SWIFTLY GENERATING WHITTLE INDICES IN CHANGING ENVIRONMENTS: AN ENVIRONMENT PARAMETER-DRIVEN ACTOR-CRITIC APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Whittle index provides a low-complexity and adaptive solution to large-scale restless multi-armed bandit (RMAB) problems. However, in practice, when the environment faced by each arm constantly changes, the Whittle index has to be repeatedly recomputed, which incurs exceedingly high computational complexity. In this paper, we develop a new environment parameter-driven neural-network approach for generating the Whittle index for constantly-changing environments. Our neural network only needs to be trained once, and it can instantly produce the Whittle index for a new environment in real-time, as soon as the new environment can be identified by its environment parameter (which can be estimated by either a hand-crafted estimator or a trained environment encoder). We further provide a new Whittle Actor-Critic (WAC) pipeline to train the index network. Compared to the prior Whittle Q-Learning pipeline, our new WAC pipeline is less sensitive to the neural-network size and can produce meaningful index-like policies even for non-indexable problems. Numerical results show that our WAC outperforms other baselines in a broad range of RMAB settings in terms of both index accuracy and control performance.

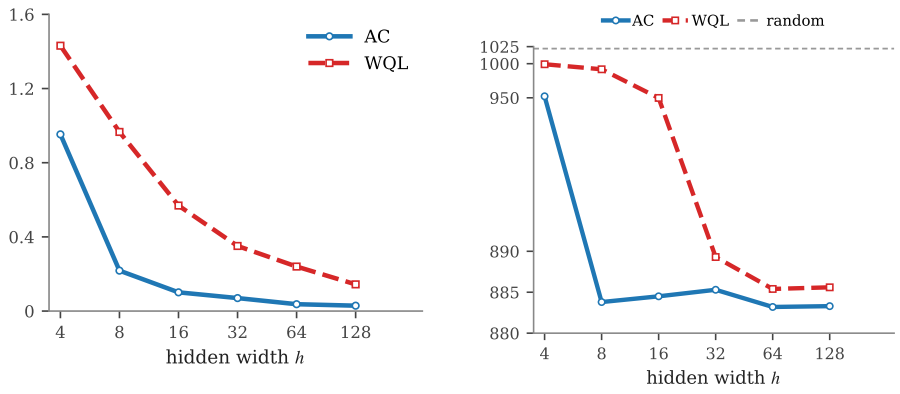
1 INTRODUCTION

Whittle index (Whittle, 1988) is an effective method to develop low-complexity and adaptive solutions for RMAB problems. Many stochastic resource allocation problems that can be modelled as RMAB problems, e.g., communication networks (Chen et al., 2022), healthcare (Biswas et al., 2021), and finance (Liu et al., 2012). In a typical RMAB problem, there are multiple arms competing for resources, where each arm evolves according to a Markov process. Unfortunately, RMAB is known to suffer from the curse of dimensionality when the number of arms is large. Whittle index can decompose an otherwise intractable RMAB problem into per-arm problems, and thus leads to significantly lower complexity. Specifically, the Whittle index for each arm can be computed based on its current state, independently of other arms. Then, the Whittle index policy simply assigns the resources to arms with higher indices. In this way, the computational complexity is at most linear in the number of arms. Further, the resulting solution is highly adaptive: if some arms change, the indices for other arms do not need to be re-calculated. As a result, Whittle index is especially appealing for large-scale and dynamic real-world problems modeled as RMAB.

However, when the environment faced by each arm changes constantly, a significant challenge to the application of the Whittle index policy is the high overhead in re-calculating the Whittle index. In wireless systems, for example, a mobile user's transmission success probability changes as the user moves. Existing model-based index methods require exact transition probabilities and cost functions and substantial recomputation time (Akbarzadeh & Mahajan, 2022; Gast et al., 2023). Model-free RL methods avoid exact model knowledge (Fu et al., 2019; Avrachenkov & Borkar, 2022; Nakhleh et al., 2022; 2021; Xiong & Li, 2023), but need many samples to relearn the index after each change.

An open question is therefore: "When the environment constantly changes, can we come up with a method that can generate the Whittle indices quickly so that we can apply index policy even when environment constantly changes?" In this paper, we provide the first positive answer for this question. Our key observation is that, for many practical problems, the environment faced by an arm can be

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Index error

(b) RMAB control performance

Figure 1: Whittle Actor-Critic is more robust to network width h than Whittle Q-Learning. At $h = 8$, WAC has $4\times$ lower index error (Figure 1a) than WQL and 12.2% lower control cost than WQL (Figure 1b). WQL needs $h = 128$ to match WAC at $h = 8$ on both metrics. See Section 5.1 for more discussion.

uniquely determined by a few parameters z. For example, in wireless AoI minimization problems (Hsu et al., 2017; Hsu, 2018; Kadota et al., 2018), the parameters z faced by each user/arm include the transmission success probability and the packet arrival rate. Further, even when environment changes, such parameters z can be estimated from the new environment relatively quickly, using either a hand-crafted estimator or an environment encoder (see Section 3.4). Motivated by this observation, in this paper we thus propose a new *environment parameter-driven* approach that can swiftly produce the Whittle index in changing environments. Specifically, we train a neural network to learn a mapping $\phi^*(s, z)$ from the state s to the index ϕ^* not just for a single environment, but across *all* environments z at the same time. Once this neural network is trained, in real time we only need to estimate the new environment parameter z, based on which our neural network can *instantly* produce the Whittle index for each arm in real-time, without any need of retraining. To make our approach more practical, we consider a standard model-free setting in which we only have access to a simulator that generates trajectories under different unknown environments (Rakelly et al., 2019; Finn et al., 2017; Yu et al., 2020b). This setting allows our pipeline to operate without prior knowledge of the state-transition matrix or cost function, where model-based approaches cannot be used to precompute the Whittle index for each environment.

While this idea of a parameter-driven index network may sound natural, developing an effective training pipeline is highly challenging. A natural approach is to extend the prior model-free Whittle Q-learning (WQL) (Avrachenkov & Borkar, 2022; Xiong & Li, 2023), which was designed for a single environment to our multi-environment setting. However, we have found that the WQL pipeline is sensitive to the neural-network size. This is because the WQL pipeline needs to first learn a Q-network, which is then used to train the index network. If the Q-network cannot produce accurate optimal Q-values (e.g., due to limited network capacity), the resulting index values will be even poorer (even though the index network may have enough capacity). This capacity issue becomes even more challenging in the multi-environment setting (Yu et al., 2020a) as we have to use one Q-network to learn the Q-values for all environments. As a result, the WQL pipeline either has to suffer poor index accuracy or has to adopt a much larger neural network size, which increases training cost.

To address these limitations, we propose a new Whittle actor-critic (WAC) pipeline for index network learning that is much more robust to neural-network size, even with multiple environments. Unlike WQL, which needs to first learn the Q-values of the *unknown* optimal policy, our WAC pipeline directly updates the index network based on the value functions of the policies given by the *current* index values. While these value functions can also be provided by a (trained) critic network (which therefore suffers from the same sensitivity to network size as a Q-network), we supplement the critic network by rollouts to further reduce the bias in the value-function estimates. As a result, the final index accuracy is mostly determined by the index-network capacity rather than by that of an intermediate Q-network. However, a naive implementation of WAC with a uniform initial-state distribution μ can be prohibitively slow and unstable. We therefore propose in section 3.1 to modify

μ to improve training stability and efficiency. As shown in Figure 1a, 1b, WAC reduces index error by a factor of 4, and reduces the average cost by 12% at $h = 8$. To achieve the same performance as WAC at $h = 8$, WQL needs $16\times$ network width (i.e., $h = 128$) and $150\times$ larger network parameters. Complementing these empirical advantages, we also provide a theoretical characterization of the optimization landscape underlying WAC. In particular, under a sufficient condition for indexbality, we show that the Whittle index is the *unique* stationary point of the WAC objective (Theorem 4.1).

Our WAC pipeline also has another unexpected, but highly desirable benefit: it can be used to learn good index-like policies even for *non-indexable* problems. Note that there is no Whittle index for non-indexable problems. However, it is still highly desirable to learn an index-like policy that assigns each state an index independently of other arms and activates arms with higher indices even for non-indexable problems. The resulting policy will then share similar benefits of low-complexity and high adaptability as the Whittle index policy. Since our WAC objective directly optimizes the performance, it is meaningful even for non-indexable arms! In contrast, WQL will have significant difficulty in non-indexable problems because the Q-network can only learn the Q-value for the optimal policy which in this case is no longer an index-like policy. We provide a preliminary theoretical result explaining why the resulting index-like policy might be effective in Theorem 4.2. The control performance gain of WAC over WQL is even higher for non-indexable problems: even with sufficient network capacity, WAC outperforms WQL by 17.5% (see Section 5). Finally, our numerical results in Section 5 shows that our environment parameter-driven approach can quicky adapt to dynamically changing environments, achieving both lower index errors and lower control costs.

2 Problem formulation

2.1 Preliminaries on RMAB and Whittle index

Consider a Restless Multi-Armed Bandit (RMAB) problem with N arms. Let the state of arm n at time t be $S_n(t) \in \mathcal{S} := \{s_1, \ldots, s_d\}$, where $d := |\mathcal{S}|$, which evolves according to the time-homogeneous transition kernel $P_n(s' \mid s, a)$, where $a \in \{0, 1\}$ is the binary activation and $s, s' \in \mathcal{S}$. Upon taking action $A_n(t)$ at state $S_n(t)$, the decision-maker incurs an immediate cost $c_n(S_n(t), A_n(t))$. The RMAB problem is to minimize the expected total discounted cost over a per-step constraint C, i.e.,

$$\min_{\bar{\pi}} \ \mathbb{E}_{\bar{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \sum_{n=1}^{N} c_n(S_n(t), A_n(t))\right] \quad \text{s.t.} \quad \sum_{n=1}^{N} A_n(t) \le C, \ \forall t, \tag{1}$$

where $\bar{\pi}$ is the policy mapping: $\mathcal{S} \times \cdots \times \mathcal{S} \to \{0, 1\}^N$ and $\gamma \in (0, 1)$ is the discount factor.

This problem suffers from the curse of dimensionality and is generally intractable when N is large. Whittle's seminal approach (Whittle, 1988) relaxes the hard constraint in (1) to a time-averaged one. Introducing a Lagrange multiplier λ to the time-averaged constraint yields the Lagrangian: $\min_{\bar{\pi}} \ \mathbb{E}_{\bar{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \sum_{n=1}^{N} \big(c_n(S_n(t), A_n(t)) + \lambda A_n(t)\big)\right]$ (omitting the constant here). Minimizing this Lagrangian can be decoupled to N independent single-arm MDPs,

$$\min_{\bar{\pi}} \ \mathbb{E}_{\bar{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \big(c_n(S_n(t), A_n(t)) + \lambda A_n(t)\big)\right], \tag{2}$$

where $1 \le n \le N$. Note that (2) is also a MDP, but is much smaller. We denote this MDP by $\mathcal{E}_n(\lambda)$, where the transition dynamics is the same as the original single-arm MDP, but the cost function is modified to $c_n(s, a) + \lambda a$. Here, λ can be interpreted as the additional price of activating the arm.

For each arm n, define the *passive set* as $\mathcal{P}_n(\lambda) = \big\{s \in \mathcal{S}_n \,\big|\, \exists\, \pi^*_\lambda \in \Pi^*_\lambda \text{ s.t. } \pi^*_\lambda(s) = 0\big\}$, where Π^*_λ is the set of optimal policies for Problem (2). The arm n is said to be *Whittle indexable* if $\mathcal{P}_n(\lambda)$ is monotonically non-decreasing in λ (Whittle, 1988). When indexability holds, the Whittle index for arm n at state s is defined as $w_n(s) = \inf\big\{\lambda \mid s \in \mathcal{P}_n(\lambda)\big\}$, which is the lowest price λ at which the optimal policy for the arm becomes passive in state s. The Whittle index policy then activates the C arms with the highest indices at each time. It has been shown that the Whittle index policy is asymptotically optimal (Weber & Weiss, 1990). Unless otherwise specified, we assume that all arms are Whittle indexable in this paper.

2.2 Formulation of the parameter-driven Whittle-Index Learning Problem

In the remainder of the paper, since the Whittle index is defined for an individual arm, we omit the arm subscript n for notational simplicity. The Whittle index of an arm depends only on its state-transition matrix and cost functions, which we collectively refer to as the *environment* of the arm. The discussion in Section 2.1 assumes that this environment is fixed. In practice, however, the environment may vary over time. To model such nonstationarity, we assume that each environment is associated with a representation vector $z \in \mathcal{Z}$ that uniquely identifies it. For brevity, we sometimes refer to the environment represented by z simply as environment z. We denote the Whittle index of state s under environment z by $\phi^*(s, z)$.

Our proposed pipeline can then be summarized as follows. At each time step, we first infer an environment representation z from a window of recently observed transitions. We then feed z, together with the current state s, into the trained index network $\phi_\theta(s, z)$ to estimate the corresponding Whittle index for each arm. The resulting indices are subsequently used by the Whittle-index policy.

However, this formulation gives rise to two key challenges:

- How can we train an index network $\phi_\theta(s, z)$ that accurately approximates $\phi^*(s, z)$ across all environments $z \in \mathcal{Z}$, while maintaining robust and efficient training?
- How can we infer an informative environment representation z from observed transitions such that different environments can be reliably distinguished?

Section 3 addresses both challenges by introducing a new Whittle actor-critic pipeline for training the index network and two approaches for inferring the environment parameters.

3 Whittle Actor Critic Index Learning

In this section, we introduce our Whittle actor-critic (WAC) based index learning method for learning $\phi_\theta(s, z)$. The WAC pipeline assumes access to the environment representation z, which in practice is inferred from a set of transitions collected through interaction with the environment. We defer the details of this inference procedure to Section 3.4 and, for now, treat z as given. We first introduce the class of λ-threshold policies induced by index vectors, which constitutes the policy space of WAC.

Definition 3.1 (λ-Threshold Policy)**.** Let $\lambda \in \mathbb{R}$ and $\phi \in \mathbb{R}^d$ be an index vector. The λ-threshold policy induced by ϕ, denoted π^ϕ_λ, selects action 1 in state s_i if $\phi(s_i) > \lambda$ and action 0 otherwise.

Let $V_\lambda^{\pi^\phi_\lambda}(\mu, z)$ denote the (true) value function under the λ-threshold policy induced by ϕ, π^ϕ_λ, for the per-arm MDP $\mathcal{E}(\lambda)$ defined in (2), with a uniform initial-state distribution μ. Suppose M is a sufficiently large constant such that $\phi^*(s, z) \in (-M, M)$ for all $s \in \mathcal{S}$ and $z \in \mathcal{Z}$. We formulate our index learning problem as minimizing the following objective

$$h(\theta) = \int_{z \in \mathcal{Z}} g(\phi_\theta(\cdot, z), z)\, \mathrm{d}z, \quad \text{where} \quad g(\phi, z) := \int_{-M}^{M} V_\lambda^{\pi^\phi_\lambda}(\mu, z)\, \mathrm{d}\lambda. \tag{3}$$

Here, $\phi_\theta(\cdot, z)$ stands for $(\phi_\theta(s_1, z), \ldots, \phi_\theta(s_d, z)) \in \mathbb{R}^d$, $g(\cdot, z)$ is a function with variable $\phi \in \mathbb{R}^d$ and $g(\phi_\theta(\cdot, z), z)$ stands for $g(\phi, z)$ evaluated at $\phi = \phi_\theta(\cdot, z)$. Intuitively, for any given environment z, the objective $g(\phi, z)$ measures the performance of the λ-threshold policies induced by ϕ over $\lambda \in [-M, M]$, where $\phi \in \mathbb{R}^d$. Correspondingly, $h(\theta)$ measures the average performance over all environments $z \in \mathcal{Z}$ with the index vector $\phi_\theta(\cdot, z)$ for each z.

The next theorem shows that, for any z, the true Whittle index $\phi^*(s, z)$ corresponds to the global minimum of $g(\phi, z)$ (w.r.t. ϕ). It further derives $g(\phi, z)$'s gradient with respect to ϕ and $h(\theta)$'s gradient with respect to θ. For notational convenience, for any passive set $P \subset \mathcal{S}$, let $V_\lambda^P(\mu, z)$ denote the value function of the MDP $\mathcal{E}(\lambda)$ with uniform initial state distribution μ, under environment z and under the policy defined by the passive set P, i.e., the action is 0 for all states in P, and 1 for states outside P. For a fixed environment z and index vector ϕ, write $\phi_i := \phi(s_i)$ and define the following passive sets $R_{0,\phi} := \emptyset$, $R_{i,\phi} := \{s_j \mid \phi_j \le \phi_i\}$, and $L_{i,\phi} := \{s_j \mid \phi_j < \phi_i\}$ for $1 \le i \le d$. When z varies, R^z_{i,ϕ_θ} and L^z_{i,ϕ_θ} to denote these passive sets evaluated at $\phi = \phi_\theta(\cdot, z)$.

Theorem 3.2. *For any environment z, the Whittle index $\phi^*(s, z)$ is the global minimizer of $g(\phi, z)$ with respect to ϕ. Moreover, at any interior point $\phi \in (-M, M)^d$, for $1 \le j \le d$, the right and left*

partial derivatives of $g(\phi, z)$ are

$$\frac{\partial g(\phi, z)}{\partial \phi_j^+} = V_{\phi_j}^{R_{j,\phi}\setminus\{s_j\}}(\mu, z) - V_{\phi_j}^{R_{j,\phi}}(\mu, z), \tag{4}$$

$$\frac{\partial g(\phi, z)}{\partial \phi_j^-} = V_{\phi_j}^{L_{j,\phi}}(\mu, z) - V_{\phi_j}^{L_{j,\phi}\cup\{s_j\}}(\mu, z). \tag{5}$$

Further, g is differentiable on the set $\mathcal{D}_M = \{\phi \in (-M, M)^d : \phi_i \neq \phi_j \text{ for all } i \neq j\}$, i.e., all components of ϕ are distinct. Thus, by chain rule, the gradient of $h(\theta)$ on the differentiable region is

$$\nabla_\theta h(\theta) = \int_{z\in\mathcal{Z}} \sum_j \Big(V_{\phi_\theta(s_j,z)}^{R^z_{j,\phi_\theta}\setminus\{s_j\}}(\mu, z) - V_{\phi_\theta(s_j,z)}^{R^z_{j,\phi_\theta}}(\mu, z)\Big)\nabla_\theta\phi_\theta(s_j, z)\,\mathrm{d}z, \tag{6}$$

Theoerem 3.2 implies that if there exists θ^* such that $\phi_{\theta^*}(s, z) = \phi^*(s, z)$ for all s and z, then θ^* is a global minimizer of $h(\theta)$. Also, to minimize $h(\theta)$, we can leverage the gradient expression (6). For ease of exposition, below we denote $\lambda_i(z) := \phi_\theta(s_i, z)$ for $1 \leq i \leq d$. We can then apply a critic network V_ψ to approximate the true value functions $V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}\setminus\{s_j\}}(\mu, z)$ and $V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}}(\mu, z)$. To approximate $V_\lambda^R(s, z)$, the critic network $V_\psi(s, R, \lambda, z)$ takes as input the state s, a multi-hot embedding of the passive set R, activation cost λ and environment z. Then we can use the idea of stochastic gradient: instead of computing the exact gradient, we sample a batch of environments z and states s_j to estimate the integral and summation in (6).

We note that both our objective $g(\phi, z)$ and its gradient resemble that of (Nakhleh et al., 2022) (Eq. 2). However, the subsequent neural-network training procedure in (Nakhleh et al., 2022) corresponds to training the optimal Q-function, making their approach fundamentally an instance of Q-learning (see the detailed discussion in Appendix C). To the best of our knowledge, our approach below is the first attempt in the literature that applies WAC to index learning.

3.1 Stabilizing training with a different initial-state distribution

Unfortunately, the above training pipeline does not work right away. The first difficulty is the vanishing gradient issue, making the training unstable. Specifically, when we use the uniform distribution μ for the initial state, the gradient in (6) can be very small, stalling the optimization process. The reason is that for most initial state s, the value difference $V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}\setminus\{s_j\}}(s, z) - V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}}(s, z)$ is typically very small. To see this, note that the policies corresponding to the passive sets R^z_{j,ϕ_θ} and $R^z_{j,\phi_\theta} \setminus \{s_j\}$ differ only at state s_j. Starting from state $s \neq s_j$, the trajectories under these two policies coincide, until they first visit state s_j. With a large state space, this hitting time to s_j is long. After discounting, the value difference (contributed by the divergence after the hitting time) will thus be very small. Since this phenomenon occurs for nearly every $s \neq s_j$, even though we average over μ, the resulting $V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}\setminus\{s_j\}}(\mu, z) - V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}}(\mu, z)$ is close to zero and easily overwhelmed by noise.

To address this difficulty, we apply a different initial state distribution: for each j, we evaluate the value difference *with initial state being exactly s_j*, rather than under the initial distribution μ. That is, we replace (6) by

$$\int_{z\in\mathcal{Z}} \sum_j \Big(V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}\setminus\{s_j\}}(s_j, z) - V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}}(s_j, z)\Big)\nabla_\theta\phi_\theta(s_j, z)\,\mathrm{d}z. \tag{7}$$

As shown in Appendix E, the two value differences $V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}\setminus\{s_j\}}(s, z) - V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}}(s, z)$ and $V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}\setminus\{s_j\}}(s_j, z) - V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}}(s_j, z)$ have the same sign. The resulting update is therefore a reweighting of the original gradient. Such reweighting of the gradient share similar ideas as in example reweighting (Ren et al., 2018) and adaptive gradient balancing for multi-task learning Chen et al. (2018). We acknowledge that, after applying the chain rule, this update in general is not guaranteed to be a descent direction for the original objective. Empirically, however, this modification appears to work correctly and substantially improves the stability of the optimization process.

3.2 USING ROLLOUTS FOR VALUE-FUNCTION ESTIMATION

The pipelines of both WQL and WAC require estimating the optimal Q-functions/value functions of given policies. However, unlike WQL, which needs to approximate the optimal Q-function purely with a neural network, our WAC could enhance the value function estimation with rollout, which accelerates our training and improves final accuracy and performance of our index learning. Specifically, for every $1 \leq j \leq d$, starting from state s_j, it follows the policy with passive set being R^z_{j,ϕ_θ}, for T ($T \geq 1$) steps, accumulates the observed discounted costs, and bootstraps the remaining cost-to-go with the critic network. We then average the estimates over M ($M \geq 1$) independent trajectories to reduce variance. In contrast, the Q-function in WQL is defined w.r.t the unknown optimal policy, precluding such on-policy rollout. This way, we can obtain a more accurate estimate of the value function $V^{R^z_{j,\phi_\theta}}_{\lambda_j(z)}(s_j, z)$. The same applies to enhancing the estimate of $V^{R^z_{j,\phi_\theta}\setminus\{s_j\}}_{\lambda_j(z)}(s_j, z)$.

3.3 NON-INDEXABLE PROBLEMS

An unexpected, but highly desirable benefit of the WAC pipeline is its ability to learn effective index-like policies even for non-indexable problems. Note that one limitation of the Whittle index is that it can only be defined for problems that satisfy the Whittle indexability (Whittle, 1988). Although there are other approaches to handle general RMAB problems without indexability (Bertsimas & Niño-Mora, 2000; Gast et al., 2022; Gast & Narasimha, 2024; Chen et al., 2024), their solutions usually require prior knowledge of the optimal dual variable λ^* first, and as a result are not as "robust" as the Whittle index policy. That is, whenever any arms change, the whole policy (along with the optimal λ^*) has to be recomputed again. In contrast, we believe that, even for non-indexable problems, it would be highly desirable in practice to learn an "index-like" policy, which will share similar benefits of low complexity and adaptivity as the Whittle index policy. The question is how.

Our WAC provides such a solution. Even for non-indexable problems, WAC can train a suitable index ϕ_θ by minimizing the cost of its induced λ-threshold policies integrated over activation prices $\lambda \in [-M, M]$. In contrast, WQL is inherently problematic in the non-indexable setting: its Q-network is trained to learn the Q-values of the optimal policy, which itself need not be index-like, making it difficult to extract a meaningful index-like policy from the learned Q-values. We further provide theoretical evidence (Theorem 4.2) and empirical results in Section 5.2 to support the effectiveness of WAC in non-indexable problems.

3.4 ESTIMATING THE PARAMETERS FOR EACH ENVIRONMENT ONLINE

Our method assumes that a few observations suffice to estimate the environment-identifying parameter z online. We use either a hand-crafted estimator based on domain knowledge or a trained encoder.

If we are familiar with the underlying model structure of the environment, then such an estimator can be designed by hand. For example, in a wireless scheduling problem, the key parameters are the transmission success probability p_s and the packet arrival rate p_g. Both can be easily estimated from the number of successful transmissions and the number of packet arrivals.

However, in most cases, we do not have prior knowledge of the environment and are therefore unable to track it manually. In these cases, inspired by (Rakelly et al., 2019), we propose to train an environment encoder that receives the environment transition observations (e.g., $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^{30}$) and outputs a (roughly-unique) representation vector z. A key challenge is how to make this encoder learn representations that distinguish environments as clearly as possible. We find that using the KL regularization as in (Rakelly et al., 2019) has poor performance in our case. Instead, we find that the VICReg objective (Bardes et al., 2022) to be effective in preventing the representation collapse (see Appendix D.1 for more details). Note that the training of such an encoder does not require any domain-specific knowledge, and is thus applicable to any RMAB problems. For more details and investigation of our approach, see section 5 and Appendix D.1.

4 THEORETICAL GUARANTEES FOR WAC

In this section, we provide some theoretical guarantees for the methods that we proposed in Section 3. Our first result is for indexable problems. Unfortunately, once neural-network based function

approximation is used in an RL pipeline similar to ours, it would be very challenging to provide any optimality guarantees for RL (Baird, 1995; Tsitsiklis & Van Roy, 1996). Thus, similar to (Watkins & Dayan, 1992; Tsitsiklis, 1994; Singh et al., 2000; Avrachenkov & Borkar, 2022), our result below is restricted to a fixed environment and a tabular setting. Even under this setting though, the optimality of WAC is non-trivial to establish because the single-environment objective g in (3) is in general non-convex. (See Appendix F for why a stationary point of $g(\phi, z)$ may not correspond to the correct Whittle index.) Still, we are able to derive the following positive result.

To prove our theorem, we need an active time condition that is a sufficient condition for indexability and is similar to the GCL condition (Bertsimas & Nino-Mora, 1993). This assumption holds for broad classes of arms, including stochastic-reset bandits with arbitrary passive dynamics, homogeneous birth–death intervention models, and all two-state arms (see Appendix H). We now introduce our result in the following theorem and defer the complete statement as well as its proof to Appendix G.

Theorem 4.1. *Suppose the active-time condition Assumption G.1 holds. Then, on $\mathcal{D}_M$, $g(\phi)$ has no stationary point other than the Whittle-index vector ϕ^*. On $(-M, M)^d \setminus \mathcal{D}_M$, it has no local minimum other than ϕ^*.*

Theorem 4.1 implies that optimizing $g(\phi, z)$ could lead to the correct Whittle index. We acknowledge that Theorem 4.1 is limited to the fixed environment and tabular setting. Nonetheless, to the best of our knowledge, our result provides the first correctness guarantee for the WAC pipeline. Further, it gives us hope that WAC could still work well in multiple environments and with function approximation.

For non-indexable problems, theoretical guarantees will be more difficult to obtain as there is no Whittle index as benchmark. Readers may question why our training objective g in (3) is still meaningful. Here, we provide the following result under the average cost setting as a justification. Suppose that, similar to other Whittle optimality proofs (Weber & Weiss, 1990), a global attractor assumption holds (see Appendix I). Further, there is a threshold value $\hat{\lambda}$ that corresponds to the lowest index among those arms activated by our index-like policy at this global attractor. We say a policy is ϵ-optimal if its cost differs by at most ϵ from the optimal cost.

Theorem 4.2 (Non-indexable guarantee)**.** *Under global attractor assumption, suppose $\widehat{\lambda}$-threshold policies induced by the learned index is ϵ-optimal in the single-arm problem with activation cost $\widehat{\lambda}$, then the induced index-like policy is ϵ-optimal in per-arm average cost for the overall RMAB.*

Therefore, Theorem 4.2 implies that any suboptimality gap ϵ in the single-arm problem directly translates into a corresponding gap in the system cost. Intuitively, our objective g in (3) which minimizes the per-arm costs of the induced threshold policies integrated over all $\lambda \in [-M, M]$ provides meaningful training signals. But we acknowledge that we are still unable to control the per-arm gap ϵ through optimizing g: (i) The optimized intergral of the per-arm cost does not imply the per-arm gap ϵ for $\widehat{\lambda}$ and (ii) the gap is in terms of average setting while we optimize the discounted setting in (3). Nonetheless, our numerical experiments in n section 5.2 further show that the index-like policies learned by WAC substantially outperform other baselines.

5 Experiments

We fix the size of the encoder mlp as a three layer MLP with output dimension 5 (implying that the dimension of the learned representation z is 5), trained for 8000 iterations. Unless otherwise specified, both the value/Q-networks and the index network are three-layer MLPs with hidden width $h = 8$ trained for 50000 iterations. See Appendix D for the training pseudocode of our WAC, and Appendix D.1 for training details.

5.1 Sensitivity to the model size

We compare WAC and WQL across network widths on an indexable wireless age-of-information (AoI) problem with stochastic arrivals (Sun et al., 2019). Each arm/user has two state components: the age of its freshest buffered packet and the receiver AoI reduction from a successful transmission. Its environment is specified by transmission success probability p_s and packet arrival rate p_g, both in $[0, 1]$ and estimable by a hand-crafted estimator (Appendix J). The critic/Q and index networks are

Table 1: Index error by transition count K used to estimate z: **enc-30** is trained on 30 transitions per environment without domain priors; **manual** is a hand-designed counting estimator.

K	20	30	200	1024
enc-30	0.767	0.626	0.271	0.176
manual (prior)	0.616	0.492	0.215	0.124

three-layer MLPs of hidden width h. We report index error relative to exact Whittle indices across states and environments, and mean AoI cost across environments.

From Figure 1, we can see that WAC has a significantly lower index error and also a lower average cost than WQL across different h. More importantly, the gap is even larger when the neural network size is small (h being small). Specifically, in Figure 1a, the index error of WAC is smaller than that of WQL by a factor of at least 2 for all $h \leq 8$ (when $h = 8$, the index error of WAC is $4\times$ smaller). This result confirms that WAC is much more robust to neural-network size than WQL. Similar trends are observed in the control performance in Figure 1b, though the gap is not as large as that of the index error. We find this phenomenon to be common for indexable problems, i.e., the control performance is less sensitive to index errors: even if the predicted indices are not exactly correct, as long as the order of the indices is preserved, the induced index policy can still be near-optimal. We observe WAC outperforms WQL by 12% in terms of the time-averaged AoI when $h = 8$. However, as we will see below, the gain on non-indexable problems is even more significant, reaching 17.5%.

We also exmine more closely on the underlying reason for the performance gap between WAC and WQL. In Figure 6, we find that the function-approximation error of WAC is significantly smaller than that of WQL, especially when h is small. We discuss it in more detail in Appendix L

5.2 Non-indexable problems

We next turn to a non-index RMAB problem in (Charles-Rebuffé et al., 2026). We find that the control performance of the index policies learned by WAC significantly outperforms that of WQL. To exlude the network size issue we mentioned in section 5.1, we choose a large network with hidden dimension 128. We compare WAC and WQL against the random policy over a grid of fixed environment parameters. We find an even bigger (17.5%) improvement of the control performance of WAC over that of WQL: WAC attains a cost of -0.67 ± 0.02, whereas WQL attains -0.57 ± 0.02, i.e., a $+17.5\%$ improvement (lower cost means better). Furthermore, the control performance of WQL is even higher than that of a random policy, suggesting that the WQL pipeline does not really learn a useful index-like policy in this setting. For more details of this non-indexable RMAB and the evaluation, please refer to Appendix J.4 and K.1.

5.3 Hand-Crafted Estimator vs. Environment Encoder

As described in Section 3.4, we compare WAC using a hand-crafted estimator with that using an encoder trained to compress 30 transitions on the AoI problem (Section 5.1). Table 1 reports index error across environments for different observation counts K.

The manual estimator achieves uniformly lower error by leveraging prior model knowledge, but the gap is modest: the encoder nearly matches its performance without relying on domain-specific knowledge. Moreover, although the encoder is trained with a context length of 30, it generalizes well to other context lengths. We provide a more detailed discussion in Appendix M, where we also ablate the rollout horizon T and the number of rollout trajectories M.

5.4 Performance over constantly changing environments

In previous subsections, we report the performance of WAC across many environments, but the environment is fixed for each experiment. Next, we evaluate the performance of WAC in constantly-changing environments. We compare WAC with WQL and DeepTOP (Nakhleh et al., 2022). Note that DeepTOP does not use a parameter-driven approach. Instead, it learns the index from interacting with each environment from scratch. We evaluate all methods on the AoI problem described in Section 5.1. The RMAB consists of 600 arms and an activation budget of 200. Each arms' parameter

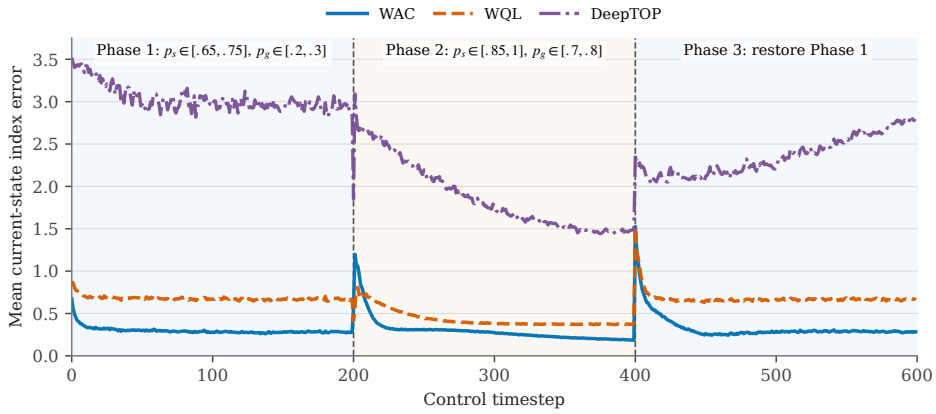


Figure 2: Per-timestep index error for WAC, WQL, and DeepTOP in a changing AoI environment. Lower is better.

Table 2: Mean per-step total cost across arms in changing RMABs (lower is better). Green/blue shading marks the lowest/second-lowest costs per row; ties share a color.

Environment	WAC	WQL	DeepTOP	Random
AoI	690	693	734	733
Healthcare	123	127	127	149
Maintenance 1D	981	1049	1287	1269
Maintenance 2D	658	735	946	1093

(p_s, p_g) changes at $t = 200$ and $t = 400$. We use a hand-crafted estimator here, estimating the transmission success rate $\hat{p}_s$ and arrival rate $\hat{p}_g$ with exponential moving averages (EMAs) to adapt to recent environmental changes. Figure 2 reports the index error at each time step, averaged over the current states of all arms.

We find from Figure 2 that our parameter-driven pipeline (both WAC and WQL) is more stable and adapts more quickly than Deeptop when environment changes. Further, WAC attains a lower index error than WQL. Indeed, every time the environment changes, the index errors of WAC and WQL immediately increase (as the current estimates $\hat{p}_s$ and $\hat{p}_g$ no longer reflect the new environment), but quickly recover to a low level as the estimates are updated. In contrast, the index error of DeepTOP does not improve to an acceptable level even when the next environment change occurs. Averaged over all timesteps and seeds, WAC achieves an index error of 0.309, compared with 0.597 for WQL and 2.422 for DeepTOP.

We also evaluate RMABs without simple hand-crafted estimators: resource-constrained healthcare interventions (Biswas et al., 2021) and preventive maintenance of deteriorating machines, including its 2D extension (Glazebrook et al., 2005; Ruiz-Hernandez et al., 2020; Appendix J). A learned encoder supplies environment representations to the index networks. Table 2 reports average costs for these problems and AoI.. WAC performs well across all four RMABs and improves on WQL by over 10% in Maintenance 2D. See Appendix K.2 for the evluation details.

6 Conclusion

In this paper, we address the index computation problem in RMAB in dynamically changing environments. We propose a parameter-driven neural-network pipeline that can swiftly generate Whittle indices for new environments in real time. We introduce a novel WAC method that can effectively train our index networks in both indexable and non-indexable settings. Experiments show that our pipeline trained with WAC achieve superior performance in terms of both index accuracy and control performance, and is robust to limited model capacity.

AI USE STATEMENT

In this work, we used generative AI tools to clean and reformat dataset, support qualitative and thematic data analysis. We did not use generative AI tools for the other research tasks requiring disclosure under the ICLR 2027 AI Policy for Authors. We also used generative AI tools to creat and edit software code, identify relevant literature, refine the writing of the paper, and assist with experimental code. We reviewed all AI-assisted work and checked AI-generated code against the intended experimental setup. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nima Akbarzadeh and Aditya Mahajan. Conditions for indexability of restless bandits and an algorithm to compute whittle index. *Advances in Applied Probability*, 54(4):1164–1192, 2022.

Konstantin E Avrachenkov and Vivek S Borkar. Whittle index based q-learning for restless bandits with average reward. *Automatica*, 139:110186, 2022.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. Technical Report CMU-CS-95-103, Carnegie Mellon University, 1995.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Dimitris Bertsimas and José Niño-Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90, 2000.

Dimitris J Bertsimas and Jose Nino-Mora. Conservation laws, extended polymatroids and multi-armed bandit problems; a unified approach to indexabel systems. 1993.

Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 4039–4046, 2021. doi: 10.24963/ijcai.2021/556.

Joël Charles-Rebuffé, Nicolas Gast, and Bruno Gaujal. Model-based learning of whittle indices. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 2026.

Gongpu Chen, Soung Chang Liew, and Yulin Shao. Uncertainty-of-information scheduling: A restless multiarmed bandit framework. *IEEE Transactions on Information Theory*, 68(9):6151–6173, 2022.

Yan Chen, Weina Wang, and Lei Ying. Achieving $\tilde{\gamma}(1/n)$ optimality gap in restless bandits through gaussian approximation. *arXiv preprint arXiv:2410.15003*, 2024.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 794–803, 2018.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Jing Fu, Yoni Nazarathy, Sarat Moka, and Peter G Taylor. Towards q-learning the whittle index for restless bandits. In *2019 Australian & New Zealand Control Conference (ANZCC)*, pp. 249–254. IEEE, 2019.

Nicolas Gast and Dheeraj Narasimha. Model predictive control is almost optimal for restless bandit. *arXiv preprint arXiv:2410.06307*, 2024.

Nicolas Gast, Bruno Gaujal, and Yan Chen. Reoptimization nearly solves weakly coupled markov decision processes. *arXiv preprint arXiv:2211.01961*, 2022.

Nicolas Gast, Bruno Gaujal, and Kimang Khun. Testing indexability and computing whittle and gittins index in subcubic time. *Mathematical Methods of Operations Research*, 97(3):391–436, 2023.

K. D. Glazebrook, H. M. Mitchell, and P. S. Ansell. Index policies for the maintenance of a collection of machines by a set of repairmen. *European Journal of Operational Research*, 165(1):267–284, 2005. doi: 10.1016/j.ejor.2004.01.036.

Yu-Pin Hsu. Age of information: Whittle index for scheduling stochastic arrivals. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 2634–2638. IEEE, 2018.

Yu-Pin Hsu, Eytan Modiano, and Lingjie Duan. Age of information: Design and analysis of optimal scheduling algorithms. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 561–565. IEEE, 2017. doi: 10.1109/ISIT.2017.8006590.

Igor Kadota, Abhishek Sinha, Elif Uysal-Biyikoglu, Rahul Singh, and Eytan Modiano. Scheduling policies for minimizing age of information in broadcast wireless networks. *IEEE/ACM Transactions on Networking*, 26(6):2637–2650, 2018. doi: 10.1109/TNET.2018.2873606.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Pat Langley. Crafting papers on machine learning, 2000.

Haoyang Liu, Keqin Liu, and Qing Zhao. Learning in a changing world: Restless multiarmed bandit with unknown dynamics. *IEEE Transactions on Information Theory*, 59(3):1902–1916, 2012.

Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I Hou, Srinivas Shakkottai, et al. Neurwin: Neural whittle index network for restless bandits via deep rl. *Advances in Neural Information Processing Systems*, 34:828–839, 2021.

Khaled Nakhleh, I Hou, et al. Deeptop: Deep threshold-optimal policy for mdps and rmabs. *Advances in Neural Information Processing Systems*, 35:28734–28746, 2022.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pp. 5331–5340. PMLR, 2019.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4334–4343, 2018.

Diego Ruiz-Hernandez, Jesus M. Pinar-Perez, and David Delgado-Gomez. Multi-machine preventive maintenance scheduling with imperfect interventions: A restless bandit approach. *Computers and Operations Research*, 119:104927, 2020. doi: 10.1016/j.cor.2020.104927.

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

Jingzhou Sun, Zhiyuan Jiang, Sheng Zhou, and Zhisheng Niu. Optimizing information freshness in broadcast network with unreliable links and random arrivals: An approximate index policy. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 115–120. IEEE, 2019.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In M.C. Mozer, M. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/e00406144c1e7e35240afed70f34166a-Paper.pdf.

John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16(3): 185–202, 1994.

I. M. Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *The Annals of Applied Probability*, 26(4):1947–1995, 2016. doi: 10.1214/15-AAP1137.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of applied probability*, 27(3):637–648, 1990.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

Guojun Xiong and Jian Li. Finite-time analysis of whittle index based q-learning for restless multi-armed bandits with neural network function approximation. *Advances in Neural Information Processing Systems*, 36:29048–29073, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020a.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020b.

Sixiang Zhou and Xiaojun Lin. An easier-to-verify sufficient condition for whittle indexability and application to aoi minimization. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*, pp. 1741–1750. IEEE, 2024.

A PARAMETER-DRIVEN WHITTLE Q-LEARNING

In this section, we show how to extend prior Whittle Q-learning methods (Avrachenkov & Borkar, 2022; Xiong & Li, 2023) to the parameter-driven setting. Specifically, we define $Q^*_\lambda(s, a, z)$ as the optimal Q-function under environment z with activation price λ. Let ϕ_θ be the index neural network. We formulate the index learning problem as minimizing $f(\theta)$, which is defined as follows

$$f(\theta) := \int_{z\in\mathcal{Z}} \int_{-M}^{M} \sum_k Q^*_\lambda\big(s_k,\, \mathbf{1}\{\lambda < \phi_\theta(s_k, z)\}, z\big)\, \mathrm{d}\lambda\, \mathrm{d}z. \tag{8}$$

The theorem below shows that the correct Whittle indices correspond to the global minimum of $f(\theta)$, and it also gives its gradient with respect to θ.

Theorem A.1. *If θ^* satisfies that $\phi_{\theta^*}(s, z)$ gives the Whittle index for state s under environment z, then θ^* is a global minimum of $f(\theta)$.*

Furthermore, if $\phi_\theta(s, z) \in (-M, M)$ and the network outputs are differentiable in θ for every state and almost every z, and if differentiation may be interchanged with integration over z, then f is differentiable at θ and its gradient is given by Here $\lambda_j(z) := \phi_\theta(s_j, z)$.

$$\nabla_\theta f(\theta) = \int_{z\in\mathcal{Z}} \sum_j \big[Q^*_{\lambda_j(z)}(s_j, 1, z) - Q^*_{\lambda_j(z)}(s_j, 0, z)\big] \nabla_\theta \phi_\theta(s_j, z)\, \mathrm{d}z. \tag{9}$$

Proof. We first show that if θ^* satisfies that $\phi_{\theta^*}(s, z)$ gives the correct Whittle index for state s and environment z, then θ^* is the global minimum of f. Fix the environment z. Denote the true Whittle index for state s_k by $\phi^*_k(z)$. By the definition of the Whittle index, $\mathbf{1}\{\lambda < \phi^*_k(z)\}$ gives the optimal action for state s_k under per-arm MDP $\mathcal{E}(\lambda)$. As a result, for any real number ϕ_k, it holds that

$$Q^*_\lambda\big(s_k,\, \mathbf{1}\{\lambda < \phi^*_k(z)\}, z\big) \le Q^*_\lambda\big(s_k,\, \mathbf{1}\{\lambda < \phi_k\}, z\big)\,.$$

Thus, if we sum over all states and integrate both sides from $-M$ to M, we have

$$\int_{-M}^{M} \sum_k Q^*_\lambda\big(s_k,\, \mathbf{1}\{\lambda < \phi^*_k(z)\}, z\big)\, \mathrm{d}\lambda \le \int_{-M}^{M} \sum_k Q^*_\lambda\big(s_k,\, \mathbf{1}\{\lambda < \phi_k\}, z\big)\, \mathrm{d}\lambda.$$

Integrating over $z \in \mathcal{Z}$ on both sides, we obtain that $f(\theta)$ is minimized when $\phi_\theta(s_k, z)$ gives the Whittle index for all states s_k.

To derive the gradient expression, we note that for any fixed s_k and z, we have

$$\int_{-M}^{M} Q^*_\lambda\big(s_k,\, \mathbf{1}\{\lambda < \phi_\theta(s_k, z)\}, z\big)\, \mathrm{d}\lambda = \int_{-M}^{\phi_\theta(s_k,z)} Q^*_\lambda(s_k, 1, z)\, \mathrm{d}\lambda + \int_{\phi_\theta(s_k,z)}^{M} Q^*_\lambda(s_k, 0, z)\, \mathrm{d}\lambda.$$

Taking the gradient on both sides, we have

$$\nabla_\theta \int_{-M}^{M} Q^*_\lambda\big(s_k,\, \mathbf{1}\{\lambda < \phi_\theta(s_k, z)\}, z\big)\, \mathrm{d}\lambda = \Big(Q^*_\lambda\big(s_k, 1; z\big)\big|_{\lambda=\phi_\theta(s_k,z)} - Q^*_\lambda\big(s_k, 0; z\big)\big|_{\lambda=\phi_\theta(s_k,z)}\Big) \nabla_\theta \phi_\theta(s_k, z).$$

Summing over all states s_k and integrating over $z \in \mathcal{Z}$, we obtain the desired gradient expression. □

B PROOF OF THEOREM 3.2

Proof. Fix z and write $\phi_i := \phi(s_i)$. In this proof only, abbreviate $V^\pi_\lambda(\mu, z)$ and $V^R_\lambda(\mu, z)$ as $V^\pi(\mu, \lambda)$ and $V^R(\mu, \lambda)$, respectively, while retaining z in $g(\phi, z)$. Here π is the policy and R is the passive set.

By the definition of the Whittle index, we note that for any fixed λ, the λ-threshold policy induced by ϕ, π^ϕ_λ, is optimal under $\mathcal{E}(\lambda)$ if ϕ gives the Whittle index for all states.

To show the right partial derivative expression, we calculate by definition,

$$\frac{\partial g(\phi, z)}{\partial \phi^+_j} = \lim_{h\to 0^+} \frac{g(\phi + he_j, z) - g(\phi, z)}{h},$$

where e_j is the j-th standard basis vector. Note that when we increase ϕ_j by h, compared with π^{ϕ}_{λ}, the λ-threshold policy induced by $\phi + he_j$, $\pi^{\phi+he_j}_{\lambda}$, only changes when $\lambda \in (\phi_j, \phi_j + h)$: the action for state s_j changes from passive to active, while the actions for all other states remain unchanged. Thus, we have

$$g(\phi+he_j, z)-g(\phi, z) = \int_{-M}^{M} \Big(V^{\pi^{\phi+he_j}_{\lambda}}(\mu, \lambda)-V^{\pi^{\phi}_{\lambda}}(\mu, \lambda)\Big)\,\mathrm{d}\lambda = \int_{\phi_j}^{\phi_j+h} \Big(V^{R_j\setminus\{s_j\}}(\mu, \lambda)-V^{R_j}(\mu, \lambda)\Big)\,\mathrm{d}\lambda.$$

Dividing both sides by h and taking the limit as $h \to 0^+$, we obtain the desired expression for the right partial derivative.

The left partial derivative expression can be derived similarly.

We now discuss the differentiable region and gradient. If $\phi_k = \phi_j$ for some $k \neq j$, the right- and left-derivative expressions in (4) and (5) need not agree. If all components of ϕ are distinct, however, then $R_{j,\phi} = L_{j,\phi} \cup \{s_j\}$ for every j. Hence the two expressions agree, and the partial derivatives are continuous on $\mathcal{D}_M$. Therefore, g is differentiable on $\mathcal{D}_M$. □

C Further discussion on (Nakhleh et al., 2022)

As we explained in our main text, (Nakhleh et al., 2022) also studies an actor-critic method for index learning. In this section, we explain the similarity to our work in terms of the RL objective and gradient, and more importantly, why the RL method of (Nakhleh et al., 2022) is NOT an actor-critic method.

Similarity: We first compare the objective of (Nakhleh et al., 2022) with ours in (3). Throughout this comparison, we fix an environment z and omit it from the notation. We use our notation $\phi = \phi_\theta(\cdot, z)$, $\phi_i = \phi(s_i)$, and π^{ϕ}_{λ} from Definition 3.1, and convert their reward-maximization objective to our cost-minimization convention by negating rewards. Their Eq. (10), expressed in this convention, is

$$\int_{-M}^{M} \sum_{i=1}^{d} Q^{\pi^{\phi}_{\lambda}}_{\lambda}(s_i, \pi^{\phi}_{\lambda}(s_i))\,\mathrm{d}\lambda.$$

Here $Q^{\pi^{\phi}_{\lambda}}_{\lambda}$ is the action-value function under π^{ϕ}_{λ}, so $Q^{\pi^{\phi}_{\lambda}}_{\lambda}(s_i, \pi^{\phi}_{\lambda}(s_i)) = V^{\pi^{\phi}_{\lambda}}_{\lambda}(s_i)$. Thus, the objective becomes

$$\int_{-M}^{M} \sum_{i=1}^{d} V^{\pi^{\phi}_{\lambda}}_{\lambda}(s_i)\,\mathrm{d}\lambda. \tag{10}$$

Recall that our objective (3), with z fixed, is

$$g(\phi) := \int_{-M}^{M} V^{\pi^{\phi}_{\lambda}}_{\lambda}(\mu)\,\mathrm{d}\lambda. \tag{11}$$

Since μ is uniform over states, (10) equals $dg(\phi)$. The factor d arises because we average over initial states, whereas (Nakhleh et al., 2022) sums over them. The policy π^{ϕ}_{λ} varies with λ inside both integrals.

Next, we compare the gradients when the components of ϕ are distinct, as assumed in their Theorem 2. Recall our definition $R_{i,\phi} = \{s_j : \phi_j \leq \phi_i\}$, the passive set of $\pi^{\phi}_{\phi_i}$. Their Theorem 2 involves the discounted visitation measure and the Q-function, but does not explicitly specify the visitation measure's policy (visitation measure depends on its policy) and the meaning of Q-function (whether its optimal or not; if not, what is the specific policy used?). If the visitation measure in their Eq. (11) is interpreted as that under the policy with passive set $R_{i,\phi} \setminus \{s_i\}$, while the Q-functions are evaluated under the policy with passive set $R_{i,\phi}$, then their expression, in our notation and cost convention, becomes

$$d \sum_{i=1}^{d} \rho_{\mu}^{R_{i,\phi}\setminus\{s_i\}}(s_i)\big(Q^{R_{i,\phi}}_{\phi_i}(s_i, 1) - Q^{R_{i,\phi}}_{\phi_i}(s_i, 0)\big)\nabla_\theta \phi_i, \tag{12}$$

Algorithm 1 Online environment-representation oracle

Require: Recent transitions $\mathcal{T}_{t,K}$ and estimator $\mathcal{X}_K$
1: **if** domain knowledge is available **then**
2: $\quad \widehat{z}_t \leftarrow \mathcal{X}_K^{\text{manual}}(\mathcal{T}_{t,K})$
3: **else**
4: $\quad \widehat{z}_t \leftarrow \mathcal{X}_K^{\text{enc}}(\mathcal{T}_{t,K})$
5: **end if**
6: **return** $\widehat{z}_t$

where $\rho_\mu^{R_{i,\phi}\setminus\{s_i\}}(s_i)$ denotes the unnormalized discounted visitation measure from initial distribution μ, as in Lemma E.1. In particular, the visitation measure and Q-functions here use policies that differ at s_i. By Lemma E.1,

$$\rho_\mu^{R_{i,\phi}\setminus\{s_i\}}(s_i)\big(Q_{\phi_i}^{R_{i,\phi}}(s_i,1)-Q_{\phi_i}^{R_{i,\phi}}(s_i,0)\big)=V_{\phi_i}^{R_{i,\phi}\setminus\{s_i\}}(\mu)-V_{\phi_i}^{R_{i,\phi}}(\mu).$$

Thus, the gradient in (12) can be rewritten as

$$d\sum_{i=1}^{d}\big(V_{\phi_i}^{R_{i,\phi}\setminus\{s_i\}}(\mu)-V_{\phi_i}^{R_{i,\phi}}(\mu)\big)\nabla_\theta\phi_i.$$

Under this interpretation of the visitation measure, the expression equals $d\,\nabla_\theta g(\phi_\theta(\cdot,z),z)$, i.e., d times the fixed-environment integrand of (6). Note that this equivalence requires the interpretation above, which is not explicit in their theorem statement.

Difference: After deriving this gradient, the RL method developed in (Nakhleh et al., 2022) is no longer an actor-critic method because it uses the optimal Q-function rather than the Q-function for a specific policy. Specifically, the RL pipeline in (Nakhleh et al., 2022) does not learn the action-value function $Q_{\phi_i}^{R_{i,\phi}}(s_i,1)$ and $Q_{\phi_i}^{R_{i,\phi}}(s_i,0)$ under a specific policy. Instead, it uses the difference between the optimal Q-values $Q_{\phi_i}^*(s_i,1)$ and $Q_{\phi_i}^*(s_i,0)$ to update the parameters. This is evident in Equation (12) of (Nakhleh et al., 2022), where the update is based on the optimal Bellman equation rather than an approximation of the policy-specific action-value functions $Q_{\phi_i}^{R_{i,\phi}}(s_i,\cdot)$. Moreover, in its released code, it also uses the optimal Q-learning to learn the algorithm[1]. Replacing $Q_{\phi_i}^{R_{i,\phi}}(s_i,1)$ and $Q_{\phi_i}^{R_{i,\phi}}(s_i,0)$ in (12) by $Q_{\phi_i}^*(s_i,1)$ and $Q_{\phi_i}^*(s_i,0)$ will produce a value that is different from the policy gradient based on the current policy. Consequently, the algorithm in (Nakhleh et al., 2022) is not an actor-critic method based on the policy gradient (12), but is instead essentially another instance of Whittle Q-learning.

D WAC TRAINING PIPELINE

For WAC, we abstract online environment identification by an oracle. Given the most recent K transitions,

$$\mathcal{T}_{t,K}:=\{(s_\tau,a_\tau,c_\tau,s_{\tau+1})\}_{\tau=t-K}^{t-1},\qquad \widehat{z}_t=\mathcal{X}_K(\mathcal{T}_{t,K}).$$

The oracle can be implemented either by a hand-crafted estimator based on domain knowledge or by the permutation-invariant environment encoder described in Section 3.4. The following algorithm uses only $\widehat{z}_t$ and does not require direct access to the underlying environment.

For $s_j\in\mathcal{S}$, let

$$\lambda_j:=\phi_\theta(s_j,\widehat{z}),\qquad \lambda_{i,j}:=\phi_\theta(s_j,\widehat{z}_i),\qquad R_{j,\phi_\theta}^{\widehat{z}}:=\{s\in\mathcal{S}:\phi_\theta(s,\widehat{z})\le\lambda_j\}.$$

For each s_j, WAC estimates the state-initialized values

$$V_{i,j}^1:=V_{\lambda_{i,j}}^{R_{j,\phi_\theta}^{\widehat{z}_i}\setminus\{s_j\}}(s_j,\widehat{z}_i),\qquad V_{i,j}^0:=V_{\lambda_{i,j}}^{R_{j,\phi_\theta}^{\widehat{z}_i}}(s_j,\widehat{z}_i).$$

[1] In the released code at `https://github.com/khalednakhleh/deeptop`, the critic is updated by optimal Q-learning with a greedy target: in `RMAB/DeepTOP.py`, `update_policy()`, the next action is chosen as $\arg\max_a Q(s',a)$ (line 101).

Algorithm 2 Parameter-driven Whittle actor-critic

Require: Index network ϕ_θ, value network V_ψ, environment batch size B, oracle $\mathcal{X}_K$, learning rates η_ψ, η_θ, and discount factor γ

for each training iteration **do**
  Sample environments $\{\mathcal{E}(z_i)\}_{i=1}^{B}$
  for $i = 1$ to B **do**
    $\widehat{z}_i \leftarrow \mathcal{X}_K(\mathcal{T}_{i,t,K})$
    for $s_j \in \mathcal{S}$ **do**
      collect on-policy transitions under both policies with passive sets $R^{\widehat{z}_i}_{j,\phi_\theta}$ and $R^{\widehat{z}_i}_{j,\phi_\theta} \setminus \{s_j\}$
      Estimate $V^1_{i,j}$ and $V^0_{i,j}$ under passive sets $R^{\widehat{z}_i}_{j,\phi_\theta} \setminus \{s_j\}$ and $R^{\widehat{z}_i}_{j,\phi_\theta}$
      Update ψ using on-policy TD data
    end for
  end for
  Enhance $V^1_{i,j}$ and $V^0_{i,j}$ with rollout estimates as in (14)
  $g_\theta \leftarrow \frac{1}{B}\sum_{i=1}^{B}\sum_{s_j\in\mathcal{S}}(V^1_{i,j} - V^0_{i,j})\nabla_\theta\phi_\theta(s_j, \widehat{z}_i)$
  $\theta \leftarrow \theta - \eta_\theta g_\theta$
end for

To combine on-policy temporal-difference updates with rollout estimates, we start each rollout from s_j and follow the policy with passive set $R^{\widehat{z}_i}_{j,\phi_\theta}$ for T steps. For the m-th trajectory, the truncated rollout estimate accumulates the observed discounted costs and uses the critic network to bootstrap the remaining cost-to-go:

$$\widehat{V}^{R^z_{j,\phi_\theta},m}_{\lambda_j(z)}(s_j, z) := \sum_{t=0}^{T-1} \gamma^t \big(c_z(s^m_t, a^m_t) + \lambda_j(z) a^m_t\big) + \gamma^T V_\psi(s^m_T, R^z_{j,\phi_\theta}, \lambda_j(z), z), \qquad s^m_0 = s_j. \tag{13}$$

The observed-cost portion is unbiased for the first T rollout steps, whereas the bootstrap contribution is discounted. We average M independent trajectories to reduce variance:

$$\widehat{V}^{R^z_{j,\phi_\theta}}_{\lambda_j(z)}(s_j, z) := \frac{1}{M}\sum_{m=1}^{M} \widehat{V}^{R^z_{j,\phi_\theta},m}_{\lambda_j(z)}(s_j, z). \tag{14}$$

The same construction gives the estimate for the passive set $R^z_{j,\phi_\theta} \setminus \{s_j\}$. The resulting estimates are used in Algorithm 2.

D.1 Training details

For the training of the environment encoder, we train a separate encoder for each RMAB task. The encoder consists of two hidden layers with width 64, followed by mean pooling and a linear projection to a five-dimensional latent representation. We train it with Adam using learning rate 10^{-4} for 8000 iterations. We use the VICReg objective in (Bardes et al., 2022) to prevent the representation collapse. Specifically, we minimize the following loss function

$$\mathcal{L}_{\text{VICReg}} = \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{var}},$$

where

$$\mathcal{L}_{\text{inv}} = \text{MSE}(z_1, z_2),$$

$$\mathcal{L}_{\text{var}} = \frac{1}{d_z}\sum_{j=1}^{d_z}\left[\text{ReLU}(1 - \text{std}(z_{1,j})) + \text{ReLU}(1 - \text{std}(z_{2,j}))\right],$$

Here d_z denotes the dimension of the environment representation. Here, z_1 and z_2 are representations of two different contexts of the same environment, and $std(z_{1,j})/std(z_{2,j})$ is the standard deviation of the j-th components of z_1/z_2 within a batch of environments. After training, we freeze the encoder and use it to generate environment representations. Here we do not use the covariance regularization term in (Bardes et al., 2022). We make this choice empirically, as adding the covariance regularization term leads to a slight performance degradation in our experiments.

As for the training of the index and value/Q networks, we sample 5 environments per iteration, Adam learning rates 3×10^{-4} for the index network and 10^{-4} for the value/Q network, and an EMA coefficient of 0.999. WAC uses two value updates per policy iteration with batch size 16, then followed by a rollout with horizon 24, and 32 rollout trajectories. WQL uses 25 Q-function updates per policy iteration and batch size 32. The discount factor is set to 0.9.

E CHANGING THE INITIAL DISTRIBUTION

We will show that $V_{\lambda_j(z)}^{R_{j,\phi_\theta}^z \setminus \{s_j\}}(\mu, z)$ and $V_{\lambda_j(z)}^{R_{j,\phi_\theta}^z}(\mu, z)$ have the same sign. We first need the following lemma, which quantifies the difference between the value functions of two policies whose passive sets differ at a single state. We will frequently invoke this lemma in the subsequent proofs. This result follows directly from the performance difference lemma (Kakade & Langford, 2002, Lemma 6.1). Let $Q_\lambda^P(s,a)$ denote the Q-function starting from state s and action a under per-arm MDP $\mathcal{E}(\lambda)$ and the policy with passive set P. Let $V^M(\nu, \lambda)$ be the value function with passive set M and under MDP $\mathcal{E}(\lambda)$ with the initial state follows the distribution ν. For completeness, we present its proof here.

Lemma E.1. *Let $P \subseteq S$ and $x \notin P$. Then,*

$$V^{P\cup\{x\}}(\nu,\lambda) - V^P(\nu,\lambda) = \rho_\nu^{P\cup\{x\}}(x)\left(Q_\lambda^P(x,0) - Q_\lambda^P(x,1)\right), \tag{15}$$

where $\rho_\nu^{P\cup\{x\}}(x) := \sum_{t=0}^\infty \gamma^t \Pr(s_t = x \mid s_0 \sim \nu, \pi^{P\cup\{x\}})$, i.e., the unnormalized discounted state visitation distribution of x under the policy with passive set $P \cup \{x\}$. Suppose $y \in P$, then similarly

$$V^{P\setminus\{y\}}(\nu,\lambda) - V^P(\nu,\lambda) = \rho_\nu^{P\setminus\{y\}}(y)\left(Q_\lambda^P(y,1) - Q_\lambda^P(y,0)\right). \tag{16}$$

Proof. Suppose there are two policies $\tilde{\pi}$ and π, then the performance difference lemma (Kakade & Langford, 2002, Lemma 6.1) states that

$$V^{\tilde{\pi}}(\nu) - V^\pi(\nu) = \sum_s \rho_\nu^{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s)\left(Q^\pi(s,a) - V^\pi(s)\right),$$

where $V^\pi(\nu)$ is the value function under policy π with initial distribution ν, $Q^\pi(s,a)$ is the Q-function under policy π, $V^\pi(s)$ is the value function at state s under policy π, and $\rho_\nu^{\tilde{\pi}}(s)$ is the unnormalized discounted state visitation distribution of s with initial distribution ν under policy $\tilde{\pi}$.

Let $\tilde{\pi}$ be the policy with passive set $P \cup \{x\}$ and π be the policy with passive set P. Then, under the MDP $\mathcal{E}(\lambda)$, we have

$$\begin{aligned}
V^{P\cup\{x\}}(\nu,\lambda) - V^P(\nu,\lambda) &= \sum_s \rho_\nu^{P\cup\{x\}}(s) \sum_a \tilde{\pi}(a|s)\left(Q_\lambda^P(s,a) - V^P(s)\right) \\
&= \rho_\nu^{P\cup\{x\}}(x) \sum_a \tilde{\pi}(a|x)\left(Q_\lambda^P(x,a) - V^P(x)\right) \\
&= \rho_\nu^{P\cup\{x\}}(x)\left(Q_\lambda^P(x,0) - Q_\lambda^P(x,1)\right),
\end{aligned}$$

where the second equality follows from the fact that for any $s \neq x$, $\tilde{\pi}(a|s) = \pi(a|s)$, and thus $\sum_a \tilde{\pi}(a|s)\left(Q_\lambda^P(s,a) - V^P(s)\right) = 0$. The last equality follows from the fact that under policy $\tilde{\pi}$, action 0 is always chosen at state x and $V^P(x) = Q_\lambda^P(x,1)$ since under policy π, action 1 is always chosen at state x. Thus, (15) is proved. The proof of (16) is similar and we omit it.

□

We are now ready to show that $V_{\lambda_j(z)}^{R_{j,\phi_\theta}^z \setminus \{s_j\}}(\mu, z)$ and $V_{\lambda_j(z)}^{R_{j,\phi_\theta}^z}(\mu, z)$ have the same sign. For fixed z and s_j, let $P = R_{j,\phi_\theta}^z$. For either initial distribution $\nu \in \{\mu, \delta_{s_j}\}$, define

$$\Delta_j^\nu(z) := V_{\lambda_j(z)}^{P\setminus\{s_j\}}(\nu, z) - V_{\lambda_j(z)}^P(\nu, z). \tag{17}$$

By Lemma E.1,

$$\Delta_j^\nu(z) = \rho_\nu^{P\setminus\{s_j\}}(s_j)\left(Q_{\lambda_j(z)}^P(s_j,1) - Q_{\lambda_j(z)}^P(s_j,0)\right). \tag{18}$$

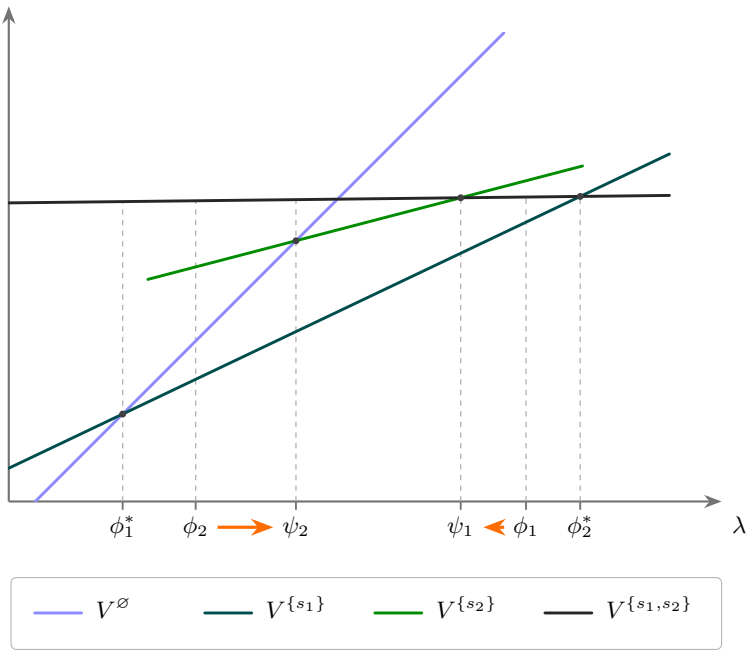


Figure 3: ϕ_i^* is the Whittle index of state s_i, $i = 1, 2$. Starting from (ϕ_1, ϕ_2) with $\phi_2 < \phi_1$, the iterates converge to (ψ_1, ψ_2), which is not the Whittle-index vector.

Thus, the value difference under the uniform initial distribution and the value difference under the state-conditioned initial distribution contain exactly the same Q-value difference. They differ only through the discounted visitation factor, which is positive. Consequently, $V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}\setminus\{s_j\}}(\mu, z)$ and $V_{\lambda_j(z)}^{R^z_{j,\phi_\theta}}(\mu, z)$ have the same sign.

F Why a Stationary Point Need Not Be the Whittle Index

This appendix motivates why the stationary-point landscape of the WAC objective is nontrivial, even for an indexable problem.

To better understand the problem, let us first assume that the initial distribution is all uniform and omit this dependence in notation, i.e., we write $V_\lambda^R(\mu, z)$ as $V^R(\lambda)$. Also, we focus on a fixed environment z and its objective $g(\phi, z)$. Suppose that the Whittle index of state s_1 is smaller than that of s_2, i.e., $\phi_1^* < \phi_2^*$. However, suppose that the initial predictions ϕ_1, ϕ_2 have the wrong order $\phi_1 > \phi_2$. Then, our WAC method will update ϕ_1 based on $\phi_1 \leftarrow \phi_1 - \alpha(V^{\{s_2\}}(\lambda) - V^{\{s_1,s_2\}}(\lambda))$, update ϕ_2 based on $\phi_2 \leftarrow \phi_2 - \alpha(V^{\emptyset}(\lambda) - V^{\{s_2\}}(\lambda))$. However, if these value functions are illustrated in Fig. 3, so that $V^{\emptyset}$ and $V^{\{s_2\}}$ intersect at ψ_2, and $V^{\{s_2\}}$ and $V^{\{s_1,s_2\}}$ intersect at ψ_1. Then, our WAC method will at best push (ϕ_1, ϕ_2) to this stationary point (ψ_1, ψ_2), even though the latter are the wrong *Whittle* indices. An illustration of this is in Figure 3.

Fortunately, the scenario depicted in Fig. 3—though plausible at first glance—cannot occur in a two-state MDP, as proved in the following subsection. For general state spaces, under the active-time condition, Theorem 4.1 establishes an analogous landscape guarantee.

F.1 Two-state Case

For a general two-state MDP, the following result rules out the problematic case illustrated in Figure 3. Recall that $V^M(\lambda)$ denotes the value function of the MDP $\mathcal{E}(\lambda)$ under the policy with passive set M and a uniform initial-state distribution. We use the following convention throughout: when the initial distribution is uniform, we either omit it from the notation or denote it by μ; when it is non-uniform, we always display it explicitly.

Theorem F.1. *For any two-state MDP, let u denote the intersection point of $V^{\emptyset}$ and $V^{\{s_1,s_2\}}$. Then one of the following must hold:*

$$V^{\{s_1\}}(u) < V^{\emptyset}(u) = V^{\{s_1,s_2\}}(u) \quad \textit{and} \quad V^{\{s_2\}}(u) > V^{\emptyset}(u) = V^{\{s_1,s_2\}}(u), \tag{19}$$

or

$$V^{\{s_1\}}(u) > V^{\emptyset}(u) = V^{\{s_1,s_2\}}(u) \quad \textit{and} \quad V^{\{s_2\}}(u) < V^{\emptyset}(u) = V^{\{s_1,s_2\}}(u), \tag{20}$$

or

$$V^{\{s_1\}}(u) = V^{\emptyset}(u) = V^{\{s_1,s_2\}}(u) = V^{\{s_2\}}(u). \tag{21}$$

We illustrate these three cases in Figure 4. By Theorem F.1, the problematic scenario discussed in Section 4 cannot occur in any two-state MDP. Moreover, for each of the three cases shown, it is not hard to see that any initialization (ϕ_1, ϕ_2) converges to the correct Whittle index; that is, the Whittle index is the unique stationary point.

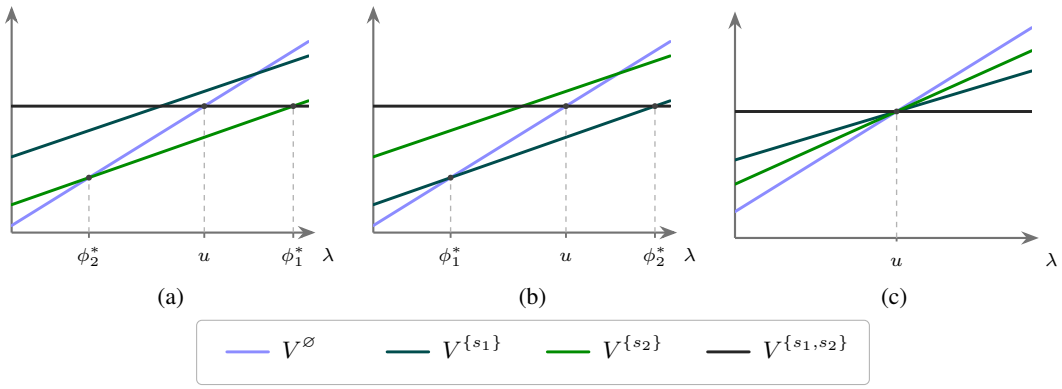


Figure 4: Visualizations of the three possible cases. In panels (a) and (b), ϕ_1^* and ϕ_2^* are the Whittle indices for state s_1 and s_2, respectively, and u is the intersection point of $V^{\varnothing}$ and $V^{\{s_1,s_2\}}$. In panel (c), s_1 and s_2 share the same Whittle index u.

We now proceed to prove Theorem F.1.

Proof of Theorem F.1. Let s_1 and s_2 denote the two states. Suppose that $V^{s_1}(u) < V^{\emptyset}(u) = V^{\{s_1,s_2\}}(u)$. Then, by Lemma E.1, we have

$$0 > V^{s_1}(u) - V^{\emptyset}(u) = \rho^{s_1}(s_1)(Q_u^{\emptyset}(s_1,0) - Q_u^{\emptyset}(s_1,1)),$$

Since $\rho^{s_1}(s_1) > 0$ under the uniform initial distribution, we obtain

$$Q_u^{\emptyset}(s_1,0) - Q_u^{\emptyset}(s_1,1) < 0. \tag{22}$$

Using Lemma E.1 again, we have

$$0 = V^{s_1,s_2}(u) - V^{\emptyset}(u) = \rho^{s_1,s_2}(s_2)(Q_u^{\emptyset}(s_2,0) - Q_u^{\emptyset}(s_2,1)) + \rho^{s_1,s_2}(s_1)(Q_u^{\emptyset}(s_1,0) - Q_u^{\emptyset}(s_1,1)).$$

Combining this with (22), we have

$$Q_u^{\emptyset}(s_2,0) - Q_u^{\emptyset}(s_2,1) > 0. \tag{23}$$

We also use Lemma E.1 to rewrite $V^{s_2}(u) - V^{\emptyset}(u)$ as

$$V^{s_2}(u) - V^{\emptyset}(u) = \rho^{s_2}(s_2)(Q_u^{\emptyset}(s_2,0) - Q_u^{\emptyset}(s_2,1)).$$

Combining this with (23), we have $V^{s_2}(u) - V^{\emptyset}(u) > 0$. Thus, we have shown that if $V^{s_1}(u) < V^{\emptyset}(u) = V^{\{s_1,s_2\}}(u)$, then $V^{s_2}(u) > V^{\emptyset}(u) = V^{\{s_1,s_2\}}(u)$, which is exactly case (19).

The other two cases $V^{s_1}(u) > V^{\emptyset}(u) = V^{\{s_1,s_2\}}(u)$ and $V^{s_1}(u) = V^{\emptyset}(u) = V^{\{s_1,s_2\}}(u)$ can be proved similarly, and we omit the details here. □

G PROOF OF THEOREM 4.1

We first give the precise statement of Theorem 4.1. The assumption we need is an active-time condition. To start with, the active time for a deterministic policy ξ is defined as

$$\mathrm{AT}(\xi) \;:=\; \mathbb{E}_{s_0\sim\mathrm{Unif}(\mathcal{S})}\left[\sum_{t=0}^{\infty}\gamma^t\,\mathbf{1}\{\xi(s_t)=1\}\right], \tag{24}$$

where the trajectory $\{s_t\}_{t\geq 0}$ follows the dynamics induced by the policy ξ. We now introduce the active time condition.

Assumption G.1. The MDP is said to satisfy the *active time condition* if, for every deterministic policy π, the policy $\tilde{\pi}$ that coincides with π except at a single state s (where $\tilde{\pi}(s)=1$ and $\pi(s)=0$) strictly increases the active time:

$$\mathrm{AT}(\tilde{\pi}) \;>\; \mathrm{AT}(\pi). \tag{25}$$

It is a sufficient condition for indexability (see Appendix H), and we have shown in Appendix H that it holds for several classes of arms.

The proof follows the same spirit as the two-state case presented in Section F.1 and readers are encouraged to consult it first for intuition.

We divide the proof into two parts. The first part shows that on the differentiable region $\mathcal{D}_M$, there is no stationary point other than the Whittle index. The second part shows that on $(-M,M)^d\setminus\mathcal{D}_M$, there is no local minimum other than the Whittle index. Combining the two parts together leads to Theorem 4.1. Recall that $V^P(\mu,\lambda)$ denotes the value function under the MDP $\mathcal{E}(\lambda)$ with passive set P and uniform initial distribution μ. When the context is clear, we may omit the explicit dependence on μ.

G.1 PROOF OF THE DIFFERENTIABLE REGION

Recall that the differentiable region is $\mathcal{D}_M=\{\phi\in(-M,M)^d:\phi_i\neq\phi_j \text{ for all } i\neq j\}$. We first introduce an auxiliary definition.

Definition G.2. Given any ϕ, an induced ordering $\mathcal{C}:\{1,2,\dots,d\}\to\mathcal{S}$ is a bijection such that $\phi(\mathcal{C}(1))\leq\cdots\leq\phi(\mathcal{C}(d))$.

Intuitively, $\mathcal{C}$ simply orders the states by non-decreasing ϕ-value.

We next introduce the notion of consistent point. Recall that $R_{0,\phi}:=\emptyset$ and $R_{i,\phi}:=\{s_j\mid\phi_j\leq\phi_i\}$ for $1\leq i\leq d$. When there is no confusion, we will drop the subscript ϕ for simplicity. Thus, we simply use R_i to denote $R_{i,\phi}$ for $0\leq i\leq d$.

Definition G.3. Under Assumption G.1, given any ϕ and one of its induced orderings $\mathcal{C}$, define

$$P_0^{\mathcal{C},\phi}:=\emptyset,\qquad P_k^{\mathcal{C},\phi}:=P_{k-1}^{\mathcal{C},\phi}\cup\{\mathcal{C}(k)\}=\{\mathcal{C}(1),\dots,\mathcal{C}(k)\},\qquad 1\leq k\leq d. \tag{26}$$

We refer to the sequence $\{P_k^{\mathcal{C},\phi}\}_{k=0}^d$ as the *passive set sequence* of ϕ and $\mathcal{C}$. We say that ϕ is a *consistent point* if the following holds: for any of its induced orderings $\mathcal{C}$ and any $1\leq k\leq d$, the intersection point of the two lines (with respect to λ) $V^{P_{k-1}^{\mathcal{C},\phi}}$ and $V^{P_k^{\mathcal{C},\phi}}$ is $\phi(\mathcal{C}(k))$.

Remark G.4. Note that under Assumption G.1, V^{P_k} and $V^{P_{k+1}}$ intersect at exactly one point. Thus, each intersection used in the definition is well defined.

For simplicity, when there is no confusion, we will simply write $\{P_k\}_{k=1}^d$ to denote the passive set sequence $\{P_k^{\mathcal{C},\phi}\}_{k=1}^d$ of ϕ and $\mathcal{C}$. We provide an illustration of the passive set sequence in Fig. 5.

The next lemma shows why we study consistent points.

Lemma G.5. *Every stationary point in $\mathcal{D}_M$ is a consistent point.*

Proof. Let $\phi\in\mathcal{D}_M$ be stationary. Since its components are distinct, ϕ has a unique induced ordering $\mathcal{C}$. For each j, write $\mathcal{C}(j)=s_{i_j}$. Applying (4) to the state index i_j gives

$$V^{R_{i_j,\phi}\setminus\{s_{i_j}\}}(\mu,\phi_{i_j})-V^{R_{i_j,\phi}}(\mu,\phi_{i_j})=0,\qquad 1\leq j\leq d.$$

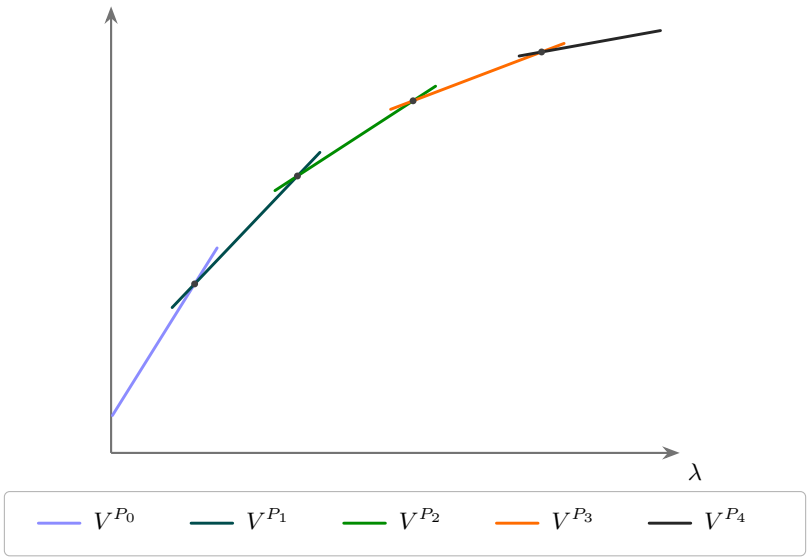


Figure 5: An illustration of the passive set sequence. As k increases, the slope of V^{P_k} decreases.

Because $\mathcal{C}$ orders the distinct components increasingly,

$$R_{i_j,\phi} = P_j^{\mathcal{C},\phi}, \qquad R_{i_j,\phi} \setminus \{s_{i_j}\} = P_{j-1}^{\mathcal{C},\phi}.$$

Thus $V^{P_{j-1}^{\mathcal{C},\phi}}$ and $V^{P_j^{\mathcal{C},\phi}}$ intersect at $\phi_{i_j} = \phi(\mathcal{C}(j))$ for every j. Hence ϕ is consistent. □

The proof strategy now is clear: to prove that on the differentiable region $\mathcal{D}_M$, there is no stationary point other than the Whittle index, it suffices to show that there is no consistent point other than the Whittle index. Before presenting the main proof, we first prove some supporting lemmas that will be used in the proof.

Building on Lemma E.1, we now prove a vital technical lemma.

Lemma G.6. *Let $A \subset \mathcal{S}$ be an arbitrary passive set and let $b, c \in \mathcal{S} \setminus A$ be two distinct states. Let u satisfy $V^{A}(u) = V^{A\cup\{b,c\}}(u)$. If $V^{A\cup\{b\}}(u) \le V^{A\cup\{b,c\}}(u)$, then $V^{A\cup\{c\}}(u) \ge V^{A\cup\{b,c\}}(u)$. Moreover, if $V^{A\cup\{b\}}(u) = V^{A\cup\{b,c\}}(u)$, then $V^{A\cup\{c\}}(u) = V^{A\cup\{b,c\}}(u)$.*

Proof. Suppose not, then $V^{A\cup\{c\}}(u) < V^{A\cup\{b,c\}}(u)$, by performance difference lemma, we have

$$Q_u^{A\cup\{b,c\}}(b,1) - Q_u^{A\cup\{b,c\}}(b,0) < 0.$$

Since $V^{A\cup\{b\}}(u) \le V^{A\cup\{b,c\}}(u)$. Thus by performance difference lemma, we have

$$Q_u^{A\cup\{b,c\}}(c,1) - Q_u^{A\cup\{b,c\}}(c,0) \le 0.$$

Again, by Lemma E.1,

$$\begin{aligned} V^{A}(u) - V^{A\cup\{b,c\}}(u) &= \rho^{A}(b)\big(Q_u^{A\cup\{b,c\}}(b,1) - Q_u^{A\cup\{b,c\}}(b,0)\big) \\ &\quad + \rho^{A}(c)\big(Q_u^{A\cup\{b,c\}}(c,1) - Q_u^{A\cup\{b,c\}}(c,0)\big) \\ &< 0, \end{aligned}$$

where the strict inequality follows from the preceding two inequalities and the positivity of the visitation factors. Indeed, because the initial distribution is uniform and has full support, $\rho^{\pi}(z) \ge \Pr(s_0 = z) > 0$ for every state z and every policy π. This contradicts the definition of u.

For the equality statement, suppose $V^{A\cup\{b\}}(u) = V^{A\cup\{b,c\}}(u)$. Define

$$\Delta_z := Q_u^{A\cup\{b,c\}}(z,1) - Q_u^{A\cup\{b,c\}}(z,0), \qquad z \in \{b,c\}.$$

Lemma E.1 gives

$$0 = \rho^{A\cup\{b\}}(c)\Delta_c.$$

The visitation factor is positive, so $\Delta_c = 0$. Using $V^A(u) = V^{A\cup\{b,c\}}(u)$ and applying the performance-difference lemma to the two states b, c now yields

$$0 = \rho^A(b)\Delta_b + \rho^A(c)\Delta_c = \rho^A(b)\Delta_b.$$

Thus $\Delta_b = 0$, and a final application of Lemma E.1 gives

$$V^{A\cup\{c\}}(u) - V^{A\cup\{b,c\}}(u) = \rho^{A\cup\{c\}}(b)\Delta_b = 0.$$

This proves the equality statement. □

We now define the notion of a triangle triple, which will be used in deriving the contradiction.

Definition G.7. We call a tuple of three straight lines (l_1, l_2, l_3) a triangle triple if, at the intersection point i of ℓ_1 and l_3, we have $l_2(i) \le l_1(i) = l_3(i)$.

The following lemma states that if we know one triangle triple, we can induce another triangle triple based on it.

Lemma G.8 (Consecutive-chain triangle)**.** *Let ϕ be consistent with passive-set sequence $P_0 \subset P_1 \subset \cdots \subset P_n$. For every $0 \le k \le d-2$,*

$$\left(V^{P_k}, V^{P_{k+1}}, V^{P_{k+2}}\right)$$

is a triangle triple.

Proof. Let a and b be the respective intersections of the first two and last two lines. Consistency gives $a \le b$. Assumption G.1 strictly orders their slopes. At a, the first line equals the middle line and lies below the third; at b, the first line lies above the equal middle and third lines. Hence the intersection u of the first and third lines lies in $[a, b]$. Since $u \ge a$, the middle line is no larger than the first line at u, which is exactly the triangle-triple property. □

Lemma G.9. *Under Assumption G.1, let ϕ be a consistent point with passive set sequence $\{P_k\}_{k=0}^d$. For any $1 \le k \le d-2$ and $x \in \mathcal{S} \setminus P_{k+1}$, let u_1 and u_2 satisfy $V^{P_k}(u_1) = V^{P_{k+1}\cup\{x\}}(u_1)$ and $V^{P_{k-1}}(u_2) = V^{P_k\cup\{x\}}(u_2)$, respectively. If $V^{P_{k+1}}(u_1) \le V^{P_{k+1}\cup\{x\}}(u_1)$, then $V^{P_k}(u_2) \le V^{P_k\cup\{x\}}(u_2)$.*

In other words, for $1 \le k \le d-2$ and $x \in \mathcal{S} \setminus P_{k+1}$, if $(V^{P_k}, V^{P_{k+1}}, V^{P_{k+1}\cup\{x\}})$ is a triangle triple, then so is $(V^{P_{k-1}}, V^{P_k}, V^{P_k\cup\{x\}})$.

Proof. Write

$$\{y\} := P_{k+1} \setminus P_k.$$

Let s be the intersection of V^{P_k} and $V^{P_{k+1}}$, and let t be the intersection of $V^{P_{k-1}}$ and V^{P_k}. Consistency gives $t \le s$.

At u_1, the assumed triangle inequality gives

$$V^{P_{k+1}}(u_1) \le V^{P_k}(u_1) = V^{P_{k+1}\cup\{x\}}(u_1).$$

Since $V^{P_k} - V^{P_{k+1}}$ is strictly increasing and vanishes at s, we have $s \le u_1$. Hence $t \le u_1$.

Applying Lemma G.6 with $A = P_k$, $b = y$, and $c = x$ yields

$$V^{P_k}(u_1) = V^{P_k\cup\{x,y\}}(u_1) \le V^{P_k\cup\{x\}}(u_1).$$

Let v be the unique intersection of V^{P_k} and $V^{P_k\cup\{x\}}$. Since their difference is strictly increasing, the preceding inequality implies $u_1 \le v$. Thus $t \le v$.

Using the definition of v,

$$\begin{aligned} V^{P_{k-1}}(v) - V^{P_k\cup\{x\}}(v) &= V^{P_{k-1}}(v) - V^{P_k}(v) \\ &\ge 0, \end{aligned}$$

where the inequality follows from $v \geq t$. Since $V^{P_{k-1}} - V^{P_k \cup \{x\}}$ is strictly increasing and vanishes at u_2, it follows that $u_2 \leq v$.

Finally, $V^{P_k} - V^{P_k \cup \{x\}}$ is strictly increasing and vanishes at v. Combining this with the previous $u_2 \leq v$, we have

$$V^{P_k}(u_2) \leq V^{P_k \cup \{x\}}(u_2) = V^{P_{k-1}}(u_2),$$

which proves the desired triangle-triple inequality. □

Repeatedly applying Lemma G.9, we now obtain the following lemma.

Lemma G.10. *Under Assumption G.1, let ϕ be a consistent point with passive set sequence $\{P_k\}_{k=0}^{d}$. Let r and l be integers such that $0 \leq r \leq l \leq d-2$, and define the state x by $\{x\} := P_{l+2} \setminus P_{l+1}$. If $(V^{P_l}, V^{P_{l+1}}, V^{P_{l+1} \cup \{x\}})$ is a triangle triple, then so is $(V^{P_r}, V^{P_{r+1}}, V^{P_{r+1} \cup \{x\}})$.*

Proof. We proceed by backward induction on the index from l down to r. By Lemma G.9, the fact that $(V^{P_l}, V^{P_{l+1}}, V^{P_{l+1} \cup \{x\}})$ is a triangle triple implies that $(V^{P_{l-1}}, V^{P_l}, V^{P_l \cup \{x\}})$ is also a triangle triple. Applying Lemma G.9 repeatedly, we obtain a chain of implications:

$$(V^{P_l}, V^{P_{l+1}}, V^{P_{l+1} \cup \{x\}}) \Rightarrow (V^{P_{l-1}}, V^{P_l}, V^{P_l \cup \{x\}}) \Rightarrow \cdots \Rightarrow (V^{P_r}, V^{P_{r+1}}, V^{P_{r+1} \cup \{x\}}).$$

Since $l - r$ applications of Lemma G.9 suffice to reach index r, the desired triple is indeed a triangle triple. □

In the following two lemmas, we show that if a consistent point ϕ has an induced ordering that is the same as that of the Whittle index, then ϕ must be the Whittle index itself.

Lemma G.11 (Whittle-compatible chains)**.** *Call an ordering $\mathcal{C}$ Whittle-compatible if $I(\mathcal{C}(1)) \leq \cdots \leq I(\mathcal{C}(d))$. Suppose Assumption G.1 holds. Let $\mathcal{C}$ be any Whittle-compatible ordering and let $P_k^{\mathcal{C}} := \{\mathcal{C}(1), \ldots, \mathcal{C}(k)\}$. For every $k = 1, \ldots, d$, the unique intersection of $V^{P_{k-1}^{\mathcal{C}}}$ and $V^{P_k^{\mathcal{C}}}$ is $I(\mathcal{C}(k))$.*

Proof. Put $w = I(\mathcal{C}(k))$. For a finite discounted MDP, the optimal action-value functions are continuous in λ. Indexability and the definition $I(s) = \inf\{\lambda : s \in \mathcal{P}(\lambda)\}$ therefore imply that at $\lambda = I(s)$ both actions are optimal at s: passivity is optimal by closedness of the optimal-action correspondence, while activity remains optimal by continuity from prices below $I(s)$. Consequently, at price w, states with index strictly below w can be chosen passive, states with index strictly above w must be chosen active, and either action may be chosen at states with index exactly w. Any statewise selection among Bellman-optimal actions defines an optimal deterministic policy. In particular, the policies with passive sets $P_{k-1}^{\mathcal{C}}$ and $P_k^{\mathcal{C}}$ have the same value at w. Assumption G.1 makes their difference a strictly monotone affine function of the activation price, so this intersection is unique. □

Lemma G.12 (Correct ordering forces correct values)**.** *Let ϕ be a consistent point. If some induced ordering $\mathcal{C}$ of ϕ is also Whittle-compatible, then $\phi = I$ coordinatewise.*

Proof. For every k, consistency says that the unique intersection of $V^{P_{k-1}^{\mathcal{C}}}$ and $V^{P_k^{\mathcal{C}}}$ equals $\phi(\mathcal{C}(k))$. Lemma G.11 says that the same intersection equals $I(\mathcal{C}(k))$. Therefore $\phi(\mathcal{C}(k)) = I(\mathcal{C}(k))$ for every k, and hence $\phi = I$. □

By Lemma G.5, to show that there is no stationary point other than the Whittle index on $\mathcal{D}_M$, it suffices to prove the following stronger theorem.

Theorem G.13. *Under Assumption G.1, there is no consistent point other than the original Whittle index vector I.*

Proof. Suppose, for contradiction, that ϕ is a consistent point and $\phi \neq I$. By Lemma G.12, no induced ordering of ϕ can be Whittle-compatible.

Choose an induced ordering $\mathcal{C}$ of ϕ that orders states inside every ϕ-tie block by nondecreasing Whittle index. Let k be the first position at which $\mathcal{C}(k)$ does not have the smallest Whittle index among the

states not yet selected. Such a position exists, because otherwise $\mathcal{C}$ would be Whittle-compatible. Define

$$B := P^{\mathcal{C}}_{k-1}, \qquad y := \mathcal{C}(k),$$

and choose $x \notin B \cup \{y\}$ with the smallest Whittle index among the remaining states. Then

$$I(x) < I(y).$$

Furthermore, every state in B was selected with the smallest remaining Whittle index. Thus B is a Whittle-compatible prefix, and $B \cup \{x\}$ is the next Whittle-compatible prefix. In particular, by Lemma G.11, the unique intersection of V^B and $V^{B\cup\{x\}}$ is $I(x)$.

For brevity, write $P_j := P^{\mathcal{C}}_j$. Let $m = \mathcal{C}^{-1}(x) > k$, put $r = k-1$, and put $l = m-2$. Then $0 \le r \le l \le d-2$,

$$P_r = B, \qquad P_{r+1} = B \cup \{y\}, \qquad \{x\} = P_{l+2} \setminus P_{l+1}.$$

Consequently, $P_{l+1} \cup \{x\} = P_{l+2}$, and Lemma G.8 shows that

$$\left(V^{P_l}, V^{P_{l+1}}, V^{P_{l+1}\cup\{x\}}\right)$$

is a triangle triple. Applying Lemma G.10 with these r, l, x therefore gives that

$$\left(V^{B}, V^{B\cup\{y\}}, V^{B\cup\{x,y\}}\right)$$

is a triangle triple. If u is the intersection of V^B and $V^{B\cup\{x,y\}}$, then

$$V^{B\cup\{y\}}(u) \le V^B(u) = V^{B\cup\{x,y\}}(u). \tag{27}$$

Let q be the unique intersection of V^B and $V^{B\cup\{y\}}$. At price $I(x)$, the policy with passive set B is optimal, whereas passivating y is strictly suboptimal because $I(y) > I(x)$. Because the initial distribution is uniform and therefore has full support, the value inequality is strict:

$$V^B(I(x)) < V^{B\cup\{y\}}(I(x)).$$

By Assumption G.1, $V^B(\lambda) - V^{B\cup\{y\}}(\lambda)$ is strictly increasing in λ, and consequently $I(x) < q$. Equation (27) similarly implies $q \le u$. Therefore $I(x) < u$.

Since V^B and $V^{B\cup\{x\}}$ intersect at $I(x)$, another use of Assumption G.1 yields

$$V^{B\cup\{x\}}(u) < V^B(u) = V^{B\cup\{x,y\}}(u).$$

On the other hand, applying Lemma G.6 with $A = B$, $b = y$, and $c = x$ to (27) gives

$$V^{B\cup\{x\}}(u) \ge V^{B\cup\{x,y\}}(u),$$

a contradiction. Thus every consistent point equals the Whittle-index vector. □

G.2 Proof of the non-differentiable region

For each state index k, with a slight abuse of notation, we define the weak and strict threshold sets as

$$P_k := \{s_j \in \mathcal{S} : \phi_j \le \phi_k\} = R_{k,\phi}, \qquad O_k := \{s_j \in \mathcal{S} : \phi_j < \phi_k\} = L_{k,\phi}.$$

These state-indexed threshold sets should not be confused with the rank-prefix sets $P^{\mathcal{C},\phi}_k$ in Definition G.3. When several components are tied, P_k contains the entire tie block, whereas O_k contains none of that block.

Definition G.14. A component ϕ_k is a *tied component* if there exists $l \neq k$ such that $\phi_k = \phi_l$; otherwise it is an *untied component*.

Note that if ϕ satisfies the following

$$\frac{\partial g(\phi)}{\partial \phi_j^+} < 0 \quad \text{or} \quad \frac{\partial g(\phi)}{\partial \phi_j^-} > 0, \quad \exists 1 \le j \le d, \tag{28}$$

then it is not a local minimum. To see why, observe that at any such point, there exists at least one coordinate j for which either the right partial derivative is negative or the left partial derivative is positive. In either case, a sufficiently small perturbation of ϕ_j in the corresponding direction strictly decreases the objective value, contradicting local minimality.

Thus, it suffices to show that any non-differentiable point that is not the Whittle index satisfies (28). The proof proceeds by contradiction. Suppose there exists a non-differentiable point ϕ, distinct from the Whittle index, that satisfies:

$$\frac{\partial g(\phi)}{\partial \phi_j^+} \geq 0 \quad \text{and} \quad \frac{\partial g(\phi)}{\partial \phi_j^-} \leq 0, \quad \forall\, j \in \{1, \ldots, d\}. \tag{29}$$

We show that this leads to a contradiction.

We first translate the characterization (29) into the language of value function.

Lemma G.15. *If $\phi \in (-M, M)^d \setminus \mathcal{D}_M$ satisfies* (29)*, then at least one component is tied. Moreover, for every $1 \leq k \leq d$,*

$$V^{P_k \setminus \{s_k\}}(\phi_k) - V^{P_k}(\phi_k) \geq 0, \qquad V^{O_k}(\phi_k) - V^{O_k \cup \{s_k\}}(\phi_k) \leq 0.$$

Thus, if ϕ_k is an untied component, then $P_k \setminus \{s_k\} = O_k$, so both quantities above are equal and hence are zero.

Proof. Non-differentiability means that at least two components are tied. The two inequalities follow by substituting $P_k = R_{k,\phi}$ and $O_k = L_{k,\phi}$ into (4) and (5). □

We then show a technical lemma.

Lemma G.16. *Let $A \subseteq \mathcal{S}$ and let $\emptyset \neq D \subseteq \mathcal{S} \setminus A$. Let u be an intersection point of V^A and $V^{A \cup D}$. Define*

$$x_1 \in \arg\min_{x \in D} \big\{ V^{(A \cup D) \setminus \{x\}}(u) - V^{A \cup D}(u) \big\} \tag{30}$$

and

$$x_2 \in \arg\max_{x \in D} \big\{ V^{A}(u) - V^{A \cup \{x\}}(u) \big\}. \tag{31}$$

Then

$$V^{(A \cup D) \setminus \{x_1\}}(u) - V^{A \cup D}(u) \leq 0, \tag{32}$$

and

$$V^{A}(u) - V^{A \cup \{x_2\}}(u) \geq 0. \tag{33}$$

Moreover, if either (32) *or* (33) *holds with equality, then for every $E \subseteq D$,*

$$V^{A \cup E}(u) = V^{A}(u) = V^{A \cup D}(u). \tag{34}$$

Proof. For $x \in D$, put

$$\Delta_x := Q_u^{A \cup D}(x, 1) - Q_u^{A \cup D}(x, 0).$$

Lemma E.1 gives

$$V^{(A \cup D) \setminus \{x\}}(u) - V^{A \cup D}(u) = \rho^{(A \cup D) \setminus \{x\}}(x) \Delta_x.$$

All visitation factors are positive because the initial distribution is uniform and has full support. If the minimum in (30) were positive, then every Δ_x would be positive. The performance-difference identity applied simultaneously to the states in D would then give

$$V^{A}(u) - V^{A \cup D}(u) = \sum_{x \in D} \rho^{A}(x) \Delta_x > 0,$$

contradicting the definition of u. This proves (32).

For (33), put $\widehat{\Delta}_x := Q_u^{A}(x, 1) - Q_u^{A}(x, 0)$. Then

$$V^{A}(u) - V^{A \cup \{x\}}(u) = \rho^{A \cup \{x\}}(x) \widehat{\Delta}_x.$$

If the maximum in (31) were negative, then every $\widehat{\Delta}_x$ would be negative, and hence

$$V^{A\cup D}(u) - V^A(u) = -\sum_{x\in D} \rho^{A\cup D}(x)\widehat{\Delta}_x > 0,$$

again a contradiction. This proves (33).

If equality holds in (32), the minimizing property of x_1 implies that all the displayed single-state differences are nonnegative. Thus $\Delta_x \geq 0$ for all $x \in D$. Since their positive weighted sum is $V^A(u) - V^{A\cup D}(u) = 0$, every Δ_x is zero. Therefore, for every $E \subseteq D$,

$$V^{A\cup E}(u) - V^{A\cup D}(u) = \sum_{x\in D\setminus E} \rho^{A\cup E}(x)\Delta_x = 0.$$

If equality instead holds in (33), the maximizing property of x_2 gives $\widehat{\Delta}_x \leq 0$ for all $x \in D$; the zero weighted-sum identity above forces every $\widehat{\Delta}_x = 0$, and comparison with the baseline policy A yields $V^{A\cup E}(u) = V^A(u)$ for all $E \subseteq D$. Together with $V^A(u) = V^{A\cup D}(u)$, this proves (34). □

We are now ready to provide the proof that a non-differentiable point ϕ that is not the Whittle index cannot be a local minimum.

Proof. Argue by contradiction. Suppose that $\phi \in (-M, M)^d \setminus \mathcal{D}_M$, $\phi \neq I$, satisfies (29). Consider any tie block

$$D := \{s_{j_1}, \ldots, s_{j_m}\}, \qquad m \geq 2, \qquad \phi_{j_1} = \cdots = \phi_{j_m} =: \phi_s,$$

and define

$$O_s := \{s_i : \phi_i < \phi_s\}, \qquad P_s := \{s_i : \phi_i \leq \phi_s\} = O_s \cup D.$$

Lemma G.15 gives, for every $1 \leq k \leq m$,

$$V^{P_s\setminus\{s_{j_k}\}}(\phi_s) - V^{P_s}(\phi_s) \geq 0, \tag{35}$$

$$V^{O_s}(\phi_s) - V^{O_s\cup\{s_{j_k}\}}(\phi_s) \leq 0. \tag{36}$$

Let u be the unique intersection of V^{O_s} and V^{P_s}, and choose

$$l \in \arg\min_{1\leq k\leq m}\big\{V^{P_s\setminus\{s_{j_k}\}}(u) - V^{P_s}(u)\big\},$$

$$r \in \arg\max_{1\leq k\leq m}\big\{V^{O_s}(u) - V^{O_s\cup\{s_{j_k}\}}(u)\big\}.$$

The intersection is unique because repeated applications of Assumption G.1 show that $V^{O_s} - V^{P_s}$ has strictly positive slope. Lemma G.16, with $A = O_s$, yields

$$V^{P_s\setminus\{s_{j_l}\}}(u) - V^{P_s}(u) \leq 0, \tag{37}$$

$$V^{O_s}(u) - V^{O_s\cup\{s_{j_r}\}}(u) \geq 0. \tag{38}$$

If both inequalities are strict, then the first difference, which is strictly increasing in the activation price by Assumption G.1, is negative at u but nonnegative at ϕ_s by (35), which implies $\phi_s > u$. The second difference is also strictly increasing, but is positive at u and nonpositive at ϕ_s by (36). Hence, we have $\phi_s < u$, a contradiction.

Thus at least one of (37) and (38) is an equality. By Lemma G.16, for every $E \subseteq D$,

$$V^{O_s\cup E}(u) = V^{O_s}(u) = V^{P_s}(u). \tag{39}$$

For any $s_{j_k} \in D$, (39) with $E = D \setminus \{s_{j_k}\}$ and $E = \{s_{j_k}\}$ shows that both differences in (35)–(36) vanish at u. Their strict monotonicity, together with (35)–(36), gives $\phi_s \geq u$ and $\phi_s \leq u$, respectively. Therefore $\phi_s = u$.

The tie block was arbitrary, so (39) holds at the common ϕ-value of every tie block. For an untied component ϕ_k, Lemma G.15 says directly that $V^{O_k}(\phi_k) = V^{P_k}(\phi_k)$. Consequently, along every induced ordering of ϕ, each pair of consecutive passive-set lines intersects at the corresponding component of ϕ, including every possible order within a tie block. Hence ϕ is a consistent point. Theorem G.13 then gives $\phi = I$, contradicting $\phi \neq I$. □

H CLASSES SATISFYING THE ASSUMPTION G.1

We establish Assumption G.1 for three classes of arms. Throughout this section, the environment is fixed, the state space is finite, and $\gamma \in (0,1)$. All statements allow arbitrary one-stage costs and all deterministic policies.

Throughout this section, a deterministic policy is a stationary map $\pi : \mathcal{S} \to \{0,1\}$. For $s, s' \in \mathcal{S}$ and $a \in \{0,1\}$, let $P_a(s,s')$ be the probability of moving from s to s' after taking action a, and set $P_\pi(s,s') = P_{\pi(s)}(s,s')$. Write $\boldsymbol{a}_\pi = (\pi(s))_{s\in\mathcal{S}}$ for the column vector of actions. The discounted active-time vector T^π has entries

$$T^\pi(s) = \mathbb{E}_s^\pi\left[\sum_{t=0}^{\infty} \gamma^t a_t\right], \qquad T^\pi = (I - \gamma P_\pi)^{-1}\boldsymbol{a}_\pi, \tag{40}$$

where I is the identity matrix and $\mathbb{E}_s^\pi$ denotes expectation for a trajectory starting at $s_0 = s$, taking $a_t = \pi(s_t)$ at every time, and evolving according to P_{a_t}. Define the auxiliary action-value function by

$$Q_T^\pi(s,a) = a + \gamma \sum_{s'\in\mathcal{S}} P_a(s,s')T^\pi(s'). \tag{41}$$

Thus, $Q_T^\pi(s,a)$ counts discounted activations when the initial action is forced to be a and policy π is followed from time 1 onward. The activation advantage is then

$$w^\pi(s) = Q_T^\pi(s,1) - Q_T^\pi(s,0) = 1 + \gamma\big(P_1(s,\cdot) - P_0(s,\cdot)\big)T^\pi, \tag{42}$$

where $P_a(s,\cdot)$ is a row vector and T^π is a column vector.

For any policy π and any state s with $\pi(s) = 0$, let $\widetilde{\pi}(s) = 1$ and $\widetilde{\pi}(x) = \pi(x)$ for every $x \neq s$. The performance difference identity of Lemma E.1 then gives

$$\mathrm{AT}(\widetilde{\pi}) - \mathrm{AT}(\pi) = \rho_\mu^{\widetilde{\pi}}(s) w^\pi(s). \tag{43}$$

Note that the visitation weight is strictly positive (since μ is the uniform distribution), Assumption G.1 is equivalent to requiring

$$w^\pi(s) > 0 \quad \text{for every deterministic policy } \pi \text{ and every } s \in \mathcal{S} \text{ such that } \pi(s) = 0.$$

The restriction $\pi(s) = 0$ reflects the passive-to-active changes in the assumption, not a restriction to a particular state or policy. Proving positivity for every policy and every state, including states where $\pi(s) = 1$, is also sufficient; the latter states need not be checked separately to establish the assumption.

Sufficiency for Indexability Note that the above condition is a sufficient condition for indexability. Indeed, from Theorem 3 of (Zhou & Lin, 2024), it suffices to show that $w^\pi(s) > 0$ for all deterministic policies π and all states s with $\pi(s) = 0$ to establish indexability. If $\pi(s) = 0$, then let $\widetilde{\pi}$ be identical to π except that $\widetilde{\pi}(s) = 1$. Then $w^\pi(s) > 0$ follows directly from (43) and the active-time condition $\mathrm{AT}(\widetilde{\pi}) - \mathrm{AT}(\pi) > 0$. If $\pi(s) = 1$, let $\bar{\pi}$ be identical to π except that $\bar{\pi}(s) = 0$. Applying the same performance difference identity in the reverse direction gives

$$0 > \mathrm{AT}(\bar{\pi}) - \mathrm{AT}(\pi) = -\rho_\mu^{\bar{\pi}}(s) w^\pi(s),$$

where the first inequality is the active-time condition applied to $\bar{\pi}$ and π. The visitation weight is strictly positive, so the inequality implies $w^\pi(s) > 0$ as well.

H.1 ALL TWO-STATE ARMS

Proposition H.1 (Two-state arms)**.** *Every arm with $d = 2$ satisfies Assumption G.1, for arbitrary transition kernels and every $\gamma \in (0,1)$.*

Proof. Write $\mathcal{S} = \{x_0, x_1\}$ and let $\pi_A(x) = \mathbf{1}\{x \in A\}$ denote the policy associated with an activation set $A \subseteq \mathcal{S}$. The all-passive policy $\pi_\varnothing$ has active time 0, and the all-active policy $\pi_\mathcal{S}$ has active time $H = (1-\gamma)^{-1}$. Under the uniform initial distribution, any singleton activation set satisfies

$$\frac{1}{2} \le \mathrm{AT}(\pi_A) \le H - \frac{1}{2}, \qquad |A| = 1.$$

Indeed, the initial action is active with probability $1/2$, which gives the lower bound. It is passive with probability $1/2$, which loses at least $1/2$ in expected active time relative to taking the active action at every time, giving the upper bound. In particular, $0 < \mathrm{AT}(\pi_A) < H$.

Note that every addition of one state to the activation set is either from the empty set to a singleton or from a singleton to the full set. Since all passive has active time 0 and all active has active time H, both changes strictly increase active time, proving the assumption. □

In particular, this result covers our healthcare adherence environment in Appendix J.2 throughout its valid parameter range.

H.2 Stochastic-Reset Bandits

Consider an arm whose passive dynamics follow an arbitrary Markov kernel K. Activation resets its state to a fixed probability distribution ν with a state-independent probability p; otherwise, the state evolves according to K:

$$P_0(s,\cdot) = K(s,\cdot), \qquad P_1(s,\cdot) = (1-p)K(s,\cdot) + p\nu(\cdot), \qquad p \in [0,1]. \tag{44}$$

The kernel K, distribution ν, and probability p may differ across arms and environments.

Proposition H.2 (Stochastic-reset arms)**.** *Every arm satisfying* (44) *satisfies Assumption G.1. In particular, for every deterministic stationary policy π and every state s,*

$$w^\pi(s) \geq \frac{1-\gamma}{1-\gamma+\gamma p} > 0. \tag{45}$$

Proof. Fix π and write $T = T^\pi$, $M = \max_x T(x)$, $B = 1 + \gamma p \nu T \geq 1$. For every state s, we have

$$w^\pi(s) = 1 + \gamma p\big(\nu T - K(s,\cdot)T\big) \geq B - \gamma p M = B(1 - (\gamma p M)/B) \tag{46}$$

We now show that $w^\pi(s) > 0$ for all s. If π is everywhere passive, then $T = 0$ and $w^\pi(s) = 1$. Otherwise, $M > 0$, and any state x attaining M must be active: if it were passive, its Bellman equation would imply $M = \gamma K(x,\cdot)T \leq \gamma M < M$. At this active state, the Bellman equation gives

$$M = 1 + \gamma P_1(x,\cdot)T = B + \gamma(1-p)K(x,\cdot)T$$

Then, by definition of M,

$$M \leq B + \gamma(1-p)M$$

This immediately gives

$$M/B \leq 1/(1-\gamma+\gamma p). \tag{47}$$

And hence we have

$$w^\pi(s) \geq B(1-(\gamma p M)/B) \geq B\left(1 - \frac{\gamma p}{1-\gamma+\gamma p}\right) = B\frac{1-\gamma}{1-\gamma+\gamma p} > 0,$$

where the first inequality is from (46), the second inequality is from (47), and the last inequality is from $B \geq 1$. □

This class includes controlled-restart models (Akbarzadeh & Mahajan, 2022), replacement and reset-maintenance models, and queues cleared upon successful service. The passive dynamics and the dimension of the state can be arbitrary, provided that a successful activation resets the entire arm state according to ν. In particular, the generate-at-will AoI model in (Zhou & Lin, 2024, Section IV-A) has

$$K(h,\cdot) = \delta_{\min\{h+1,K_d\}}, \qquad \nu = \delta_1, \qquad p = p_s,$$

where δ_x denotes the point mass at x. Proposition H.2 therefore verifies the assumption for every deterministic stationary policy, for every finite AoI truncation and every transmission success probability.

H.3 Homogeneous Birth–Death Intervention Models

Let $\mathcal{S} = \{0, \dots, L\}$, where $L \geq 1$. Passive operation increases the state by one with probability p, whereas activation decreases it by one with probability r. All other transitions remain at the current state, and outward moves remain at the boundary:

$$\begin{aligned} P_0(x, \cdot) &= (1-p)\delta_x + p\delta_{\min\{x+1, L\}}, \\ P_1(x, \cdot) &= (1-r)\delta_x + r\delta_{\max\{x-1, 0\}}, \qquad p, r \in [0, 1]. \end{aligned} \tag{48}$$

Proposition H.3 (Homogeneous birth–death arms)**.** *Every arm satisfying* (48) *satisfies Assumption G.1. For every deterministic stationary policy π and every state s with $\pi(s) = 0$,*

$$w^\pi(s) \geq \frac{1-\gamma}{1-\gamma+\gamma p} > 0. \tag{49}$$

Proof. Fix any policy π and a state s with $\pi(s) = 0$, and write $T = T^\pi$. Since the passive action agrees with π at s, its Bellman equation gives $Q_T^\pi(s, 0) = T(s)$. For $s > 0$, the active transition in (48) gives

$$Q_T^\pi(s, 1) = 1 + \gamma\big[(1-r)T(s) + rT(s-1)\big],$$

whereas at $s = 0$ the active transition stays at 0, so $Q_T^\pi(0, 1) = 1 + \gamma T(0)$. Subtracting $Q_T^\pi(s, 0) = T(s)$ in each case yields the quantity we need to bound:

$$w^\pi(s) = \begin{cases} 1 - (1-\gamma)T(0), & s = 0, \\ 1 + \gamma r T(s-1) - (1-\gamma+\gamma r)T(s), & s > 0. \end{cases} \tag{50}$$

We now upper bound $T(s)$ for s being passive and lower bound $T(s)$ for s being active. To do so, set

$$\alpha = \frac{\gamma p}{1-\gamma+\gamma p}, \quad \beta = \frac{\gamma r}{1-\gamma+\gamma r}, \\ b = \frac{1}{1-\gamma+\gamma r}, \quad H = \frac{1}{1-\gamma}.$$

Here $\alpha, \beta < 1$ and $b = (1-\beta)H$. The Bellman equations of T take the form

$$\begin{cases} T(x) = \alpha T(x+1), & \pi(x) = 0,\ x < L, \\ T(L) = 0, & \pi(L) = 0, \\ T(x) = b + \beta T(x-1), & \pi(x) = 1,\ x > 0, \\ T(0) = H, & \pi(0) = 1. \end{cases} \tag{51}$$

We first establish the active-time values at the boundary. For any $1 \leq j \leq L$ such that $\pi(j-1) = 0$ and $\pi(j) = 1$, the two Bellman equations give

$$T(j-1) = \alpha T(j), \qquad T(j) = b + \beta T(j-1).$$

Solving this system gives

$$T(j) = \frac{b}{1-\alpha\beta}, \qquad T(j-1) = \frac{\alpha b}{1-\alpha\beta}.$$

We are now ready to upper bound the active time for passive states. Starting from any passive state x, follow consecutive passive states to the right. If they reach the passive state L, then $T(x) = 0$. Otherwise, they reach a passive–active boundary at states $j-1$ and j, where $j > x$. Applying the passive-state Bellman equation along this sequence gives

$$T(x) = \frac{\alpha^{j-x} b}{1-\alpha\beta} \leq \frac{\alpha b}{1-\alpha\beta},$$

since $0 \leq \alpha < 1$ and $j - x \geq 1$. Thus,

$$T(x) \leq \alpha b/(1-\alpha\beta)$$

for every passive state x.

We are now ready to lower bound the active time for active states. Starting from any active state, we either follow consecutive active states until reaching the active boundary (the leftmost active state), or the active state is 0 itself. In the first case, the active time is at least that at the boundary. Thus, $T(x) \geq b/(1-\alpha\beta)$. Also, the active time at state 0 is at least $T(x) \geq b/(1-\alpha\beta)$. To see this, note that

$$H \geq \frac{(1-\beta)H}{1-\alpha\beta} = \frac{b}{1-\alpha\beta},$$

where the inequality follows from $\alpha \leq 1$ and the equality holds since $b = (1-\beta)H$. Thus, the active time at state 0 is at least $b/(1-\alpha\beta)$. Thus, for any active state x, its active time satisfies

$$T(x) \geq b/(1-\alpha\beta)$$

for all active states x.

We can now return to (50). If $s = 0$, the passive-state bound gives

$$w^\pi(0) = 1 - (1-\gamma)T(0) \geq 1 - (1-\gamma)\frac{\alpha b}{1-\alpha\beta} \geq 1 - \alpha.$$

For $s > 0$, the second case of (50) can be written as

$$w^\pi(s) = (1-\gamma+\gamma r)\big[b + \beta T(s-1) - T(s)\big].$$

If $\pi(s-1) = 0$, then $T(s-1) = \alpha T(s)$, and hence

$$w^\pi(s) = (1-\gamma+\gamma r)\big[b - (1-\alpha\beta)T(s)\big] \geq (1-\gamma+\gamma r)\left[b - (1-\alpha\beta)\frac{\alpha b}{1-\alpha\beta}\right] = 1-\alpha.$$

If $\pi(s-1) = 1$, the two bounds established above give $T(s-1) \geq b/(1-\alpha\beta)$ and $T(s) \leq \alpha b/(1-\alpha\beta)$. Consequently,

$$w^\pi(s) \geq (1-\gamma+\gamma r)\left[b + \frac{\beta b}{1-\alpha\beta} - \frac{\alpha b}{1-\alpha\beta}\right] = \frac{(1-\alpha)(1+\beta)}{1-\alpha\beta} \geq 1-\alpha.$$

In all cases, $w^\pi(s) \geq 1-\alpha = (1-\gamma)/(1-\gamma+\gamma p) > 0$. □

This proposition covers the entire parameter range of our one-dimensional maintenance environment in Appendix J.3, with $L = 7$, $p = p_{\text{up}}$, and $r = p_{\text{rep}}$. In this model, a successful intervention repairs only one degradation level, so a complete reset is not required.

I Proof of Theorem 4.2

Let $\phi_\theta(s)$ be the learned index and ρ be such that $C_N/N \to \rho$ as $N \to \infty$. For simplicity, we prove the average-cost guarantee for N identical arms with a common finite state space $\mathcal{S}$, transition kernel $P(s' \mid s, a)$, and bounded cost $c(s, a)$. Also, instead of assuming $C_N/N \to \rho$ as $N \to \infty$, we assume $C_N/N = \rho$ for all N for simplicity (our augument also holds for this limit form). The argument can be extended to multiple classes of arms with different costs and transition kernels, as in (Verloop, 2016). The budget is $C_N = \rho N$, where $0 < \rho < 1$ is fixed and C_N is an integer.

For the *coupled RMAB*, the average cost per arm under policy π is

$$J_N^\pi := \limsup_{T\to\infty} \frac{1}{NT}\mathbb{E}_\pi\left[\sum_{t=0}^{T-1}\sum_{n=1}^{N} c(S_n(t), A_n(t))\right].$$

Let J_N^{opt} be the minimum subject to $\sum_n A_n(t) \leq C_N$ at every time (the true optimal cost for the RMAB with hard constraint), and let J_N^ϕ be the cost of the learned policy that activates the C_N arms with highest indices. For the *relaxed RMAB*, J_N^{rel} is the minimum of the same cost subject only to $\limsup_{T\to\infty}(NT)^{-1}\mathbb{E}_\pi[\sum_{t<T}\sum_n A_n(t)] \leq \rho$.

For an *independent single arm*, let η^π be its stationary state distribution under policy π. Suppressing the arm subscript, its average base cost, activation rate, and penalized cost are

$$j^\pi := \sum_{s,a}\eta^\pi(s)\pi(a \mid s)c(s,a), \qquad \alpha^\pi := \sum_s \eta^\pi(s)\pi(1 \mid s), \qquad g^\pi(\lambda) := j^\pi + \lambda\alpha^\pi.$$

Write $d(\lambda) := \inf_\pi g^\pi(\lambda)$ for the optimal single-arm penalized cost.

Let $x(s)$ denote *expected* state occupacy of s. Let $u(s;x)$ be the fraction of all arms in state s. The learned index policy that prioritize the arms with the higher learned indices can be analogously defined to prioritize states with higher indices such that $\sum_s u(s;x) = \rho$ (with randomization for tie states). Intuitively, let $s_{\text{boundary}} = \arg\min\{\phi_\theta(s) : u(s;x^*) > 0\}$ be the state that has the lowest activated index. Then the learned index policy basically only activates the arms in states with indices higher than $\widehat{\lambda}$ and randomizes the activation for the arms in state s_{boundary} to satisfy the budget constraint.

How $x(s)$ will evolve under the learned index policy is then described by the following:

$$f(s';x) := \sum_s \left[u(s;x)P(s' \mid s,1) + \big(x(s) - u(s;x)\big)P(s' \mid s,0)\right].$$

We now introduce the global attractor assumption (Weber & Weiss, 1990; Verloop, 2016). It basically assumes that there is a occupacy x^* that is invariant and starting from any initial occupancy $x(0)$, the occupacy will converge to x^* under the learned index policy. Let $\mathcal{X} := \{x \geq 0 : \sum_s x(s) = 1\}$.

Assumption I.1 (Global attractor)**.** The priority-policy fluid dynamics have a globally asymptotically stable equilibrium $x^* \in \mathcal{X}$:

$$f(x^*) = x^*, \qquad \widetilde{x}(t) \longrightarrow x^* \quad \text{for every } \widetilde{x}(0) \in \mathcal{X}.$$

At the attractor, define the cutoff as the lowest activated index:

$$\widehat{\lambda} := \min\{\phi_\theta(s) : u(s;x^*) > 0\}.$$

Let $\widehat{\pi}$ be the $\widehat{\lambda}$-threshold policy for per-arm problem, where for s such that $\phi_\theta(s) > \widehat{\lambda}$, $\widehat{\pi}(1 \mid s) = 1$; for s such that $\phi_\theta(s) < \widehat{\lambda}$, $\widehat{\pi}(1 \mid s) = 0$; and for s such that $\phi_\theta(s) = \widehat{\lambda}$ (i.e., $s = s_{\text{boundary}}$), $\widehat{\pi}(1 \mid s)$ is random. We immediately see that we can choose $\widehat{\pi}(1 \mid s_{\text{boundary}})$ such that

$$x^*(s)\widehat{\pi}(1 \mid s) = u(s;x^*) \qquad \text{for every } s. \tag{52}$$

This is because our $\widehat{\lambda}$-threshold policy shares exactly the same dynamics as that of the learned index policy defined above, so it suffices to choose the same randomization as that of the learned index policy at the boundary state s_{boundary}.

Theorem I.2 (Non-indexable average-cost guarantee)**.** *Under Assumption I.1, if the independent cutoff policy satisfies*

$$g^{\widehat{\pi}}(\widehat{\lambda}) - d(\widehat{\lambda}) \leq \epsilon,$$

then the learned priority policy satisfies

$$\limsup_{N\to\infty}\big(J_N^\phi - J_N^{\text{opt}}\big) \leq \epsilon.$$

Proof. Relaxing the hard constraint and applying weak duality to the separable single-arm Lagrangian give, for every $\eta \geq 0$,

$$J_N^{\text{opt}} \geq J_N^{\text{rel}} \geq d(\eta) - \eta\rho. \tag{53}$$

Taking $\eta = \widehat{\lambda}$ and using single-arm ϵ-optimality,

$$\begin{aligned} J_N^{\text{opt}} \geq J_N^{\text{rel}} &\geq d(\widehat{\lambda}) - \widehat{\lambda}\rho \\ &\geq g^{\widehat{\pi}}(\widehat{\lambda}) - \widehat{\lambda}\rho - \epsilon \\ &= j^{\widehat{\pi}} - \epsilon. \end{aligned} \tag{54}$$

Here the first inequality follows from relaxing the instantaneous budget, the second from weak duality, and the third from the assumed single-arm ϵ-optimality. We justify the final equality next.

To see why the equality in (54) holds, substituting (52) into $f(x^*) = x^*$ yields

$$x^*(s') = \sum_{s,a} x^*(s)\widehat{\pi}(a \mid s)P(s' \mid s,a).$$

Consequently,

$$\alpha^{\widehat{\pi}} = \sum_s x^*(s)\widehat{\pi}(1 \mid s) = \sum_s u(s; x^*) = \rho \tag{55}$$

Thus, we have $j^{\widehat{\pi}} = g^{\widehat{\pi}}(\widehat{\lambda}) - \widehat{\lambda}\rho$.

We now connect this single-arm cost to J_N^ϕ. Under the fluid limit scaling as in (Weber & Weiss, 1990; Verloop, 2016) and the global attractor assumption, the occupancy measure converges to x^*, we have $X_N \to x^*$ as $N \to \infty$. Thus, we can write the average cost of the learned policy J_N^ϕ as

$$J_N^\phi \longrightarrow \sum_s \big[u(s; x^*)c(s, 1) + \big(x^*(s) - u(s; x^*)\big)c(s, 0)\big] \tag{56}$$

Subsitituting (52) into the right-hand side of (56), we have

$$\sum_s \big[u(s; x^*)c(s, 1) + \big(x^*(s) - u(s; x^*)\big)c(s, 0)\big] = \sum_{s,a} x^*(s)\widehat{\pi}(a \mid s)c(s, a) = j^{\widehat{\pi}}.$$

Thus, we have $J_N^\phi \to j^{\widehat{\pi}}$ as $N \to \infty$.

Now writing (54) as $J_N^{\text{opt}} + \epsilon \geq (j^{\widehat{\pi}} - J_N^\phi) + J_N^\phi$. Since $J_N^\phi \to j^{\widehat{\pi}}$, taking $N \to \infty$ on both sides gives the desired. □

J RMAB Environments

J.1 Two-dimensional AoI minimization with unreliable links and random arrivals

This environment models a base station that schedules status updates over unreliable wireless links when fresh packets arrive randomly (Sun et al., 2019). Each user is one arm. Its state is

$$s = (a, \Delta) \in \mathcal{S} = \{(a, \Delta) : 1 \leq a \leq K_a,\ 0 \leq \Delta \leq K_\Delta\},$$

where a is the age of the freshest packet currently buffered at the transmitter and Δ is the reduction in receiver AoI that a successful transmission would provide. The finite limits K_a and K_Δ truncate the two age coordinates. The binary action $u \in \{0, 1\}$ indicates whether the user is scheduled. In the experiments, we let $K_a = K_\Delta = 3$, so each arm's state space has 12 states.

The underlying environment parameter is $\zeta = (p_s, p_g)$, where a fresh packet arrives with probability p_g in each slot and a scheduled transmission succeeds with probability p_s. Let $\bar{a} = \min\{a + 1, K_a\}$ and $\bar{\Delta} = \min\{a + \Delta, K_\Delta\}$. If $u = 0$, only packet arrival is possible, so the next state satisfies

$$s' = \begin{cases} (1, \bar{\Delta}), & \text{with probability } p_g, \\ (\bar{a}, \Delta), & \text{with probability } 1 - p_g. \end{cases}$$

If $u = 1$, the arrival and transmission events are independent, so the next state satisfies

$$s' = \begin{cases} (1, a \wedge K_\Delta), & \text{with probability } p_g p_s, \\ (\bar{a}, 0), & \text{with probability } (1 - p_g)p_s, \\ (1, \bar{\Delta}), & \text{with probability } p_g(1 - p_s), \\ (\bar{a}, \Delta), & \text{with probability } (1 - p_g)(1 - p_s). \end{cases}$$

Let $Y \sim \text{Bernoulli}(p_s)$ denote channel success. The realized one-step cost is the AoI *after the scheduling decision*,

$$C_\zeta((a, \Delta), u, Y) = a + \Delta(1 - uY).$$

The MDP stage cost is $c_\zeta((a, \Delta), u) := \mathbb{E}[C_\zeta((a, \Delta), u, Y) \mid (a, \Delta), u] = a + (1 - p_s u)\Delta$. Thus, the passive and active costs are $a + \Delta$ and $a + (1 - p_s)\Delta$, respectively.

J.2 HEALTHCARE ADHERENCE

This environment models scarce interventions intended to prevent disengagement from a healthcare program, following the RMAB formulation of preventive healthcare in (Biswas et al., 2021). The state is $s \in \{0, 1\}$, where 0 denotes engaged and 1 denotes disengaged, and the action $u \in \{0, 1\}$ indicates whether an intervention is made. The underlying environment parameter is $\zeta = (p_{\text{drop}}, p_{\text{recover}})$. Its passive and active transition matrices, with rows indexed by the current state and columns by the next state, are

$$P_0 = \begin{bmatrix} 1 - p_{\text{drop}} & p_{\text{drop}} \\ 0 & 1 \end{bmatrix}, \qquad P_1 = \begin{bmatrix} 1 - \frac{1}{2}p_{\text{drop}} & \frac{1}{2}p_{\text{drop}} \\ p_{\text{recover}} & 1 - p_{\text{recover}} \end{bmatrix}.$$

An intervention therefore halves the probability that an engaged individual disengages and gives a disengaged individual probability p_{recover} of re-engaging. The one-step cost is

$$c_\zeta(s, u) = s + 0.5u.$$

J.3 PREVENTIVE MAINTENANCE

These environments model allocating a limited number of maintenance interventions among deteriorating machines, an established RMAB application (Glazebrook et al., 2005; Ruiz-Hernandez et al., 2020). In both versions, $u = 1$ denotes maintenance and $u = 0$ denotes passive operation.

One-dimensional birth–death maintenance. The state $x \in \{0, \dots, 7\}$ is the degradation level and the underlying environment parameter is $\zeta = (p_{\text{up}}, p_{\text{rep}})$. With passive operation,

$$x' = \begin{cases} \min\{x + 1, 7\}, & \text{with probability } p_{\text{up}}, \\ x, & \text{otherwise}, \end{cases}$$

whereas active maintenance gives

$$x' = \begin{cases} \max\{x - 1, 0\}, & \text{with probability } p_{\text{rep}}, \\ x, & \text{otherwise.} \end{cases} \qquad c_\zeta(x, u) = x + u.$$

This is the imperfect-repair birth–death version used in the reported experiment, rather than a reset-to-zero repair model.

Two-dimensional degradation and operating stress. We also consider a more complex two-dimensional version of the preventive maintenance problem. We truncate the state space to be $(x, q) \in \{0, \dots, 3\} \times \{0, \dots, 2\}$, where x is degradation and q is operating stress. The underlying environment parameter is $\zeta = (p_{\text{deg}}, p_{\text{stress}})$. Conditional on (x, q, u), the two coordinates transition independently as bounded birth–death chains. Define

$$r_x^+ = \text{clip}(p_{\text{deg}} + 0.06x + 0.04q, 0.03, 0.82), \quad r_x^- = \text{clip}(0.76 - 0.06x, 0.35, 0.80),$$
$$r_q^+ = \text{clip}(p_{\text{stress}} + 0.04q, 0.04, 0.65), \quad r_q^- = 0.18.$$

For $u = 0$, degradation increases by one with probability r_x^+ and otherwise remains unchanged; stress increases by one with probability r_q^+, decreases by one with probability r_q^-, and otherwise remains unchanged. For $u = 1$, degradation decreases by one with probability r_x^- and otherwise remains unchanged; the stress increase and decrease probabilities become $0.55r_q^+$ and 0.38, respectively. All moves are clipped at the state-space boundaries, so probability assigned to an outward move remains at the boundary. The one-step cost is

$$c_\zeta((x, q), u) = 0.70x + 0.35q + 0.55xq + 10\,\mathbf{1}\{x = 3, q = 2\} + u(1.40 + 0.25q).$$

J.4 A NON-INDEXABLE FOUR-STATE RMAB

We use the four-state MDP family constructed by (Charles-Rebuffé et al., 2026) to show that the set of indexable MDPs is not open. Each arm has state space $\mathcal{S} = \{0, 1, 2, 3\}$, binary action $u \in \{0, 1\}$,

and environment parameter $\zeta = \varepsilon \in [0, 0.24]$. With rows and columns ordered as $(0, 1, 2, 3)$, the passive and active transition matrices in our implementation are

$$P_0^\varepsilon = \begin{bmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} - \varepsilon & 0 & \frac{1}{2} + \varepsilon & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad P_1^\varepsilon = \begin{bmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} + 2\varepsilon & 0 & \frac{1}{2} - 2\varepsilon & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

The upper bound 0.24 keeps every transition probability nonnegative. Both actions have the same reward vector,

$$r^0 = r^1 = (0, 1, 1, 1)^\top.$$

Equivalently, our cost-minimization formulation uses $c(s, u) = -r_s$; the Whittle-relaxed single-arm problem adds the activation price λu. The family is indexable at the boundary $\varepsilon = 0$, but is non-indexable for every $\varepsilon > 0$ (Charles-Rebuffé et al., 2026).

K Evaluation protocol

K.1 Setup for Section 5.2

We evaluate the three methods (WAC, WQL, and random) on the four-state non-indexable RMAB with fixed environment parameter ε. We set the number of arms to 12, the activation budget to 6 per time step, and the horizon to 200 timesteps. For each fixed

$$\varepsilon \in \{0.02, 0.04, 0.06, \ldots, 0.24\},$$

all arms use the same ε. Then the same index network is applied independently at every fixed-ε setting to implement the index policy, and the reported cost is averaged over the 12 different ε settings and 40 random seeds per setting.

K.2 Setup for Section 5.4

We evaluate each method on the RMAB with constantly changing environment parameters for each arm. Each evaluation contains 200 arms and permits exactly 30 activations per time step. WAC and WQL's environment encoders receive a context containing the most recent $K = 50$ observed transitions. For fair comparison, the DeepTOP baseline uses the same 50 pre-control transitions for initialization. All methods share the parameter schedules, initial states, pre-control transitions, and transition randomness within a seed. We evaluate for 600 time steps, with each arm's underlying environment parameters changing at $t = 0, 200, 400$. We resample the arm's underlying environment parameters from the same distribution at $t = 0, 200, 400$. We specify them as follows:

AoI: each phase uses $(p_s, p_g) \in [0.5, 1] \times [0.5, 1]$.

Healthcare: each phase uses $(p_{\text{drop}}, p_{\text{recover}}) \in [0.05, 0.35] \times [0.25, 0.85]$.

Maintenance 1D: each phase uses $(p_{\text{up}}, p_{\text{rep}}) \in [0, 1] \times [0, 1]$.

Maintenance 2D, train range: each phase uses $(p_{\text{deg}}, p_{\text{stress}}) \in [0.05, 0.30] \times [0.05, 0.30]$.

L Additional experiment results

We also evaluate the underlying function-approximation error, i.e., the gap between the learned value (or optimal Q-) function and the exact value (or optimal Q-) function, to better understand the two methods.

We can further see why WAC is more robust to limited NN size than WQL. Specifically, WAC is bottlenecked only by the index network capacity, but WQL is further bottlenecked by the Q-network capacity. Indeed in Figure 6, the function approximation error is flat for WAC across different h, while WQL's Q-function approximation error decreases substantially as the network size increases. Intuitively, this is because WAC uses roll-out to assist the value-function estimates, and thus the error for the value-function estimate is low even when the critic network has a small size. In contrast, WQL cannot use rollout. Since representing the optimal Q-function across multiple environments

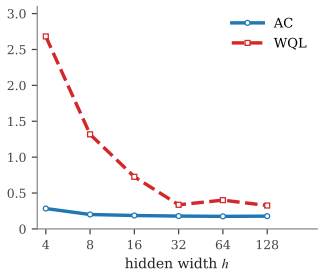


Figure 6: Function approximation error

Table 3: Rollout ablations for WAC. Both tables report the ℓ_1 index error.

(a) Rollout horizon.

Method	**Index error**
WQL ($J = 20$)	0.42 ± 0.02
WAC ($T = 0, J = 20$)	0.40 ± 0.19
WAC ($T = 1, J = 19$)	$\mathbf{0.17} \pm 0.01$
WAC ($T = 10, J = 10$)	0.19 ± 0.06
WAC ($T = 20, J = 0$)	0.17 ± 0.04

(b) Number of trajectories.

Method	**Index error**
WQL ($J = 20$)	0.42 ± 0.02
WAC ($M = 1$)	0.28 ± 0.03
WAC ($M = 8$)	0.25 ± 0.14
WAC ($M = 16$)	0.24 ± 0.12
WAC ($M = 32$)	$\mathbf{0.19} \pm 0.06$

is generally more complex, the Q-network becomes a bottleneck even after the index network is sufficiently expressive. The index error of WAC has already reached a low level at $h = 64$ and remains essentially unchanged thereafter.

M Ablation Studies

We study the effect of the rollout horizon, the number of rollout trajectories, and the number of transitions used to estimate the environment representation z. For both WAC and WQL, we train for 5k iterations with networks hidden dimension being $h = 16$.

Rollout horizon T. We vary the rollout horizon T among $\{0, 1, 10, 20\}$ while fixing the number of rollout trajectories to $M = 32$ in (14) ($T = 0$ stands for not using rollout while $T = 20$ stands for a no critic approach). We keep the batchsize of Q/value network update as 32 and the number of trajectories averaged as $M = 32$. To keep the total number of value-estimator updates comparable, we adjust the number J of value-network updates per index-network update: a larger T is paired with a smaller J. The results in Table 3a show that even one-step rollout greatly enhance the performance of WAC. Specifically, $T = 1$ and $T = 20$ obtain the lowest mean index error, while $T = 1$ has the smaller variation.

Number of rollout trajectories M. We vary the number of trajectories averaged in (14), fixing the rollout horizon and the number of value-network updates to $T = 10$ and $J = 10$, respectively. The WQL baseline uses $J = 20$. Table 3b shows that increasing M generally reduces the index error. Note that even $M = 1$ substantially outperforms WQL.

Number of transitions used to estimate z. We study the sensitivity to the number of transitions used to estimate the environment representation z of our two approaches. We still apply them on the AoI RMAB from Section 5.1. We compare the learned encoder and the hand-crafted estimator for four representative inference-context lengths, $K \in \{20, 30, 200, 1024\}$. The encoder is trained with a context of 30 transitions. The resulting index errors for different numbers of transitions provided at inference time are reported in Table 1.

For both representation methods, the index error decreases as the number of transitions used to estimate z increases across the reported values. Moreover, even when the number of transitions is small, both approaches produce a reasonable index estimate.

We also note that even though the encoder is trained to compress a context of 30 transitions, it generalizes well to longer contexts. In particular, the encoder is still able to achieve a low index error when the number of transitions is 1024. We attribute this robustness in part to the permutation-invariant mean pooling, which naturally allows additional transitions to be incorporated at inference time without retraining.