrow k$
    break
  end if
end loop
▷ *Progressively refine the output to Bellman optimal policies of order* n
for $m = -1, 0, \ldots, n$ **do**
  loop
    Compute the bias vector $(y^{\pi_k}_{m'})_{-1 \leq m' \leq m+2}$
    if there is $s \in \mathcal{S}$ with $\pi_k(s) \notin \operatorname{soft}_\varepsilon \operatorname{argmax}_{a \in \mathcal{A}_{m-1}(s)} \{r_{m+1}(s,a) + p(s,a) y^{\pi_k}_{m+1}\}$ **then**
      ▷ *First Stage Policy Improvement*
      $\pi_{k+1}(s) \leftarrow \operatorname{soft}_\varepsilon \operatorname{argmax}_{a \in \mathcal{A}_{m-1}(s)} \{r_{m+1}(s,a) + p(s,a) y^{\pi_k}_{m+1}\}$ (break ties arbitrarily)
      $\pi_{k+1}(s') \leftarrow \pi_k(s')$ for $s' \neq s$
      $k \leftarrow k+1$
    else
      ▷ *Second Stage Policy Improvement*
      $\mathcal{A}^k_n(s) \leftarrow \operatorname{soft}_\varepsilon \operatorname{argmax}_{a \in \mathcal{A}_{m-1}(s)} \{r_{m+1}(s,a) + p(s,a) y_{m+1}\}$
      if there is $s \in \mathcal{S}$ with $\pi_k(s) \notin \operatorname{soft}_\varepsilon \operatorname{argmax}_{a \in \mathcal{A}^k_m(s)} \{r_{m+2}(s,a) + p(s,a) y^{\pi_k}_{m+2}\}$ **then**
        $\pi_{k+1}(s) \leftarrow \operatorname{soft}_\varepsilon \operatorname{argmax}_{a \in \mathcal{A}^k_m(s)} \{r_{m+2}(s,a) + p(s,a) y^{\pi_k}_{m+2}\}$ (break ties arbitrarily)
        $\pi_{k+1}(s') \leftarrow \pi_k(s')$ for $s' \neq s$
        $k \leftarrow k+1$
      else
        $\mathcal{A}_m(s) = \mathcal{A}^k_m(s)$, $k_m \leftarrow k$
        break
      end if
    end if
  end loop
end for
return $\mathcal{A}_n$

desired order n. That is, at stage $m \geq -2$, the goal of the algorithm is to construct a set of state-action pairs $\mathcal{Z}_m \subseteq \mathcal{Z}$ such that the associated set of policies

$$\Pi(\mathcal{Z}_m) := \{\pi \in \Pi : \forall s \in \mathcal{S}, (s, \pi(s)) \in \mathcal{Z}_m\} \tag{C.1}$$

satisfies $\Pi^*_{m+1}(M) \subseteq \Pi(\mathcal{Z}_m) \subseteq \Pi^*_m(M)$ (see Corollary C.6). To do so, the algorithm proceeds as follows.

At stage $m \geq -2$ and round k, the algorithm tries to refine the current policy π_k with a two-staged process. Write $\mathcal{A}_\ell(s) := \{a : (s, a) \in \mathcal{Z}_\ell\}$ the right-projection of $\mathcal{Z}_\ell$. Those are the actions that have been found to be optimal at order ℓ.

- (*First Stage*) The algorithm looks at the Bellman optimal equations of order $m+1$ clipped to $\mathcal{Z}_{m-1}$, the pairs that are known to be optimal at order $m-1$, and takes a soft argmax:

$$\mathcal{A}^k_m(s) := \left\{a \in \mathcal{A}_{m-1}(s) : r_{m+1}(s, a) + p(s, a)y^{\pi_k}_{m+1} \geq \max_{a' \in \mathcal{A}_m(s)} \left\{r_{m+1}(s, a') + p(s, a')y^{\pi_k}_{m+1}\right\} - \varepsilon\right\}. \tag{C.2}$$

  If the algorithm finds an improving state, i.e., $s \in \mathcal{S}$ such that $\pi_k(s) \notin \mathcal{A}^k_m(s)$, it improves the policy into π_{k+1} and starts over to round $k+1$. Otherwise, it tries to further improve π_k by moving to the second stage.

- (*Second stage*) The algorithm looks at the Bellman optimal equations of order $m+2$ clipped to $\mathcal{A}^k_m(s)$, the ε-optimal actions computed in the first stage, see Equation (C.2), and takes a soft argmax:

$$\mathcal{A}^k_{m+1}(s) := \left\{a \in \mathcal{A}^k_m(s) : r_{m+2}(s, a) + p(s, a)y^{\pi_k}_{m+2} \geq \max_{a' \in \mathcal{A}_m(s)} \left\{r_{m+2}(s, a') + p(s, a')y^{\pi_k}_{m+2}\right\} - \varepsilon\right\}. \tag{C.3}$$

  Then it does the same as in the first stage: If the algorithm finds an improving state, i.e., $s \in \mathcal{S}$ such that $\pi_k(s) \notin \mathcal{A}^k_{m+1}(s)$, it improves the policy into π_{k+1} and starts over to round $k+1$. Otherwise, $\mathcal{A}^k_{m+1}$ is considered correct, and the algorithm sets

$$\mathcal{Z}_m := \bigcup_{s \in \mathcal{S}} \{s\} \times \mathcal{A}^k_m(s). \tag{C.4}$$

  The algorithm sets $\pi_{k+1} \equiv \pi_k$, increases m to $m+1$ and moves to the next round $k+1$, starting the new phase $m+1$, to compute $\mathcal{Z}_{m+1}$.

On a side note, the case $m = -2$ is treated slightly differently than the others. Indeed, the usual criterion for Bellman optimality is that the policy must have constant gain (see Definition I.9 from Boone (2024)), so this is what we use here.

The next few sections are dedicated to proving the correctness of `HOPI`.

C.2 Links Between Optimalities and Gaps

To begin with, note that we can work on the aperiodic transform of the underlying Markov decision process $M \equiv (\mathcal{Z}, p, r)$, given by $M' \equiv (\mathcal{Z}, p', r')$ where $p'(s'|s, a) := \frac{1}{2}(p(s'|s, a) + \mathbb{1}(s' = s))$ and $r'(s, a) := \frac{1}{2}r(s, a)$, see (Puterman, 1994, §8.5.4). Indeed, we have $y^\pi_n(M') = 2^{n+2}y^\pi_n(M)$ (proof by induction on n), so that the optimality classes Π^*_n and Bellman optimalities are unchanged. Overall, we can assume freely that $p(s|s, a) \geq \frac{1}{2}$. In particular, all policies are aperiodic.

The following lemma quantifies the differences between bias vectors of two policies with regards to their gaps. Depending on whether some state-action pairs are transient or recurrent under π_2, a different order of gap will appear.

Lemma C.1. *Let $M \equiv (\mathcal{Z}, p, r)$ be a Markov decision process such that $p(s|s, a) \geq \frac{1}{2}$. Let $\pi_1, \pi_2 \in \Pi$ be two policies such that $y^{\pi_1}_n = y^{\pi_2}_n$ for $n \geq -1$. Then, for all $s \in \mathcal{S}$ and $T \geq 1$, we have*

$$Ty^{\pi_2}_{n+1}(s) + y^{\pi_2}_{n+2}(s) = Ty^{\pi_1}_{n+1}(s) + y^{\pi_1}_{n+2}(s) - \mathbb{E}^{\pi_2}_s\left[y^{\pi_1}_{n+2}(S_{T+1}) + \sum_{t=1}^{T}\left(\Delta^{\pi_1}_{n+2}(Z_t) + (T-t)\Delta^{\pi_1}_{n+1}(Z_t)\right)\right] + \mathrm{o}(1).$$

For $n=-2$, we have for any $s\in\mathcal{S}$ that

$$g^{\pi_2}(s)=g^{\pi_1}(s)-\lim_{T\to+\infty}\frac{1}{T}\mathbb{E}_s^{\pi_2}\left[\sum_{t=1}^{T}\big(\Delta_0^{\pi_1}(Z_t)+(T-t)\Delta_{-1}^{\pi_1}(Z_t)\big)\right].$$

Proof. Given a policy π_1, recall that the gap is defined by $\Delta_{n+1}^{\pi_1}(s,a)\coloneqq y_{n+1}^{\pi_1}(s)+y_n^{\pi_1}(s)-p(s,a)y_{n+1}^{\pi_1}-r_{n+1}(s,a)$. Therefore, for any pair of policies π_1,π_2, we have

$$r_{n+1}(s,a)-y_n^{\pi_1}(s)=y_{n+1}^{\pi_1}(s)-p(s,a)y_{n+1}^{\pi_1}-\Delta_{n+1}^{\pi_1}(s,a)$$
$$\mathbb{E}_s^{\pi_2}\left[\sum_{t=1}^{T'-1}(r_{n+1}(Z_t)-y_n^{\pi_1}(S_t))\right]=y_{n+1}^{\pi_1}(s)-\mathbb{E}_s^{\pi_2}\big[y_{n+1}^{\pi_1}(S_{T'})\big]-\mathbb{E}_s^{\pi_2}\left[\sum_{t=1}^{T'-1}\Delta_{n+1}^{\pi_1}(Z_t)\right].$$

Moreover, $r_{n+2}(s,a)-y_{n+1}^{\pi_1}(s)=y_{n+2}^{\pi_1}(s)-p(s,a)y_{n+2}^{\pi_1}-\Delta_{n+2}^{\pi_1}(s,a)$, so that

$$\begin{aligned}&\mathbb{E}_s^{\pi_2}\left[\sum_{t=1}^{T'-1}(r_{n+1}(Z_t)-y_n^{\pi_1}(S_t))\right]\\&\quad=y_{n+1}^{\pi_1}(s)+\mathbb{E}_s^{\pi_2}\big[y_{n+2}^{\pi_1}(S_{T'})-y_{n+2}^{\pi_1}(S_{T'+1})-\Delta_{n+2}^{\pi_1}(Z_{T'})-r_{n+2}(Z_{T'})\big]-\mathbb{E}_s^{\pi_2}\left[\sum_{t=1}^{T'-1}\Delta_{n+1}^{\pi_1}(Z_t)\right].\end{aligned}\tag{C.5}$$

For $n>-2$, $r_{n+2}=0$. Summing over $T'=1,\dots,T$, we get

$$\begin{aligned}&\mathbb{E}_s^{\pi_2}\left[\sum_{T'=1}^{T}\sum_{t=1}^{T'-1}(r_{n+1}(Z_t)-y_n^{\pi_1}(S_t))\right]\\&\quad=Ty_{n+1}^{\pi_1}(s)+y_{n+2}^{\pi_1}(s)-\mathbb{E}_s^{\pi_2}\big[y_{n+2}^{\pi_1}(S_{T+1})\big]-\mathbb{E}_s^{\pi_2}\left[\sum_{T'=1}^{T}\left(\Delta_{n+2}^{\pi_1}(Z_{T'})+\sum_{t=1}^{T'-1}\Delta_{n+1}^{\pi_1}(Z_t)\right)\right].\end{aligned}$$

Since $y_n^{\pi_1}=y_n^{\pi_2}$, the left hand term is equal to $\mathbb{E}_s^{\pi_2}\left[\sum_{T'=1}^{T}\sum_{t=1}^{T'-1}(r_{n+1}(Z_t)-y_n^{\pi_2}(S_t))\right]$. Take the equation for $\pi_1=\pi_2$. Because for every state s, policy π and integer m, we have $\Delta_m^{\pi}(s,\pi(s))=0$, it follows that

$$\begin{aligned}&Ty_{n+1}^{\pi_2}(s)+y_{n+2}^{\pi_2}(s)\\&\quad=Ty_{n+1}^{\pi_1}(s)+y_{n+2}^{\pi_1}(s)+\mathbb{E}_s^{\pi_2}\big[y_{n+2}^{\pi_2}(S_{T+1})-y_{n+2}^{\pi_1}(S_{T+1})\big]-\mathbb{E}_s^{\pi_2}\left[\sum_{T'=1}^{T}\left(\Delta_{n+2}^{\pi_1}(Z_{T'})+\sum_{t=1}^{T'-1}\Delta_{n+1}^{\pi_1}(Z_t)\right)\right]\\&\quad\overset{\circ 1}{=}Ty_{n+1}^{\pi_1}(s)+y_{n+2}^{\pi_1}(s)+\mathbb{E}_s^{\pi_2}\big[y_{n+2}^{\pi_2}(S_{T+1})-y_{n+2}^{\pi_1}(S_{T+1})\big]-\mathbb{E}_s^{\pi_2}\left[\sum_{t=1}^{T}\big(\Delta_{n+2}^{\pi_1}(Z_t)+(T-t)\Delta_{n+1}^{\pi_1}(Z_t)\big)\right]\\&\quad\overset{\circ 2}{=}Ty_{n+1}^{\pi_1}(s)+y_{n+2}^{\pi_1}(s)-\mathbb{E}_s^{\pi_2}\big[y_{n+2}^{\pi_1}(S_{T+1})\big]-\mathbb{E}_s^{\pi_2}\left[\sum_{t=1}^{T}\big(\Delta_{n+2}^{\pi_1}(Z_t)+(T-t)\Delta_{n+1}^{\pi_1}(Z_t)\big)\right]+\mathrm{o}(1)\end{aligned}$$

where ($\circ$1) is obtained by reindexing the left-most sum and ($\circ$2) follows from $\mathbb{E}_s^{\pi_2}[\mathbb{1}(S_{T+1}=s')]\to\mu^{\pi_2}(s'|s)$ where $\mu^{\pi_2}(-|s)$ is a probability invariant measure of π_2 (by aperiodicity), combined with $\sum_{s'\in\mathcal{S}}\mu^{\pi_2}(s'|s)y_{n+2}^{\pi_2}(s')=0$. This concludes the proof for $n>-2$.

For $n=-2$, Equation (C.5) becomes

$$0=g^{\pi_1}(s)+\mathbb{E}_s^{\pi_2}\left[b^{\pi_1}(S_{T'})-b^{\pi_1}(S_{T'+1})-\Delta_0^{\pi_1}(Z_{T'})-r(Z_{T'})-\sum_{t=1}^{T'-1}\Delta_{-1}^{\pi_1}(Z_t)\right]$$

$$= Tg^{\pi_1}(s) + b^{\pi_1}(s) - \mathbb{E}_s^{\pi_2}[b^{\pi_1}(S_{T+1})] - \mathbb{E}_s^{\pi_2}\left[\sum_{T'=1}^{T} r(Z_{T'})\right] - \mathbb{E}_s^{\pi_2}\left[\sum_{T'=1}^{T}\left(\Delta_0^{\pi_1}(Z_{T'}) + \sum_{t=1}^{T'-1}\Delta_{-1}^{\pi_1}(Z_t)\right)\right]$$

by summing over $T' = 1, \dots, T$. This yields the result as $T \to +\infty$ for $\pi_1, \pi_2 = \pi, \pi'$. □

With this, we can prove the link between n-Bellman optimality and $(n-1)$-optimality, extending a result known for $n = 0$.

Proposition C.2. *Let $M \equiv (\mathcal{Z}, p, r)$ be a communicating Markov decision process such that $p(s|s,a) \geq \frac{1}{2}$. For $m \geq -1$, a policy that is Bellman optimal at order m is $(m-1)$-optimal.*

Proof. For $m = -1$, there is nothing to say because $\Pi_{-2}^* = \Pi$. For $m = 0$, this is a well-known result, that a policy $\pi \in \Pi$ satisfying the pair of optimality equations

$$\forall s \in \mathcal{S}, \quad g^\pi(s) = \max_{a \in \mathcal{A}(s)} \{p(s,a) g^\pi\}$$
$$\forall s \in \mathcal{S}, \quad g^\pi(s) + b^\pi(s) = \max_{a \in \mathcal{A}(s)} \{r(s,a) + p(s,a) b^\pi\}$$

has optimal gain. See for example Proposition I.4 from Boone (2024). We focus on the case $m \geq 1$. By induction, π is $(m-2)$-optimal, i.e., $y_k^* = y_k^\pi$ for $k = -2, -1, \dots, m-2$. In particular and taking π^* a Blackwell optimal policy, we have $\Delta_k^\pi = \Delta_k^{\pi^*}$ for $k = -1, 0, \dots, m-2$, so that $\Delta_k^\pi(s, \pi^*(s)) = 0$ for all $s \in \mathcal{S}$. Since π is $(m-1)$-Bellman optimal, it follows that $\Delta_{m-1}^\pi(s, \pi^*(s)) \geq 0$ for all $s \in \mathcal{S}$. In all cases, invoking Lemma C.1 for $n \equiv m-1$, $\pi_1 = \pi$ and $\pi_2 = \pi^*$, we find

$$y_{m-1}^*(s) = y_{m-1}^\pi(s) - \lim_{T \to \infty} \mathbb{E}_s^{\pi^*}\left[\frac{1}{T}\sum_{t=1}^{T}\Big(\Delta_m^{\pi^*}(Z_t) + (T-t)\Delta_{m-1}^{\pi^*}(Z_t)\Big)\right]. \tag{C.6}$$

The crucial additional observation is the following: Since π is m-optimal, for every $z \in \mathcal{Z}$ such that $\Delta_k^\pi(z) = 0$ for $k \leq m-2$, either (i) $\Delta_{m-1}^\pi(z) > 0$; or (ii) $\Delta_{m-1}^\pi(z) = 0$ and $\Delta_m^\pi(z) \geq 0$. Accordingly, we have (i) $\Delta_{m-1}^\pi(Z_t) > 0$ or (ii) $\Delta_{m-1}^\pi(Z_t) = 0$ and $\Delta_m^\pi(Z_t) \geq 0$; $\mathbb{P}_s^{\pi^*}$-a.s. So,

$$\lim_{T \to \infty} \mathbb{E}_s^{\pi^*}\left[\frac{1}{T}\sum_{t=1}^{T}\Big(\Delta_m^{\pi^*}(Z_t) + (T-t)\Delta_{m-1}^{\pi^*}(Z_t)\Big)\right] \geq 0. \tag{C.7}$$

Combining Equations (C.6) and (C.7), we conclude that $y_{m-1}^* \leq y_{m-1}^\pi$. So π is $(m-1)$-optimal. □

C.3 `HOPI` Without Soft Argmax

In this subsection, we use the previous results to prove the correctness of `HOPI`$(n, 0)$ for any n, that is, of `HOPI` running without raw argmax rather than a soft argmax. The idea of the proof is the following. First, we show that the two-phased policy improvement of `HOPI` improves the policy indeed, by showing that $(y_m^{\pi_k})_{m \geq -1}$ is increasing with $k \geq 1$ for the lexicographic order, see Lemma C.3. Second, we show in Lemma C.5 that the stopping rule of phase m is correct: At the end of phase m, the policy π_k is $(m+1)$-Bellman optimal. Combining the two, we deduce that `HOPI` runs in finite time, and runs a set of pairs $\mathcal{Z}_n$ such that $\Pi_{n+1}^* \subseteq \Pi(\mathcal{Z}_n) \subseteq \Pi_n^*$.

First of all, we build a policy improvement lemma: changing the first non-zero gap to be null improves the policy.

Lemma C.3 (Local Policy Improvement works)**.** *Let $n \geq -1$. Let π and π' be two policies, such that*

1. *π and π' coincide on all states but one, denoted s';*
2. *For $m \leq n$, we have $\Delta_m^\pi(s', \pi'(s')) = 0$;*
3. *$\Delta_{n+1}^\pi(s', \pi'(s')) < 0$.*

Then $y_m^\pi = y_m^{\pi'}$ for $m \leq n-1$, and $(y_n^{\pi'}, y_{n+1}^{\pi'}) > (y_n^\pi, y_{n+1}^\pi)$ in lexicographic order.

Proof. By induction, applying Lemma C.1 at orders $-2, \dots, n-2$ yields $y_m^\pi = y_m^{\pi'}$ for $m = -1, \dots, n-1$.

Let us first assume that there exists a state s such that s' is recurrent under π' starting from s, meaning the asymptotic visitation frequency $\mu_s^{\pi'}(s') > 0$. Then for any such state s, Lemma C.1 yields

$$y_n^{\pi'}(s) = y_n^\pi(s) - \Delta_{n+1}^\pi(s', \pi'(s'))\mu_s^{\pi'}(s') > y_n^\pi(s)$$

and for any other state s, $y_n^{\pi'}(s) = y_n^\pi(s)$, yielding the result.

If there exists no such state, then $y_n^{\pi'} = y_n^\pi$, and we can apply Lemma C.1 at order n.

$$y_{n+1}^{\pi'}(s) = y_{n+1}^\pi(s) - \lim_{T\to+\infty} \mathbb{E}_s^{\pi'}\left[\sum_{T'=1}^{T}\sum_{t=1}^{T'-1} \Delta_{n+1}^\pi(Z_t)\right]$$

where $\lim_{T\to+\infty} \mathbb{E}_s^{\pi'}[\sum_{T'=1}^{T}\sum_{t=1}^{T'-1} \Delta_{n+1}^\pi(Z_t)] \le 0$, and $\lim_{T\to+\infty} \mathbb{E}_{s'}^{\pi'}[\sum_{T'=1}^{T}\sum_{t=1}^{T'-1} \Delta_{n+1}^\pi(Z_t)] < 0$ since starting in s' guarantees that s' is visited. Therefore, $y_{n+1}^{\pi'} > y_{n+1}^\pi$. □

This means that (π_k) will have increasing bias vectors (in lexicographic order), which entails that algorithm `HOPI`$(n, 0)$ necessarily stops (Corollary C.4).

Corollary C.4. *`HOPI`$(n, 0)$ stops in finite time.*

Even if the sequence of policies (π_k) is forever improving until the algorithm stops, we have yet to prove that the output policy is indeed optimal at order $n+1$. We thus push Lemma C.3 once step further, and show that once stage n finishes with the policy π_{k_n}, we have found an optimal policy of order $n+1$, because no policy can best it.

Lemma C.5 (Phase m stops with a m-Bellman optimal policy)**.** *For any $n \ge -1$, if $\Pi_n^* \subseteq \Pi(\mathcal{Z}_{n-1}) \subseteq \Pi_{n-1}^*$ and if $\pi_{k_{n-1}}$ is $(n+1)$-Bellman-optimal, then π_{k_n} is $(n+2)$-Bellman optimal and $\Pi_{n+1}^* \subseteq \Pi(\mathcal{Z}_n) \subseteq \Pi_n^*$.*

Proof. Assume that $\Pi_n^* \subseteq \Pi(\mathcal{Z}_{n-1}) \subseteq \Pi_{n-1}^*$ and that $\pi_{k_{n-1}}$ is $(n+1)$-Bellman optimal.

We start by proving that

$$\pi_{k_n} \text{ is } (n+2)\text{-Bellman optimal.} \tag{C.8}$$

First of all, we claim that π_k is n-optimal for all $k \ge k_{n-1}$. This is shown by induction on $k \ge k_{n-1}$. By assumption, we know that π_k is n-optimal for $k = k_{n-1}$. For the other inductive case, assume that π_k is n-optimal. For any state s, π_{k+1} satisfies $\Delta_{n+1}^{\pi_k}(s, \pi_{k+1}(s)) \le 0$ and $\Delta_n^{\pi_k}(s, \pi_{k+1}(s)) = \Delta_n^{\pi_{k_{n-1}}}(s, \pi_{k+1}(s)) = 0$ by definition of $\mathcal{Z}_{n-1}$ and by induction hypothesis. So, invoking Lemma C.1 up to order $n-1$, we see that π_{k+1} has better bias $y_m^{\pi_{k+1}}$ than π_k for $m = -1, 0, \dots, n$. As π_k is n-optimal, we conclude that π_{k+1} is n-optimal as well.

In particular, π_{k_n} is n-optimal, so $(y_m^{\pi_{k_n}})_{m\le n} = (y_m^{\pi_{k_{n-1}}})_{m\le n}$, and $(\Delta_m^{\pi_{k_n}})_{m\le n} = (\Delta_m^{\pi_{k_{n-1}}})_{m\le n}$. Because $\pi_{k_{n-1}}$ is $(n+1)$-Bellman optimal, it follows that π_{k_n} is $(n+1)$-Bellman optimal. To see that it is $(n+2)$-Bellman optimal, pick $z \in \mathcal{Z}$ such that $\Delta_m^{\pi_{k_n}}(z) = 0$ for $m = -1, 0, \dots, n+1$. Because π_{k_n} and $\pi_{k_{n-1}}$ have the same gaps up to order $m = n$, we have $z \in \mathcal{Z}_{n-1}$. Then, by definition of $\mathcal{Z}_n$ (see Algorithm 2), every $z' \in \mathcal{Z}_{n-1}$ either satisfies (1) $\Delta_{n+1}^{\pi_{k_n}}(z') \ge 0$, or (2) $\Delta_{n+1}^{\pi_{k_n}}(z') = 0$ and $\Delta_{n+2}^{\pi_{k_n}}(z') \ge 0$. By choice of z, (1) doesn't hold; so (2) does for $z' \equiv z$. So by Definition 1, π_{k_n} is $(n+2)$-Bellman optimal, proving (C.8).

Now, we prove that $\Pi_{n+1}^* \subseteq \Pi(\mathcal{Z}_n)$. Let $\pi \in \Pi_{n+1}^*$. In particular, we have $\pi \in \Pi_n^*$, so $\pi \in \Pi(\mathcal{Z}_{n-1})$ by assumption. Recall that $(n+2)$-Bellman optimal policies are $(n+1)$-optimal (Proposition C.2), so that $y_{n+1}^\pi = y_{n+1}^{\pi_{k_n}}$. Therefore, for any state $s \in \mathcal{S}$, $0 = \Delta_{n+1}^\pi(s, \pi(s)) = \Delta_{n+1}^{\pi_{k_n}}(s, \pi(s))$. Hence $\pi \in \Pi^*(\mathcal{Z}_n)$.

Lastly, we prove that $\Pi(\mathcal{Z}_n) \subseteq \Pi_n^*$. Let $\pi \in \Pi(\mathcal{Z}_n)$. Since $\mathcal{Z}_n \subseteq \mathcal{Z}_{n-1}$, $\pi \in \Pi(\mathcal{Z}_{n-1})$. So π is $(n-1)$-optimal by assumption. Furthermore, by defintion of $\mathcal{Z}_n$, we have $\Delta_{n+1}^{\pi_{k_n}}(s, \pi(s)) = 0$, $\Delta_n^{\pi_{k_n}}(s, \pi(s)) = 0$ for every state $s \in \mathcal{S}$. So, by invoking Lemma C.1 at order $n-1$, we obtain $y_n^\pi \ge y_n^{\pi_{k_n}} = y_n^*$. Hence $\pi \in \Pi_n^*$. □

Corollary C.6 (Correctness of `HOPI`)**.** *For any $n \ge -2$, $\Pi_{n+1}^* \subseteq \Pi(\mathcal{Z}_n) \subseteq \Pi_n^*$.*

Proof. The proof goes by induction on $n = -2, -1, \dots$.

At $n = -2$, for all state $s \in \mathcal{S}$, we have $\mathcal{Z}_{-2}(s) = \mathcal{Z}$, and therefore $\Pi = \Pi(\mathcal{Z}_{-2})$. Moreover, $\Pi_{-2}^* = \Pi$, so therefore we have $\Pi_{-1}^* \subseteq \Pi(\mathcal{Z}_{-2}) \subseteq \Pi_{-2}^*$. The fact that π_{k-2} is 0-Bellman optimal is a consequence of well-known results

Puterman (1994), because the first loop of $\texttt{HOPI}(n, 0)$ consists in running the vanilla (first order) Policy Iteration algorithm, that returns a policy satisfying the Bellman equations; hence a 0-Bellman optimal policies.

Therefore, we can propagate the property to larger n using Lemma C.5. □

With the previous result, we have the correctness of the $\texttt{HOPI}(n, 0)$ algorithm.

We now have to look into its approximate version, that replace the use of the raw argmax$(-)$ operator by a soft argmax during the execution of the algorithm.

C.4 HOPI Running with Non-Trivial Soft Argmax

The point of this section is to prove that, when both $\varepsilon > 0$ and $d_\infty(M, M')$ are small enough, then $\texttt{HOPI}(n, 0, M) = \texttt{HOPI}(n, \varepsilon, M')$. As our main result, we show that when ε and $d_\infty(M, M')$ are well-tuned with respect to each other, every step that $\texttt{HOPI}(n, \varepsilon, M')$ takes during its execution is the same as $\texttt{HOPI}(n, 0, M)$.

Given a set $\mathcal{X} \subseteq \mathbb{R}$, introduce

$$\operatorname{dgap}(\mathcal{X}) := \min(\{|x - y| : x \in \mathcal{X}, y \in \mathcal{X}, x \neq y\}) \tag{C.9}$$

the minimal gap between two distints elements of $\mathcal{X}$.

Lemma C.7. *Let $M \equiv (\mathcal{Z}, p, r)$ a communicating Markov decision process. Fix $n \geq 0$, a policy $\pi \in \Pi$, a state $s \in \mathcal{S}$ and a set of actions $\mathcal{A}^*(s) \subseteq \mathcal{A}(s)$. Denote $\alpha_n(M) > 0$ defined in Lemma B.11 and $D_*(M)$ the worst-diameter of M (Definition B.2). For all $\varepsilon > 0$ and instance $\hat{M} \equiv (\mathcal{Z}, \hat{p}, \hat{r})$ that satisfy*

$$d_\infty(M, \hat{M}) \leq \min\left\{ \frac{1}{D^*(M)}, \frac{\varepsilon}{2\alpha_n(M)}, \frac{\operatorname{dgap}(\{\Delta_n^\pi(s, a) : a \in \mathcal{A}^*(s)\}) - \varepsilon}{2\alpha_n(M)} \right\}, \tag{C.10}$$

we have

$$\operatorname*{argmax}_{a \in \mathcal{A}^*(s)} \{r_n(s, a) + p(s, a) y_n^\pi\} = \operatorname*{soft}_{\varepsilon} \operatorname*{argmax}_{a \in \mathcal{A}^*(s)} \{\hat{r}_n(s, a) + \hat{p}(s, a) \hat{y}_n^\pi\}.$$

Proof. We proceed by double inclusion.

Let $a \in \operatorname{argmax}_{a \in \mathcal{A}^*(s)} \{r_n(s, a) + p(s, a) y_n^\pi\}$. We have

$$\begin{aligned}
\hat{r}_n(s, a) + \hat{p}(s, a) \hat{y}_n^\pi &\overset{\circ 1}{\geq} r_n(s, a) + p(s, a) y_n^\pi - \alpha_n(M) d_\infty(M, \hat{M}) \\
&= \max_{a' \in \mathcal{A}^*(s)} \{r_n(s, a') + p(s, a') y_n^\pi\} - \alpha_n(M) d_\infty(M, \hat{M}) \\
&\overset{\circ 1}{\geq} \max_{a' \in \mathcal{A}^*(s)} \{\hat{r}_n(s, a') + \hat{p}(s, a') \hat{y}_n^\pi\} - 2\alpha_n(M) d_\infty(M, \hat{M})
\end{aligned}$$

where both instances of ($\circ$1) stem from Lemma B.11, since $d_\infty(M, \hat{M}) \leq 1/D^*(M)$ by Equation (C.10). We deduce that, since $2\alpha_n(M) d_\infty(M, \hat{M}) \leq \varepsilon$ by Equation (C.10) again, $a \in \operatorname{soft}_\varepsilon \operatorname{argmax}_{a \in \mathcal{A}^*(s)} \{\hat{r}_n(s, a') + \hat{p}(s, a') \hat{y}_n^\pi\}$.

Conversely, fix $a \in \operatorname{soft}_\varepsilon \operatorname{argmax}_{a \in \mathcal{A}^*(s)} \{\hat{r}_n(s, a') + \hat{p}(s, a') \hat{y}_n^\pi\}$. Similarly,

$$r_n(s, a) + p_n(s, a) y_n^\pi \geq \max_{a' \in \mathcal{A}^*(s)} \{r_n(s, a') + p(s, a') y_n^\pi\} - 2\alpha_n(M) d_\infty(M, \hat{M}) - \varepsilon$$

Since $2\alpha_n(M) d_\infty(M, \hat{M}) \leq \operatorname{dgap}(\{\Delta_n^\pi(s, a') : a' \in \mathcal{A}^*(s)\})$ by Equation (C.10) always, we conclude that $a \in \operatorname{argmax}_{a \in \mathcal{A}^*(s)} \{r_n(s, a) + p(s, a) y_n^\pi\}$, which concludes the proof. □

Strong of Lemma C.7, we may conclude that $\texttt{HOPI}(n, 0, M) = \texttt{HOPI}(n, \varepsilon, M')$ under the right tuning of ε and $d_\infty(M, M')$.

Corollary C.8 (Bissimulation)**.** *Let $M \equiv (\mathcal{Z}, p, r)$ a communicating Markov decision process and fix $n \geq 0$. Denote $\alpha_n(M) > 0$ defined in Lemma B.11 and $D_*(M)$ the worst-diameter of M (Definition B.2). For all $\varepsilon > 0$ and instance $M' \equiv (\mathcal{Z}, p', r')$ that satisfy*

$$d_\infty(M, \hat{M}) \leq \min\left\{ \frac{1}{D^*(M)}, \frac{\varepsilon}{2\alpha_n(M)}, \min_{\pi \in \Pi} \min_{m \leq n+2} \min_{s \in \mathcal{S}} \frac{\operatorname{dgap}(\{\Delta_m^\pi(s, a) : a \in \mathcal{A}(s)\}) - \varepsilon}{2\alpha_m(M)} \right\}, \tag{C.11}$$

we have $\texttt{HOPI}(n,0,M) = \texttt{HOPI}(n,\varepsilon,M')$.

Proof. This is essentially a bissimulation argument. Let (π_k) and (π'_k) the respective sequences of policies computed by $\texttt{HOPI}(n,0,M)$ and $\texttt{HOPI}(n,\varepsilon,M')$. Let $\mathcal{Z}_m, k_m$ (resp. $\mathcal{Z}'_m, k'_m$) the respective sequences of pairs and phases starting times of $\texttt{HOPI}(n,0,M)$ (resp. of $\texttt{HOPI}(n,0,M')$). By induction on $k \geq 1$, we show that $\pi_k = \pi'_k$, so that $\mathcal{Z}_m = \mathcal{Z}'_m$ and $k_m = k'_m$; In other words, both algorithms takes the same decisions at every point.

Indeed, assume that $\pi'_k = \pi_k$, and that both $\texttt{HOPI}(n,0,M)$ and $\texttt{HOPI}(n,\varepsilon,M')$ have the same pair mask $\mathcal{Z}_0$ and are at the same state $m_0 \leq n$. Write $\mathcal{A}_0(s) := \{a \in \mathcal{A}(s) : (s,a) \in \mathcal{Z}_0\}$. Because Equation (C.11) holds, by Lemma C.7, we have

$$\forall s \in \mathcal{S}, \quad \operatorname*{argmax}_{a\in\mathcal{A}_0(s)}\big\{r_{m_0}(s,a) + p(s,a)y^{\pi_k}_{m_0}\big\} = \operatorname*{soft}_{\varepsilon}\operatorname*{argmax}_{a\in\mathcal{A}_0(s)}\big\{r'_{m_0}(s,a) + p'(s,a)y'^{\pi_k}_{m_0}\big\}.$$

Therefore, the current policy will be improved at the first stage during the execution of $\texttt{HOPI}(n,0,M)$ if, and only if it is improved so for $\texttt{HOPI}(n,\varepsilon,M')$, and $\pi_{k+1} = \pi'_{k+1}$. If the policy enters the second stage of policy improvement, we conclude that $\pi_k = \pi'_k$ with the same arguments. □

C.5 The Learning Variant of HOPI: HOPE

Our algorithm HOPE (Algorithm 1) feeds the empirical model $\hat{M}_t$ and a precision $\varepsilon_t \equiv t^{-1/4}$ to HOPI, and outputs a policy generated by $\texttt{HOPI}(n,\varepsilon_t,\hat{M}_t)$ without modification. So, it remains to show that, at some point, HOPE does yield small enough ε and a good enough approximation of the underlying instance M so that Lemma C.7 can be applied. As Corollary C.8 foreshadows, the empirical MDP must converge to the real one faster than ε_t goes to 0.

In other words, we prove Theorem 1 from the main text, restated below for convenience.

Theorem 1. *For all $n \geq -1$, $\texttt{HOPE}(n)$ (Algorithm 1) is consistent for Π^*_n.*

Proof of Theorem 1. By definition of consistency, we want to show that, for each model M, $\mathbb{P}(\hat{\pi}_t \notin \Pi^*_n(M)) = \mathrm{o}(1)$. For that, we fix a Markov decision process M and, for any $\delta > 0$, we will show the existence of a t_δ satisfying

$$\mathbb{P}(\exists t \geq t_\delta, \hat{\pi}_t \notin \Pi^*_n(M)) \leq \delta.$$

To achieve the above, the point is to invoke Corollary C.8, to state that $\texttt{HOPI}(n,0,M)$ and $\texttt{HOPI}(n,\varepsilon_t,\hat{M}_t)$ return the same thing, and conclude by correctness of $\texttt{HOPI}(n,0,M)$, see Corollary C.6. We want t_δ to be such that, $\varepsilon_{t_\delta} \leq \min_{m\leq n+2} \frac{1}{2}\mathrm{dgap}(m;M)$, where

$$\mathrm{dgap}(m;M) := \min_{k\leq m}\min_{\pi\in\Pi}\mathrm{dgap}(\{\Delta^\pi_k(z); z \in \mathcal{Z}\}) \tag{C.12}$$

This is possible, since $\varepsilon_t \equiv t^{-1/4}$ is decreasing in t: it is satisfied for $t_\delta \geq t_0 := \left(\frac{2}{\min_{m\leq n+2}\mathrm{dgap}(m;M)}\right)^4$.

We also want t_δ to be such that $d_\infty(M,\hat{M}_t)$ is smaller than $D_*(M)^{-1}$ and ε_t with high probability when $t \geq t_\delta$. For any $t_1 \geq 1$,

$$\begin{aligned}
&\mathbb{P}\left(\exists t \geq t_\delta, d_\infty(M,\hat{M}_t) > \min\left\{\frac{1}{D_*(M)}, \varepsilon_t\right\}\right)\\
&\quad\leq \mathbb{P}\left(\exists t \geq t_1, \exists z \in \mathcal{Z}, N_t(z) < \beta t\right) + \mathbb{P}\left(\begin{array}{c}\exists t \geq t_1, d_\infty(M,\hat{M}_t) > \min\{D^*(M)^{-1}, \varepsilon_t\}\\ \text{and } \forall t \geq t_1, \forall z \in \mathcal{Z}, N_t(z) \geq \beta t\end{array}\right)\\
&\quad\overset{\circ 1}{\leq} \exp(-\gamma t_1 + \lambda) + \mathbb{P}\left(\begin{array}{c}\exists t \geq t_1, d_\infty(M,\hat{M}_t) > \min\{D^*(M)^{-1}, \varepsilon_t\}\\ \text{and } \forall t \geq t_1, \forall z \in \mathcal{Z}, N_t(z) \geq \beta t\end{array}\right)
\end{aligned}$$

where ($\circ$1) comes from Lemma B.6. Suppose now that t_1 is such that, for $t \geq t_1$,

$$\min\left\{\frac{1}{D_*(M)}, \varepsilon_t\right\} \geq \sqrt{\frac{|\mathcal{S}|}{\beta t}\log\left(\frac{8|\mathcal{Z}|\sqrt{1+\beta t}}{\delta}\right)}$$

(there exists such a t_1 since ε_t is of the form $\varepsilon_t = \frac{1}{t^y}$ with $y < 1/2$). Then,

$$\begin{aligned}
&\mathbb{P}\Big(\exists t \geq t_1, \|M - \hat{M}_t\| > \min\{D^*(M)^{-1}, \varepsilon_t\}\Big) \\
&\quad \leq \exp(-\gamma t_1 + \lambda) + \mathbb{P}\left(\exists t \geq t_1, \|M - \hat{M}_t\| > \sqrt{\frac{|\mathcal{S}|}{\beta t}\log\left(\frac{8|\mathcal{Z}|\sqrt{1+\beta t}}{\delta}\right)} \cap \forall t \geq t_1, \min(N_t) \geq \beta t\right) \\
&\quad \overset{\circ 1}{\leq} \exp(-\gamma t_1 + \lambda) + \mathbb{P}\left(\exists t \geq t_1, \|M - \hat{M}_t\| > \sqrt{\frac{|\mathcal{S}|}{\beta t}\log\left(\frac{8|\mathcal{S}||\mathcal{A}|\sqrt{1+\min(N_t)}}{\delta}\right)} \cap \forall t \geq t_1, \min(N_t) \geq \beta t\right) \\
&\quad \overset{\circ 2}{\leq} \exp(-\gamma t_1 + \lambda) + \frac{\delta}{2}
\end{aligned}$$

where ($\circ$1) is due to $x > 0 \mapsto \frac{\log y\sqrt{1+x}}{x}$ being decreasing when $y > 1$, and ($\circ$2) invokes Corollary B.4.

Defining $t_2 := \left\lceil \frac{1}{\gamma}\left(\ln\frac{\delta}{2} - \lambda\right)\right\rceil$, we thus have that when $t_\delta \geq \max\{t_1, t_2\}$,

$$\mathbb{P}\left(\exists t \geq t_\delta, d_\infty(M, \hat{M}_t) > \min\left\{\frac{1}{D_*(M)}, \varepsilon_t\right\}\right) \leq \delta.$$

Let us put it all together now. With $t_\delta \geq \max\{t_0, t_1, t_2\}$,

$$\begin{aligned}
&\mathbb{P}(\exists t \geq t_\delta, \hat{\pi}_t \notin \Pi_n^*(M)) \\
&\quad \overset{\circ 1}{\leq} \mathbb{P}\left(\exists t \geq t_\delta, \exists m \leq n+2, d_\infty(M, \hat{M}_t) > \min\left\{\frac{1}{D_*(M)}, \frac{\varepsilon_t}{2\alpha_m(M)}, \frac{\mathrm{dgap}(m; M) - \varepsilon_t}{2\alpha_m(M)}\right\}\right) \\
&\quad \leq \mathbb{P}\left(\exists t \geq t_\delta, \exists m \leq n+2, d_\infty(M, \hat{M}_t) > \min\left\{\frac{1}{D_*(M)}, \frac{\varepsilon_t}{2\alpha_m(M)}\right\} \cup \varepsilon_t > \frac{\mathrm{dgap}(m; M)}{2}\right) \leq \delta
\end{aligned}$$

where ($\circ$1) stems from Corollary C.6, the correctness of $\texttt{HOPI}(n, 0, M)$, and Equation (C.12). □

D DEGENERATE MDPS AND PROOFS OF SECTION 4

D.1 Proofs of Section 4.1

In this paragraph, we prove Proposition 4 from the main text, of which we recall the statement below.

Proposition 4. *Let $\Pi^* : \mathcal{M} \to 2^\Pi$ be a measurable optimality map. If $M \in \mathcal{M}$ is non-degenerate, then there is a policy $\pi \in \Pi^*(M)$ that remains optimal in a neighbourhood of M, i.e.,*

$$\exists \varepsilon > 0, \quad \bigcap\{\Pi^*(M') : d_\infty(M, M') < \varepsilon\} \neq \emptyset.$$

Proof of Proposition 4. Let $M \in \mathcal{M}$ be a non-degenerate model. By definition of non-degeneracy (Definition 4), there exists an identification algorithm $(\Lambda, \hat{\pi})$ that is consistent and such that, for all $\delta > 0$, there is a stopping time τ_δ making $(\Lambda, \hat{\pi}, \tau_\delta)$ δ-PC and such that $\mathbb{P}^{M,\Lambda}(\tau_\delta < \infty) = 1$. Fix $M' \in \mathcal{M}$ with $d_\varepsilon(M, M') < \varepsilon$ for some $\varepsilon > 0$ to be specified later. Given $M'' \in \{M, M'\}$, the probability operator $\mathbb{P}^{M'',\Lambda}(-)$ on $\mathcal{H}$ can be restricted to the sub-space $\mathcal{H}_T$ of possible histories of length at time T; It is written $\mathbb{P}_T^{M'',\Lambda}(-)$. Now, for all $T \geq 1$, we have

$$\begin{aligned}
\mathbb{P}^{M,\Lambda}(\hat{\pi}_{\tau_\delta} \in \Pi^*(M')) &\overset{\circ 1}{=} \sup_{T \geq 1}\left\{\mathbb{P}^{M,\Lambda}(\hat{\pi}_{\tau_\delta} \in \Pi^*(M') \text{ and } \tau_\delta \leq T)\right\} \\
&\overset{\circ 2}{=} \sup_{T \geq 1}\left\{\mathbb{P}_T^{M,\Lambda}(\hat{\pi}_{\tau_\delta} \in \Pi^*(M') \text{ and } \tau_\delta \leq T)\right\} \\
&\overset{\circ 3}{\geq} \sup_{T \geq 1}\left\{\mathbb{P}_T^{M',\Lambda}(\hat{\pi}_{\tau_\delta} \in \Pi^*(M') \text{ and } \tau_\delta \leq T) - d_{\mathrm{TV}}\left(\mathbb{P}_T^{M,\Lambda}, \mathbb{P}_T^{M',\Lambda}\right)\right\} \\
&\overset{\circ 4}{\geq} \sup_{T \geq 1}\left\{1 - \delta - \mathbb{P}^{M',\Lambda}(\tau_\delta > T) - d_{\mathrm{TV}}\left(\mathbb{P}_T^{M,\Lambda}, \mathbb{P}_T^{M',\Lambda}\right)\right\}
\end{aligned}$$

$$\overset{\circ 5}{\geq} 1-\delta-\inf_{T\geq 1}\left\{\mathbb{P}^{M,\Lambda}(\tau_\delta>T)+2d_{\mathrm{TV}}\left(\mathbb{P}_T^{M,\Lambda},\mathbb{P}_T^{M',\Lambda}\right)\right\}$$

where ($\circ$1) uses that $(\hat{\pi}_{\tau_\delta}\in\Pi^*(M'))$ is the increasing union of $(\hat{\pi}_{\tau_\delta}\in\Pi^*(M')$ and $\tau_\delta\leq T)$; ($\circ$2) uses that the event $(\hat{\pi}_{\tau_\delta}\in\Pi^*(M')$ and $\tau_\delta\leq T)$ is $\sigma(H_T)$-measurable; ($\circ$3) follows by definition of the total-variation distance; ($\circ$4) follows by definition of the δ-PC property; ($\circ$5) follows by definition of the total-variation distance again.

Fix $\delta,x>0$ with $|\Pi|(\delta+2x)<1$, e.g., $\delta<|\Pi|^{-1}$ and $x<\frac{1}{2}(|\Pi|^{-1}-\delta)$.

Because $\mathbb{P}^{M,\Lambda}(\tau_\delta<\infty)=1$, there exists $T\geq 0$ such that $\mathbb{P}^{M,\Lambda}(\tau_\delta>T)\leq x$. By induction on $T\geq 1$, we find that

$$d_{\mathrm{TV}}\left(\mathbb{P}_T^{M,\Lambda},\mathbb{P}_T^{M',\Lambda}\right)\leq Td_\infty(M,M').$$

So, for $d_\infty(M,M')<\frac{x}{2T}=:\varepsilon$, we get $\mathbb{P}^{M,\Lambda}(\hat{\pi}_{\tau_\delta}\in\Pi^*(M'))\geq 1-\delta-2x$. So

$$\mathbb{P}^{M,\Lambda}(\hat{\pi}_{\tau_\delta}\notin\Pi^*(M'))\leq\delta+2x. \tag{D.1}$$

Now, let $\Pi_\varepsilon:=\bigcap\Pi^*(M'):d_\infty(M,M')<\varepsilon$ be the set of policies that are universally optimal in the ε-neighborhood of M. For every policy $\pi\notin\Pi_\varepsilon$, there is a model $M'_\pi\in\mathcal{M}$ with $d_\infty(M,M'_\varepsilon)<\varepsilon$ that rejects π in the sense that $\pi\notin\Pi^*(M'_\varepsilon)$. Following from this observation, we obtain

$$\begin{aligned}\mathbb{P}^{M,\Lambda}(\hat{\pi}_{\tau_\delta}\in\Pi_\varepsilon)&=\mathbb{P}^{M,\Lambda}(\forall\pi\notin\Pi_\varepsilon,\hat{\pi}_{\tau_\delta}\in\Pi^*(M'_\pi))\\&=1-\mathbb{P}^{M,\Lambda}(\exists\pi\notin\Pi_\varepsilon,\hat{\pi}_{\tau_\delta}\notin\Pi^*(M'_\pi))\\&\geq 1-\sum_{\pi\notin\Pi_\varepsilon}\mathbb{P}^{M,\Lambda}(\hat{\pi}_{\tau_\delta}\notin\Pi^*(M'_\pi))\\&\overset{\circ 1}{\geq}1-|\Pi\setminus\Pi_\varepsilon|(\delta+2x)\geq 1-|\Pi|(\delta+2x)\overset{\circ 2}{>}0\end{aligned}$$

where ($\circ$1) follows from (D.1); and ($\circ$2) follows by definition of δ and x. We conclude that $\Pi_\varepsilon\neq\emptyset$. □

D.2 Proofs of Section 4.2

In this paragraph, we prove Proposition 5 and Lemma 6 from the main text (both re-stated below). Recall that the point of Lemma 6 is only to serve in the proof of Proposition 5.

Lemma 6. *Let $M\equiv(\mathcal{Z},\nu,p)$ be a communicating instance. For every unichain Bellman optimal policy $\pi\in\Pi^{\mathrm{B}}(M)$ and $\varepsilon>0$, there exists $M'\equiv(\mathcal{Z},r',p)$ with $d_\infty(M,M')<\varepsilon$ for which*

(1) *π is the unique Bellman optimal policy;*
(2) *$g^*(M')=g^*(M)$;*
(3) *$b^*(M')=b^*(M)$.*

Proof of Lemma 6. Let $\pi\in\Pi$ be a Bellman optimal policy of M and $\varepsilon>0$ be a precision parameter. Consider the Markov decision process $M_\varepsilon\equiv(\mathcal{Z},r_\varepsilon,p)$ with mean reward function $r_\varepsilon(z):=\varepsilon+(1-2\varepsilon)r(z)$. Accordingly, we have $\varepsilon\leq r_\varepsilon(z)\leq 1-\varepsilon$ for all $z\in\mathcal{Z}$. The policy π is still Bellman optimal in M_ε and we have $d_\infty(M,M_\varepsilon)\leq\varepsilon$.

Consider the Markov decision process $M'_\varepsilon\equiv(\mathcal{Z},r'_\varepsilon,p)$ with mean reward function given by

$$r'_\varepsilon(s,a):=r_\varepsilon(s,a)-\varepsilon\mathbb{1}(a\neq\pi(s)).$$

Note that $g^\pi(M'_\varepsilon)=g^\pi(M_\varepsilon)$ and $b^\pi(M'_\varepsilon)=b^\pi(M_\varepsilon)$ because the reward and kernel functions of π are the same in M_ε and M'_ε. Now, for $a\neq\pi(s)$, we have

$$\begin{aligned}r'_\varepsilon(s,a)+p(s,a)b^\pi(M'_\varepsilon)&\overset{\circ 1}{=}r_\varepsilon(s,a)+p(s,a)b^\pi(M_\varepsilon)-\varepsilon\\&\overset{\circ 2}{\leq}g^\pi(s;M_\varepsilon)+b^\pi(s;M_\varepsilon)-\varepsilon\overset{\circ 3}{=}g^\pi(M'_\varepsilon)+b^\pi(M'_\varepsilon)-\varepsilon\end{aligned}$$

where ($\circ$1) follows by definition of r'_ε and $b^\pi(M'_\varepsilon)=b^\pi(M_\varepsilon)$; ($\circ$2) holds because π is bias optimal in M_ε; and ($\circ$3) follows from $g^\pi(M'_\varepsilon)=g^\pi(M_\varepsilon)$ and $b^\pi(M'_\varepsilon)=b^\pi(M_\varepsilon)$. Together with $\mathrm{sp}(g^\pi(M'_\varepsilon))=\mathrm{sp}(g^\pi(M_\varepsilon))=0$, we conclude that π is the unique Bellman optimal policy of M'_ε. By triangular inequality, we further have $d_\infty(M,M'_\varepsilon)\leq 2\varepsilon$, concluding the proof. □

Lemma 6 will be combined to the following well-known policy improvement result.

Lemma D.1 (First order policy improvement; Puterman (1994))**.** *Fix $M \equiv (\mathcal{Z}, r, p)$ a Markov decision process. Let $\pi \in \Pi$ satisfying $\mathrm{sp}(g^\pi) = 0$. If $\pi' \in \Pi$ is such that $\Delta^\pi(s, \pi'(s)) \leq 0$ for all $s \in \mathcal{S}$, then $(g^\pi, b^\pi) \leq_{\mathrm{lex}} (g^{\pi'}, b^{\pi'})$ where $\leq_{\mathrm{lex}}$ is the lexicographic order.*

Proof. The proof is essentially the same as the one of Lemma C.3, by changing $<$ to $\leq$. □

Now, we can prove Proposition 5.

Proposition 5. *Let M be a communicating instance. For every unichain Bellman optimal policy $\pi \in \Pi^{\mathrm{B}}(M)$ and precision $\varepsilon > 0$, there exists $M' \in \mathcal{M}$ with $d_\infty(M, M') < \varepsilon$ for which π is the unique gain optimal policy, i.e., $\Pi^*_{-1}(M') = \{\pi\}$.*

Proof of Proposition 5. Let π be a Bellman optimal policy of M and fix a precision $\varepsilon > 0$. By Lemma 6, there exists $M' \equiv (\mathcal{Z}, r', p)$ such that $d_\infty(M, M') \leq \varepsilon$ where (1) π is the unique Bellman optimal policy, (2) $g^*(M') = g^*(M)$ and (3) $b^*(M') = b^\pi(M)$. Consider the model $M'' \equiv (\mathcal{Z}, r'', p'')$ with reward and kernel functions given by

$$\begin{aligned} r''(s,a) &= r'(s,a) + (p''(s,a) - p'(s,a))b^\pi(M), \\ p''(s'|s,a) &= (1-\varepsilon)p'(s,a) + \frac{\varepsilon}{|\mathcal{A}(s)|}. \end{aligned}$$

Note that $\|p''(z) - p(z)\|_1 \leq |\mathcal{S}|\varepsilon$, so that $\|r'' - r'\|_\infty \leq \frac{1}{2}\mathrm{sp}(b^\pi(M))|\mathcal{S}|\varepsilon$ by Lemma B.5. Further note that M'' is uniformly ergodic, since all transition probabilities are positive.

Consider another policy $\pi' \in \Pi$. Since π' is ergodic in M'', there exists $c(\pi') > 0$ such that every state is visited at least $c(\pi')$ in average in M''. For $s \in \mathcal{S}$, we have

$$\begin{aligned} g^{\pi'}(s; M'') &= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_s^{\pi', M''}\left[\sum_{t=1}^T r''(Z_t)\right] \\ &\overset{\circ 1}{=} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_s^{\pi', M''}\left[\sum_{t=1}^T (r'(Z_t) + (p'(Z_t) - p''(Z_t))b^\pi(M'))\right] \\ &\overset{\circ 2}{=} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_s^{\pi', M''}\left[\sum_{t=1}^T (g^\pi(S_t; M') + (e_{S_t} - p''(Z_t))b^\pi(M') - \Delta^\pi(Z_t; M'))\right] \\ &\overset{\circ 3}{=} g^\pi(s; M') + \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_s^{\pi', M''}\left[b^\pi(s; M') - b^\pi(S_{T+1}; M') - \sum_{t=1}^T \Delta^\pi(Z_t; M')\right] \\ &\overset{\circ 4}{\leq} g^\pi(s; M') - c(\pi') \max_{s\in\mathcal{S}} \Delta^*(s, \pi'(s); M') \end{aligned}$$

where ($\circ$1) unfolds the definition of r'' and uses that $b^\pi(M) = b^\pi(M')$; ($\circ$2) follows by definition of the gaps of π, i.e., $\Delta^\pi(s,a) = g^\pi(s) + b^\pi(s) - r(s,a) - p(s,a)b^\pi$; ($\circ$3) rewrites the telescopic sum $\mathbb{E}[\sum_{t=1}^T (e_{S_t} - p''(Z_t))b^\pi(M')]$; and ($\circ$4) follows by definition of $c(\pi)$ and uses that $\Delta^\pi(M') = \Delta^*(M')$ that holds because π is Bellman optimal in M'.

From ($\circ$3), note that for $\pi = \pi'$, we obtain

$$g^\pi(M'') = g^\pi(M')$$

since pairs played by π satisfy $\Delta^\pi(z; M') = 0$ by construction of M'.

Let $\mathrm{dmin}(\Delta^*(M')) := \min\{\Delta^*(z; M') : z \in \mathcal{Z} \text{ and } \Delta^*(z; M') > 0\}$ the minimum positive gap of M'. Because π is the unique Bellman optimal policy in M', for every $\pi' \neq \pi$, there is $s \in \mathcal{S}$ such that $\Delta^\pi(s, \pi'(s); M') > 0$—Indeed, if, ad absurdum $\Delta^\pi(s, \pi'(s); M') = 0$ for all $s \in \mathcal{S}$, then because π is the unique Bellman optimal policy, it is bias optimal. So, by Lemma D.1, we see that π' is bias optimal as well, so is Bellman optimal in particular; A contraction.

Continuing, for every policy $\pi' \neq \pi$, we get

$$g^{\pi'}(M'') \leq g^{\pi}(M) - \min_{\pi''} c(\pi'')\mathrm{dmin}(\Delta^*) < g^{\pi}(M') = g^{\pi}(M'').$$

Accordingly, π is the unique gain optimal policy of M''. Furthermore, the distance between M'' and M is $\mathrm{O}(\varepsilon)$. This concludes the proof. □

E STOPPING OF HOPE AND PROOFS OF SECTION 5

E.1 Proof of Lemma 9

In this paragraph, we prove Lemma 9, restated below for convenience.

Lemma 9. *Let M be a communicating instance. Then*

(1) *If M has a unique Bellman optimal policy, then this policy is unichain.*
(2) *Conversely, if there exists a unichain policy $\pi \in \Pi$ such that $g^{\pi}(s) + b^{\pi}(s) > r(s,a) + p(s,a)b^{\pi}$ for all $a \neq \pi(s)$ and $s \in \mathcal{S}$, then π is the unique Bellman optimal policy of M.*

Proof of Lemma 9. We prove both directions separately.

The first assertion (1) is proved by contradiction. Assume that M has a unique Bellman optimal policy π that is multi-chain. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two disjoint recurrent classes of states of π. For $i \in \{1,2\}$ and $\varepsilon > 0$, consider the Markov decision processes $M_i^{\varepsilon} \equiv (\mathcal{Z}, p, r_{\varepsilon}^i)$ with reward functions given by $r_i^{\varepsilon}(s,a) = r(s,a) - \varepsilon \mathbb{1}(s \notin \mathcal{S}_i)$. Then, for M_1^{ε} for instance, we see that π does not satisfy $\mathrm{sp}(g^{\pi}(M_1^{\varepsilon})) = 0$—Yet M_1^{ε} is communicating, so $\mathrm{sp}(g^*(M_1^{\varepsilon})) = 0$. So π is sub-optimal in M_1^{ε}. Therefore, M_1^{ε} must admit a Bellman optimal policy π_1^{ε} that is not π. Because Π is finite, we see that there exists $\varepsilon_n \to 0$ and $\pi' \neq \pi$ such that $\pi' \in \Pi_0^{\mathrm{B}}(M_1^{\varepsilon_n})$ for all $n \geq 0$. Accordingly, for all $n \geq 1$ and $s \in \mathcal{S}$,

$$g^{\pi'}(s; M_1^{\varepsilon_n}) + b^{\pi'}(s; M_1^{\varepsilon_n}) = \max_{a \in \mathcal{A}(s)} \left\{ r_1^{\varepsilon_n}(s,a) + p(s,a) b^{\pi'}(M_1^{\varepsilon_n}) \right\} \tag{E.1}$$

with $\mathrm{sp}(g^{\pi'}(M_1^{\varepsilon_n})) = 0$. Because the gain $g^{\pi'}$ and bias functions $b^{\pi'}$ of π' are linear functions of the mean reward vector, so by making $n \to \infty$, we see that (E.1) and $\mathrm{sp}(g^{\pi'}) = 0$ are satisfied for M as well. In other words, π' is Bellman optimal in M; A contradiction.

The second assertion (2) follows from a standard argument that is reminiscent of the analysis of Policy Iteration. Let $\pi \in \Pi$ be a unichain policy such that $\Delta^{\pi}(s,a) := g^{\pi}(s) + b^{\pi}(s) - r(s,a) - p(s,a)b^{\pi} > 0$ for all $a \neq \pi(s)$. Because it is unichain, we have $\mathrm{sp}(g^{\pi}) = 0$ so π is Bellman optimal in M.

Let π' a Bellman optimal policy. We show that $\pi = \pi'$. We have

$$\mathbb{E}_s^{\pi'}[R_1 + \ldots + R_T] = T g^{\pi}(s) + \mathbb{E}_s^{\pi'}[b^{\pi}(s) - b^{\pi}(S_{T+1})] - \mathbb{E}_s^{\pi'}\left[\sum_{t=1}^{T} \Delta^{\pi}(Z_t)\right] \tag{E.2}$$

and since $g^{\pi}(s) = g^{\pi'}(s)$, we conclude that every state $s \in \mathcal{S}$ such that $\Delta^{\pi}(s, \pi'(s)) > 0$ must be transient under π'. So, π' must coincide with π on its recurrent states. In particular, every recurrent class of π' is a recurrent class of π. Because π is unichain, we conclude that π' is unichain as well as both policies have the unique recurrent class of states $\mathcal{S}_0$. In particular, $b^{\pi} = b^{\pi'}$ in $\mathcal{S}_0$. Now,

$$\mathbb{E}_s^{\pi'}[R_1 + \ldots + R_T] = T g^{\pi}(s) + \mathbb{E}_s^{\pi'}\Big[b^{\pi'}(s) - b^{\pi'}(S_{T+1})\Big]. \tag{E.3}$$

Combining Equations (E.2) and (E.3), we find that for all $s \in \mathcal{S}$, we have

$$\begin{aligned} b^{\pi'}(s) - \mathbb{E}_s^{\pi'}\Big[b^{\pi'}(S_{T+1})\Big] &= b^{\pi}(s) - \mathbb{E}_s^{\pi'}[b^{\pi}(S_{T+1})] - \mathbb{E}_s^{\pi'}\left[\sum_{t=1}^{T} \Delta^{\pi}(Z_t)\right] \\ &= b^{\pi}(s) - \mathbb{E}_s^{\pi'}\Big[b^{\pi'}(S_{T+1})\Big] + \mathbb{E}_s^{\pi'}\Big[b^{\pi'}(S_{T+1}) - b^{\pi}(S_{T+1})\Big] - \mathbb{E}_s^{\pi'}\left[\sum_{t=1}^{T} \Delta^{\pi}(Z_t)\right] \end{aligned}$$

$$\overset{\circ 1}{\leq} b^{\pi}(s) - \mathbb{E}_s^{\pi'}\Big[b^{\pi'}(S_{T+1})\Big] + \Big\|b^{\pi'} - b^{\pi}\Big\|_{\infty} \mathbb{P}_s^{\pi'}(S_{T+1} \notin \mathcal{S}_0) - \mathbb{E}_s^{\pi'}\left[\sum_{t=1}^{T} \Delta^{\pi}(Z_t)\right]$$
$$= b^{\pi}(s) - \mathbb{E}_s^{\pi'}\Big[b^{\pi'}(S_{T+1})\Big] - \mathbb{E}_s^{\pi'}\left[\sum_{t=1}^{T} \Delta^{\pi}(Z_t)\right] + \mathrm{o}(1)$$

where ($\circ$1) follows from $b^{\pi} = b^{\pi'}(s)$. We conclude that $b^{\pi'}(s) \leq b^{\pi}(s)$ with strict inequality if $\pi(s) \neq \pi'(s)$. Yet, symmetrically, we find that $b^{\pi} \leq b^{\pi'}$. So $b^{\pi} = b^{\pi'}$ and $\pi = \pi'$ necessarily. □

E.2 Proof of Section 5.2

In this paragraph, we prove Lemma 10 and Proposition 11 from the main text.

Lemma 10. *Let $M \equiv (\mathcal{Z}, \nu, p)$ be a communicating instance with unique Bellman optimal policy π^*. For every $M' \equiv (\mathcal{Z}, \nu', p')$ such that $\mathrm{supp}(p') \supseteq \mathrm{supp}(p)$, if*

$$d_{\infty}(M, M') < \beta(M)$$

then π^* is the unique Bellman optimal policy of M'.

Proof of Lemma 10. Since $\mathrm{supp}(p') \supseteq \mathrm{supp}(p)$, π^* is unichain in M' as well. So, by Lemma 9, we see that $\Pi_0^{\mathrm{B}}(M') = \{\pi^*\}$ if, and only if $\mathrm{supp}(\Delta^{\pi^*}(M')) = \mathrm{supp}(\Delta^{\pi^*}(M))$. Because $\Delta^{\pi^*}(s, a; M) > 0$ for all $a \neq \pi^*(s)$, we deduce that $\Pi_0^{\mathrm{B}}(M') = \{\pi^*\}$ holds in particular when

$$\Big\|\Delta^{\pi^*}(M') - \Delta^{\pi^*}(M)\Big\|_{\infty} < \mathrm{dmin}(\Delta^*(M)).$$

We conclude with the explicit bound on the variations of the gap function of Lemma B.10. □

Proposition 11. *Upon running with the stopping rule*

$$\tau_{\delta} := \inf\Big\{t \geq 1 : \xi_{\delta}(t) \leq \beta(\hat{M}_t) \text{ and } \Pi^{\mathrm{B}}(\hat{M}_t) = \{\hat{\pi}_t\}\Big\}$$

where $\xi_{\delta}(t)$ and $\beta(t)$ are respectively given by Equations (3) and (4), `HOPE` is δ-PC for Π_n^*, whatever $n \geq 1$. Moreover, for every instance M with unique Bellman optimal policy, we have $\mathbb{P}^{M,\texttt{HOPE}}(\tau_{\delta} < \infty) = 1$.

Proof of Proposition 11. By Corollary B.4, we know that $\mathbb{P}^{M,\Lambda}(\exists t \geq 1, d_{\infty}(\hat{M}_t, M) \geq \xi_{\delta}(t)) \leq \delta$. Combined with Lemma 10, we know that if $\pi_t^* = \hat{\pi}_t$ is the unique Bellman optimal policy of $\hat{M}_t$, then every model $M' \gg \hat{M}_t$ that satisfies $d_{\infty}(\hat{M}_t, M') \leq \xi_{\delta}(t)$ further satisfies $\Pi_0^{\mathrm{B}}(M') = \{\pi_t^*\}$. We deduce that

$$\mathbb{P}^{M,\Lambda}(\tau_{\delta} < \infty \text{ and } \Pi^*(M) \neq \{\hat{\pi}_{\tau_{\delta}}\}) \leq \delta.$$

In particular, the algorithm is δ-PC.

To conclude, we show that $\mathbb{P}^{M,\Lambda}(\tau_{\delta} < \infty) = 1$. Pick M a model with a unique Bellman optimal policy π^*. By consistency, we have

$$\mathbb{P}^{M,\Lambda}(\hat{\pi}_t = \pi^*) = 1 + \mathrm{o}(1) \tag{E.4}$$

when $t \to \infty$. As actions are played uniformly, we have $N_t(z) \to \infty$ a.s. So $\hat{M}_t \to M$ a.s. and together with the deviations results of Lemmas B.7 to B.10, we see that $\beta(\hat{M}_t) \to \beta(M)$ a.s. Combined with Equation (E.4), we obtain in particular

$$\mathbb{P}^{M,\Lambda}\left(\hat{\pi}_t = \pi_t^* = \pi^* \text{ and } \beta(\hat{M}_t) \geq \frac{1}{2}\beta(M)\right) = 1 + \mathrm{o}(1) \tag{E.5}$$

when $t \to \infty$. Moreover, we know that there exists $\alpha > 0$ such that $\mathbb{P}(\lim_{t\to\infty} \frac{1}{t} N_t(z) > \alpha) = 1$. It follows that $\xi_{\delta}(t) \to 0$ a.s. In tandem with Equation (E.5), we conclude that $\mathbb{P}^{M,\Lambda}(\hat{\pi}_t = \pi_t^* = \pi^*, \beta(\hat{M}_t) \geq \frac{1}{2}\beta(M)$ and $\xi_{\delta}(t) \leq \frac{1}{2}\beta(M)) = 1 + \mathrm{o}(1)$. Hence $\mathbb{P}^{M,\Lambda}(\tau_{\delta} < \infty) = 1$. □