NEAR-OPTIMALITY FOR MULTI-ACTION MULTI-RESOURCE RESTLESS BANDITS WITH MANY ARMS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Xiangyu Zhang

August 2022

NEAR-OPTIMALITY FOR MULTI-ACTION MULTI-RESOURCE RESTLESS BANDITS WITH MANY ARMS

Xiangyu Zhang, Ph.D.

Cornell University 2022

We consider multi-action restless bandits with multiple resource constraints, also referred to as weakly coupled Markov decision processes. This problem is important in recommender systems, active learning, revenue management, and many other areas. An optimal policy can be theoretically found by solving a Markov decision process, but the computation required scales exponentially in the number of arms N. Thus, scalable approximate policies are important for problems with large N. We study the optimality gap, i.e., the loss in expected performance vs. that of the optimal policy, of such scalable policies. The tightest previous theoretical bounds, which apply only for a handful of carefully-designed policies, show that this optimality gap is $O(\sqrt{N})$ for the finite-horizon case and $o(N)$ for the infinite-horizon case. This dissertation significantly improves these bounds by characterizing a much wider class of novel practically-computable policies for which the optimality gap is $O(\sqrt{N})$ for both finite- and infinite-horizon restless bandits. Furthermore, for the finite-horizon case including time-varying environmental variables that affect transitions and rewards, we characterize a non-degeneracy condition under which the optimality gap is surprisingly $O(1)$. We demonstrate that our policies offer state-of-the-art empirical performance in numerical experiments.

BIOGRAPHICAL SKETCH

Xiangyu Zhang was born in Xi'an, Shaanxi Province in the middle of China. When he first read about Euler's story in the fifth grade at his elementary school, he was shocked by the extraordinary intelligence a human being could possibly achieve and started to dream of being a mathematician in the future.

During high school, Xiangyu Zhang participated in the mathematical olympiad competition. These days were still one of the happiest times in his life by now. After winning two silver medals in the Chinese Mathematical Olympiad, he was admitted to Tsinghua University to continue his mathematical study journey.

Four years at Tsinghua University gave Xiangyu Zhang opportunities for getting exposure to advanced mathematics. And thanks to Professor Zongxia Liang's guide, Xiangyu Zhang got interested in modern probability theory and stochastic control problems. So after graduating from Tsinghua, Xiangyu Zhang came to Cornell University to study operation research under the supervision of Professor Peter I. Frazier.

With a great advisor and great friends around, life at Cornell was another happiest time for Xiangyu Zhang. In May 2022, Xiangyu Zhang successfully defended his thesis and came to New York City to be a quantitative researcher.

This document is dedicated to all Cornell graduate students.

ACKNOWLEDGEMENTS

I want to say thank you to many people.

First, I can not express my gratitude enough to my advisor Peter I. Frazier. Peter is one of the most successful researchers I have ever seen in both academia and industry. However, although with great achievements, Peter is the most humble and egoless person I have ever met. Under Peter's guidance, I learned how to become a mature researcher, but this is not the most valuable thing I learn from him. The most valuable thing I learned from Peter is how to be a humble person, who is always ready to drop his ego, admit his limitation, and improve himself.

Second, I would like to say thank you to other committee members. I am very honored to have Professor Yudong Chen, Professor Jim Dai, and Professor Huseyin Topaloglu on my committee board.

Third, I would like to thank my parents, from whom I learn what is unconditional love. Their unconditional love deeply and fundamentally shaped my personality, although I did not realize it until recently.

Lastly, I would like to say thank you to all my friends. I would like to thank friends I know at Cornell, without whom life at Ithaca would be much less interesting. Also, I would like to thank some old friends who I started to know back in Tsinghua, in my high school, and even in my elementary school. They are always there ready for my phone call whenever I have any questions to ask or feelings to share.

CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Contents . . . vi
List of Figures . . . ix

1 Introduction **1**
1.1 Motivation and Contributions . . . 1
1.2 Literature Review . . . 5

2 Binary-Action Finite-Horizon Restless Bandit **12**
2.1 Literature Review and Contributions . . . 12
2.2 System Model . . . 14
2.3 Background: Preliminary Results and Notation . . . 17
2.4 Sufficient Conditions for Achieving an $o(N)$ Opt Gap . . . 20
2.5 Sufficient Conditions for Achieving an $O(\sqrt{N})$ Opt Gap . . . 22
2.6 Fluid-priority Policies . . . 24
2.7 Non-degeneracy Condition: Achieving an O(1) Opt Gap . . . 27
2.7.1 Budget-relaxed fluid-priority policies . . . 28
2.7.2 Non-degeneracy . . . 29
2.7.3 Main result . . . 32
2.7.4 The best fluid-priority policy is at least as good as the best index policy . . . 34
2.8 Numerical Experiments . . . 35
2.8.1 Bayesian bandit with Bernoulli rewards . . . 36
2.8.2 Crowdsourced labeling . . . 40
2.8.3 Dynamic assortment optimization . . . 42

3 Multi-Action Multi-Resource Finite-Horizon Restless Bandit with Markovian Environmental Variables **45**
3.1 Literature Review and Contributions . . . 46
3.2 System Model . . . 48
3.3 Background: Preliminary Results and Notations . . . 50
3.4 Sufficient Conditions for Achieving an $o(N)$ Opt Gap . . . 54
3.5 Sufficient Conditions for Achieving an $O(\sqrt{N})$ Opt Gap . . . 55
3.6 Fluid-priority policies . . . 56
3.7 Non-degeneracy Condition: Achieving an O(1) Opt Gap . . . 58
3.7.1 Budget-relaxed fluid-priority policies . . . 60
3.7.2 Non-degeneracy . . . 61
3.7.3 Main result . . . 62

4 Binary-Action Infinite-Horizon Restless Bandit **66**
4.1 Literature Review and Contributions 67
4.2 System Model 72
4.3 Background: Preliminary Results and Notation 74
4.4 Diffusion Regular Conditions 76
4.5 Fluid-balance Policy 77
4.6 Numerical Experiment 79
4.6.1 Does steady-and-slow win the race? 80
4.6.2 Whittle index: not a good benchmark 82

A Appendix: Binary-action finite-horizon restless bandit **85**
A.1 Proof for Lemma 2.1 85
A.2 Discussion of the rounding error in budget constraints 86
A.3 Proof of Lemma 2.2 88
A.4 Proof of Lemma 2.3 89
A.5 Proof of Lemma 2.4 90
A.6 Proof of Lemma 2.5 94
A.7 Proof of Theorem 2.3 96
A.8 Proof of Lemma 2.6 98
A.9 Proof of Lemma 2.7 98
A.10 Proof of Lemma 2.8 103
A.11 Proof for Proposition 2.1 105
A.12 Proof for Proposition 2.2 107
A.13 Discussion of policies in previous literature 107
A.14 Choice of Occupation Measure 110

B Appendix: Multi-Action Multi-Resource Finite-Horizon Restless Bandit **113**
B.1 Proof for Lemma 3.1 113
B.2 Proof of Theorem 3.1 114
B.3 Proof of Theorem 3.2 117
B.4 Proof of Lemma B.2 118
B.5 Proof of Theorem 3.3 120
B.6 Proof of Lemma 3.2 121
B.7 Proof of Lemma 3.3 121
B.8 Proof of Lemma 3.4 125

C Appendix: Binary-action infinite-horizon restless bandit **127**
C.1 Proof for Lemma 4.1 127
C.2 Discussion of the rounding error in budget constraints 128
C.3 Proof of Lemma 4.2 129
C.4 Proof of Lemma 4.3 132
C.5 Proof of Proposition 4.1 134
C.6 Proof of Proposition 4.2 135

C.7 Proof of Proposition 4.3 136
C.8 Proof of Proposition 4.4 137

LIST OF FIGURES

1.1 Relationships between learning bandits, Bayesian bandits, frequentist bandits, adversarial bandits and Markov decision process bandits. 6

2.1 Bayesian bandit with Bernoulli rewards. An upper bound on the opt gap (relaxed problem's expected total reward minus a simulation-based estimate of reward) vs number of arms N, for the finite-horizon Bayesian multi-armed bandit with horizons $T = 15$ (left) and $T = 20$ (right). The fluid-priority policy has its opt gap bounded above by a constant while UCB and Thompson sampling have opt gaps that grow linearly with the number of arms. 39

2.2 Crowdsourced labeling. An upper bound on the opt gap (relaxed problem's expected total reward minus a simulation-based estimate of reward) vs. number of arms N. The left and right panel show the same data but use different scales for the y-axis. Both Knowledge Gradient and Optimistic Knowledge Gradient have opt gaps that seem to grow linearly. The fluid-priority policy has an opt gap that is $O(\sqrt{N})$ because the non-degeneracy condition does not hold in this problem. 42

4.1 Performance comparison between Whittle index and fluid-balance policy. The left panel shows the average reward per arm versus number of arms, where we compare the relaxation upper bound, the whittle index and the fluid-balance policy. The right panel shows an upper bound on the opt gap (relaxation upper bound minus a simulation-based estimate of reward) versus number of arms N. As we can see, opt gap of Whittle index grows linearly while opt gap of fluid-balance policy grows sublinearly with respect to N. 84

CHAPTER 1

INTRODUCTION

This Chapter describes the restless bandit problems that this dissertation studies at a high level, reviews related past work, and summarizes the contributions of this dissertation.

1.1 Motivation and Contributions

We study the restless bandit, a sequential decision-making problem in which a decision maker manages N Markov processes (colloquially called "arms"), each with finite state space and finite action space. The transition kernels and state-action-dependent rewards of the arms are known. Arms produce rewards and evolve independently but are coupled through constraints in each period (called "budget" constraints), since each action is associated with some state-dependent resource consumption and there is a total amount of resources shared by all arms in each period. Subject to budget constraints, the decision-maker seeks to maximize the expected total reward.

We consider three versions of this problem.

- Chapter 2 begins by considering a finite horizon, binary actions, with a single resource and state-independent resource consumption.
- Chapter 3 generalizes previous setting to consider multiple actions, multiple resources, and a stochastically changing environmental variable that affects rewards and resource consumption. This setting includes problems studied under the names multi-action restless bandit [57], weakly-coupled

Markov decision process [29], optionally in the presence of environmental variables [14].

- Finally, Chapter 4 generalize Chapter 2 to its infinite-horizon version, continuing to assume binary actions, a single resource, and state-independent resource consumption.

The above problem arises in various fields. For example, when classifying images with crowd workers [17], we treat each image as an arm. Each arm is associated with two actions: asking a new worker to label this image or do nothing. The state of an arm is modeled as the Bayesian posterior distribution on the corresponding image's class given past noisy labels. A limited supply of crowd workers imposes budget constraints: the new label we can request each period is no greater than the number of crowd workers. Another example arises in dynamic assortment optimization [13]. A sales manager selects products to display subject to limited display space. Each product generates revenue when displayed at an unknown rate, which can be learned from its revenue history. In this example, we can treat each product as an arm, and each arm is associated with two actions: whether we display the product or not. The arm's state is the Bayesian posterior distribution on the product's revenue-generation rate, and the budget constraints are imposed via the limited display space. Problems in target search by unmanned aerial vehicles [34, 46], online advertising [28, 16], network communication [38, 4], and sensor management [30] also fit into the restless bandit framework.

We study a regime in which the number of arms grows large with a fixed per-arm budgets in each period. This regime was first proposed in [55], where a binary-action single-resource special case is studied: each arm is associated with

two actions ("pull" or "idle") and pulling the arm consumes state-independent amounts of resources. Since its introduction, this problem has been of long-standing theoretical interest. Moreover, it is practically important in many settings. In the examples above, crowdsourced labeling is most challenging when there are many images to label and assortment optimization is most challenging when many products are available.

Despite its importance, this regime presents substantial algorithmic difficulties. While, in principle, one can compute the optimal policy for restless bandit problems via dynamic programming, the state of this dynamic program includes the state of each arm and so its dimension grows linearly with N. Thus, solving this dynamic program requires computation exponential in N.

As a result, there has been substantial interest (e.g., [55, 31, 57, 13]) in developing approximate policies with strong performance but affordable computation overhead. To measure the performance quality, people define "opt gap" of a policy as the difference of achieved rewards between itself and the optimal policy. Since the introduction of the restless bandit, substantial effort are spent in developing policies with slowly-growing opt gap with respect to N while keeping the computation overhead not depending on N. For example, [55] proposed the Whittle index policy and conjectured its opt gap is $o(N)$ for the infinite-horizon case and [13] proposed a heuristic policy with showing its opt gap is $O(\sqrt{N})$ for the finite-horizon case. Despite substantial interest and effort, current understanding is still limited in several important aspects.

First, the theoretical bound on proposed policies' performance is usually looser than the practical performance observed. For example, simulation shows that the index policy proposed in [13] achieve $O(1)$ opt gap in some experi-

ments, while its performance bound provided is $O(\sqrt{N})$.

Second, there may be better policies with a tighter theoretical performance bound. In particular, the best known opt gap growing regime is $o(N)$ for infinite-horizon bandits and $O(\sqrt{N})$ for finite-horizon bandits. However, intuitively speaking, the infinite-horizon case should not be fundamentally harder than finite-horizon case. Thus, a natural question to ask is whether we can design a policy for the infinite-horizon problem with $O(\sqrt{N})$ opt gap.

Third, all the performance analysis for infinite-horizon bandits requires hard-to-verify conditions. For example, the most popular approach, the Whittle index proposed in [55], requires a hard-to-verify indexability condition to be well-defined and another hard-to-verify condition to guarantee a $o(N)$ optimality gap. Thus, a natural question to ask is whether we can find a policy both well-defined and guaranteed to achieve strong performance for all problem instances.

Summary of Contributions This thesis address these holes in the literature. Chapter 2 address the first limitation: it studies finite-horizon binary-action single-resource restless bandit and proposes a class of policies always achieving $O(\sqrt{N})$ opt gap, and achieving $O(1)$ opt gap when a non-degeneracy condition holds true. Chapter 3 generalizes settings in Chapter 2 and addresses the first limitation but in the setting of finite-horizon multi-action multi-resource restless bandit with an external environmental variable, proposing a class of policies always achieving $O(\sqrt{N})$ opt gap, and achieving $O(1)$ opt gap when a non-degeneracy condition holds true. Finally, adapting the technique developed in Chapter 2 to infinite-horizon cases, Chapter 4 address the second and the third

limitation: it proposes a class of policies well-defined for every infinite-horizon problem instances and always achieving $O(\sqrt{N})$ opt gap.

1.2 Literature Review

There is a large literature on multi-armed bandits. When one considers closely-related problems, such as weakly coupled MDPs, the literature becomes even larger. We think of the problems within this larger literature as being divided into four different classes. The first three classes all focus on what we call "learning bandits". They are motivated by the problem of choosing slot machine arms to pull sequentially over time to maximize the amount of money earned, despite uncertainty about the payoffs associated with each arm. One can "learn" about an arm's payoffs by pulling it and observing the resulting payoff. These three classes are Bayesian bandits, frequentist bandits, and adversarial bandits and are described in more detail below. The fourth class, Markov decision process (MDP) bandits, generalizes Bayesian bandits to settings beyond slot machines, including those where there is nothing to learn and instead the goal is to manage resources over time subject to stochasticity. We summarize and explain these four classes below and in the Venn diagram pictured in Figure 1.1. Our work falls within the MDP bandit setting and thus also applies to the Bayesian bandit setting.

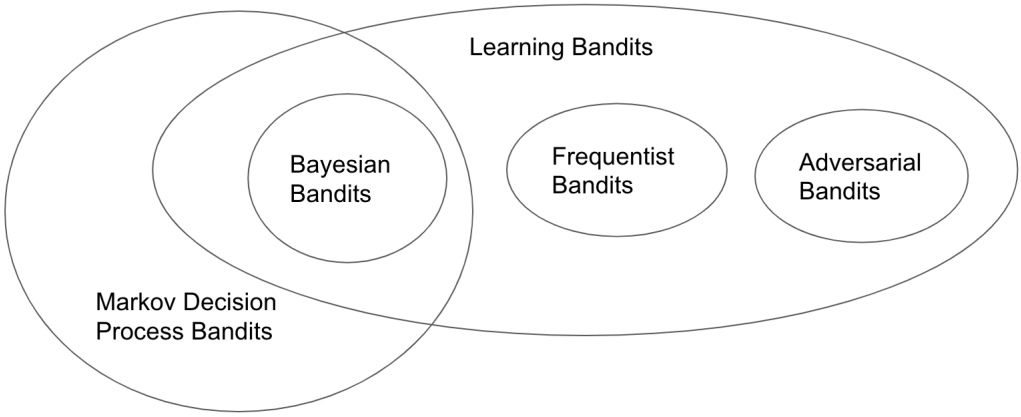


Figure 1.1: Relationships between learning bandits, Bayesian bandits, frequentist bandits, adversarial bandits and Markov decision process bandits.

In all four of these classes, much of the literature considers a canonical setting where one arm is pulled per time period, nothing ties arms together, the constraint on the number of arms pulled is the only resource constraint, and there is no additional information available beyond the arms' past rewards. There is, however, a great deal of work generalizing beyond these assumptions. Generalizations considered include:

- pulling more than one arm per period [55];
- multiple types of resources, sometimes referred to as a weakly coupled MDP [29] or bandits with knapsacks [7];
- bandits with more than the two actions (idle or activate) associated with each arm in the classical literature, called multi-action bandits [57] or bandit superprocesses [55] and also included in weakly coupled MDPs [29];
- bandits whose state changes even when it isn't pulled, called restless bandits in both the MDP bandit [55] and learning settings [48];

- stochastic environmental variables affecting all arms' reward, resource consumption, and total resources available [14];
- dueling bandits [56], where are a kind of learning bandit in which one activates pairs of arms and sees outcomes comparing their quality;
- contextual bandits [36], where each arm can be presented as a list of features which affects this arm's reward by unknown parameters.

This thesis focuses on MDP bandits only, including Bayesian bandits as a special case. In particular, Chapter 2 focuses on the binary-action finite-horizon case, Chapter 3 focuses on the multiple-action finite-horizon case with stochastic environmental variables, and Chapter 4 focuses on the binary-action infinite-horizon case. Detailed discussion of the most recent literature and their limitation in each cases can be found in their own chapter's literature review section.

We now discuss the four classes of bandit problems (MDP bandits and the three classes of learning bandits: Bayesian bandits, frequentist bandits, and adversarial bandits) in more details.

Markov decision process (MDP) bandits In an MDP bandit formulation, there is a decision maker, in charge of managing N Markov processes. Colloquially, we call each process an "arm". Each arm is associated with a finite state space and finite action space. Applying an action to an arm would consume resources, and after action applied, the state of the arm may transition and certain amount of reward would be generated. The state-transition kernels, state-action-dependent rewards and state-action-dependent resource consumption of each arm are known to the decision maker. Given a total amount of resources shared by all arms in each period, the decision maker seeks to maximize the

expected total reward. Depending on whether the problem is of finite horizon (Chapter 2) or infinite horizon (Chapter 4), we study them separately. We also study a generalization in Chapter 3 where there is an stochastic environment variable affecting each arm's reward and resource consumption.

Our approach follows the tradition of a long literature stream. The first seminal result in this stream may be the work by Gittins [22]. In this work, Gittins derived the optimal policy for the infinite-horizon problem where only a single arm can be activated in each period. Gittins designed an index for each state (called Gittins index), and in each period, the optimal policy just activates the arm whose state is assigned with the largest index. However, Gittins index is only defined for problems where an arm's state does not transition if the arm is idled. For many restless real-world problems where the state of an arm transitions even when idled, Gittins does not provide a solution. Thus, later Whittle [55] generalizes the Gittins setting and formulates the restless bandit problem. Also motivated by the Gittins index, Whittle proposed so-called Whittle index as a solution for the infinite-horizon restless bandit problem. Different from Gittins index, Whittle index is not the optimal policy but is shown to be asymptotically optimal under certain conditions in the regime of number of arms grows large. Thus, Whittle's work opens the gate of finding better policies for restless bandit problems with stronger and stronger performance but less and less computation overhead. For example, in the finite-horizon cases, there is a stream of literature [31, 57, 13, 58] proposing policies for restless bandit problems with stronger and stronger performance guarantee.

Learning bandits Moving away from MDP bandits, the term "bandit" often implicitly refers to a class of problems involving learning, where limited re-

sources need to be allocated to a set of competing objects (colloquially called "arm"s) to maximize the total generated reward. However, the reward distribution of each arm is unknown to the decision maker initially and the underlying distribution may only become more and more understood if more resources are allocated to the arm as time passes by. In these problems one often needs to tackle the so-called "exploration vs. exploitation tradeoff" (Kaelbling or Sutton and Barto) by balancing the amount of resources used to explore the less-understood arms to better understand their reward distribution and amount of resources used to exploit high-reward arms to maximize the objective.

We refer to these problems as "learning bandits". There are three predominant mathematical formulations of learning bandits: Bayesian bandits, frequentist bandits and adversarial bandits. As we will see below, Bayesian bandits are a special case of MDP bandits, while frequentist and adversarial bandits are not.

Bayesian bandits The Bayesian approach would model the expected reward of each arm as an unknown parameter sampled from a known prior, and tries to design policies in a Bayesian-optimal sense.

Bayesian bandit can be viewed as a special case of our MDP approach. First, each arm is separately modeled as a MDP, where its state is the posterior distribution over the parameter we are interested for this arm. Each time we pull the arm, a new reward is sampled and the posterior is updated according to the Bayesian rule. Then, all arms together would form a joint MDP with the state space as the Cartesian product of each arm's state space. And our aim is to design policies to maximize the total reward generated from this joint large MDP.

Frequentist bandits This approach would model the expected reward of each arm as an unknown parameter, and tries to design policies with acceptable worst-case performance bound.

To characterize the worst-case performance, this approach introduces the notion of regret for a policy: the difference between the policy's expected reward and the optimal policy's expected reward with full information. Apparently, the optimal policy with full information would activate the arm with highest reward in each period, so its total reward is just the reward of best arm times the decision horizon. Thus, the smaller the regret, the better the policy.

This approach is quite different from ours for two reasons. First, it uses a different performance measure. Rather than worst-case expected regret, we maximize average case expected reward where unknown parameters are drawn at random from a known prior. This difference in performance measure creates significant differences in achievable performance. As we show in Chapter 2, policies with $O(1)$ average case opt gap in the number of arms N exist for finite-horizon restless bandit. In frequentist bandits, however, the (worst-case) regret would grow linearly with N [33].

Second, the asymptotic regime on which the frequentist approach is focused is different from ours. Rather than focusing on large N regime, most of this literature focuses on the regime where the horizon T increases to infinity with the number of arms N fixed. [33] bounds the regret by a factor proportional to $\log(T)$. Celebrated algorithms such as UCB (Upper Confidence Bound [2]) and Thompson Sampling [3] are proved to achieve this lower bound asymptotically. This stream of work relies on the fact that a long horizon permits many pulls per arm, which distinguish the "best" arm from others with high probability. In

our setting where the number of arms is large enough to permit only a small number of pulls per arm and the horizon remains fixed, asymptotic guarantees focusing on large T may not be relevant. Thus, although there is a large literature demonstrating that variants of UCB, Thompson Sampling, epsilon greedy [50], and other related algorithms have provably small regret in the large T setting, these results do not imply good performance in the large N setting.

Adversarial bandits The adversarial bandit approach is more conservative than either the Bayesian approach or frequentist approach. It imagines there is one adversary who is able to not only foresee your decision strategy but also choose the reward realization against your strategy. The decision maker should design strategies in this permissive setting to minimize the "weak regret" [6]: the difference of the total expected reward generated by the strategy and the maximal reward achievable by consistently activating a single arm.

Any deterministic strategy performs badly in the adversarial setting, but surprisingly sublinear weak regret can be achieved by randomized strategies. For example, the EXP3 algorithm [6] achieves $O(\sqrt{NT\log N})$ regret.

We now begin the technical contributions of this thesis, starting with the study of binary-action finite-horizon restless bandits in Chapter 2. Then, Chapter 3 generalizes the methodology to multi-action multi-resources restless bandit with stochastic environmental variables. Finally, Chapter 4 adapts the methodology developed in Chapter 2 to the binary-action infinite-horizon restless bandits.

CHAPTER 2

BINARY-ACTION FINITE-HORIZON RESTLESS BANDIT

This Chapter formally describe the binary-action finite-horizon restless bandit problem, and proposes a novel class of policies called "fluid-priority". Fluid-priority policies would always achieve $O(\sqrt{N})$ opt gap and even achieve $O(1)$ opt gap when a non-degeneracy condition holds. At the end of this Chapter, we also illustrate the state-of-the-art performance of the fluid-priority policies via numerical experiments.

2.1 Literature Review and Contributions

The asymptotic regime we study where the number of arms and the budget per period grow proportionally with a fixed horizon is first introduced in [55]. However, when first introduced, only the infinite-horizon case is studied. In this pioneering work, [55] introduced a time-homogeneous Lagrangian relaxation of the budget constraints and proposed the "Whittle index" policy when arms are "indexable", conjecturing that the Whittle index achieves an $o(N)$ opt gap when this indexability condition holds. However, [54] later showed that even under indexability, Whittle index policy's opt gap is $\Omega(N)$ for some problems. Though intuitively promising, the Whittle index policy suffers from the difficulty of verifying indexibility, the inability to use the policy if indexability does not hold, and, in some problems, from weak empirical performance. Nevertheless, as a pioneering work in restless bandits, the Whittle index inspired a stream of follow-up work, in both the infinite-horizon [24, 19] and finite-horizon cases. As the finite-horizon case is the focus of this chapter, we now discuss its

most recent progress in detail.

Following Whittle's earlier work, later literature (e.g., [31, 57, 13]) studies the finite-horizon restless bandit using Lagrangian relaxations. Unlike Whittle's work, they use a time-dependent Lagrange relaxation because of the non-stationary nature of finite-horizon problems. This technique yields both promising theoretical guarantees and empirical performance without the need for an indexability condition. For example, [31] studies the binary-action bandit problem and proposes an index policy achieving an $o(N)$ opt gap. [57] studies the multi-action bandit problem and proposes a policy achieving an $O(\sqrt{N}\log N)$ opt gap. [13] studies the same setting as [31] and proposes policies with an $O(\sqrt{N})$ opt gap.

Although many exciting progresses in the area of finite-horizon bandits, current understanding is still limited in many important aspects. First, simulation studies show much better performance for large N in some problems than the best existing theoretical results. Indeed, the tightest existing upper bound on the opt gap for such policies is $O(\sqrt{N})$, shown by [13]. Surprisingly, however, their simulation studies suggest that the true opt gap in some problems actually does not grow at all with the number of arms. The proof techniques used by [13], however, rely heavily on the Central Limit Theorem (CLT), and do not offer a path toward showing an $O(1)$ bound.

Second, existing results bounding the opt gap are restricted to specific policies ($o(N)$, $O(\sqrt{N}\log N)$ and $O(\sqrt{N})$, respectively in [31, 57, 13]). However, one would expect many policies to achieve such opt gaps asymptotically.

Summary of Contributions Our work addresses these holes in the litera-

ture: we propose a broad class of policies, called fluid-priority policies, which generalize the essential characteristics of policies proposed by [13]. Addressing the inconsistency between simulation studies and past theoretical results, we characterize a sufficient condition, which we call "non-degeneracy", under which any fluid-priority policy achieves an $O(1)$ opt gap, strictly better than all previous results. We show that the simulation study suggesting an $O(1)$ opt gap in [13] satisfies this non-degeneracy condition. We also address the current literature's lack of generality by providing general easy-to-verify sufficient conditions ensuring $o(N)$ and $O(\sqrt{N})$ opt gaps. All fluid-priority policies satisfy these conditions and thus always achieve an $O(\sqrt{N})$ opt gap. The policies proposed by [31] and [13] also satisfy the sufficient conditions for an $O(\sqrt{N})$ opt gap and thus our results generalize those in this previous work.

2.2 System Model

This section formulates our decision-making problem as a Markov Decision Process (MDP).

Model: There are N arms, each of which shares the same finite state space S. We use $s_{t,i}$ to indicate the state of arm i at time t. At each period t for each arm i, the decision-maker chooses whether to pull the arm ($a_{t,i} = 1$) or leave it idle ($a_{t,i} = 0$). We define $A = \{0, 1\}$ to be the action space in which $a_{t,i}$ takes values. These actions must respect a so-called "budget constraint" in which the number of arms pulled at period t is $B_t = \lfloor \alpha_t N \rfloor$, where $0 \leq \alpha_t \leq 1$ is a pre-specified budget ratio.

Based on the action applied, each arm's state transitions stochastically to

time $t+1$ according to a known transition kernel $P_t = \{p_t(s,a,s')\}_{s,s' \in S, a \in A}$ where $p_t(s,a,s') = \mathbb{P}(s_{t+1,i} = s' | s_{t,i} = s, a_{t,i} = a)$. We assume all arms share the same transition kernel, and any arm's transition is conditionally independent from others given its own state and action. At period t, each state-action pair is associated with a reward, given by a known function $r_t : S \times A \to \mathbb{R}$. The decision-maker aims to maximize the total reward from all N arms over a finite horizon subject to the budget constraint.

To complete the formal definition of our N-arm problem, we introduce some additional notation. We use $\mathbb{S} = S^N$ to denote the N-fold Cartesian product of the state space S and define $\mathbb{A} = A^N$ similarly. All N arms together form an MDP with state space $\mathbb{S}$ and action space $\mathbb{A}$. We call this the "joint MDP" to distinguish it from MDPs that we reference later involving a single arm. The state in this joint MDP at time t is $\mathbf{s}_t = (s_{t,1}, s_{t,2}, ..., s_{t,N}) \in \mathbb{S}$, which indicates that arm i has state $s_{t,i}$. The action is $\mathbf{a}_t = (a_{t,1}, a_{t,2}, ..., a_{t,N}) \in \mathbb{A}$, which indicates that action $a_{t,i}$ is applied to arm i.

The reward function of the joint MDP, $R_t : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$, is the sum of the single-arm rewards defined above, $R_t(\mathbf{s}_t, \mathbf{a}_t) = \sum_{i=1}^N r_t(s_{t,i}, a_{t,i})$.

For element $\mathbf{a} = (a_1, a_2, ..., a_N)$ in $\mathbb{A}$, we use $|\mathbf{a}| = \sum_{i=1}^N a_i$ to indicate the L^1-norm of $\mathbf{a}$, i.e, the number of pulled arms. We write our budget constraint at time t as $|\mathbf{a}_t| = B_t$.

The transition kernel for the joint MDP is the product of each arm's transition kernel,

$$\mathbb{P}[\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t] = \prod_{i=1}^N p_t(s_{t,i}, a_{t,i}, s_{t+1,i}).$$

We assume all arms start from the same initial state s^*. Our analysis can be

easily generalized to the case where arms start from different states.

A policy π is a function that maps the current state $\mathbf{s}_t \in \mathbb{S}$ and time t to an action $\mathbf{a}_t \in \mathbb{A}$. The objective of the policy is to maximize the expected total reward, subject to the budget constraint specified above. This objective can be written as,

$$\begin{aligned} &\max_{\pi} \ \mathbb{E}_{\pi} \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) \\ &\text{subject to: } |\mathbf{a}_t| = \lfloor \alpha_t N \rfloor, \ \forall t \in [T], \end{aligned} \tag{2.1}$$

where $\mathbb{E}_{\pi}$ indicates the expectation taken under policy π.

We define the value function of a policy π as $V_N(\pi) = \mathbb{E}_{\pi} \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t)$ and measure a policy's performance by comparing its value with that of an optimal policy solving (2.1). Let $V_N^* = \sup_{\pi} V_N(\pi)$ be the value of an optimal policy. Then the opt gap of a policy π is defined as $V_N^* - V_N(\pi)$. Maximizing the value function across policies is equivalent to minimizing the opt gap. We are interested in finding policies with small opt gaps when N is large.

Applications: The above model has many applications. In the most direct application, each arm corresponds to a physical process that evolves independently of the other physical processes according to a known transition kernel, e.g. network communication [38, 4] and machine maintenance [24, 18]. For example, in maintenance of military aircraft with low radar visibility (so-called "stealth" aircraft) [18], each aircraft is treated as an arm. Radar visibility (the state of the arm) increases stochastically according to a known transition kernel each time the aircraft flies as small particles in the air damage its paint. This damage can be repaired (the arm can be pulled) by pausing an aircraft's flights and performing maintenance. We must allocate limited maintenance resources to maximize an objective combining flights flown and the number of aircraft

with low radar visibility.

There are also many applications where information evolves over time. In such applications, each arm corresponds to one of several independent unknown quantities. An arm's state represents our information about this quantity. Examples include autonomous target tracking [34, 30]: each target is an arm and its state includes whether it is tracked by a sensor and physical features affecting its motion. Based on its state, the target moves to a new location and our goal is to track as many targets for as long as possible.

In perhaps the most famous restless bandit, each arm corresponds to a slot machine. Each slot machine generates payoffs according to a distribution from a parametric family (e.g., Bernoulli). The parameter governing an arm's rewards (for Bernoulli arms, the payoff probability) is drawn at random from a Bayesian prior distribution and is unobserved. The state of the arm is the Bayesian posterior distribution on its parameter, given all observed payoffs from the arm. When we pull an arm, we earn a reward (whose distribution is given by marginalizing over the posterior on the arm's uncertain parameter) and the new state is determined by Bayes' rule and the observed reward. If an arm's underlying parameter changes over time, then this causes the posterior to change even if the arm is not pulled, making the problem restless.

2.3 Background: Preliminary Results and Notation

In this section, we define a LP relaxation that provides an upper bound $\hat{V}_N^*$ for V_N^*. Although this bound is standard in the literature and is not part of our contribution, we include it to provide a self-contained presentation and to establish

notation used later.

Linear Programming Relaxation: Similar to [20, 26], we introduce this relaxation of Problem (2.1):

$$\begin{aligned} \hat{V}_N^* := \max_{\pi} \ & \mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) \\ \text{subject to } & \mathbb{E}_\pi |\mathbf{a}_t| = \alpha_t N, \ \forall t \in [T]. \end{aligned} \tag{2.2}$$

This relaxes Problem (2.1)'s almost sure cardinality constraints on the number of pulls to constraints on the expected cardinality. As we will see soon, solving relaxation (2.2) is equivalent to solving a linear program whose number of decision variables does not depend on N (see Lemma 2.1 and Problem (2.4)). For simplicity of presentation, we assume that α_t are rational and we restrict attention and limits over N to those N with integral $\alpha_t N$ for all $t \in [T]$. Our results generalize to irrational α_t and non-integral $\alpha_t N$ as discussed in Appendix A.2.

The value of this relaxed problem, $\hat{V}_N^*$, bounds V_N^* above. We use this to bound the opt gap of the policies we study. Moreover, the policies we study in §2.6 heavily leverage this relaxation in their definition. They benefit from the fact that the relaxation yields a low-dimensional problem whose number of decision variables and constraints do not scale with N. This allows the relaxation's solution to be computed and used to define practical policies, even when N is large.

The following lemma formally states this bound and also observes (via Fenchel's duality theorem, and the separability of a dualized version of Problem (2.2)) that $\hat{V}_N^*$ is determined by the solution to a single-armed problem $\hat{V}_1^*$. Its proof can be found in Appendix A.1. Later in the paper whenever we omit the technical proof for a lemma, theorem or proposition, its proof can be found in the Appendix.

Lemma 2.1. *$V_N^* \le \hat{V}_N^* = N\hat{V}_1^*$ where the quantity $\hat{V}_1^*$ is given by,*

$$\begin{aligned} &\max_{\pi} \ \mathbb{E}_\pi \sum_{t=1}^{T} r_t(s_t, a_t) \\ &\textit{subject to } \mathbb{E}_\pi |a_t| = \alpha_t, \ \forall t \in [T]. \end{aligned} \tag{2.3}$$

Later analysis and computation is supported by the following equivalent version of Problem (2.3). Defining the occupation measure, $x_t(s,a) := \mathbb{P}[s_t = s, a_t = a]$, Problem (2.3) is equivalent to

$$\max \sum_{s\in S, a\in A} \sum_{t=1}^{T} r_t(s,a) x_t(s,a)$$

subject to

$$\begin{aligned} &\sum_{a\in A} x_t(s,a) = \sum_{a\in A}\sum_{s'\in S} x_{t-1}(s',a) p_{t-1}(s',a,s), \ \forall s \in S, 2 \le t \le T; \\ &\sum_{s\in S} x_t(s,1) = \alpha_t, \ t \in [T]; \ \sum_{a\in A} x_1(s^*,a) = 1; \\ &\sum_{a\in A}\sum_{s\in S} x_1(s,a) = 1, \ \forall s \in S; \ x_t(s,a) \ge 0, \ \forall s \in S, a \in A, t \in [T]. \end{aligned} \tag{2.4}$$

The first constraint ensures that flows are balanced; the second ensures that the budget constraint is met; the third enforces the initial occupation measure; and the fourth and fifth ensure that x_t is a probability distribution for each t. We let $x_t(s,a)$ denote the entries in an optimal occupation measure, i.e., one that solves Problem (2.4). Then, we can compute, $\hat{V}_1^* = \sum_{s\in S, a\in A} \sum_{t=1}^{T} r_t(s,a) x_t(s,a)$.

The class of policies we analyze depend on solving Problem (2.4) computationally using a LP solver. As noted above, this is possible, even when N is large, because the dimensionality of Problem (2.4) does not depend on the number of arms N.

Additional Notation: Here we introduce some additional notation used in the following sections. Given the optimal occupation measure, we use $z_t(s) :=$

$\sum_{a\in A} x_t(s,a)$ to denote the probability that an arm is in state s at time t under this measure. We use z_t and x_t to refer to the corresponding vector (or matrix), i.e., $z_t := (z_t(s), s \in S)$ or $x_t := (x_t(s,a) : s \in S, a \in A)$.

In the joint MDP with N arms, we let $X_t^N(s,a)$ be the number of arms in state s for which we take action a at time t. We let $Z_t^N(s)$ be the number of arms in state s at time t. We use Z_t^N, X_t^N to refer to the vectors $(Z_t^N(s) : s \in S)$ and matrix $(X_t^N(s,a) : s \in S, a \in A)$. Using this notation, a policy π of the joint MDP is a map from Z_t^N to X_t^N.

§2.5 will study deviations between the realization of (Z_t^N, X_t^N) and (Nz_t, Nx_t), and how these deviations impact the joint MDP's reward. To support this analysis, we define *diffusion statistics* $\tilde{Z}_t^N = \frac{Z_t^N - Nz_t}{\sqrt{N}}$ and $\tilde{X}_t^N = \frac{X_t^N - Nx_t}{\sqrt{N}}$. Using this notation, a policy π of the joint MDP naturally induces a class of maps $\tilde{\pi}_{t,N}$ indexed by t and N, from diffusion $\tilde{Z}_t^N$ to diffusion $\tilde{X}_t^N$, such that

$$\pi(t, Z_t^N) = X_t^N \iff \tilde{\pi}_{t,N}(\tilde{Z}_t^N) = \tilde{X}_t^N. \tag{2.5}$$

2.4 Sufficient Conditions for Achieving an $o(N)$ Opt Gap

This section establishes the first of our contributions: general sufficient conditions for an $o(N)$ opt gap. This result allows us to directly verify that the policy in [57] has an $o(N)$ opt gap. We build on the results here in the next section, where we give stronger conditions sufficient for an $O(\sqrt{N})$ gap and apply it to the policies in [31] and [13]. This is in preparation for our main contribution in §2.6.

The main idea in this section is, essentially, that as long as the number of

arms we pull in each state, X_t^N, is approximately proportional to the optimal occupation measure x_t (a property we call "fluid consistency"), the number of arms in the next period Z_{t+1}^N in each state will be approximately proportional to z_{t+1}. This will cause the reward of the joint MDP to scale proportionally with $\hat{V}_1^*$. While random fluctuations cause proportionality to hold only approximately, their resulting loss in reward is $o(N)$. We begin by formally defining fluid consistency.

Definition 2.1. *Under a policy π, if $\pi(t, Z_t^N)/N \to x_t$ for all $t \in [T]$ and sequences $(Z_t^N : N)$ satisfying $Z_t^N/N \to z_t$, then we say the policy π is fluid consistent.*

Based on this definition, we have the following lemma.

Lemma 2.2. *If a policy π is fluid consistent, then $\frac{Z_t^N}{N} \to z_t$ and $\frac{X_t^N}{N} \to x_t$ a.s. for any $t \in [T]$ as $N \to \infty$.*

Using Lemma 2.2, we now show the main result of this section: that fluid consistency implies the opt gap is $o(N)$.

Theorem 2.1. *If a policy π is fluid consistent, then $V_N^* - V_N(\pi) = o(N)$.*

Proof of Theorem 2.1. Because the policy π is fluid consistent, Lemma 2.2 shows $\frac{Z_t^N}{N} \to z_t$ and $\frac{X_t^N}{N} \to x_t$. Then the total reward of the joint MDP, divided by N, is

$$\frac{1}{N}\mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) = \frac{1}{N}\mathbb{E}_\pi \sum_{t=1}^{T} \sum_{s\in S, a\in A} r_t(s,a) X_t^N(s,a) = \mathbb{E}_\pi \sum_{t=1}^{T} \sum_{s\in S, a\in A} r_t(s,a) \frac{X_t^N(s,a)}{N}.$$

This has a limit of $\mathbb{E}_\pi \sum_{t=1}^{T} \sum_{s\in S, a\in A} r_t(s,a) x_t(s,a)$ as $N \to \infty$, where we leverage the dominated convergence theorem, the fact that rewards are bounded, and $0 \le X_t^N(s,a) \le N$. □

One can show that the policies in [31, 57, 13] are all fluid consistent and thus have $o(N)$ opt gaps. We show this for [57] in Appendix A.13. Below, we show that [31, 13] meet a stronger condition and thus have $O(\sqrt{N})$ opt gaps.

2.5 Sufficient Conditions for Achieving an $O(\sqrt{N})$ Opt Gap

This section establishes our second contribution: a substantially more general result than in the literature showing sufficient conditions for an $O(\sqrt{N})$ opt gap. Using this result, we directly verify that policies in [31] and [13] have $O(\sqrt{N})$ opt gaps. This section also provides stepping stones towards our main contribution, described in §2.6.

The main idea in this section is that, as long as the diffusion statistic $\tilde{X}_t^N$ is bounded by $O(1)$, then $\tilde{Z}_{t+1}^N$ will also be bounded by $O(1)$. Thus, the deviation between the reward of the joint MDP and the relaxation's upper bound $\hat{V}_N^*$ will be bounded by $\sqrt{N} \cdot O(1) = O(\sqrt{N})$.

Recall Equation (2.5), that a policy π naturally induces a class of maps $\tilde{\pi}_{t,N}$. Using this idea, we say a policy π is "diffusion regular" if all induced maps $\tilde{\pi}_{t,N}$ keep the diffusion $\tilde{X}_t^N$ bounded by $O(1)$. We define this formally here.

Definition 2.2. *A policy π is called diffusion regular if its induced maps $\tilde{\pi}_{t,N}$ satisfy the following conditions, where $|\cdot|$ indicates the L^1-norm in Euclidean space.*

1. *There exists $C_1 > 0$ s.t. $|\tilde{\pi}_{t,N}(\theta_1) - \tilde{\pi}_{t,N}(\theta_2)| \leq C_1|\theta_1 - \theta_2|$ for all t, N, θ_1 and θ_2.*
2. *There exists $C_2 > 0$ s.t. $|\tilde{\pi}_{t,N}(0)| \leq C_2$ for all t and N.*
3. *There exists a map $\tilde{\pi}_{t,\infty}$ s.t. $\tilde{\pi}_{t,N}(\theta) \to \tilde{\pi}_{t,\infty}(\theta)$ as $N \to \infty$ for all θ.*

We briefly note the following useful fact.

Lemma 2.3. *If a policy is diffusion regular then it is also fluid consistent.*

We now show that if a policy π is diffusion regular, the diffusion statistics $\tilde{X}_t^\infty$ and $\tilde{Z}_t^\infty$ converge in distribution (Lemma 2.4) and their second moments are uniformly bounded (Lemma 2.5).

Lemma 2.4. *If a policy π is diffusion regular, then for any $t \in [T]$, there exists sub-Gaussian random vectors $(\tilde{Z}_t^\infty, \tilde{X}_t^\infty)$ such that $(\tilde{Z}_t^N, \tilde{X}_t^N) \to (\tilde{Z}_t^\infty, \tilde{X}_t^\infty)$ in distribution as $N \to \infty$.*

Lemma 2.5. *If a policy π is diffusion regular, then there exists a constant C such that $\mathbb{E}_\pi[||\tilde{Z}_t^N||_2^2] \le C$ and $\mathbb{E}_\pi[||\tilde{X}_t^N||_2^2] \le C$ for all $t \in [T]$ and N, where $||\cdot||_2$ indicates the L^2 norm.*

Based on Lemma 2.4 and 2.5, we can prove the following theorem.

Theorem 2.2. *If a policy π is diffusion regular, then $V_N^* - V_N(\pi) = O(\sqrt{N})$.*

Proof of Theorem 2.2. Since the policy π is diffusion regular, there exists sub-Gaussian random vectors $\tilde{Z}_t^\infty, \tilde{X}_t^\infty$, such that $\tilde{Z}_t^N \to \tilde{Z}_t^\infty$ and $\tilde{X}_t^N \to \tilde{X}_t^\infty$ in distribution as $N \to \infty$ by Lemma 2.4.

Also, the opt gap is bounded above by

$$V_N^* - V_N(\pi) \le N\hat{V}_1^* - V_N(\pi) = -\sqrt{N}\mathbb{E}_\pi \sum_{t=1}^{T} \sum_{s\in S, a\in A} r_t(s,a)\tilde{X}_t^N(s,a).$$

Divide both sides of this bound by $\sqrt{N}$ and take $N \to \infty$. Then, since $\tilde{X}_t^N$ and $\hat{Y}_t^N$ are uniformly integrable (Lemma 2.5),

$$\limsup_N \frac{1}{\sqrt{N}}(V_N^* - V_N(\pi)) \le \limsup_N -\mathbb{E}_\pi \sum_{t=1}^{T} \sum_{s\in S, a\in A} r_t(s,a)\tilde{X}_t^N(s,a) = -\mathbb{E}_\pi \sum_{t=1}^{T} \sum_{s\in S, a\in A} r_t(s,a)\tilde{X}_t^\infty(s,a).$$

To summarize, we have shown $V_N^* - V_N(\pi) = O(\sqrt{N})$. □

We verify that the policies proposed by [31] and [13] are diffusion regular (in Appendix A.13) and thus (by Theorem 2.2) have $O(\sqrt{N})$ opt gaps. Thus, Theorem 2.2 generalizes the performance guarantees shown in that previous work.

2.6 Fluid-priority Policies

This section defines fluid-priority policies and show that they are always diffusion regular and thus achieve an $O(\sqrt{N})$ opt gap. Later, in §7, we show that they achieve an $O(1)$ opt gap if an additional condition is satisfied.

Roughly speaking, a fluid-priority policy is defined by first fetching an optimal solution of the LP relaxation, then classifying states into three disjoint categories based on the solution: fluid-active, fluid-neutral and fluid-inactive. A fluid priority policy is one that pulls arms while respecting a prioritization derived from these categories: arms in fluid-active states are prioritized for pulling over those in fluid-neutral states; and arms in fluid-neutral states are prioritized in turn over arms in fluid-inactive states. Additionally, when pulling arms in fluid-neutral states, a fluid-priority policy must do so according to proportions derived from the LP relaxation.

Mathematically speaking, a fluid-priority policy is parameterized by an occupation measure $\{x_t(s,a)\}_{t,s,a}$ solving Problem (2.4) and a sequence of "priority-score" functions $\{\mathcal{P}_t(\cdot)\}_t$ assigning each state a real number. Based on the occupation measure $\{x_t(s,a)\}_{t,s,a}$, a fluid-priority policy classifies states into these

three disjoint categories:

$$\begin{aligned}
&\text{The } \textit{fluid-active} \text{ category: } C_t^+ := \{s \in S \,|\, x_t(s,1) > 0, x_t(s,0) = 0\},\\
&\text{The } \textit{fluid-neutral} \text{ category: } C_t^0 := \{s \in S \,|\, x_t(s,1) > 0, x_t(s,0) > 0\},\\
&\text{The } \textit{fluid-inactive} \text{ category: } C_t^- := \{s \in S \,|\, x_t(s,1) = 0, x_t(s,0) = 0\}.
\end{aligned} \tag{2.6}$$

We refer to an arm with its state in the fluid-active category as a *fluid-active arm*. We define a *fluid-neutral arm* and a *fluid-inactive arm* similarly. With these definitions, the fluid-priority policy corresponding to an occupation measure and priority-score function is defined by Algorithm 1.

Algorithm 1 allocates its budget by first pulling as many fluid-active arms as possible, subject to the budget constraint (Lines 5-7). If budget remains, then it pulls as many fluid-neutral arms as possible, again subject to the constraint on the remaining budget (Lines 9-17).

When there is enough budget to pull some fluid-neutral arms, but not all of them, the budget is allocated carefully across them to ensure fluid-consistency. This is related to "tie-breaking" discussed in Algorithm 2 of [31]. In particular, lines 9-13 ensure that the number of arms pulled in each fluid-neutral state is at least equal to $\lfloor N x_t(s,1) \rfloor$, the number of arms from this state pulled in the fluid relaxation, as long as the budget constraint B_t and number of available arms $Z_t(s)$ allows. If budget remains after this is achieved, additional fluid-neutral arms are pulled.

Finally, if budget remains after all fluid-neutral arms are pulled, additional fluid-inactive arms are pulled until the budget is exhausted. Within each category (fluid-active, fluid-neutral, fluid-inactive), states are prioritized based on the priority score.

Algorithm 1 Fluid-priority policy

Input: optimal occupation measure $(x_t(s,a))_{t\in[T],s\in S,a\in A}$, found by solving the LP (2.4), and priority-score functions $\{\mathcal{P}_t\}_{t\in[T]}$.

for $t = 1, 2, ..., T$ **do**
  Classify states into fluid-active (C_t^+), fluid-neutral (C_t^0) and fluid-inactive (C_t^-) categories based on the occupation measure, according to equation (2.6).
  Observe there are $Z_t(s)$ arms in state s and remaining budget $B_t = \lfloor \alpha_t N \rfloor$.
  Each of the for loops below iterates over states in decreasing order of $\mathcal{P}_t(s)$
  for state s in C_t^+ **do**
    Plan to pull $X_t^N(s,1) \leftarrow \min\{B_t, Z_t(s)\}$ arms out of the $Z_t(s)$ arms in state s.
    Update remaining budget $B_t \leftarrow B_t - \min\{B_t, Z_t(s)\}$.
  end for
  for state s in C_t^0 **do**
    Plan to pull (at least) $X_t^N(s,1) \leftarrow \min\{B_t, Z_t(s), \lfloor N x_t(s,1) \rfloor\}$ arms in state s.
    Store the number of undecided arms $U_t^N(s) \leftarrow Z_t^N(s) - X_t^N(s,1)$.
    Update the remaining budget $B_t \leftarrow B_t - \min\{B_t, Z_t(s), \lfloor N x_t(s,1) \rfloor\}$.
  end for
  for state s in C_t^0 **do**
    Plan to pull $\min\{B_t, U_t(s)\}$ additional undecided arms in state s.
    Update $X_t^N(s,1) \leftarrow X_t^N(s,1) + \min\{B_t, U_t(s)\}$.
    Update $B_t \leftarrow B_t - \min\{B_t, U_t(s)\}$.
  end for
  for state s in C_t^- **do**
    Plan to pull $X_t^N(s,1) \leftarrow \min\{B_t, Z_t(s)\}$ arms out of the $Z_t(s)$ arms in state s.
    Update remaining budget $B_t \leftarrow B_t - \min\{B_t, Z_t(s)\}$.
  end for
  For each s, pull $X_t^N(s,1)$ arms in state s (as planned above)
end for

With this definition in place, we now state the main result of this section: that fluid-priority policies are diffusion regular, implying they have an $O(\sqrt{N})$ opt gap by Theorem 2.2.

Theorem 2.3. *Any fluid-priority policy π is diffusion regular and its optimality gap is $O(\sqrt{N})$.*

2.7 Non-degeneracy Condition: Achieving an O(1) Opt Gap

This section presents our main contribution: that fluid-priority policies achieve an $O(1)$ opt gap under a non-degeneracy condition. We define and discuss this condition before showing this result.

To motivate this non-degeneracy condition, consider a fluid-priority policy and another policy motivated by the relaxed problem (2.2) in which the almost-sure budget constraint ($|\mathbf{a}_t| = \alpha_t N$) has been relaxed. This so-called "budget-relaxed" policy first categorizes states into fluid-active, fluid-neutral, and fluid-inactive categories in the same way as its corresponding fluid-priority policy. It pulls all fluid-active arms (even if this would exceed the budget). If budget remains, it then pulls fluid-neutral arms in the same way as its corresponding fluid-priority policy. It does not pull any fluid-inactive arms, even if budget remains after fluid-active and fluid-neutral arms are pulled.

Pulling all fluid-active arms and idling fluid-inactive arms is exactly the property required for any feasible policy to be optimal in the LP relaxation (2.2). Thus, this budget-relaxed policy's reward is close to the relaxed problem's optimal reward (Lemma 2.8). Moreover, it behaves identically to its corresponding fluid-priority policy (Lemma 2.6) except on a specific "budget violation" event:

that the number of fluid-active arms exceeds the budget, or the number of fluid-active and fluid-neutral arms together fail to exceed the budget. The probability of budget-violation allows us to bound the opt gap for fluid-priority policies by comparing them with their budget-relaxed versions.

The non-degeneracy condition (Definition 2.3) characterizes the probability of budget violation: when it is met, the expected number of fluid-active arms is *strictly* below the budget and the expected number of fluid-active and fluid-neutral arms is *strictly* above the budget. Thus, using concentration bounds, problems meeting the non-degeneracy condition are ones in which the probability of budget violation vanishes exponentially fast as N grows (Lemma 2.7). As a result, in such problems, the fluid-priority policy behaves the same as its budget-relaxed version with high probability for large N. We use this fact to show an $O(1)$ opt gap in Theorem 2.4.

In the rest of this section, we first formally introduce budget-relaxed policies, then define the non-degeneracy condition, and finally prove fluid-priority policies achieve an $O(1)$ opt gap when this condition holds. We also show that no index policy can strictly outperform all fluid-priority policies. Since the LP relaxation can have multiple optimal occupation measures, some being degenerate and others non-degenerate, we provide an algorithm in Appendix A.14 to identify a non-degenerate occupation measure when one exists.

2.7.1 Budget-relaxed fluid-priority policies

Given a fluid-priority policy π_F, its budget-relaxed version π_R is defined by Algorithm 2. Similar to π_F, π_R first classifies states into three categories: fluid-

active, fluid-neutral and fluid-inactive, using the same occupation measure as π_F. Then, π_R sorts states in each category in order of decreasing priority-score (line 4), using the same priority score as π_F. Afterwards, π_R pulls all arms in the fluid-active category (lines 5 - 8), exceeding the budget if necessary. If there is still budget remaining, π_R iterates over each state s in the fluid-neutral category C_t^0. It pulls arms in this state until the number pulled reaches the quantity $\lfloor Nx_t(s,1) \rfloor$ derived from the optimal occupation measure, no arms remain in this state, or no budget remains. Unpulled arms in each such state are called "undecided" (lines 12 - 16). Finally, π_R iterates over each state in the fluid-neutral category C_t^0 again and pulls undecided arms until either no budget remains or all undecided arms are pulled (lines 17 - 21). Notice π_R idles all arms in the fluid-inactive category, even if budget remains.

Policy π_R behaves the same as its corresponding fluid-priority policy π_F, except π_R pulls all arms in the fluid-active category and idles all arms in the fluid-inactive category regardless of the budget constraint. This is formally stated as Lemma 2.6.

Lemma 2.6. *Define event* $\Delta_t := \left\{ \sum_{s \in C_t^+} Z_t^N(s) \leq \alpha_t N \leq \sum_{s \in C_t^+} Z_t^N(s) + \sum_{s \in C_t^0} Z_t^N(s) \right\}$. *Then on the event* Δ_t, $\pi_R(t, Z_t^N) = \pi_F(t, Z_t^N)$.

We write the complement of Δ_t as Δ_t^c and refer to this as a "budget violation" event.

2.7.2 Non-degeneracy

The non-degeneracy condition (stated formally below) states that the fluid-neutral category is not empty, which is sufficient to prove the budget-violation

Algorithm 2 Budget-relaxed fluid-priority policies

Input: optimal occupation measure $(x_t(s,a))_{t\in[T],s\in S,a\in A}$, priority-score function $\{\mathcal{P}_t\}_{t\in[T]}$.

for $t = 1, 2, ..., T$ **do**
  Classify states into fluid-active (C_t^+), fluid-neutral (C_t^0) and fluid-inactive (C_t^-) categories based on the occupation measure, according to equation (2.6).
  Observe there are $Z_t(s)$ arms in state s, and remaining budget $B_t = \lfloor \alpha_t N \rfloor$.
  Each of the for loops below iterates over states in decreasing order of $\mathcal{P}_t(s)$.
  for state s in C_t^+ **do**
    Plan to pull $X_t^N(s,1) \leftarrow Z_t(s)$ arms out of $Z_t(s)$ arms in state s.
    Update the remaining budget $B_t \leftarrow B_t - Z_t(s)$.
  end for
  if $B_t \leq 0$ **then**
    continue
  end if
  for state s in C_t^0 **do**
    Plan to pull (at least) $X_t^N(s,1) \leftarrow \min\{B_t, Z_t(s), \lfloor N x_t(s,1) \rfloor\}$ in state s.
    Store the number of undecided arms $U_t^N(s) \leftarrow Z_t^N(s) - X_t^N(s,1)$.
    Update the remaining budget $B_t \leftarrow B_t - \min\{B_t, Z_t^N(s), N x_t(s,1)\}$.
  end for
  for state s in C_t^0 **do**
    Plan to pull $\min\{B_t, U_t(s)\}$ additional undecided arms in state s.
    Update $X_t^N(s,1) \leftarrow X_t^N(s,1) + \min\{B_t, U_t(s)\}$,
    Update $B_t \leftarrow B_t - \min\{B_t, U_t(s)\}$.
  end for
  For each $s \in C_t^+ \cup C_t^0$, pull $X_t^N(s,1)$ arms in state s (as planned above). Idle all arms in C_t^-.
end for

events, Δ_t^c, are probabilistically negligible. Roughly speaking, non-emptiness of C_t^0 guarantees that the occupation measure satisfies, $\sum_{s\in C_t^+} z_t(s) < \alpha_t < \sum_{s\in C_t^+} z_t(s) + \sum_{s\in C_t^0} z_t(s)$. Since both the budget-relaxed fluid-priority policy and the fluid-priority policy are fluid consistent, the number of arms in state s, $Z_t^N(s)$, is roughly proportional to $z_t(s)$, with excursions described by a central limit theorem. Thus the probability of event Δ_t approaches 1 exponentially fast as N grows by concentration inequalities.

Definition 2.3 formally defines non-degeneracy and Lemma 2.7 states that non-degeneracy implies that budget violations are probabilistically negligible for large N.

Definition 2.3. *We say an optimal occupation measure $(x_t(s,a))_{t\in[T],s\in S,a\in A}$ is non-degenerate if $\forall t \in [T]$, $|C_t^0| \geq 1$. Otherwise, we call it degenerate. We also call a fluid-priority policy non-degenerate (degenerate) when its associated occupation measure is non-degenerate (degenerate).*

Lemma 2.7. *If an optimal occupation measure $(x_t(s,a))_{t\in[T],s\in S,a\in A}$ is non-degenerate, then for any priority-score functions $\{\mathcal{P}_t\}_{t\in[T]}$ and the corresponding fluid-priority policy π_F and budget-relaxed policy π_R, there exists a constant $\delta > 0$ and a constant L such that for all $t \in [T]$ and all N,*

$$\max\{\mathbb{P}_{\pi_R}(\Delta_t^c), \mathbb{P}_{\pi_F}(\Delta_t^c)\} \leq L\exp(-\delta N).$$

Empirically, one can check the non-degeneracy condition for a given optimal occupation measure x^* returned by solving the LP relaxation (2.4) with a commercial LP solver. Recalling from (2.6) that states s in the fluid-neutral category are those with both $x_t(s,1) > 0$ and $x_t(s,0) > 0$, we check whether x^* is non-degenerate by assessing whether there is at least one such state for each t. If there are multiple optimal occupation measures, then some may be degenerate and others non-degenerate. Since a fluid-priority policy's optimality guarantee is significantly stronger when it uses a non-degenerate optimal occupation measure, it is important to be able to find one if it exists. We present an algorithm for doing so in Appendix A.14.

2.7.3 Main result

We now state and prove this section's main result: a fluid-priority policy achieves an $O(1)$ opt gap when it is non-degenerate. Before that, we need one last building block: the budget-relaxed policies' reward deviates from the relaxed problem's optimal reward by $O(1)$ under non-degeneracy.

We show this in the following lemma. There are two main ideas in the proof. First, recall that the budget-relaxed fluid-priority policy π_R pulls all arms in C_t^+ and no arms in C_t^-. Thus, its decisions are optimal under a Lagrangian relaxation of Problem 2.2 in which the budget constraint on the expected number of arms pulled is replaced by a well-chosen linear penalty (in this Lagrangian relaxation, pulling fluid-neutral arms does not affect optimality as the incremental reward is offset by the linear penalty). Second, the fact that π_R pulls a number of arms equal to the almost-sure budget constraint with high probability ensures that it nearly satisfies the constraint in Problem 2.2 on the expected budget. This fact causes the linear penalty to be nearly 0. Finally, the fact that the Lagrangian relaxation is the sum of this penalty and the unpenalized reward, which we call $V_N(\pi_R)$, implies that $V_N(\pi_R)$ is within a constant of $\hat{V}_N^*$.

Lemma 2.8. *Let $V_N(\pi_R) = \max_\pi \mathbb{E}_\pi \sum_{t=1}^T R_t(\mathbf{s}_t, \mathbf{a}_t)$ for a budget-relaxed fluid priority policy π_R. If an optimal occupation measure $(x_t(s,a))_{t\in[T],s\in S,a\in A}$ is non-degenerate, then for any priority-score functions $\{\mathcal{P}_t\}_{t\in[T]}$, the corresponding budget-relaxed fluid-priority policy π_R satisfies $|\hat{V}_N^* - V_N(\pi_R)| \leq m$, where m is a constant not depending on N.*

Now we are ready to state and prove our main result: that a fluid-priority policy achieves an $O(1)$ opt gap when it is non-degenerate. A fluid-priority policy π_F's opt gap can be bounded by first comparing the reward $V_N(\pi_F)$ with

the reward of its corresponding budget-relaxed policy π_R. Combining the fact that π_F deviates from π_R with negligible probability (Lemma 2.7) and that π_R's reward deviates by $O(1)$ from $\hat{V}_N^*$ (Lemma 2.8), $V_N(\pi_F)$ is at most $O(1)$ away from $\hat{V}_N^*$.

Theorem 2.4. *If an optimal occupation measure $(x_t(s,a))_{t\in[T],s\in S,a\in A}$ is non-degenerate, then for any priority-score functions $\{\mathcal{P}_t\}_{t\in[T]}$, the corresponding fluid-priority policy π_F satisfies $\hat{V}_N^* - V_N(\pi_F) \le m$, where m is a constant not depending on N.*

Proof of Theorem 2.4. Under π_F, the reward is $V_N(\pi_F) = \mathbb{E}_{\pi_F} \sum_{t=1}^{T} \sum_{s\in S,a\in A} r_t(s,a) X_t^N(s,a)$. Under π_R, the reward is $V_N(\pi_R) = \mathbb{E}_{\pi_R} \sum_{t=1}^{T} \sum_{s\in S,a\in A} r_t(s,a) X_t^N(s,a)$.

Denote $\Omega_T := \Delta_1 \cap \Delta_2 \cap ... \cap \Delta_T$. On this event, π_R and π_F produce identical decisions by Lemma 2.6. Using this in the second line below, we have:

$$\begin{aligned}
V_N(\pi_R) - V_N(\pi_F) &= \mathbb{E}_{\pi_R}\Big[1_{\Omega_T} \sum_{t=1}^{T} \sum_{s\in S,a\in A} r_t(s,a) X_t^N(s,a)\Big] + \mathbb{E}_{\pi_R}\Big[1_{\Omega_T^c} \sum_{t=1}^{T} \sum_{s\in S,a\in A} r_t(s,a) X_t^N(s,a)\Big] \\
&\quad - \mathbb{E}_{\pi_F}\Big[1_{\Omega_T} \sum_{t=1}^{T} \sum_{s\in S,a\in A} r_t(s,a) X_t^N(s,a)\Big] - \mathbb{E}_{\pi_F}\Big[1_{\Omega_T^c} \sum_{t=1}^{T} \sum_{s\in S,a\in A} r_t(s,a) X_t^N(s,a)\Big] \\
&= \mathbb{E}_{\pi_R}\Big[1_{\Omega_T^c} \sum_{t=1}^{T} \sum_{s\in S,a\in A} r_t(s,a) X_t^N(s,a)\Big] - \mathbb{E}_{\pi_F}\Big[1_{\Omega_T^c} \sum_{t=1}^{T} \sum_{s\in S,a\in A} r_t(s,a) X_t^N(s,a)\Big] \\
&\le \mathbb{E}_{\pi_R}\Big[1_{\Omega_T^c} \sum_{t=1}^{T} \sum_{s\in S,a\in A} |r_t(s,a)||X_t^N(s,a)|\Big] + \mathbb{E}_{\pi_F}\Big[1_{\Omega_T^c} \sum_{t=1}^{T} \sum_{s\in S,a\in A} |r_t(s,a)||X_t^N(s,a)|\Big].
\end{aligned}$$

Inequalities $0 \le X_t^N(s) \le N$ and $0 \le Y_t^N(s) \le N$ then imply

$$\begin{aligned}
V_N(\pi_R) - V_N(\pi_F) &\le \mathbb{E}_{\pi_R}\Big[1_{\Omega_T^c} \sum_{t=1}^{T} \sum_{s\in S,a\in A} |r_t(s,a)|N\Big] + \mathbb{E}_{\pi_F}\Big[1_{\Omega_T^c} \sum_{t=1}^{T} \sum_{s\in S,a\in A} |r_t(s,a)|N\Big] \\
&\le 2\left(\mathbb{E}_{\pi_R}[1_{\Omega_T^c}] + \mathbb{E}_{\pi_F}[1_{\Omega_T^c}]\right) T|S|r^*N,
\end{aligned}$$

where $r^* := \max_{t,s,a} |r_t(s,a)|$. Then, applying Lemma 2.7 and $\mathbb{P}_\pi[\Omega_T^c] \le \sum_{t=1}^{T} \mathbb{P}_\pi[\Delta_t^c]$, we have:

$$V_N(\pi_R) - V_N(\pi_F) \le 2T|S|r^*N \sum_{t=1}^{T} \mathbb{P}_{\pi_R}[\Delta_t^c] + \mathbb{P}_{\pi_F}[\Delta_t^c] \le 4T^2|S|r^*NL\exp(-\delta N).$$

Finally, applying Lemma 2.8 concludes the proof. □

2.7.4 The best fluid-priority policy is at least as good as the best index policy

Here we compare fluid-priority policies against index policies [23]. An index policy assigns each state an "index" and prioritizes each arm based on the index of its current state from high to low, pulling arms until we exhaust the current period's budget.

A policy can be both a fluid-priority policy and index policy. This occurs if there is at most one fluid-neutral state in any period and indices of all fluid-active states are higher than those of all fluid-neutral states, which are higher in turn than the indices of all fluid-inactive states. There are, however, index policies that are not fluid priority policies, and vice versa. If the indices do not respect the ordering implied by the fluid-active, fluid-neutral, and fluid-inactive categories then this index policy is not a fluid priority policy. Also, if multiple fluid-neutral states can be occupied in one period, a fluid-priority policy will allocate pulls across these arms in accordance with an occupation measure and in a way that is different from an index policy's strict prioritization.

Since index policies are widely known and used, it is instructive to compare them with fluid-priority policies. The discussion above shows that the difference between $\hat{V}_N^*$ (the optimal objective of the relaxation) and the value of a fluid-priority policy $V_N(\pi_F)$ is bounded above by a constant when π_F is non-degenerate, i.e., that $\sup_N \hat{V}_N^* - V_N(\pi_F)$ is finite. The following proposition shows

that the best fluid priority is at least as good as the best index policy, when measured by $\sup_N \hat{V}_N^* - V_N(\pi) \in \mathbb{R} \cup \{\infty\}$, regardless of whether non-degeneracy holds.

Proposition 2.1. *Consider an index policy π_I such that $\sup_N \hat{V}_N^* - V_N(\pi_I) < \infty$. Then, there exists a fluid priority policy π_F such that $\sup_N \hat{V}_N^* - V_N(\pi_F) \leq \sup_N \hat{V}_N^* - V_N(\pi_I)$.*

2.8 Numerical Experiments

This section evaluates the performance of fluid-priority and other policies on three problems, leveraging both simulation experiments, computational investigations and our earlier theoretical results.

§2.8.1 studies a classical problem: the finite-horizon Bayesian bandit with Bernoulli rewards. It compares a fluid-priority policy against UCB and TS policies. We first show the fluid-priority policy substantially outperforms both methods numerically. Then we show that both UCB and TS's opt gaps are $\Omega(N)$, while the opt gap is $O(1)$ for the fluid-priority policy as it is non-degenerate.

§2.8.2 considers an active learning problem based on [17] in which an algorithm allocates crowd workers to image labeling tasks to support learning an accurate classifier. We show via numerical experiments that fluid-priority policies outperform a state-of-the-art policy (Optimistic Knowledge Gradient [17]) specifically designed for this problem.

§2.8.3 shows via direct computation that the dynamic assortment problem previously studied in [13] satisfies the non-degeneracy condition. This and our main theoretical result shows that fluid-priority policies have an $O(1)$ opt gap,

explaining the poorly understood performance of Lagrangian index policies in [13]'s experiments.

2.8.1 Bayesian bandit with Bernoulli rewards

This section evaluates fluid-priority policies' performance on the Bayesian bandit problems, which is a standard benchmark in the bandit literature. While the problem is not restless, it allows us to study benchmarks designed for non-restless settings.

We compare the performance of a fluid-priority policy against UCB and TS policies and show that the fluid-priority policy achieves an $O(1)$ opt gap while the opt gaps of both UCB and TS grow linearly with the number of arms. While UCB and TS are well-known for having a logarithmic asymptotic performance guarantee of $O(N\log(T))$, this is linear in N. (It also applies to a slightly different problem setting than the one we study here: a stochastic frequentist setting with one pull period and where regret is measured with respect to the policy that pulls the best arm.) Thus, the classical regret guarantee for these policies is not inconsistent with our finding that these policies have a $\Omega(N)$ opt gap in a Bayesian analysis with multiple pulls per period.

This suggests that when T is small and N is large, and where prior information supports the use of a Bayesian analysis, there is significant value in using fluid-priority policies over UCB or TS.

Problem Setup: There are N arms, of which we may pull $\lfloor N/3 \rfloor$ in each of T periods. Before any arms are pulled, each arm i has a parameter θ_i sampled

independently from the Bayesian prior distribution on the arm's reward. This prior distribution is $U[0,1]$. Then, conditioning on θ_i, each arm i's rewards are generated when pulled as conditionally independent Bernoulli random variables with the parameter θ_i. Our objective is to maximize the expected total reward.

This problem is similar to the more widely-studied stochastic bandit, except that the arm's reward is drawn at random from the prior. The expected reward calculated can be understood as the average-case reward over stochastic bandit problem instances, i.e., over $(\theta_i : i)$, where the weight on a particular instance $(\theta_i : i)$ is proportional to its density under the prior.

The non-degeneracy condition holds in this problem for both horizons considered, $T = 15$ and $T = 20$. We verified this numerically by solving the LP (2.4) and confirming that there is at least one state with a strictly positive occupation measure in the fluid-neutral category in each period.

Policy Implementation: We describe implementation of UCB, TS and our fluid-priority policy.

UCB tracks the posterior belief on θ_i for each arm i based on the arm's past reward realization, and calculates an upper confidence bound for θ_i as $\mu_i + \delta\sigma_i$, where δ is a fixed parameter, μ_i is the mean of the posterior belief and σ_i is the standard deviation. The top $\lfloor N/3 \rfloor$ arms ranked by their upper confidence bound are pulled. We run UCB with δ varying from 0.1 to 1 and report results for the one with the best expected reward ($\delta^* = 0.5$) in both experiments.

TS also tracks the posterior belief on θ_i for each arm i. At each period, TS samples a value from each arm's posterior belief on θ_i, then pulls the $\lfloor N/3 \rfloor$ arms

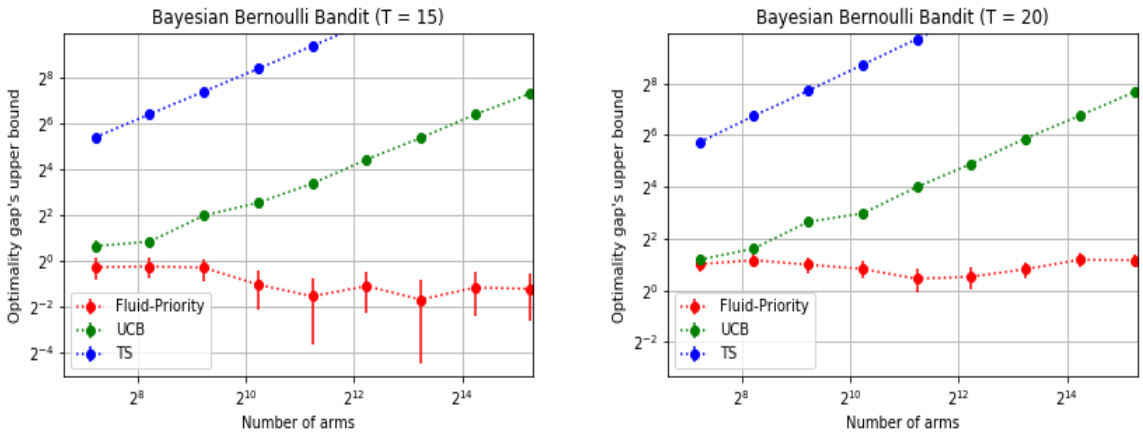
with the highest sampled values.

The fluid-priority policy is constructed as follows. First, we solve Problem (2.4) to fetch an optimal occupation measure. Second, to construct the priority-score function, we use a Lagrangian-relaxation approach similar to [31]: we solve the min-max problem

$$(\lambda_1^*, ..., \lambda_T^*) \leftarrow \min_{\lambda_1,...,\lambda_T} \max_{\pi} \mathbb{E}_\pi \sum_{t=1}^{T} r_t(s_t, a_t) + \lambda_t(\alpha_t - a_t), \tag{2.7}$$

where the inner max can be solved via dynamic programming and the outer min can be solved via the subgradient method. Then we compute the $Q-$function based on the optimal Lagrangian multiplier $(\lambda_1^*, \lambda_2^*, ..., \lambda_T^*)$ iteratively: $Q_t(s,a) = r_t(s,a) - \lambda_t a + \sum_{s'} p_t(s,a,s') \max_{a'} Q_{t+1}(s',a')$ for $0 \le t \le T-1$, with $Q_T(s,a) = r_T(s,a) - \lambda_T a$, and construct the priority-score function as $\mathcal{P}_t(s) = Q_t(s,1) - Q_t(s,0)$. Finally, we plug the optimal occupation measure and the score-function into Algorithm 1 to construct the fluid-priority policy.

Numerical Experiments: We compare the just-described fluid-priority policy against UCB and TS using two different time horizons T of 15 and 20. Figure 1a and 1b display results for $T = 15$ and $T = 20$ respectively. In both experiments, we iteratively double the number of arms (from $N = 300$ to 38400) and plot an upper bound on the opt gap. This bound on the opt gap is computed by first computing the value of the relaxed problem $\hat{V}_N^*$ (which is an upper bound on the value of the optimal policy) and then subtracting the value of the UCB, TS or fluid-priority policy estimated via simulation. We compare this upper bound across policies instead of the exact opt gap because computing the exact opt gap would require knowing the value of the optimal policy, which would take time exponential in N, as discussed in §2.3. We use $50N$ replications to estimate a policy's value when there are N arms. We use more samples when there are more



(a) $T = 15$ (b) $T = 20$

Figure 2.1: Bayesian bandit with Bernoulli rewards. An upper bound on the opt gap (relaxed problem's expected total reward minus a simulation-based estimate of reward) vs number of arms N, for the finite-horizon Bayesian multi-armed bandit with horizons $T = 15$ (left) and $T = 20$ (right). The fluid-priority policy has its opt gap bounded above by a constant while UCB and Thompson sampling have opt gaps that grow linearly with the number of arms.

arms because having more arms increases the variance of a policy's reward. We also compute a confidence interval on this upper bound, computed as the difference between $\hat{V}_N^*$ and the upper and lower limits of a confidence interval on the policy's expected reward.

Figure 2.1 compares the performance of fluid-priority, UCB and TS policies. For both time horizons T of 15 and 20, the fluid-priority policy performs significantly better than UCB and TS, especially for large N. The fluid-priority policy's reward differs from the optimal policy's reward by at most 1 for $T = 15$ and at most 2 for $T = 20$ even when there are 38400 arms available. UCB outperforms TS in both time horizons, perhaps due to the tuning of UCB's hyperparameter. These results are consistent with Theorem 2.4 and our numerical validation that the non-degeneracy condition is satisfied, which implies that the fluid-priority policy's opt gap is $O(1)$. In contrast, Figure 2.1 suggests that the opt gap for

UCB and TS are $\Omega(N)$.

Proposition 2.2 provides a theoretical analysis to confirm that UCB and TS have $\Omega(N)$ opt gaps. Its proof defines an iterative algorithm over t to calculate the occupation measure for UCB and TS in the large N limit. We then use this algorithm to compute occupation measures for specific values for T and compare it to the optimal occupation measure. We find that the occupation measures are suboptimal for the values of T used in these experiments, implying that UCB and TS are not fluid-consistent and their opt gaps are $\Omega(N)$. These values of T are representative, and UCB and TS have $\Omega(N)$ opt gap for other T as well.

Proposition 2.2. *The opt gap for both UCB and TS is* $\Omega(N)$ *for* $T = 15$ *and* $T = 20$.

2.8.2 Crowdsourced labeling

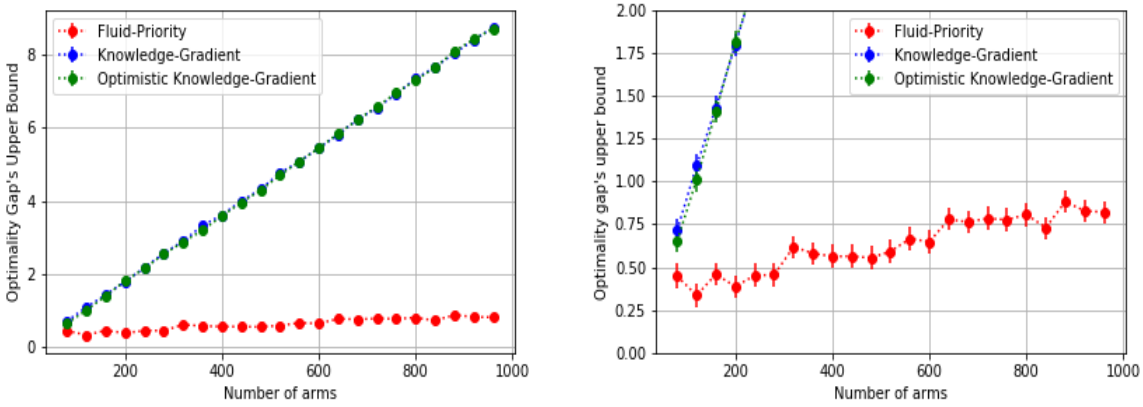
This section evaluates a fluid-priority policy's performance in an active learning problem [17] on the allocation of crowd workers for accurate image classification. We construct this policy by searching over all fluid-priority policies to find the one with the best empirical performance. There are roughly 30 fluid-priority policies in this problem, resulting from different priority score functions and a unique optimal occupation measure. We compare the performance of the fluid-priority policy selected against the Optimistic Knowledge-Gradient [17], a method specifically designed for this problem, and the Online Knowledge-Gradient [49]. The fluid-priority policy outperforms both methods significantly.

We formulate the crowdsourced labeling problem following [17]. There are N images needing binary labels (e.g., whether this is a picture of a pedestrian or not) to support training an automatic image classifier to be built later. We ask

crowd workers to label these images. Each image i has a true underlying binary class, along with an associated probability p_i that a crowd worker will label the image with the correct class. A crowd worker may provide an incorrect label because, e.g., the image is blurry or the worker is distracted. We assume $p_i > 1/2$, i.e. the majority of crowd workers give the correct label. We use an independent prior belief for each image's $p_i \sim U[1/2, 1]$. We may request $T = 7$ batches of labels from crowd workers, with up to $\lfloor N/4 \rfloor$ images per batch. After the last batch, we estimate each image's class via majority vote, which is also the class with maximum probability under posterior.

Figure 2.2 compares the fluid-priority policy against the Online Knowledge-Gradient and Optimistic Knowledge-Gradient methods as we vary the number of arms N, reporting an upper bound on the opt gap for each policy computed in the same way as §2.8.1. The fluid-priority policy performs extremely well and incorrectly classifies at most 1 more image on average than the optimal policy even when there are 1000 images' labels to be learned. Online Knowledge-Gradient and Optimistic Knowledge-Gradient perform similarly in our experiment and they both underperform the fluid-priority policy by misclassifying at least 8 more images on average with 1000 images' labels to be learned. Even though the Optimistic Knowledge-Gradient was designed specifically for this problem, the fluid-priority policy has a significantly better performance.

Figure 2.2 is consistent with our theoretical results. We can verify the non-degeneracy condition does not hold in this example by implementing the algorithm in Appendix A.14. The lack of non-degeneracy implies that the opt gap of the fluid-priority policy is $O(\sqrt{N})$. Its performance in the plot is consistent with this scaling. The Online Knowledge Gradient and the Optimistic Knowledge



(a) (b)

Figure 2.2: Crowdsourced labeling. An upper bound on the opt gap (relaxed problem's expected total reward minus a simulation-based estimate of reward) vs. number of arms N. The left and right panel show the same data but use different scales for the y-axis. Both Knowledge Gradient and Optimistic Knowledge Gradient have opt gaps that seem to grow linearly. The fluid-priority policy has an opt gap that is $O(\sqrt{N})$ because the non-degeneracy condition does not hold in this problem.

Gradient, however, seem to have suboptimality that scales linearly with N.

2.8.3 Dynamic assortment optimization

This section discusses a dynamic assortment optimization problem proposed in §6.2.1 in [13]. [13] observes empirically that the opt gap of their proposed policy, shown there to be $O(\sqrt{N})$, seems to stay constant with N, suggesting that the $O(\sqrt{N})$ bound is loose. We first describe the problem setting and then confirm that our theoretical results provide the tighter bound suggested by these empirical results.

A retailer repeatedly chooses products to display in a selling season. The retailer has N products but a shelf-space constraint allows only showing $\lfloor N/4 \rfloor$

of them in each time period. Each product, if sold, generates profit of \$1. The demand rate for each product i is unknown to the retailer but follows a Poisson process with intensity γ_i. The retailer holds a Bayesian prior belief on γ_i, which is Gamma-distributed with shape parameter m_i and inverse scale parameter k_i, $\gamma_i \sim \text{Gamma}(m_i, k_i)$. All products share the same prior belief $(m_i, k_i) = (1, 0.1)$. The retailer updates these prior beliefs after observing demand realizations for displayed products using Bayes rule.

The Gamma distribution is a conjugate prior distribution when observations are Poisson-distributed, which causes the posterior to remain Gamma-distributed. More specifically, the posterior on γ_i in time period t is $\text{Gamma}(m_{t,i}, k_{t,i})$ where $m_{t,i}$ and $k_{t,i}$ can be computed recursively. For a product i that was displayed in time period t, letting $d_{t,i}$ be the demand for the product in the period, $m_{t+1,i} = m_{t,i} + d_{t,i}$ and $k_{t+1,i} = k_{t,i} + 1$. For a product i that was not displayed in t, $m_{t+1,i} = m_{t,i}$ and $k_{t+1,i} = k_{t,i}$. At $t = 0$, $m_{0,i} = 1$ and $k_{0,i} = 0.1$.

The retailer's objective is to adaptively choose which products to display in each period subject to the shelf-space constraint to maximize the expected total profit over a T-period selling season. This is formulated as a restless bandit with time horizon T where each product i is an arm whose state at time t is $(m_{t,i}, k_{t,i})$. A good policy must balance exploration and exploitation by showing products that observed sales and the prior suggest have large γ_i (exploitation) and also showing those for which we have little observed sales data to support learning γ_i (exploration).

[13] study performance of their proposed Lagrangian policy when $T = 8$. They find their policy "perform(s) very well for large N", and produces profit "within \$6 of the optimal value!" when $N = 16,384$. They do not, however, offer

an explanation for why the performance would be so good for a policy with their shown $O(\sqrt{N})$ opt gap.

Our results explain this phenomenon. By solving the LP relaxation (2.4) for this problem, we confirm that the set of fluid-neutral states is non-empty in each period, thus confirming that the problem is non-degenerate. Moreover, there is exactly one state in each period's fluid-neutral category. This is also observed by [13], as they mention that "there are no scenarios where products in different states have the same priority indices". Thus, for this optimal occupation measure, all fluid-priority policies are index policies and the Lagrangian policy is a specific example. This explains why the Lagrangian policy achieves an $O(1)$ opt gap.

CHAPTER 3

MULTI-ACTION MULTI-RESOURCE FINITE-HORIZON RESTLESS BANDIT WITH MARKOVIAN ENVIRONMENTAL VARIABLES

This Chapter studies the multi-action finite-horizon multiple-resource restless bandit problem with a Markovian environmental variable. This problem generalizes the restless bandit problem considered in Chapter 2. Generalizing our analysis in that previous chapter, we propose a generalized version of fluid-priority policies appropriate for these more general problems. We then show that these generalized fluid-priority policies always achieve a $O(\sqrt{N})$ opt gap, and achieve $O(1)$ opt gap if a generalized non-degeneracy condition holds.

The problem we study here generalizes the restless bandit in several directions:

- It allows for multiple actions associated with an arm per period, rather than the simple "activate" or "idle" binary action assumed by a restless bandit
- It allows for multiple resources. In the restless bandit, there was implicitly a single resource: the number of arms that we could activate per period. In the problem we consider here, an action applied to a state in a time period consumes some known amount of a collection of resources. We have a constraint in each period associated with each resource that the amount consumed does not exceed this constraint.
- It allows for a global Markovian environmental variable, whose transitions are independent of the arms' states but which influences arms' rewards.

We first provide a literature review, covering topics specific to the more general problem we study here. We then describe the model setup and sequentially show how to achieve $o(N)$, $O(\sqrt{N})$ and $O(1)$ opt gap. Some of the results and techniques are similar to those from Chapter 2 and these are presented briefly with less discussion than in Chapter 2. Others, such as the definition of our fluid policy to handle multiple actions per arm, are significant generalizations.

3.1 Literature Review and Contributions

The model we studied here can be viewed as the intersection of weakly coupled MDP and MDP with environmental variables. In particular, the weakly coupled MDP can be viewed as a generalization of the classic restless bandit first proposed by Whittle in [55]. Since the classic restless bandit literature is well-introduced in the previous two chapters, we focus on weakly coupled MDP and MDP with environmental variables in this section.

Weakly coupled MDP Since its introduction, weakly coupled MDP has draw strong theoretical interests. As a natural generalization of restless bandit, instead of assuming there is a single-type of resource and each arm only permits two actions, weakly coupled MDP allows multiple resource types and multiple actions for each arm. Under the same regime with restless bandit where total resources grows proportionally with number of arms in each period, decision maker seeks policies to maximize the total reward generated from all arms. Starting from [29], [1, 10] generalized Lagrangian technique developed in [55] for the weakly coupled MDP and proposed heuristic strategies mimicking Whit-

tle index for weakly coupled MDP.

Weakly coupled MDP is also of strong practical interest due to its wide applicability. For example, [9] models text marketing as weakly coupled MDP where each potential customer is treated as an arm, and each type of message is treated as an action. [8] models rover control problems as weakly coupled MDP, where each Mars rover is treated as an arm, and its state incorporating information of its current site, amount of time remaining, its current condition etc. Each action corresponds to a scientific mission, such as taking a photo or doing some experiments, and the decision maker needs to optimally assign missions to rovers optimally so that we can maximize our scientific discovery. Other applications includes network revenue management [51], dynamic assortment [15], multi-location inventory management [41], etc.

MDP with environmental variables Different from weakly coupled MDP, MDP with environmental variables is a relatively new topic in literature. [12] first formulated the problem and generalized standard value iteration and policy iteration technique for this setting. Later on, [14] studied the same setting with ours, the weakly coupled MDP with environmental variables, using technique developed in [12]. In this work, [14] proposed an index policy with $O(\sqrt{N})$ opt gap.

Although [14] has already proposed policy with sublinear opt gap for weak coupled MDP with environmental variables, current understanding is still limited in some important aspects. One of the most important questions is, whether it is possible to achieve $O(1)$ opt gap in this setting, and if so, under which conditions. The answer for restless bandit is shown in Chapter 2, where non-

degeneracy guarantee $O(1)$ opt gap.

Summary of Contributions This chapter considers the same setting with [13]. By generalizing fluid-priority policies in Chapter 2, we propose a class of policies which are guaranteed to always achieve $O(\sqrt{N})$ opt gap, and achieve $O(1)$ opt gap when some non-degeneracy conditions hold true.

3.2 System Model

This section formulates our decision-making problem in a Markov Decision Process (MDP) form.

Model We have N arms, each of which share the same finite state space S and action space A. We use $s_{i,t} \in S$ to indicate the state of i-th arm at time t, and $a_{i,t} \in A$ to indicate the activation action for i-th arm at time t. Based on the state and activation action, the arm's state transitions to time $t+1$ according to a known transition kernel

$$p_t(s, a, s') = \mathbb{P}(s_{t+1,i} = s' | s_{t,i} = s, a_{t,i} = a).$$

All arms share the same transition kernel, and any arm's transition is conditionally independent from others given its own state and action.

At each period t, a signal ω_t is generated, which affects the total resource available, reward for each arm and the resources consumption for each arm. We assume signal $\{\omega_t\}_{t\in[T]}$ is a finite-state Markov process with known initial state $\omega_1 = \omega^*$ and transition kernel $K_t(\omega, \omega') = \mathbb{P}[\omega_{t+1} = \omega' | \omega_t = \omega]$. Assume the state space is W with $|W|$ number of elements within it.

At each period t, the total budget available per arm is $b_t(\omega_t)$, where $b_t : \Omega \to \mathbb{R}^p$ is a known mapping and p indicates the number of different resources' type. The resources consumption for an arm in state s with action a taken receiving ω signal is $c_t(s, a, \omega)$, where $c_t : S \times A \times \Omega \to \mathbb{R}^p$. Similarly, we can define the generated reward as $r_t(s, a, \omega)$, where $r_t : S \times A \times \Omega \to \mathbb{R}$. Furthermore, we assume there is a null action a^*, s.t. $c_t(s, a^*, \omega) = 0$ for all $(s, \omega, t) \in S \times \Omega \times [T]$. Action a^* is essential and different from all the other actions since its existence guarantees the feasibility of our budget constraint. Sometimes we also refer to taking action a^* on an arm as "idling the arm", and taking any other non-null action as "activating the arm".

To define the objective function mathematically, we introduce some additional notation. We use $\mathbb{S} = S^N$ to denote the N-product of state space S, and we use $\mathbb{A} = A^N$ to denote the N-product of action space A. All N arms together form a new MDP with state space $\mathbb{S}$ and action space $\mathbb{A}$. At time t, with i-th arm's state $s_{t,i}$ and action $a_{t,i}$, we use $\mathbf{s}_t = (s_{t,1}, s_{t,2}, ..., s_{t,N})$ in $\mathbb{S}$ and $\mathbf{a}_t = (a_{t,1}, a_{t,2}, ..., a_{t,N})$ in $\mathbb{A}$ to indicate the state and action of the joint MDP.

The reward function of the joint MDP, $R_t : \mathbb{S} \times \mathbb{A} \times \Omega \to \mathbb{R}$, is defined as the summation across arms of the single-arm reward. At time t with state $\mathbf{s}_t = (s_{t,1}, s_{t,2}, ..., s_{t,N})$, action $\mathbf{a}_t = (a_{t,1}, a_{t,2}, ..., a_{t,N})$ and signal realization ω_t, we use $R_t(\mathbf{s}_t, \mathbf{a}_t, \omega_t) = \sum_{i=1}^N r_t(s_{t,i}, a_{t,i}, \omega_t)$ to denote this total reward. The transition kernel of the joint MDP is the product of the arms' individual transition kernels,

$$\mathbb{P}[\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t] = \prod_{i=1}^N p_t(s_{t,i}, a_{t,i}, s_{t+1,i}),$$

which models the transitions as happening independently across arms given the action taken.

A policy π is a function that maps the state $(\mathbf{s}_t, \omega_t)$ of the joint MDP to a

feasible action $\mathbf{a}_t \in \mathbb{A}$. The objective of the policy is to maximize the expected total reward with respect to a total budget $B_t = b_t(\omega_t)N$ per time period. This objective can be written as,

$$\begin{aligned} &\sup_{\pi} \; V_N(\pi) = \mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t, \omega_t) \\ &\text{subject to } \sum_{i=1}^{N} c_t(s_{i,t}, a_{i,t}, \omega_t) \le b_t(\omega_t)N, \; \forall t \in [T]. \end{aligned} \tag{3.1}$$

Without loss of generality, we assume each arm starts at the same state $s^* \in S$ at $t = 1$.

Performance Measurement We measure the performance of the policy π by comparing it with optimal policies to maximize (3.1). Let

$$V_N^* = \sup_{\pi} V_N(\pi)$$

be the expected total reward obtained by an optimal policy. Then the optimality gap of the policy π is defined as

$$V_N^* - V_N(\pi).$$

The smaller the optimality gap, the better the policy.

We are interested in the asymptotic regime where the total budget B_t and total number of arms N grow to infinity proportionally.

3.3 Background: Preliminary Results and Notations

In this section, we use a linear programming relaxation to provide an easy-to-compute upper bound $\hat{V}_N^*$ for V_N^*. Some of results shown in the section is similar

to Section 2.3. We describe these results briefly and leave their proofs to the Appendix.

Linear Programming Relaxation We relax Problem (3.1)'s almost sure cardinality constraints on resource use to constraints on their expectation:

$$\max_{\pi} \ \mathbb{E}_{\pi} \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t, \omega_t)$$
$$\text{subject to } \mathbb{E}_{\pi,\omega_1,\omega_2,...,\omega_t} \sum_{i=1}^{N} c_t(s_{i,t}, a_{i,t}, \omega_t) \le b_t(\omega_t)N, \ \forall t \in [T], \omega_1 \in \Omega, \omega_2 \in \Omega, ..., \omega_t \in \Omega.$$

Here, the expectation in the constraint is taken over the joint distribution of all arms' state conditioned on the policy π and all previously realized signals $\omega_1, \omega_2, ..., \omega_t$.

To write the above relaxed problem in a more compact form, we introduce some notation first. We denote all signals before period t as a vector $\boldsymbol{\omega}_t \in \Omega^t$:

$$\boldsymbol{\omega}_t = (\omega_1, ..., \omega_t).$$

We denote the s-th component of $\boldsymbol{\omega}_t$ as $\boldsymbol{\omega}_{t,s}$. Sometimes it is useful to truncate $\boldsymbol{\omega}_t$ to the history of signals before $s \le t$. To do this, we define truncation operators $P_s : \Omega^t \to \Omega^s$ by $P_s(\boldsymbol{\omega}_t) := \boldsymbol{\omega}_s = (\omega_1, ..., \omega_s)$.

With this new notation, we can rewrite the above relaxed problem as

$$\begin{aligned} &\max_{\pi} \ \mathbb{E}_{\pi} \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t, \omega_t) \\ &\text{subject to } \mathbb{E}_{\pi,\boldsymbol{\omega}_t} \sum_{i=1}^{N} c_t(s_{i,t}, a_{i,t}, \omega_t) \le b_t(\omega_t)N, \ \forall \boldsymbol{\omega}_t. \end{aligned} \tag{3.2}$$

Similar to Lemma 2.1, we can show

Lemma 3.1. $V_N^* \le \hat{V}_N^* = N\hat{V}_1^*$.

To calculate $\hat{V}_1^*$, we only need to solve

$$\begin{aligned} &\max_{\pi} \ \mathbb{E}_{\pi} \sum_{t=1}^{T} r_t(s_t, a_t, \omega_t) \\ &\text{subject to } \mathbb{E}_{\pi,\omega_t}[c_t(s_t, a_t, \omega_t)] \le b_t(\omega_t), \ \forall t \in [T], \boldsymbol{\omega}_t \in \Omega^t. \end{aligned} \tag{3.3}$$

In terms of the occupation measure, $\mu_t(s, a, \boldsymbol{\omega}_t) = \mathbb{P}[s_t = s, a_t = a|\boldsymbol{\omega}_t]\ \mathbb{P}[\boldsymbol{\omega}_t]$, Problem (3.3) is equivalent to

$$\max \sum_{t=1}^{T} \sum_{\boldsymbol{\omega}_t \in \Omega^t} \sum_{a \in A} \sum_{s \in S} r_t(s, a, \omega_t) \mu_t(s, a, \boldsymbol{\omega}_t)$$

subject to

$$\begin{aligned} &\sum_{a \in A} \mu_t(s, a, \boldsymbol{\omega}_{t+1}) = \sum_{a \in A} \sum_{s' \in S} \mu_t(s', a, P_t(\boldsymbol{\omega}_{t+1})) p_t(s', a, s) K_t(\omega_t, \omega_{t+1}), \\ &\qquad\qquad \forall s \in S, \boldsymbol{\omega}_{t+1} \in \Omega^{t+1}, t \le T-1. \\ &\sum_{s \in S, a \in A} \mu_t(s, a, \boldsymbol{\omega}_t) c_t(s, a, \omega_t) \le b_t(\omega_t) \sum_{s \in S, a \in A} \mu_t(s, a, \boldsymbol{\omega}_t), \ \forall \boldsymbol{\omega}_t \in \Omega^t, t \in [T]. \\ &\sum_{a \in A} \mu_1(s^*, a, \omega^*) = 1. \\ &\mu_t(s, a, \boldsymbol{\omega}_t) \ge 0, \ \forall s \in S, a \in A, \boldsymbol{\omega}_t \in \Omega^t, t \in [T]. \end{aligned} \tag{3.4}$$

Notation We introduce some notation for simplicity of analysis. First, we let

$$z_t(s, \boldsymbol{\omega}_t) := \frac{\sum_{a \in A} \mu_t(s, a, \boldsymbol{\omega}_t)}{\sum_{a \in A} \sum_{s \in S} \mu_t(s, a, \boldsymbol{\omega}_t)}$$

denote the conditional probability distribution of an arm in state s at time t given signal realization history $\boldsymbol{\omega}_t$. Without loss of generality, we assume $\mathbb{P}[\boldsymbol{\omega}_t] = \sum_{a \in A} \sum_{s \in S} \mu_t(s, a, \boldsymbol{\omega}_t) > 0$ for all $\boldsymbol{\omega}_t$ to provide a generic analysis of all periods. If not, we can remove the $\boldsymbol{\omega}_t$ with $\mathbb{P}[\boldsymbol{\omega}_t] = 0$ and treat these $\boldsymbol{\omega}_t$ separately.

Second, we let

$$x_t(s, a, \boldsymbol{\omega}_t) := \frac{\mu_t(s, a, \boldsymbol{\omega}_t)}{\sum_{a \in A} \sum_{s \in S} \mu_t(s, a, \boldsymbol{\omega}_t)}$$

denote the conditional probability distribution of an arm in state s with activation action a at time t given signal realization history $\boldsymbol{\omega}_t$.

Sometimes, we will use $z_t(\boldsymbol{\omega}_t)$ (or $x_t(\boldsymbol{\omega}_t)$) to refer to the corresponding vector (or matrix), i.e., $z_t(\boldsymbol{\omega}_t) := (z_t(s, \boldsymbol{\omega}_t), s \in S)$ (or $x_t(\boldsymbol{\omega}_t) := (x_t(s, a, \boldsymbol{\omega}_t) : s \in S, a \in A)$).

To be consistent with $z_t(s, \boldsymbol{\omega}_t)$ and $x_t(s, a, \boldsymbol{\omega}_t)$, we let $Z_t^N(s, \boldsymbol{\omega}_t)$ be a random variable whose distribution is the same as the conditional distribution of the number of arms in state s at time t given the signal realization history $\boldsymbol{\omega}_t$. Similarly, $X_t^N(s, a, \boldsymbol{\omega}_t)$ is a random variable distributed according to the number of arms in state s with action a at time t given signal realization history $\boldsymbol{\omega}_t$. As in the definition of $Z_t^N(s, \boldsymbol{\omega}_t)$, we sometimes need to consider random variables restricted to a realization of the signal history. Thus, we use $1(\boldsymbol{\omega}_t)$ (or $\mathbb{P}[\boldsymbol{\omega}_t]$) to denote the indicator (or probability) of the event that the realization signal history is $\boldsymbol{\omega}_t$.

Similarly to $z_t(\boldsymbol{\omega}_t)$ and $x_t(\boldsymbol{\omega}_t)$, we use $Z_t^N(\boldsymbol{\omega}_t)$ and $X_t^N(\boldsymbol{\omega}_t)$ to refer to vectors $(Z_t^N(s, \boldsymbol{\omega}_t) : s \in S)$ and matrix $(X_t^N(s, a, \boldsymbol{\omega}_t) : s \in S, a \in A)$.

Starting from Section 3.5, we will analyze the deviation between the realization of $Z_t^N(\boldsymbol{\omega}_t), X_t^N(\boldsymbol{\omega}_t)$ and $Nz_t(\boldsymbol{\omega}_t), Nx_t(\boldsymbol{\omega}_t)$. To support this analysis, we define diffusion statistics $\tilde{Z}_t^N(\boldsymbol{\omega}_t)$ as $\tilde{X}_t^N(\boldsymbol{\omega}_t)$ as

$$\tilde{Z}_t^N(\boldsymbol{\omega}_t) = \frac{Z_t^N(\boldsymbol{\omega}_t) - Nz_t(\boldsymbol{\omega}_t)}{\sqrt{N}}, \ \tilde{X}_t^N(\boldsymbol{\omega}_t) = \frac{X_t^N(\boldsymbol{\omega}_t) - Nx_t(\boldsymbol{\omega}_t)}{\sqrt{N}}.$$

Recall a policy π of the joint MDP is a map from $(\boldsymbol{s}_t, \omega_t)$ to $\boldsymbol{a}_t$. For analysis in the following sections, we augment the policy class a little bit. We define a policy π of the joint MDP is a map from $(\boldsymbol{\omega}_t, Z_t^N(\boldsymbol{\omega}_t))$ to $X_t^N(\boldsymbol{\omega}_t)$. Now the policy not only depends on the signal at current period, but also the signal history before. The reason we use this augmentation class is that, the history of signals determines the optimal occupation measure in the future.

Using this new notation, a policy π of the joint MDP naturally induces a class

of maps $\tilde{\pi}_{t,N}$ indexed by t, N, from diffusion $(\omega_t, \tilde{Z}_t^N(\omega_t))$ to diffusion $\tilde{X}_t^N(\omega_t)$, s.t.

$$\pi(\omega_t, Z_t^N(\omega_t)) = X_t^N(\omega_t) \Longleftrightarrow \tilde{\pi}_{t,N}(\omega_t, \tilde{Z}_t^N(\omega_t)) = \tilde{X}_t^N(\omega_t). \tag{3.5}$$

3.4 Sufficient Conditions for Achieving an $o(N)$ Opt Gap

This section establishes our first contribution: a substantially more general result showing sufficient conditions for $o(N)$ optimality gap. Results in this section are similar with Section 2.4, so we only briefly state all the results without detailed explanation. All the proofs in this section can be found in the Appendix.

Similar with Definition 2.1, we define

Definition 3.1. *Under a policy π, if $\frac{Z_t^N(\omega_t)}{N} \to z_t(\omega_t)$ implies*

$$\frac{X_t^N(\omega_t)}{N} \to x_t(\omega_t)$$

almost surely, we say the policy π is fluid consistent.

(Here, when we say that an *event A implies an event B almost surely*, we mean that the union of event B and the complement of event A holds almost surely.)

Similar to Theorem 2.1, we have Theorem 3.1 below.

Theorem 3.1. *If a policy π is fluid consistent, then $V_N^* - V_N(\pi) = o(N)$.*

3.5 Sufficient Conditions for Achieving an $O(\sqrt{N})$ Opt Gap

This section establishes our second contribution: substantially more general sufficient conditions than [14] for an $O(\sqrt{N})$ optimality gap. Results in this section are similar to Section 2.5, so we only briefly state the results without detailed explanation. All the proofs for results in this section can be found in the Appendix.

Generalizing Definition 2.2, we formulate Definition 3.2. Then, similar to Theorem 2.2, we have Theorem 3.2 below.

Definition 3.2. *A policy π is called diffusion regular if its induced maps $\tilde{\pi}_{t,N}$ satisfy the following conditions*

- *There exists $C_1 > 0$ s.t. for any t, N, θ_1, θ_2, and any fixed $\boldsymbol{\omega}_t \in \Omega^t$,*

$$|\tilde{\pi}_{t,N}(\boldsymbol{\omega}_t, \theta_1) - \tilde{\pi}_{t,N}(\boldsymbol{\omega}_t, \theta_2)| \leq C_1 |\theta_1 - \theta_2|.$$

- *There exists $C_2 > 0$, s.t. for any t, N, and any fixed $\boldsymbol{\omega}_t \in \Omega^t$,*

$$|\tilde{\pi}_{t,N}(\boldsymbol{\omega}_t, 0)| \leq C_2.$$

- *There exists a map $\tilde{\pi}_{t,\infty}$, s.t. $\tilde{\pi}_{t,N}(\boldsymbol{\omega}_t, \theta) \to \tilde{\pi}_{t,\infty}(\boldsymbol{\omega}_t, \theta)$ for any fixed $\boldsymbol{\omega}_t \in \Omega^t$ as $N \to +\infty$.*

where $|\cdot|$ stands for the L^1-norm in Euclidean space.

Theorem 3.2. *If a policy π is diffusion regular, then $V_N^* - V_N(\pi) = O(\sqrt{N})$.*

3.6 Fluid-priority policies

This section defines fluid-priority policies and show that they are always diffusion regular and thus achieve an $O(\sqrt{N})$ opt gap. Later, in §7, we show that they achieve an $O(1)$ opt gap if an additional condition is satisfied. The fluid-priority policy defined here can be viewed as a generalization in Section 2.6, with a Weight-Round-Robin subroutine to determine which action to activate for arms in the same state.

Roughly speaking, a fluid-priority policy is defined by first fetching an optimal solution of the LP relaxation, then classifying states into three disjoint categories based on the solution: fluid-active, fluid-neutral and fluid-inactive. Then a priority score function is defined to be consistent with this category classification: fluid-active states are prioritized above fluid-neutral states and fluid-neutral states are prioritized above fluid-inactive states. When deciding which action to take within each state, fluid-priority policy pulls arm in a weighted Round-Robin (RR) way respecting the optimal occupation measure.

Mathematically speaking, a fluid-priority policy is parameterized by an optimal occupation measure $\{x_t(s,a,\boldsymbol{\omega}_t)\}_{\omega_t,s,a}$ solving Problem (3.4) and a set of "priority-score" functions $\{\mathcal{P}_{\omega_t}(\cdot)\}_{\omega_t}$ assigning each state a real number. Based on the occupation measure $\{x_t(s,a,\boldsymbol{\omega}_t)\}_{t,s,a}$, a fluid-priority policy classifies states into these three disjoint categories:

The *fluid-active* category: $C^+(\boldsymbol{\omega}_t) := \{s \in S \,|\, \sum_{a\neq a^*} x_t(s,a,\boldsymbol{\omega}_t) > 0, x_t(s,a^*,\boldsymbol{\omega}_t) = 0\}$,

The *fluid-neutral* category: $C^0(\boldsymbol{\omega}_t) := \{s \in S \,|\, \sum_{a\neq a^*} x_t(s,a,\boldsymbol{\omega}_t) > 0, x_t(s,a^*,\boldsymbol{\omega}_t) > 0\}$,

The *fluid-inactive* category: $C^-(\boldsymbol{\omega}_t) := \{s \in S \,|\, \sum_{a\neq a^*} x_t(s,a,\boldsymbol{\omega}_t) = 0, x_t(s,a^*,\boldsymbol{\omega}_t) > 0\}$.

Then the priority-score function $\mathcal{P}_{\omega_t}(\cdot)$ are chosen to prioritize states in $C^+(\boldsymbol{\omega}_t)$ over $C^0(\boldsymbol{\omega}_t)$, and states in $C^+(\boldsymbol{\omega}_t)$ over $C^-(\boldsymbol{\omega}_t)$. With these definitions, the fluid-priority policy corresponding to an occupation measure and priority-score function is defined by Algorithm 3.

Algorithm 3 Fluid-priority policy

Input: an optimal occupation measure $(x_t(s,a,\boldsymbol{\omega}_t))_{\omega_t,s,a}$ solving the LP (3.4), and a priority-score functions $\{\mathcal{P}_{\omega_t}(\cdot)\}_{\omega_t}$ consistent with the category classification.

for $t = 1, 2, ..., T$ **do**
  Observe signal realization ω_t and form the full signal realization path $\boldsymbol{\omega}_t$, and observe there are $Z_t(s)$ arms in state s and remaining budget $B_t = \lfloor b_t(\omega_t)N \rfloor$.
  for state s in decreasing order given by $\mathcal{P}_{\omega_t}(\cdot)$ **do**
    Determine $X_t(s,a) \leftarrow$ Weighted-Round-Robin$(s, (x_t(s,a,\boldsymbol{\omega}_t))_{a\in A}, B_t)$
    Update $B_t \leftarrow B_t - \sum_a X_t(s,a)c_t(s,a,\omega_t)$
  end for
  Activate $X_t(s,a)$ number of arms in state s with action a.
end for

Inside Algorithm 3, the subroutine Weighted-Round-Robin is more of technical details. Roughly speaking, when determining $X_t(s,a)$, i.e. how many arms in state s we should activate with action a, we try to make $(X_t(s,a))_{a\in A}$ as proportional to the occupation measure $(x_t(s,a,\boldsymbol{\omega}_t))_{a\in A}$ as possible while respecting the budget constraint. The formal definition of Weighted-Round-Robin is defined as in Algorithm 4.

Algorithm 4 Weight-Round-Robin

Input: a state of interest s, an occupation measure $(x_t(s,a,\boldsymbol{\omega}_t))_{a\in A}$, remaining budget B_t.

Initialize $X_t(s,a) = 0$ for all $a \in A$. Normalize $x_t(s,a,\boldsymbol{\omega}_t)$ to $x_t(s,a) := x_t(s,a,\boldsymbol{\omega}_t)/\sum_a x_t(s,a,\boldsymbol{\omega}_t)$.
while $\sum_{a\in A} X_t(s,a) < Z_t(s)$ and $\{a|x_t(s,a) > 0, c_t(s,a,\omega_t) \le B_t\} \neq \varnothing$ **do**
  Take $\hat{a}$ minimizes $X_t(s,a) - Nx_t(s,a)$ within $\{a|x_t(s,a) > 0, c_t(s,a,\omega_t) \le B_t\}$.
  $X_t(s,\hat{a}) \leftarrow X_t(s,\hat{a}) + 1,\ B_t \leftarrow B_t - c_t(s,\hat{a},\omega_t)$.
end while
$X_t(s,a^*) \leftarrow Z_t(s) - \sum_{a\neq a^*} X_t(s,a)$.

Output: $(X_t(s,a))_{a\in A}$

With the algorithm in place, we now state the main result of this section: that fluid-priority policies are diffusion regular, implying they have an $O(\sqrt{N})$ opt gap by Theorem 3.2.

Theorem 3.3. *Any fluid-priority policy π is diffusion regular and its optimality gap is $O(\sqrt{N})$.*

3.7 Non-degeneracy Condition: Achieving an O(1) Opt Gap

This section presents our main contribution: that fluid-priority policies achieve an $O(1)$ opt gap under a non-degeneracy condition. We define and discuss this condition before showing this result.

To motivate this non-degeneracy condition, consider a fluid-priority policy and another policy motivated by the relaxed problem (3.2) in which the almost-sure budget constraint ($\sum_{i=1}^{N} c_t(s_{i,t}, a_{i,t}, \omega_t) \leq b_t(\omega_t)N$) has been relaxed. This so-called "budget-relaxed" policy first categorizes states into fluid-active, fluid-neutral, and fluid-inactive categories in the same way as its corresponding fluid-priority policy. For fluid-active arms, it activates all all of them in the same weighted Round-Robin way, no matter whether the budget constraint is violated. If budget constraint is not violated after activating all fluid-active arms, it then activate / idle all fluid-neutral arms in the same way as its corresponding fluid-priority policy. Same with the fluid-priority policy, it does not activate any fluid-inactive arms.

Introduction of this budget-relaxed policy builds a bridge between its corresponding fluid-priority policy and the optimal reward of the relaxed problem (3.2). First, activating all fluid-active arms and idling all fluid-inactive arms

plus some technical resource consumption property implies a policy to be optimal in the relaxation problem (3.2). Thus, this budget-relaxed policy's reward is close to the relaxed problem's optimal reward (Lemma 3.4). Second, it behaves identically to its corresponding fluid-priority policy (Lemma 3.2) except on a specific "budget violation" event: that activating all arms in fluid-active category exceeds the budget. So combing the first and the second points together, bounding the probability of budget-violation allows us to bound the opt gap for a fluid-priority policy by comparing it with its budget-relaxed version and comparing the budget-relaxed version with the relaxed problem's optimal rewards.

The non-degeneracy condition (Definition 3.3) characterizes the probability of budget violation. When it is met, the expected resource consumption amount per fluid-active is *strictly* below the budget per arm. Thus, using concentration bounds, problems meeting the non-degeneracy condition are ones in which the probability of budget violation vanishes exponentially fast as N grows (Lemma 3.3). As a result, in such problems, the fluid-priority policy behaves the same as its budget-relaxed version with high probability for large N.

The non-degeneracy condition also requires that expected resource consumption would be tight compared to the budget constraint at most for one resource type. This more technical condition is used to bound the reward of budget-relaxed policy and the optimal reward of the relaxed problem by $O(1)$. Thus, when the non-degeneracy condition is met, we can bound the opt gap of the fluid-priority policy by $O(1)$.

In the rest of this section, we first formally introduce budget-relaxed policies, then define the non-degeneracy condition, and finally prove fluid-priority

policies achieve an $O(1)$ opt gap when this condition holds.

3.7.1 Budget-relaxed fluid-priority policies

Given a fluid-priority policy π_F, its budget-relaxed version π_R is defined by Algorithm 5. Similar to π_F, π_R first classifies states into three categories: fluid-active, fluid-neutral and fluid-inactive, using the same occupation measure as π_F. Then, π_R activates all arms in the fluid-active category using the same Weighted-Round-Robin procedure, exceeding the budget if necessary. Afterwards, if budget is still not violated, π_R iterates over each state s in the fluid-neutral category C_t^0 and activate / idle arms in a weighted Round-Robin manner. Finally, π_R idles all arms in fluid-inactive category.

Algorithm 5 Budget-relaxed fluid-priority policies

Input: an optimal occupation measure $(x_t(s,a,\boldsymbol{\omega}_t))_{\omega_t,s,a}$ solving the LP (3.4), and a priority-score functions $\{\mathcal{P}_{\omega_t}(\cdot)\}_{\omega_t}$.

for $t = 1, 2, ..., T$ **do**
  Observe signal realization ω_t and form the full signal realization path $\boldsymbol{\omega}_t$, and observe there are $Z_t(s)$ arms in state s.
  Classify states into fluid-active $C_t^+(\boldsymbol{\omega}_t)$, fluid-neutral $C_t^0(\boldsymbol{\omega}_t)$ and fluid-inactive $C_t^-(\boldsymbol{\omega}_t)$ categories based on the occupation measure.
  for state $s \in C^+(\boldsymbol{\omega}_t)$ in decreasing order given by $\mathcal{P}_{\omega_t}(\cdot)$ **do**
    Determine $X_t(s,a) \leftarrow$ Weighted-Round-Robin$(s, (x_t(s,a,\boldsymbol{\omega}_t))_{a\in A}, \infty)$ by pretending unlimited budget.
  end for
  Update remaining budget $B_t \leftarrow \lfloor b_t N \rfloor - \sum_a \sum_{s\in C^+(\omega_t)} X_t(s,a) c_t(s,a,w_t)$.
  for state $s \in C^0(\boldsymbol{\omega}_t)$ in decreasing order given by $\mathcal{P}_{\omega_t}(\cdot)$ **do**
    Determine $X_t(s,a) \leftarrow$ Weighted-Round-Robin$(s, (x_t(s,a,\boldsymbol{\omega}_t))_{a\in A}, B_t)$.
    Update $B_t \leftarrow B_t - \sum_a X_t(s,a) c_t(s,a,\omega_t)$
  end for
  for state $s \in C^-(\boldsymbol{\omega}_t)$ **do**
    $X_t(s,a^*) = Z_t(s)$ and $X_t(s,a) = 0$ for all $a \neq a^*$.
  end for
  Activate $X_t(s,a)$ number of arms in state s with action a.
end for

Policy π_R behaves the same as its corresponding fluid-priority policy π_F, given sufficient resourced needed for all arms in fluid-active category. This is formally stated as Lemma 3.2.

Lemma 3.2. *Define conditional event given the realization history $\boldsymbol{\omega}_t$:*

$$\Delta(\boldsymbol{\omega}_t) := \Big\{ \sum_{s \in C^+(\boldsymbol{\omega}_t)} \sum_{a \in A} \lceil Z_t^N(s, \boldsymbol{\omega}_t) \frac{x_t(s, a, \boldsymbol{\omega}_t)}{\sum_a x_t(s, a, \boldsymbol{\omega}_t)} \rceil c_t(s, a, \omega_t) \le b_t(\omega_t) N \Big\}.$$

Then on the event $\Delta(\boldsymbol{\omega}_t)$, $\pi_R(\boldsymbol{\omega}_t, Z_t^N) = \pi_F(\boldsymbol{\omega}_t, Z_t^N)$.

$\Delta(\boldsymbol{\omega}_t)$ characterizes the scenario where sufficient resources are available so that we are not forced to idle any arm (take action a^*) in fluid-active category due to the budget constraint. So we would refer to the complement of $\Delta(\boldsymbol{\omega}_t)$ as $\Delta^c(\boldsymbol{\omega}_t)$ as a "budget violation" event.

3.7.2 Non-degeneracy

The non-degeneracy condition (stated formally below) requires that budget per arm strictly dominates the resources needed per arm in fluid-active category to be activated. This guarantees that budget-violation events are probabilistically negligible (Lemma 3.3). The high-level reason is that, since both the budget-relaxed fluid-priority policy and the fluid-priority policy are fluid consistent, the number of arms in state s is roughly proportional to the optimal occupation measure, with excursions described by the Central Limit Theorem. Thus the probability of budget-violation events approaches 0 exponentially fast as N grows by concentration inequalities.

Definition 3.3. *We say an optimal occupation measure $(x_t(s, a, \boldsymbol{\omega}_t))_{\omega_t, s, a}$ is non-*

degenerate if

$$\forall \boldsymbol{\omega}_t, \sum_{s \in C^+(\boldsymbol{\omega}_t)} \sum_a x_t(s,a,\boldsymbol{\omega}_t) c_t(s,a,\omega_t) < b_t(\omega_t)\mathbb{P}[\boldsymbol{\omega}_t] \text{ and}$$

$$\sum_{s \in S} \sum_a x_t(s,a,\boldsymbol{\omega}_t) c_t(s,a,\omega_t) \le b_t(\omega_t)\mathbb{P}[\boldsymbol{\omega}_t] \text{ with at most one equality achieved.}$$

Otherwise, we call it degenerate. We also call a fluid-priority policy non-degenerate (degenerate) when its associated occupation measure is non-degenerate (degenerate).

We notice that non-degeneracy condition also requires that the resource constraint is tight in each period for at most one resource type. This is more of technical details, and is only used to prove Lemma 3.4: a budget-relaxed policy deviates from the relaxed problem's optimal reward by at most $O(1)$.

Lemma 3.3. *If an optimal occupation measure $(x_t(s,a,\boldsymbol{\omega}_t))_{\omega_t,s,a}$ is non-degenerate, then for any priority-score functions $\{\mathcal{P}_{\omega_t}\}_{\omega_t}$ and the corresponding fluid-priority policy π_F and budget-relaxed policy π_R, there exists a constant $\delta > 0$ and a constant L such that for any $\boldsymbol{\omega}_t$ and all N,*

$$\max\{\mathbb{P}_{\pi_R}(\Delta^c(\boldsymbol{\omega}_t)), \mathbb{P}_{\pi_F}(\Delta^c(\boldsymbol{\omega}_t))\} \le L\exp(-\delta N).$$

Empirically, one can check the non-degeneracy condition for a given optimal occupation measure x^* returned by solving the LP relaxation (3.4) with a commercial LP solver.

3.7.3 Main result

We now state and prove this section's main result: a fluid-priority policy achieves an $O(1)$ opt gap when it is non-degenerate. Before that, we need one

last building block: the budget-relaxed policies' reward deviates from the relaxed problem's optimal reward by $O(1)$ under non-degeneracy.

We show this in the following lemma. There are two main ideas in the proof. First, recall that the budget-relaxed fluid-priority policy π_R activates all arms in $C^+(\boldsymbol{\omega}_t)$ regardless of the budget constraint. Thus, its decisions are optimal under a Lagrangian relaxation of Problem 3.2 in which the budget constraint on the expected number of arms pulled is replaced by a well-chosen linear penalty (in this Lagrangian relaxation, activating / idling fluid-neutral arms does not affect optimality as the incremental reward is offset by the linear penalty). Second, in the well-chosen linear penalties, only penalties of the resource type with tight budget constraints are non-zero. Thus, the linear penalty is zero under the event that the resource type with tight budget constraints are exhausted. We can show the probability of this event approaches 1 exponentially fast as number of arms N grows large.

Lemma 3.4. *Let $V_N(\pi_R) = \max_\pi \mathbb{E}_\pi \sum_{t=1}^T R_t(\mathbf{s}_t, \mathbf{a}_t, \omega_t)$ for a budget-relaxed fluid priority policy π_R. If an optimal occupation measure $(x_t(s,a,\boldsymbol{\omega}_t))_{\omega_t,s,a}$ is non-degenerate, then for any priority-score functions $\{\mathcal{P}_{\omega_t}\}_{\omega_t}$, the corresponding budget-relaxed fluid-priority policy π_R satisfies $|\hat{V}_N^* - V_N(\pi_R)| \le m$, where m is a constant not depending on N.*

Now we are ready to state and prove our main result: that a fluid-priority policy achieves an $O(1)$ opt gap when it is non-degenerate. A fluid-priority policy π_F's opt gap can be bounded by first comparing the reward $V_N(\pi_F)$ with the reward of its corresponding budget-relaxed policy π_R. Combining the fact that π_F deviates from π_R with negligible probability (Lemma 3.3) and that π_R's reward deviates by $O(1)$ from $\hat{V}_N^*$ (Lemma 3.4), $V_N(\pi_F)$ is at most $O(1)$ away from $\hat{V}_N^*$.

Theorem 3.4. *If an optimal occupation measure $(x_t(s,a,\boldsymbol{\omega}_t))_{\omega_t,s,a}$ is non-degenerate, then for any priority-score functions $\{\mathcal{P}_{\omega_t}\}_{\omega_t}$, the corresponding fluid-priority policy π_F satisfies $\hat{V}_N^* - V_N(\pi_F) \le m$, where m is a constant not depending on N.*

Proof of Theorem 3.4. Under π_F, the reward is $V_N(\pi_F) = \mathbb{E}_{\pi_F} \sum_{\omega_t,s,a} r_t(s,a,\omega_t) X_t^N(s,a,\boldsymbol{\omega}_t)$. Under π_R, the reward is $V_N(\pi_R) = \mathbb{E}_{\pi_R} \sum_{\omega_t,s,a} r_t(s,a,\omega_t) X_t^N(s,a,\boldsymbol{\omega}_t)$.

Within each event $\boldsymbol{\omega}_T$, denote $\Omega(\boldsymbol{\omega}_T) := \Delta(P_1(\boldsymbol{\omega}_T)) \cap \Delta(P_2(\boldsymbol{\omega}_T)) \cap ... \cap \Delta(P_T(\boldsymbol{\omega}_T))$. On this event, π_R and π_F produce identical decisions by Lemma 3.2. Using this in the second line below, we have:

$$
\begin{aligned}
V_N(\pi_R) - V_N(\pi_F) &= \mathbb{E}_{\pi_R}\Big[\mathbb{E}\Big[\big(1_{\Omega(\boldsymbol{\omega}_T)} + 1_{\Omega^c(\boldsymbol{\omega}_T)}\big) \sum_{t=1}^{T} \sum_{s\in S, a\in A} r_t(s,a,\omega_t) X_t^N(s,a,\boldsymbol{\omega}_t) | \boldsymbol{\omega}_T\Big]\Big] \\
&\quad - \mathbb{E}_{\pi_F}\Big[\mathbb{E}\Big[\big(1_{\Omega(\boldsymbol{\omega}_T)} + 1_{\Omega^c(\boldsymbol{\omega}_T)}\big) \sum_{t=1}^{T} \sum_{s\in S, a\in A} r_t(s,a,\omega_t) X_t^N(s,a,\boldsymbol{\omega}_t) | \boldsymbol{\omega}_T\Big]\Big] \\
&= \mathbb{E}_{\pi_R}\Big[\mathbb{E}\Big[1_{\Omega^c(\boldsymbol{\omega}_T)} \sum_{t=1}^{T} \sum_{s\in S, a\in A} r_t(s,a,\omega_t) X_t^N(s,a,\boldsymbol{\omega}_t) | \boldsymbol{\omega}_T\Big]\Big] \\
&\quad - \mathbb{E}_{\pi_F}\Big[\mathbb{E}\Big[1_{\Omega^c(\boldsymbol{\omega}_T)} \sum_{t=1}^{T} \sum_{s\in S, a\in A} r_t(s,a,\omega_t) X_t^N(s,a,\boldsymbol{\omega}_t) | \boldsymbol{\omega}_T\Big]\Big] \\
&\le \mathbb{E}_{\pi_R}\Big[\mathbb{E}\Big[1_{\Omega^c(\boldsymbol{\omega}_T)} \sum_{t=1}^{T} \sum_{s\in S, a\in A} |r_t(s,a,\omega_t)| X_t^N(s,a,\boldsymbol{\omega}_t) | \boldsymbol{\omega}_T\Big]\Big] \\
&\quad + \mathbb{E}_{\pi_F}\Big[\mathbb{E}\Big[1_{\Omega^c(\boldsymbol{\omega}_T)} \sum_{t=1}^{T} \sum_{s\in S, a\in A} |r_t(s,a,\omega_t)| X_t^N(s,a,\boldsymbol{\omega}_t) | \boldsymbol{\omega}_T\Big]\Big]
\end{aligned}
$$

Inequalities $0 \le X_t^N(s,a,\boldsymbol{\omega}_t) \le N$ then implies

$$
\begin{aligned}
&V_N(\pi_R) - V_N(\pi_F) \\
&\le \mathbb{E}_{\pi_R}\Big[\mathbb{E}\Big[1_{\Omega^c(\boldsymbol{\omega}_T)} \max_{t,s,a,\omega_t} |r_t(s,a,\omega_t)|\, TN | \boldsymbol{\omega}_T\Big]\Big] + \mathbb{E}_{\pi_F}\Big[\mathbb{E}\Big[1_{\Omega^c(\boldsymbol{\omega}_T)} \max_{t,s,a,\omega_t} |r_t(s,a,\omega_t)|\, TN | \boldsymbol{\omega}_T\Big]\Big] \\
&\le \Big(\mathbb{E}_{\pi_R}\mathbb{E}[[1_{\Omega^c(\boldsymbol{\omega}_T)} | \boldsymbol{\omega}_T]] + \mathbb{E}_{\pi_R}\mathbb{E}[[1_{\Omega^c(\boldsymbol{\omega}_T)} | \boldsymbol{\omega}_T]]\Big)\, TN \max_{t,s,a,\omega_t} |r_t(s,a,\omega_t)|.
\end{aligned}
$$

Then, applying Lemma 3.3 and $\mathbb{P}_\pi[\Omega^c(\boldsymbol{\omega}_T)] \leq \sum_{t=1}^{T} \mathbb{P}_\pi[\Delta(P_t(\boldsymbol{\omega}_T))]$, we have:

$$V_N(\pi_R) - V_N(\pi_F) \leq TN \max_{t,s,a,\omega_t} |r_t(s,a,\omega_t)| \sum_{\boldsymbol{\omega}_T} \sum_{t=1}^{T} \mathbb{P}_{\pi_R}[\Delta^c(P_t(\boldsymbol{\omega}_T))] + \mathbb{P}_{\pi_F}[\Delta^c(P_t(\boldsymbol{\omega}_T))]$$

$$\leq 2T^2 N \max_{t,s,a,\omega_t} |r_t(s,a,\omega_t)| \, |W|^T L \exp(-\delta N),$$

where W is the state space of Markov process $\{\omega_t\}_t$.

Finally, combining with Lemma 3.4 concludes the proof. □

CHAPTER 4

BINARY-ACTION INFINITE-HORIZON RESTLESS BANDIT

This Chapter formally describe the binary-action infinite-horizon restless bandit problem, and proposes a novel class of policies called "fluid-balance" policies. We show here in this Chapter that fluid-balance policies always achieve $O(\sqrt{N})$ opt gap. At the end of this Chapter, we also illustrate the state-of-the-art performance of the fluid-balance policies via numerical experiments.

The results in this section are not a straightforward extension of the results in Chapter 2. Indeed, while Chapter 2 showed that fluid-priority policies have a $O(\sqrt{N})$ optimality gap for undiscounted finite-horizon problems, this bound's dependence on T is exponential. Simply applying the bound from Chapter 2 to a truncated infinite-horizon problem (and leveraging discounting to bound the reward obtained after truncation) results in an opt gap bound that grows faster than $\sqrt{N}$. Our fluid-balance policies and their analysis are specifically adapted to the infinite-horizon setting to circumvent this challenge.

We first review the literature and explain our contributions in more detail in the context of that literature in Section 4.1. We then describe the model setup (Section 4.2) and review a standard technique: linear relaxation (Section 4.3). Fluid-balance policies are introduced in Section 4.4 and shown to achieve $O(\sqrt{N})$ opt gap. Finally, we use numerical experiments to justify the state of art performance of fluid-balance policies over other commonly used policies (Section 4.6).

4.1 Literature Review and Contributions

This section first reviews approaches specifically designed for infinite-horizon bandits. Then we describe the most recent progress in the area of finite-horizon bandits which motivates our methodology in this chapter.

Infinite-horizon restless bandits The infinite-horizon restless bandit problem was first formulated by [55]. Since then, the problem has attracted substantial research interest, both from theoretical and practical perspectives. Here we review two main streams of this research: the Whittle index and simulation-based approaches.

Whittle index When the restless bandit problem was first formulated in [55], this paper also proposed an index policy, the so-called Whittle index, as a solution. The Whittle index is defined by considering a problem with a single arm in which one can pull the arm, paying a cost, or idle it. The Whittle index for a state is the cost that makes an optimal policy indifferent between pulling the arm and idling it. This implies a ranking over states that, intuitively, is the same as ranking by a state's "marginal productivity": the difference in discounted long-run reward between activating and idling an arm in this state in the original problem [45]. Intuitively, it should be a good policy to simply pull the arms in the states with the highest marginal productivity. Then Whittle index policy does exactly this: it activates arms according to their indices, from high to low, until all resources are used.

Although intuitively promising, [55] noticed that the willingness to pull an arm in a single-arm problem is not always monotone: it may be optimal to pull the arm when the cost-per-pull is low, idle it when the cost-per-pull is in

an intermediate range, and pull it when the cost-per-pull is high. In such settings, the Whittle index is not well-defined and its link to marginal productivity is lost. [55] conjectured that indexability would imply asymptotic optimality: the difference between the Whittle index's expected performance and that of an optimal policy divided by the number of arms vanishes as the number of arms grows, allowing a constant fraction of the arms to be pulled per time period. Later, however, [54] provided a counterexample to Whittle's conjecture: the Whittle index can fail to be asymptotically optimal even when the indexability condition is satisfied.

Responding to the challenge of establishing indexability, [23, 44] establish alternate sufficient conditions for indexability and [24] characterize some indexable restless bandit families. [37, 39] and [35] show their studied system is indexable. [25, 26, 27, 5] and [32] have extended these ideas to more general settings e.g. convex reward, convex resource budget, stochastic arriving and leaving arms, etc. Nevertheless, establishing indexability remains challenging for most problems and typically entails additional theoretical work that must be done on a problem-by-problem basis.

When a problem is not indexable, multiple values satisfy the conditions that usually define the Whittle index. Thus, attempting to deploy a Whittle index policy in practice without first verifying indexability requires the implementation to explicitly handle this non-uniqueness. The use of implementations assuming a unique Whittle index value in non-indexable problems creates a risk that Whittle index computation produces errors or fails to converge. Also, the intuition for why a Whittle index policy would perform well relies on indexability. When indexability is lacking, policies prioritizing arms based on a Whittle

index computation (while handling non-uniqueness) may be less likely to perform well.

If indexability can be verified, establishing asymptotic optimality requires verifying the additional sufficient conditions discussed above. Past literature suggests that this may be even more difficult than verifying indexability. Most work using Whittle indices does not prove its asymptotic optimality in the problem studied [37, 35] or only proves it in a specific parameter regime [39, 52]. Instead, past literature often relies on numerical simulation to justify the Whittle index's performance.

Thus, despite its popularity, the difficulty of verifying indexability and the additional conditions needed for asymptotic optimality remain a challenge when applying Whittle index policy in real-world problems.

Simulation-based approaches Responding to the limitations of the Whittle index policy, simulation-based approaches have been developed. For example, [40] develop rollout-based heuristic policies and [42] and [53] develop a deep reinforcement learning strategy, using neural networks to approximate the value function. Numerical performance on a collection of benchmark problem instances is their primary concern rather than theoretical guarantees. A policy that performs well in the problem instances simulated may perform poorly in other closely-related problem instances, and so performing well in a simulation-based study may not guarantee good performance across a wider range of problem instances faced after a policy is deployed to the field.

Moreover, if all benchmark policies included in a numerical study are asymptotically suboptimal, a new asymptotically optimal policy has the poten-

tial to significantly outperform all of them. Thus, identifying new asymptotically optimal policies is of significant interest.

Finite-horizon restless bandits While the Whittle index faces challenges in verifying indexability and the additional conditions required for asymptotic optimality, recent progress on finite-horizon restless bandits provides algorithms without these drawbacks in this alternate setting.

In rapid succession, [31, 57, 13] proposed index policies and show that they have $o(N)$, $O(\sqrt{N}\log N)$ and $O(\sqrt{N})$ opt gaps respectively. Then, [58] proposed a class of index policies generalizing [13] and [31], showing that this larger class of policies have at most a $O(\sqrt{N})$ opt gap and, surprisingly, a $O(1)$ opt gap if a non-degeneracy condition is met.

Unlike the Whittle index, these index policies do not require an indexability condition to be well-defined. Moreover, they come with performance guarantees that do not require verifying extra sufficient conditions: $O(\sqrt{N})$ for [58, 13, 31], $O(\sqrt{N}\log N)$ for [57].

We argue in this paper that a key difference in the approach enabled these finite-horizon analyses to achieve asymptotic optimality and to avoid challenges in establishing indexability: their Lagrangian relaxation uses a *sequence* of Lagrange multipliers, one for each time period, while the Whittle index uses a single global Lagrange multiplier. Using a time-varying Lagrange multiplier is intuitive in the finite-horizon setting: the finite horizon causes the problem to be non-stationary, naturally inspiring a time-inhomogeneous approach.

We show in this paper that this time-inhomogeneous approach can be generalized to the infinite-horizon setting to overcome the shortcomings of the Whit-

tle index and other past approaches to the infinite-horizon restless bandit. That a time-inhomogeneous approach would be relevant to the infinite-horizon setting may, at first glance, seem surprising: the infinite-horizon problem is stationary, implying the existence of stationary optimal policies, and suggesting that asymptotically optimal policies should also be stationary. Part of our contribution is to explain why non-stationarity is an important tool for providing asymptotic optimality in stationary infinite-horizon problems.

Summary of Contribution Out work provide a novel class of computationally scalable non-stationary infinite-horizon restless bandit policies called "fluid-balance" policies. We show that they are asymptotically optimal, achieving a $O(\sqrt{N})$ opt gap. This result does not require indexability or other sufficient conditions beyond those defining the problem we study, such as arms' states belonging to a finite state space and state transitions that are conditionally independent across arms. Moreover, despite being time-inhomogeneous in an infinite horizon problem, we show that they can be computed in finite time. They are computed by considering a finite linear program formed by truncating the infinite-horizon problem. Truncating at the $O(\log N)$-th period allows fluid-balance policies to achieve a $O(\sqrt{N})$ opt gap.

This requires going substantially beyond applying a previously proposed finite-horizon policy to the truncated problem. Policies with $O(\sqrt{N})$ opt gaps in the finite-horizon setting proposed in [13] and [58] have opt gap bounds that depend exponentially on the time horizon. Simply applying one of these policies and its associated performance bound to a truncated problem (and leveraging discounting to bound the reward obtained after truncation) results in an opt gap bound that grows faster than $O(\sqrt{N})$. Our fluid-balance policies and their

analysis are specifically adapted to the infinite-horizon setting to circumvent this challenge.

4.2 System Model

This section formulates the infinite-horizon restless bandit problem as a Markov Decision Process (MDP).

Model There are N arms, each of which shares the same finite state space S. We use $s_{i,t}$ to indicate the state of i-th arm at time t. At each period t for each arm i, the decision-maker chooses whether to pull the arm ($a_{i,t} = 1$) or leave it idle ($a_{i,t} = 0$). We define $A = \{0, 1\}$ to be the space of available actions in which $a_{i,t}$ takes values. These actions must respect a so-called "budget constraint" in which the number of arms pulled at period t is $B_t = \lfloor \alpha_t N \rfloor$, where $0 \leq \alpha_t \leq 1$ is a pre-specified budget ratio.

Based on the action applied, each arm's state transitions stochastically to time $t + 1$ according to a time-homogeneous known transition kernel $P = \{p(s, a, s')\}_{s,s' \in S, a \in A}$ where $p(s, a, s') = \mathbb{P}(s_{t+1,i} = s' | s_{t,i} = s, a_{t,i} = a)$. All arms share the same transition kernel, and any arm's transition is conditionally independent from others given its own state and action. Each state-action pair is associated with a time-homogeneous reward, given by a known reward function $r : S \times A \to \mathbb{R}$. The decision-maker aims to maximize the total discounted reward collected from all N arms over the infinite horizon with discount factor γ.

To complete the formal definition of our problem involving N arms, we introduce some additional notation. We use $\mathbb{S} = S^N$ to denote the N-fold Cartesian

product of the state space S, and define $\mathbb{A} = A^N$ similarly. All N arms together form an MDP with state space $\mathbb{S}$ and action space $\mathbb{A}$. We call this the "joint MDP" to distinguish it from MDPs that we reference later involving a single arm. The state in this joint MDP at time t is $\mathbf{s}_t = (s_{t,1}, s_{t,2}, ..., s_{t,N}) \in \mathbb{S}$, which indicates that arm i has state $s_{t,i}$. The action is $\mathbf{a}_t = (a_{t,1}, a_{t,2}, ..., a_{t,N}) \in \mathbb{A}$, which indicates that action $a_{t,i}$ is applied to arm i.

The reward function of the joint MDP, $R : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$, is the summation across all the single-arm reward defined above,

$$R(\mathbf{s}_t, \mathbf{a}_t) = \sum_{i=1}^{N} r(s_{t,i}, a_{t,i}).$$

For element $\mathbf{a} = (a_1, a_2, ..., a_N)$ in $\mathbb{A}$, we use $|\mathbf{a}| = \sum_{i=1}^{N} a_i$ to indicate the L^1-norm of $\mathbf{a}$, i.e, the number of pulled arms. We write our budget constraint at time t as $|\mathbf{a}_t| = B_t$.

The transition kernel for the joint MDP is the product of each arm's transition kernel,

$$\mathbb{P}[\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t] = \prod_{i=1}^{N} p(s_{t,i}, a_{t,i}, s_{t+1,i}).$$

We assume all arms start from the same initial state s^*. Our analysis can be easily generalized to the case where arms start from different states.

A policy π is a function that maps the current state $\mathbf{s}_t \in \mathbb{S}$ and time t to an action $\mathbf{a}_t \in \mathbb{A}$. The objective of the policy is to maximize the expected total reward, subject to budget and initial states constraints specified above. This objective can be written as,

$$\begin{aligned} &\max_{\pi} \ \mathbb{E}_{\pi} \sum_{t=1}^{\infty} \gamma^t R(\mathbf{s}_t, \mathbf{a}_t) \\ &\text{subject to: } |\mathbf{a}_t| = \lfloor \alpha N \rfloor \ \forall t, \end{aligned} \tag{4.1}$$

where $\mathbb{E}_\pi$ indicates the expectation taken under policy π.

We define the value function of a policy π as $V_N(\pi) = \sum_{t=1}^{\infty} \gamma^t R(\mathbf{s}_t, \mathbf{a}_t)$ and measure a policy's performance by comparing it with the optimal policy solving (4.1). Let $V_N^* = \sup_\pi V_N(\pi)$ be the expected total reward obtained by an optimal policy. Then the optimality gap of the policy π is defined as

$$V_N^* - V_N(\pi).$$

Maximizing policies' value function is equivalent to minimizing their optimality gap. We are interested in solving (4.1) when N is large.

4.3 Background: Preliminary Results and Notation

In this section, we define a linear programming relaxation that provides an upper bound $\hat{V}_N^*$ for V_N^*. Some of results shown in the section are similar to Section 2.3. We describe these results briefly and leave their proofs to the Appendix.

Linear Programming Relaxation We relax problem (4.1)'s almost sure cardinality constraints on the number of pulls to constraints on the expectation of the number of puls:

$$\hat{V}_N^* = \max_\pi \left\{ \mathbb{E}_\pi \sum_{t=1}^{\infty} \gamma^t R(\mathbf{s}_t, \mathbf{a}_t) \middle| \mathbb{E}|\mathbf{a}_t| = \alpha_t N, \ \forall t \right\}. \tag{4.2}$$

However, solving the above infinite-horizon relaxed Problem (4.2) exactly is typically not possible. A technique common among those wishing to solve such infinite-horizon discounted problems is to truncate (4.2) up to a large horizon T and solve this finite-horizon approximation:

$$\hat{V}_N^*(T) = \max_\pi \left\{ \mathbb{E}_\pi \sum_{t=1}^{T} \gamma^t R(\mathbf{s}_t, \mathbf{a}_t) \middle| \mathbb{E}|\mathbf{a}_t| = \alpha_t N, \ \forall t \le T \right\}. \tag{4.3}$$

Computational difficulties aside, our notation explicitly allows $T = \infty$, in which case $\hat{V}_N^*(\infty) = \hat{V}_N^*$. We similarly denote the optimal reward of the truncated version of the original problem as

$$V_N^*(T) = \max_{\pi} \left\{ \mathbb{E}_\pi \sum_{t=1}^{T} \gamma^t R(\mathbf{s}_t, \mathbf{a}_t) \middle| |\mathbf{a}_t| = \alpha_t N, \ \forall t \le T \right\}. \tag{4.4}$$

For simplicity, we assume N is taken such that $\alpha_t N$ are integral for all $s \in S$. All the following results generalize if this is not true as we discuss briefly in Appendix C.2.

Similar with Lemma 2.1, we can show

Lemma 4.1. $V_N^*(T) \le \hat{V}_N^*(T) = N \, \hat{V}_1^*(T)$.

To calculate $\hat{V}_1^*$, we only need to solve

$$\max_{\pi} \left\{ \mathbb{E}_\pi \sum_{t=1}^{T} \gamma^t r(s_t, a_t) \middle| \mathbb{E}|a_t| = \alpha_t, \forall t \le T \right\}. \tag{4.5}$$

Writing Problem (4.5) in terms of the occupation measure $x_t(s,a) := \mathbb{P}[s_t = s, a_t = a]$, then Problem (4.5) is equivalent to

$$\begin{aligned}
&\max \sum_{s\in S, a\in A} \sum_{t=1}^{T} \gamma^t r(s,a) x_t(s,a) \\
&\text{subject to} \\
&\sum_{a\in A} x_t(s,a) = \sum_{a\in A} \sum_{s'\in S} x_{t-1}(s',a) p_{t-1}(s',a,s), \ \forall s \in S, \ 2 \le t \le T; \\
&\sum_{s\in S} x_t(s,1) = \alpha_t, \ t \le T; \\
&\sum_{a\in A} x_1(s^*,a) = 1; \ \sum_{a\in A, s\in S} x_1(s,a) = 1; \\
&x_t(s,a) \ge 0, \ \forall s \in S, a \in A, t \le T.
\end{aligned} \tag{4.6}$$

Notation We use additional notation borrowed from Section 2.3. To make this section more self-contained, we list this additional notation here:

- $z_t(s)$, $Z_t^N(s)$, $x_t(s,a)$ and $X_t^N(s,a)$ for all $s \in S$ and $a \in A$;
- the vector form of the above pieces of notation, z_t, Z_t^N, x_t and X_N^t;
- diffusion statistics $\tilde{Z}_t^N$ and $\tilde{X}_t^N$;
- a policy π and its induced maps $\tilde{\pi}_{t,N}$.

4.4 Diffusion Regular Conditions

This section characterize a set of properties of policies, which we refer to as diffusion regularity. Then we deduce some implications of these properties. This section builds the foundation for the following sections: fluid-balance policy proposed in section 4.5 is diffusion regular.

The insight behind diffusion regularity is, roughly speaking, as long as the diffusion statistic $\tilde{X}_t^N$ is bounded by $O(1)$, $\tilde{Z}_{t+1}^N$ will also be bounded by $O(1)$. Thus, we could control the grow of the first moments of diffusion statistics $(\tilde{Z}_t^N, \tilde{X}_t^N)$ across periods (Lemma 4.2).

Formally speaking,

Definition 4.1. *A policy π is called diffusion regular if its induced maps $\tilde{\pi}_{t,N}$ satisfy*

- *There exists $C_1 > 0$, s.t. for any $t \le T$, N, θ_1 and θ_2*

$$|\tilde{\pi}_{t,N}(\theta_1) - \tilde{\pi}_{t,N}(\theta_2)| \le C_1|\theta_1 - \theta_2|.$$

- *There exits $C_2 > 0$, s.t. for any $t \le T$ and N*

$$|\tilde{\pi}_{t,N}(0)| \le C_2.$$

- *There exists a map* $\tilde{\pi}_{t,\infty}$*, s.t.* $\tilde{\pi}_{t,N}(\theta) \to \tilde{\pi}_{t,\infty}(\theta)\ \forall \theta$ *and* $t \le T$*, as* $N \to +\infty$*.*
- *There exists* $C_3 > 0$*, s.t. for any* $t \le T$*,* N*,* θ *and* $s \in S$*,* $\sum_{a\in A} |\tilde{\pi}_{t,N}(\theta)(s,a)| \le |\theta(s)| + C_3$*.*

where $|\cdot|$ *stands for the* L^1*-norm in Euclidean space.*

If a policy π is diffusion regular, their first moments can be upper bounded by a linear-growth sequence (Lemma 4.2). Proof of Lemma 4.2 are left in the Appendix.

Lemma 4.2. *If a policy* π *is diffusion regular, then there exists constant* c_1 *and* c_2 *(neither depends on* T*), s.t. for all* $t \le T$ *and* N *(*N *could be infinity),*

$$\mathbb{E}[|\tilde{Z}_t^N|] \le c_1 + c_2 t.$$

4.5 Fluid-balance Policy

This section defines fluid-balance policies, and shows that all fluid-balance policies are diffusion regular and achieve $O(\sqrt{N}) + O(N\gamma^T)$ optimality gap.

Roughly speaking, a fluid-balance policy is defined by two pieces: an optimal solution of the LP relaxation problem and prioritization scores over states. Then a fluid-balance policy pulls arms respecting two rules: a consistency rule and a prioritization rule. The consistency rule requires the diffusion statistics $\hat{X}_t^N(s,\cdot)$ share the same sign with $\hat{Z}_t^N(s)$ for each state s. The prioritization rule requires pulling arms according to the prioritization score as much as we can with respect to the consistency rule.

Mathematically speaking, a fluid-priority policy is parameterized by an occupation measure $\{x_t(s,a)\}_{t,s,a}$ solving Problem (4.6), and "priority-score" functions $\{\mathcal{P}_t(\cdot)\}_t$ assigning each state a real number. With these notations, the fluid-balance policy is defined as in Algorithm 6.

Algorithm 6 fluid-balance policy

Input: optimal occupation measure $(x_t(s,a))_{t\geq 0, s\in S, a\in A}$, priority-score functions $\{\mathcal{P}_t\}_{t\geq 0}$.

for $t = 1, 2, ..., T$ **do**
  Classify states into three categories: fluid-active, fluid-neutral and fluid-inactive.
  Observe there are $Z_t(s)$ arms in state s, and its associated diffusion statistics $\tilde{Z}_t(s)$.
  for state s in $C_t^+ \cup C_t^0 \cup C_t^-$ **do**
    $X_t(s,1) \leftarrow \min\{Z_t(s), \lceil x_t(s,1)N + \sqrt{N}|\tilde{Z}_t(s)|\rceil\}$.
  end for
  while $\sum_s X_t(s,1) > \lfloor \alpha_t N \rfloor$ **do**
    Find state s with the lowest priority-score s.t.

$$X_t(s,1) > \max\{0, \lfloor x_t(s,1)N - \sqrt{N}|\tilde{Z}_t(s)|\rfloor\}$$

    $X_t(s,1) \leftarrow X_t(s,1) - 1$
  end while
  Pull $X_t(s,1)$ arms in state s.
end for

We can show that a fluid-balance policy is always diffusion regular, and actually yields a slightly tighter L^1−moment bound than Lemma 4.2. The proof is left in the Appendix.

Lemma 4.3. *Any fluid-balance policy π is diffusion regular, and $\mathbb{E}_\pi[|\tilde{Z}_t^N|] \leq 2t|S|^2$ for $t \leq T$.*

Now we are able to show the main results

Theorem 4.1. *Given any fluid-balance policy π, $V_N^* - V_N(\pi) = O(\sqrt{N}) + O(N\gamma^T)$.*

Proof of Theorem 4.1. Denote π^* the optimal policy maximizing the infinite-horizon Problem (4.1). Then by denoting $B := \max_{s,a} |r(s,a)|$,

$$\begin{aligned} V_N^* - V_N(\pi) &= \mathbb{E}_{\pi^*} \sum_{t=1}^{T} \gamma^t r(s_{i,t}, a_{i,t}) - \mathbb{E}_{\pi} \sum_{t=1}^{T} \gamma^t r(s_{i,t}, a_{i,t}) + \mathbb{E}_{\pi^*} \sum_{t=T+1}^{\infty} \gamma^t r(s_{i,t}, a_{i,t}) - \mathbb{E}_{\pi} \sum_{t=T+1}^{\infty} \gamma^t r(s_{i,t}, a_{i,t}) \\ &\leq \hat{V}_N(T) - \mathbb{E}_{\pi} \sum_{t=1}^{T} \gamma^t r(s_{i,t}, a_{i,t}) + 2 \sum_{t=T+1}^{\infty} \gamma^t BN \end{aligned}$$

where the last inequality is due to the definition of $\hat{V}_N(T)$.

We deal with these two terms $\hat{V}_N(T) - \mathbb{E}_{\pi} \sum_{t=1}^{T} \gamma^t r(s_{i,t}, a_{i,t})$ and $2 \sum_{t=T+1}^{\infty} \gamma^t BN$ separately. For the first term,

$$\hat{V}_N(T) - \mathbb{E}_{\pi} \sum_{t=1}^{T} \gamma^t r(s_{i,t}, a_{i,t}) = -\sqrt{N} \mathbb{E}_{\pi} \sum_{t=1}^{T} \sum_{s,a} \gamma^t r(s,a) \tilde{X}_t^N(s,a) \leq \sqrt{N} B \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} \gamma^t |\tilde{Z}_t^N| \right].$$

Recall Lemma 4.3, we have

$$\hat{V}_N(T) - \mathbb{E}_{\pi} \sum_{t=1}^{T} \gamma^t r(s_{i,t}, a_{i,t}) \leq \sqrt{N} B \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} \gamma^t |\tilde{Z}_t^N| \right] \leq \sqrt{N} B \sum_{t=1}^{T} \gamma^t 2t |S|^2 \leq \sqrt{N} \frac{2|S|^2 \gamma B}{(1-\gamma)^2} = O(\sqrt{N}).$$

For the second term,

$$2 \sum_{t=T+1}^{\infty} \gamma^t BN = \frac{2B\gamma}{1-\gamma} N\gamma^T = O(N\gamma^T).$$

Combining above analysis together, we conclude $V_N^* - V_N(\pi) = O(\sqrt{N}) + O(N\gamma^T)$. □

4.6 Numerical Experiment

This section we illustrate the performance of fluid-balance policy through two numerical experiments, especially focusing on its advantage over Whittle index policy.

In the first experiment, we construct an example which reflects a classic question in the real-world decision making: whether slow and steady could win the race. The decision maker could either choose to generate large amount of reward but only possible in short horizons or generate steady small amount of reward in a very long horizon. We show this example is not indexable, thus Whittle index is not well-defined. However, fluid-balance policy can be analytically solved and is actually the optimal policy.

In the second experiment, we compare the fluid-balance policy against the Whittle index for an discounted version of a problem studied in [21] and [11]. Although this problem is indexable, fluid-balance policy can outperform the Whittle index by over 30%. As Whittle index serves as the benchmark in these literature [21, 11], we would like to kindly remind researchers in the future that fluid-balance policy may serve as a better benchmark.

4.6.1 Does steady-and-slow win the race?

This section we construct an example which reflects the a common question in the real-world decision making: should we pursue the short-term interest or long-term interest? We will show that Whittle index is not well-defined for this example while fluid-balance policy is actually the optimal policy.

Problem setup There are 7 different states: $\{A_0, A, A', B, B', B^*, C\}$. The tran-

sition between different states are given as

$\mathbb{P}\big[s_{t+1} = A' | s_t = A_0, a_t = a\big] = 1, \forall a \in \{0, 1\}$;

$\mathbb{P}\big[s_{t+1} = A' | s_t = A', a_t = a\big] = 1, \forall a \in \{0, 1\}$;

$\mathbb{P}\big[s_{t+1} = A' | s_t = A, a_t = 1\big] = 1 - \epsilon,\ \mathbb{P}\big[s_{t+1} = C | s_t = A, a_t = 1\big] = \epsilon,\ \mathbb{P}\big[s_{t+1} = B | s_t = A, a_t = 0\big] = 1$;

$\mathbb{P}\big[s_{t+1} = B^* | s_t = B^*, a_t = a\big] = 1, \forall a \in \{0, 1\}$;

$\mathbb{P}\big[s_{t+1} = B^* | s_t = B', a_t = 1\big] = 1,\ \mathbb{P}\big[s_{t+1} = B' | s_t = B', a_t = 0\big] = 1$;

$\mathbb{P}\big[s_{t+1} = B' | s_t = B, a_t = 1\big] = 1 - \epsilon,\ \mathbb{P}\big[s_{t+1} = C | s_t = B, a_t = 1\big] = \epsilon,\ \mathbb{P}\big[s_{t+1} = A | s_t = B, a_t = 0\big] = 1$;

$\mathbb{P}\big[s_{t+1} = C | s_t = C, a_t = a\big] = 1, \forall a \in \{0, 1\}$.

Reward for most state-action pairs are zero except

$$r(A', 1) = \alpha,\ r(B', 1) = \beta.$$

We try to maximize the infinite-horizon discounted reward with discount factor $\gamma = 1 - \epsilon$. The parameters satisfy $0 < (1 + \frac{1}{\gamma^2})\ \alpha < \beta < \frac{\gamma}{1-\gamma}\alpha$ and $\epsilon < \frac{1}{8}$.

We can pull γN number of arms out of N arms in each period. And at the initial stage, there are $\phi_1 N$ arms are in state A, $\phi_2 N$ arms are in state C and $\phi_3 N$ arms are in state A_0, where $\phi_1 = 2 - \frac{1}{\gamma},\ \phi_2 = \gamma + \frac{1}{\gamma} - 2$ and $\phi_3 = 1 - \gamma$.

Indexability This problem is not indexable, thus Whittle index is not well-defined.

Proposition 4.1. *The problem is not indexable.*

Fluid-balance policy This infinite-horizon relaxation problem permits an analytical solution, as shown below.

Proposition 4.2. *The policy that pulls all arms in A_0 and A in the first period, then pulls all arms in A′ starting from the second period solves the relaxation problem.*

Based on the optimal policy for the relaxed problem, we can construct its corresponding fluid-balance policy π: π pulls all arms in A_0 and A in the first period, then pull as many arms in A' as possible starting from the second period. If there are less than γN arms in state A', pull arms in C to meet the budget.

Surprisingly, the fluid-balance policy is not only asymptotically optimal, but truly optimal.

Proposition 4.3. *The fluid-balance policy π is the optimal policy for the original problem.*

When analyzing the opt gap of a fluid-balance policy in Theorem 4.1, we compare the reward of the policy against the optimal reward of the relaxation problem rather than the original problem. Since the fluid-balance policy is optimal, the opt gap is 0. But still if we compare against the optimal reward of the relaxation problem, we still get $O(\sqrt{N})$ upper bound of opt gap.

Proposition 4.4. $\hat{V}_N^* - V_N(\pi) = \theta\sqrt{N} + O(1)$, *where* $\theta = \sqrt{\frac{(1-\gamma)(2\gamma-1)}{2\pi}}$.

Proof of above propositions are left in the appendix.

4.6.2 Whittle index: not a good benchmark

As we discuss in section 4.1, there is a stream of "learning"-style bandit literature which assumes the state transition is unknown, e.g. [21] and [11]. Rather than designing policy based on full information, these papers design algorithms to generate policies on-the-fly while estimating the state transition kernel. Thus, to benchmark the performance of their algorithm, they usually compare against

a full-information-based policy (e.g., Whittle index) on some well-known testing problems. We choose one of the most common testing problems in our second numerical experiment. In [21] and [11], they benchmark their learning algorithm against an unidiscounted version of this problem which $N/10$ arms are allowed to be pulled out of N arms per period.

In this section, we show that benchmarking against the Whittle index policy may not always be a good idea since it may not be asymptotically optimal. For the discounted setting with $\gamma = 1/2$ and $N/2$ arms are allowed to be pulled out of N arms per period, we show fluid-balance policy outperforms Whittle index policy by over 30%.

Problem setup There are 4 different states: $\{0, 1, 2, 3\}$ and transition kernels $P_a = p(s, 0, s')_{s,s' \in S}$ for action $a = 0$ and $a = 1$ are given by

$$P_0 = \begin{pmatrix} 1/2 & 0 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}, \; P_1 = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \end{pmatrix}.$$

The reward solely depends on the state and irrelevant of the action:

$$r(0, a) = -1, \; r(1, a) = 0, \; r(2, a) = 0, \; r(3, a) = 1; \forall a \in \{0, 1\}.$$

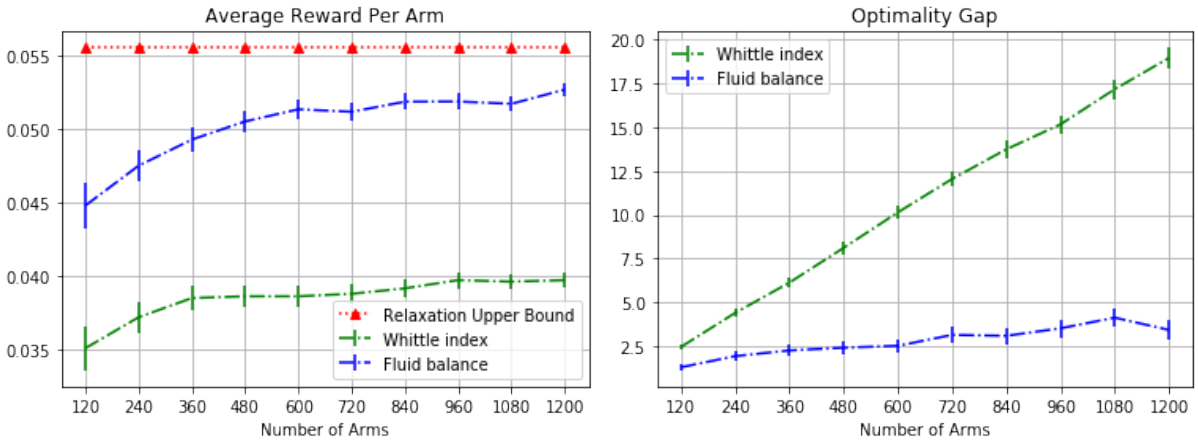
The discount factor $\gamma = 1/2$ and we are allowed to pull $N/2$ arms out of N arms per period. Initially, there are $N/6$ arms in state 0, $N/3$ arms in state 1 and $N/2$ arms in state 2.

Indexability Via direct calculation, we can show that this problem is indexable. Ranking from the highest to the lowest according to the Whittle index, state 2 > state 1 > state 0 > state 3.

Fluid-balance policy Since this infinite-horizon problem's relaxation does not permit an analytical solution, we solve the truncated version up to $T = 100$. This provides a satisfying approximation of the true relaxation upper bound because total rewards after period 100 is only of machine precision scale: $\frac{1}{2}^{100} << 10^{-17}$ where 10^{-17} is the machine precision of 64bit float number.

After solving the truncated relaxation problem, we need a final piece before implementing a fluid-balance policy: the prioritization score. We adhere to Whittle index here, where we rank states from high to low as state 2 > state 1 > state 0 > state 3.

Comparison We compare the performance of Whittle index and fluid-balance policy here.



(a) (b)

Figure 4.1: Performance comparison between Whittle index and fluid-balance policy. The left panel shows the average reward per arm versus number of arms, where we compare the relaxation upper bound, the whittle index and the fluid-balance policy. The right panel shows an upper bound on the opt gap (relaxation upper bound minus a simulation-based estimate of reward) versus number of arms N. As we can see, opt gap of Whittle index grows linearly while opt gap of fluid-balance policy grows sublinearly with respect to N.

APPENDIX A

APPENDIX: BINARY-ACTION FINITE-HORIZON RESTLESS BANDIT

This section provides all technical proofs not included in the main paper.

A.1 Proof for Lemma 2.1

In the original formulation of the restless bandit, problem (2.1), the budget constraint $|\mathbf{a}_t| = \lfloor \alpha_t N \rfloor$ applies on each sample path. The relaxed problem (2.2) is identical except that this constraint is replaced by the weaker one, $\mathbb{E}|\mathbf{a}_t| = \alpha_t N$. Recalling our assumption here that $\alpha_t N$ is an integer, the right-hand sides of these two constraints are the same. (Generalizations to non-integer $\alpha_t N$ are discussed in Appendix A.2). Thus, the set of feasible policies in (2.1) is a subset of those in (2.2), implying that the value of (2.1) is bounded above by that of (2.2), i.e.,

$$V_N^* \leq \hat{V}_N^*. \tag{A.1}$$

To prove $\hat{V}^*(N) = N\hat{V}^*(1)$, we use a Lagrangian Relaxation similar to [20, 26] as the key idea in the following argument.

Through straightforward imitation of the proof of the Fenchel Duality Theorem [47],

$$\max_{\pi} \min_{\lambda_{1:T}} \mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) + \lambda_t(\alpha_t N - |\mathbf{a}_t|) = \min_{\lambda_{1:T}} \max_{\pi} \mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) + \lambda_t(\alpha_t N - |\mathbf{a}_t|). \tag{A.2}$$

In this use of the Fenchel Duality Theorem, we note that maximization over policies π on the right-hand side of (A.2) with fixed $\lambda_{1:T}$ can be viewed as as a

linear program. More detailed discussion of this standard result can be found in [13].

The left-hand side of Equation (A.2) equals $\hat{V}^*(N)$. On the right hand side, for fixed $\lambda_{1:T}$,

$$\mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) + \lambda_t(\alpha_t N - |\mathbf{a}_t|) = \mathbb{E}_\pi \sum_{t=1}^{T} \sum_{i=1}^{N} r_t(s_{t,i}, a_{t,i}) + \lambda_t(\alpha_t - a_{t,i}).$$

Since all arms share the same transition kernel, reward function, and distribution over initial state,

$$\mathbb{E}_\pi \sum_{t=1}^{T} \sum_{i=1}^{N} r_t(s_{t,i}, a_{t,i}) + \lambda_t(\alpha_t - a_{t,i}) = N\, \mathbb{E}_\pi \sum_{t=1}^{T} r_t(s_{t,1}, a_{t,1}) + \lambda_t(\alpha_t - a_{t,1}).$$

So we conclude

$$\min_{\lambda_{1:T}} \max_{\pi} \mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) + \lambda_t(\alpha_t N - |\mathbf{a}_t|) = N \min_{\lambda_{1:T}} \max_{\pi} \mathbb{E}_\pi \sum_{t=1}^{T} r_t(s_{t,1}, a_{t,1}) + \lambda_t(\alpha_t - a_{t,1}). \tag{A.3}$$

By using Fenchel Duality again on the one-arm problem,

$$\begin{aligned} \min_{\lambda_{1:T}} \max_{\pi} \mathbb{E}_\pi \sum_{t=1}^{T} r_t(s_{t,1}, a_{t,1}) + \lambda_t(\alpha_t - a_{t,1}) &= \max_{\pi} \min_{\lambda_{1:T}} \mathbb{E}_\pi \sum_{t=1}^{T} r_t(s_{t,1}, a_{t,1}) + \lambda_t(\alpha_t - a_{t,1}) \\ &= \hat{V}^*(1). \end{aligned} \tag{A.4}$$

Summarizing, equations (A.2), (A.3) and (A.4) together imply,

$$\hat{V}^*(N) = N\, \hat{V}^*(1).$$

A.2 Discussion of the rounding error in budget constraints

The original problem (2.1) constrains the number of pulls to $\lfloor \alpha_t N \rfloor$ (almost surely), while the relaxed problem (2.2) constrains this number to $\alpha_t N$ (in expectation). We think of these differences as "rounding errors" in the relaxed

problem. Here we discuss their effect and show that they result in at most a constant difference in the optimal objective value.

Mathematically speaking, denote

$$\hat{V}_N^* = \max_{\pi} \left\{ \mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) \middle| \mathbb{E}|\mathbf{a}_t| = \alpha_t N,\ \mathbb{E} \sum_{i=1}^{N} 1(s_{1,i} = s^*) = N, \forall t \in [T] \right\},$$

$$\hat{V}_{N,R}^* = \max_{\pi} \left\{ \mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) \middle| \mathbb{E}|\mathbf{a}_t| = \lfloor \alpha_t N \rfloor,\ \mathbb{E} \sum_{i=1}^{N} 1(s_{1,i} = s^*) = N, \forall t \in [T] \right\}.$$

We claim that $|\hat{V}_N^* - \hat{V}_{N,R}^*| \leq c$, where c does not depend on N. The theoretical analysis through the rest of the paper after Lemma 1 compares policy performance against $\hat{V}_N^*$ and shows that this difference is $o(N)$, $O(\sqrt{N})$, or $O(1)$ depending on conditions. The fact that $\hat{V}_N^*$ is separated from $\hat{V}_{N,R}$ by at most a constant then implies that the difference in policy performance compared to $\hat{V}_{N,R}^*$ has the same asymptotic dependence on N. This and the fact that $\hat{V}_{N,R}$ is an upper bound on (2.1) even when $\alpha_t N$ are not integers provides opt gaps of $o(N)$, $O(\sqrt{N})$ or $O(1)$ respectively.

The proof of the claim that $|\hat{V}_N^* - \hat{V}_{N,R}^*| \leq c$ is straightforward. As seen from Lemma 2.1, there exists a single-arm strategy that pulls α_t arms per period in expectation and achieves objective value $\hat{V}_1^*$. Thus, we can pull $N - \max_t \lceil \frac{1}{\alpha_t} \rceil$ arms according to this strategy and pull each remaining arm with probability $\frac{\lfloor \alpha_t N \rfloor - \alpha_t (N - \max_t \lceil \frac{1}{\alpha_t} \rceil)}{\max_t \lceil \frac{1}{\alpha_t} \rceil} \in [0, 1]$ at period t. Thus, we show

$$\frac{N - \max_t \lceil \frac{1}{\alpha_t} \rceil}{N} \hat{V}_N^* - \hat{V}_{N,R}^* \leq\ T \max_{s,a,t} r_t(s, a).$$

Similarly, we can show

$$\frac{N - \max_t \lceil \frac{1}{\alpha_t} \rceil}{N} \hat{V}_{N,R}^* - \hat{V}_N^* \leq\ T \max_{s,a,t} r_t(s, a).$$

Combining the above two inequalities with the fact that $\hat{V}_{N,R}^*/N$ and $\hat{V}_N^*/N$

are both uniformly bounded by $T \max_{s,a,t} |r_t(s,a)|$ concludes the statement with $c = T(1 + \max_t \lceil \frac{1}{\alpha_t} \rceil) \max_{s,a,t} |r_t(s,a)|$.

A.3 Proof of Lemma 2.2

We prove Lemma 2.2 by induction on t. When $t = 1$, that all arms are in state s^* implies $\frac{Z_1^N}{N} \to z_1$. Then, by the definition of fluid consistency, $\frac{X_1^N}{N} \to x_1$. Thus Lemma 2.2 holds for $t = 1$.

Now assume Lemma 2.2 holds for t, and we will show it holds for $t + 1$. By the definition of fluid consistency, we only need to prove

$$\frac{Z_{t+1}^N}{N} \to z_{t+1}. \tag{A.5}$$

Recalling our system dynamics,

$$Z_{t+1}^N(s) = \sum_{s' \in S, a \in A} \sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s), \tag{A.6}$$

where $1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s)$ is the indicator function of the event $\{s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s\}$, we only need to show that

$$\frac{1}{N} \sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) \to x_t(s', a) p(s', a, s). \tag{A.7}$$

since the sum over s' of the right-hand side is $\sum_{s' \in S} x_t(s', a) p(s', a, s) = z_{t+1}(s)$.

If $x_t(s', a) > 0$, then $X_t^N(s', a) \to \infty$ as $N \to \infty$ by the induction hypothesis. Thus as $N \to \infty$,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) &= \frac{X_t^N(s', a)}{N} \frac{1}{X_t^N(s', a)} \sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) \\ &\to x_t(s', a) p(s', a, s), \end{aligned}$$

by the definition of fluid consistency and the strong law of large numbers.

If $x_t(s', a) = 0$,

$$\frac{1}{N}\sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) \leq \frac{X_t^N(s', a)}{N} \to 0.$$

Combining the cases $x_t(s', a) > 0$ and $x_t(s', a) = 0$, equation (A.7) is shown.

To summarize,

$$\frac{Z_{t+1}^N(s)}{N} = \sum_{s' \in S} \frac{1}{N}\sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) \to \sum_{s' \in S}\sum_{a \in A} x_t(s', a)p(s', a, s) = z_{t+1}(s).$$

A.4 Proof of Lemma 2.3

This section proves Lemma 2.3. When $Z_t^N/N \to z_t$, we have

$$\frac{\tilde{Z}_t^N}{\sqrt{N}} = \frac{Z_t^N - Nz_t}{N} \to 0.$$

To show $\frac{X_t}{N} \to x_t$, it is equivalent to show

$$\frac{\pi_{t,N}(\tilde{Z}_t^N)}{\sqrt{N}} \to 0.$$

Notice by Conditions 1 and 2 in Definition 2.2,

$$|\pi_{t,N}(\tilde{Z}_t^N)| \leq |\pi_{t,N}(0)| + C_1|\tilde{Z}_t^N| \leq C_2 + C_1|\tilde{Z}_t^N|.$$

Thus,

$$\frac{\pi_{t,N}(\tilde{Z}_t^N)}{\sqrt{N}} \leq \frac{C_2}{\sqrt{N}} + C_1\frac{|\tilde{Z}_t^N|}{\sqrt{N}} \to 0.$$

A.5 Proof of Lemma 2.4

This section proves Lemma 2.4. To begin with, we first state and prove Lemma A.1.

Lemma A.1. *If a policy π is diffusion regular and $\tilde{Z}_t^N \to \tilde{Z}_t^\infty$ in distribution, then $\tilde{X}_t^N \to \tilde{X}_t^\infty$ in distribution for some random variable $\tilde{X}_t^\infty$.*

Proof of Lemma A.1. By the Skorokhod representation Theorem, there exists a probability space $(\Omega, \mathbb{P})$ and a sequence of random variables $\{\widetilde{Z}_t^N\}_N$ and $\widetilde{Z}_t^\infty$ such that

$$\widetilde{Z}_t^N = \tilde{Z}_t^N \text{ and } \widetilde{Z}_t^\infty = \tilde{Z}_t^\infty \text{ in distribution,}$$
$$\widetilde{Z}_t^N \to \widetilde{Z}_t^\infty \text{ as } N \to \infty \quad a.s.$$

We will prove convergence in distribution of $\tilde{X}_t^N = \tilde{\pi}_{t,N}(\widetilde{Z}_t^N)$ to $\tilde{X}_t^\infty := \tilde{\pi}_{t,\infty}(\widetilde{Z}_t^\infty)$. Notice

$$\begin{aligned}|\tilde{\pi}_{t,N}(\widetilde{Z}_t^N) - \tilde{\pi}_{t,\infty}(\widetilde{Z}_t^\infty)| &\le |\tilde{\pi}_{t,N}(\widetilde{Z}_t^N) - \tilde{\pi}_{t,N}(\widetilde{Z}_t^\infty)| + |\tilde{\pi}_{t,N}(\widetilde{Z}_t^\infty) - \tilde{\pi}_{t,\infty}(\widetilde{Z}_t^\infty)| \\ &\le C_1|\widetilde{Z}_t^N - \widetilde{Z}_t^\infty| + |\tilde{\pi}_{t,N}(\widetilde{Z}_t^\infty) - \tilde{\pi}_{t,\infty}(\widetilde{Z}_t^\infty)|,\end{aligned}$$

which converges to 0 as $N \to \infty$ by almost sure convergence of $\widetilde{Z}_t^N$ to $\widetilde{Z}_t^\infty$ and the convergence of $\tilde{\pi}_{t,N}$ to $\tilde{\pi}_{t,\infty}$ required by the fact that π is diffusion regular. □

Proof of Lemma 2.4. We prove Lemma 2.4 by induction on t. When $t = 1$, $\tilde{Z}_1^N = 0$ implying $\tilde{Z}_1^\infty = 0$. Then, according to Lemma A.1, we know there exists a constant vector $\tilde{X}_1^\infty$ s.t. $\tilde{X}_1^N \to \tilde{X}_1^\infty$. Thus, Lemma 2.4 holds true for $t = 1$.

Now assume Lemma 2.4 holds for t and we will prove it holds for $t + 1$. It is sufficient to prove there exists a sub-Gaussian random vector $\tilde{Z}_{t+1}^\infty$ s.t. $\tilde{Z}_{t+1}^N \to$

$\tilde{Z}^{\infty}_{t+1}$ in distribution. This is because (1) existence of the limit $\tilde{X}^{\infty}_{t+1}$ follows from Lemma A.1 and (2) showing $\tilde{Z}^{\infty}_{t+1}$ is sub-Gaussian implies $\tilde{X}^{\infty}_{t+1}$ is sub-Gaussian because

$$|\tilde{X}^{\infty}_{t+1}| = |\tilde{\pi}_{t,\infty}(\tilde{Z}^{\infty}_{t+1})| \leq |\tilde{\pi}_{t,\infty}(\tilde{Z}^{\infty}_{t+1}) - \tilde{\pi}_{t,\infty}(0)| + |\tilde{\pi}_{t,\infty}(0)| \leq C_1|\tilde{Z}^{\infty}_{t+1}| + C_2.$$

We prove the existence of $\tilde{Z}^{\infty}_{t+1}$ by constructing an explicit formula for this limit,

$$\tilde{Z}^{N}_{t+1} \to \tilde{Z}^{\infty}_{t+1} := M_t + \sum_{s' \in S} \sum_{a \in A} p(s', a, \cdot)\tilde{X}^{\infty}_t(s', a) \tag{A.8}$$

where $M_t \sim N(0, \Sigma_t)$ is independent of $\tilde{X}^{\infty}_t(s', a)$. The covariance matrix Σ_t is defined as

$$\Sigma_t(s'', s''') = \sum_{s'} \sum_{a \in A} x_t(s', a)\mathrm{Cov}[1(s_{t+1,1} = s''), 1(s_{t+1,1} = s''')|s_{t,1} = s', a_{t,1} = a]$$

where $\mathrm{Cov}[1(s_{t+1,1} = s''), 1(s_{t+1,1} = s''')|s_t = s', a_t = a]$ is the conditional covariance of the indicators of events $\{s_{t+1,1} = s''\}$ and $\{s_{t+1,1} = s'''\}$ given $s_{t,1} = s', a_{t,1} = a$.

Once (A.8) is shown, then the fact that $\tilde{Z}^{\infty}_{t+1}$ is sub-Gaussian follows because M_t and $\tilde{X}^{\infty}_t$ are both sub-Gaussian.

To prove (A.8), by our system dynamics (A.6) with a vector form,

$$Z^{N}_{t+1} = \sum_{s' \in S, a \in A} B^{N}_t(s', a), \tag{A.9}$$

where the $B^{N}_t(s', a)$ are conditionally independent (across s' and a) multinomial distributions with parameters $X^{N}_t(s', a)$ and $p(s', a) := (p(s', a, s))_{s \in S}$, i.e.,

$$B^{N}_t(s', a) \mid X^{N}_t \sim \mathrm{Multinomial}(X^{N}_t(s', a), p(s', a)).$$

$B^{N}_t(s', a)$ is a vector counting the number of arms in each state, among those arms that were previously in state s' and for which we used action a.

Recall that $X_t^N(s',a)$ can be decomposed as $Nx_t(s',a) + \sqrt{N}\tilde{X}_t^N(s',a)$. According to Lemma A.2, there exists two random variables $C_t^N(s',a)$ and $\Delta_t^N(s',a)$, such that

$$B_t^N(s',a) = C_t^N(s',a) + \Delta_t^N(s',a), \tag{A.10}$$

and that, conditionally on $X_t^N(s',a)$, have marginal distributions:

$$C_t^N(s',a) \mid X_t^N \sim \text{Multinomial}(Nx_t(s',a), p(s',a)),$$
$$\Delta_t^N(s',a) \mid X_t^N \sim \text{sgn}(\tilde{X}_t^N(s',a))\text{Multinomial}(\sqrt{N}\left|\tilde{X}_t^N(s',a)\right|, p(s',a)).$$

By (A.9), (A.10) and the definition of our diffusion statistic $\tilde{Z}_{t+1}^N$ in terms of Z_{t+1}^N,

$$\tilde{Z}_{t+1}^N = \frac{1}{\sqrt{N}} \sum_{s'\in S, a\in A} C_t^N(s',a) - x_t(s',a)p(s',a)N + \Delta_t^N(s',a).$$

By Lemma A.3,

$$\frac{1}{\sqrt{N}}\Delta_t^N(s',a) - p(s',a)\tilde{X}_t^N(s',a) \to 0.$$

Thus,

$$\begin{aligned}\tilde{Z}_{t+1}^N &= \left[\frac{1}{\sqrt{N}} \sum_{s'\in S, a\in A} C_t^N(s',a) - x_t(s',a)p(s',a)N\right] + \left[\sum_{s'\in S, a\in A} p(s',a)\tilde{X}_t^N(s',a)\right] + \epsilon_N \\ &= \left[\frac{1}{\sqrt{N}} \sum_{s'\in S, a\in A} C_t^N(s',a) - x_t(s',a)p(s',a)N\right] + \left[\sum_{s'\in S, a\in A} p(s',a)\tilde{X}_t^\infty(s',a)\right] + \epsilon_N + \epsilon_N',\end{aligned}$$

where $\epsilon_N, \epsilon_N' \to 0$.

The first term satisfies

$$\frac{1}{\sqrt{N}} \sum_{s'\in S, a\in A} C_t^N(s',a) - x_t(s',a)p(s',a)N \to N(0,\Sigma_t),$$

where Σ_t is defined above. We define M_t to be equal to this limit. This shows (A.8) as claimed. Although it is not needed for the proof, we observe that because $\tilde{X}_t^\infty$ was constructed to be equal only in distribution to $\lim_N \tilde{X}_t^N$, we are free to construct it so that it is independent of M_t.

To summarize, we have shown $\tilde{Z}_{t+1}^N \to \tilde{Z}_{t+1}^\infty$ in distribution. □

Here we give the statement and proof of Lemma A.2 and A.3.

Lemma A.2. *Let $Y \sim \text{Multinomial}(n, p)$. Then for a given non-negative integer m, there exist random vectors Y_1 and Y_2 such that $Y = Y_1 + Y_2$ and*

$$Y_1 \sim \text{Multinomial}(m, p),\ Y_2 \sim \text{sgn}(n-m)\text{Multinomial}(|n-m|, p),$$

where $\text{sgn}(\cdot)$ *is the sign function.*

Proof of Lemma A.2. There exists a sequence of i.i.d random vectors $X_i \sim \text{Multinomial}(1, p)$ s.t.

$$Y = \sum_{i=1}^{n} X_i.$$

If $n > m$, taking $Y_1 = \sum_{i=1}^{m} X_i, Y_2 = \sum_{j=m+1}^{n} X_j$ concludes the proof. If $n \le m$, taking $Y_1 = \sum_{i=1}^{m} X_i, Y_2 = -\sum_{j=n+1}^{m} X_j$ concludes the proof. □

Lemma A.3. *Consider a sequence of random variables $X_1, X_2, ..., X_N, ...$ converging to X_∞ in distribution and a sequence of i.i.d Bernoulli random variable $B_1, B_2, ..., B_n, ...$ with $\mathbb{E}[B_1] = p$ that are also independent of sequence $X_1, X_2, ...$ Then define*

$$Y_N = \frac{1}{\sqrt{N}} \sum_{n=1}^{X_N \sqrt{N}} (B_n - p).$$

Then $Y_N \to 0$.

Proof of Lemma A.3. We calculate the characteristic function of Y_N as follows:

$$\begin{aligned}
\mathbb{E}[\exp(i\lambda Y_N)] &= \mathbb{E}[\mathbb{E}[\exp(i\lambda Y_N)|X_N]] \\
&= \mathbb{E}[\mathbb{E}[\exp(i\lambda \sum_{n=1}^{X_N\sqrt{N}} \frac{1}{\sqrt{N}}(B_n - p))|X_N]] \\
&= \mathbb{E}[\mathbb{E}[\exp(i\lambda \frac{B_1 - p}{\sqrt{N}})|X_N]^{X_N\sqrt{N}}] \\
&= \mathbb{E}[(p\exp(i\lambda\frac{1-p}{\sqrt{N}}) + (1-p)\exp(-i\lambda\frac{p}{\sqrt{N}}))^{X_N\sqrt{N}}].
\end{aligned}$$

We have

$$\begin{aligned}
&(p\exp(i\lambda\frac{1-p}{\sqrt{N}}) + (1-p)\exp(-i\lambda\frac{p}{\sqrt{N}}))^{X_N\sqrt{N}} \\
&= (p(1 + i\lambda\frac{1-p}{\sqrt{N}} + O(\frac{1}{N})) + (1-p)(1 - i\lambda\frac{p}{\sqrt{N}} + O(\frac{1}{N})))^{X_N\sqrt{N}} \\
&= (1 + O(\frac{1}{N}))^{X_N\sqrt{N}} \to 1, \text{ as } N \to \infty.
\end{aligned}$$

We would like to then argue that this almost sure convergence implies convergence of the expectations as well, i.e., that $\mathbb{E}[\exp(i\lambda Y_N)]$ converges to 1. To show this we use the dominated convergence theorem and the following bound:

$$|p\exp(i\lambda\frac{1-p}{\sqrt{N}}) + (1-p)\exp(-i\lambda\frac{p}{\sqrt{N}})| \le p|\exp(i\lambda\frac{1-p}{\sqrt{N}})| + (1-p)|\exp(-i\lambda\frac{p}{\sqrt{N}})| = p + (1-p) = 1.$$

Thus $\mathbb{E}[\exp(i\lambda Y_N)] \to 1$, which implies $Y_N \to 0$. □

A.6 Proof of Lemma 2.5

We only need to prove there exists a constant C s.t. $\mathbb{E}[||\tilde{Z}_t^N||_2^2] \le C$ for all $t \in [T]$ and N. The claim in the lemma for $\tilde{X}_t^N$ follows directly from diffusion regularity. Because

$$|\tilde{X}_t^N| = |\tilde{\pi}_{t,N}(\tilde{Z}_t^N)| \le |\tilde{\pi}_{t,N}(\tilde{Z}_t^N) - \tilde{\pi}_{t,N}(0)| + |\tilde{\pi}_{t,N}(0)| \le C_1|\tilde{Z}_t^N| + C_2,$$

we have

$$||\tilde{X}_t^N||_2^2 \le |S||\tilde{X}_t^N|^2 \le |S||C_1|\tilde{Z}_t^N| + C_2|^2 \le 2|S|C_1^2|\tilde{Z}_t^N|^2 + 2|S|C_2^2 \le 2|S|^2C_1^2||\tilde{Z}_t^N||_2^2 + 2|S|C_2^2.$$

By taking the expectation,

$$\mathbb{E}||\tilde{X}_t^N||_2^2 \le 2|S|^2C_1^2\,\mathbb{E}||\tilde{Z}_t^N||_2^2 + 2|S|C_2^2. \tag{A.11}$$

Similar to the analysis in the proof of Lemma 2.4,

$$\tilde{Z}_{t+1}^N = \frac{1}{\sqrt{N}} \sum_{s'\in S, a\in A} C_t^N(s',a) - x_t(s',a)p(s',a)N + \Delta_t^N(s',a),$$

where

$$C_t^N(s',a) \mid X_t^N \sim \text{Multinomial}(Nx_t(s',a), p(s',a)),$$
$$\Delta_t^N(s',a) \mid X_t^N \sim \text{sgn}(\tilde{X}_t^N(s',a))\text{Multinomial}(\sqrt{N}\left|\tilde{X}_t^N(s',a)\right|, p(s',a)).$$

Thus,

$$\begin{aligned}||\tilde{Z}_{t+1}^N||_2^2 &= \left|\left|\frac{1}{\sqrt{N}} \sum_{s'\in S, a\in A} C_t^N(s',a) - x_t(s',a)p(s',a)N + \Delta_t^N(s',a)\right|\right|_2^2 \\ &\le 2\left|\left|\frac{1}{\sqrt{N}} \sum_{s'\in S, a\in A} C_t^N(s',a) - x_t(s',a)p(s',a)N\right|\right|_2^2 + 2\left|\left|\frac{1}{\sqrt{N}} \sum_{s'\in S, a\in S} \Delta_t^N(s',a)\right|\right|_2^2.\end{aligned}$$

Notice from its definition as a multinomial random variable that the absolute value of each component of $\frac{1}{\sqrt{N}}\Delta_t^N(s',a)$ is bounded above by $|\tilde{X}_t^N(s',a)|$. Thus

$$\left|\left|\frac{1}{\sqrt{N}} \sum_{s'\in S, a\in S} \Delta_t^N(s',a)\right|\right|_2^2 \le |S|\left[\sum_{s'\in S, a\in S} \left|\tilde{X}_t^N(s',a)\right|\right]^2 \le 2|S|^2 \sum_{s'\in S, a\in S} \left|\tilde{X}_t^N(s',a)\right|^2 = 2|S|^2||\tilde{X}_t^N||_2^2. \tag{A.12}$$

On the other hand, noting that $C_t^N(s',a) - x_t(s',a)p(s',a)N$ has mean 0 and is

independent across different s, a to get the first equality, we have

$$\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{s'\in S,a\in A} C_t^N(s',a) - x_t(s',a)p(s',a)N\right\|_2^2 = \frac{1}{N}\sum_{s'\in S,a\in A}\mathbb{E}\left\|C_t^N(s',a) - x_t(s',a)p(s',a)N\right\|_2^2 \tag{A.13}$$

$$= \sum_{s'\in S,a\in A} x_t(s',a)(1-\sum_{s\in S} p(s',a,s)^2). \tag{A.14}$$

Combining inequality (A.12) and (A.13) together,

$$\mathbb{E}||\tilde{Z}_{t+1}^N||_2^2 \le 2C_4 + 4|S|^2\mathbb{E}||\tilde{X}_t^N||_2^2. \tag{A.15}$$

where $C_4 := \sum_{s'\in S,a\in A} x_t(s',a)(1-\sum_{s\in S} p(s',a,s)^2)$.

Recall inequality (A.11) and combine it with (A.15) to obtain,

$$\mathbb{E}||\tilde{Z}_{t+1}^N||_2^2 \le 2C_4 + 8|S|^3C_2^2 + 8|S|^4C_2^2C_1^2\mathbb{E}||\tilde{Z}_t^N||_2^2.$$

By $\tilde{Z}_1^N = 0$ and induction, there exists a constant C s.t. $\mathbb{E}[||\tilde{Z}_t^N||_2^2] \le C$ for all $t \in [T]$ and N. □

A.7 Proof of Theorem 2.3

Given a fluid-priority policy π, we directly check whether the induced map $\tilde{\pi}_{t,N}$ satisfies all three conditions in Definition 2.2.

Verification of Condition 1 Write the induced map $\tilde{\pi}_{t,N}$ as a collection of maps, $(\tilde{\pi}_{t,N}^1, ..., \tilde{\pi}_{t,N}^{|S|})$, one giving each component. That is, $\tilde{\pi}_{t,N}(\theta)$ is the vector comprised of $(\tilde{\pi}_{t,N}^i(\theta) : 1 \le i \le |S|)$.

A direct calculation shows each component function, $\tilde{\pi}^i_{t,N}$ $(1 \le i \le |S|)$, is continuous, piecewise linear, and has bounded gradients when they exist. Mathematically speaking, there exists a constant $\tilde{C}_1$, s.t., for any θ, any t and any N,

$$|\nabla_\theta \tilde{\pi}^i_{t,N}(\theta)| \le \tilde{C}_1, \text{ when } \nabla_\theta \tilde{\pi}^i_{t,N}(\theta) \text{ exists.}$$

For any θ_1 and θ_2, there exists a sequence $(\nu^0, \nu^1, ..., \nu^m)$ lying on the line segment between θ_1 and θ_2, s.t.

1. $\tilde{\pi}^i_{t,N}$ restricted on the line segment between ν^j and ν^{j+1} is linear for $j = 0, 1, ..., m-1$

2. $\nu^0 = \theta_1$ and $\nu^m = \theta_2$.

Thus

$$|\tilde{\pi}^i_{t,N}(\theta_1) - \tilde{\pi}^i_{t,N}(\theta_2)| \le \sum_{j=0}^{m-1} |\tilde{\pi}^i_{t,N}(\nu^j) - \tilde{\pi}^i_{t,N}(\nu^{j+1})| \le \sum_{j=0}^{m-1} \tilde{C}_1 |\nu^j - \nu^{j+1}| = \tilde{C}_1 |\theta_1 - \theta_2|.$$

So by taking $C_1 = |S|\tilde{C}_1$,

$$|\tilde{\pi}_{t,N}(\theta_1) - \tilde{\pi}_{t,N}(\theta_2)| \le \sum_{i=1}^{|S|} |\tilde{\pi}^i_{t,N}(\theta_1) - \tilde{\pi}^i_{t,N}(\theta_2)| \le \sum_{i=1}^{|S|} \tilde{C}_1 |\theta_1 - \theta_2| = C_1 |\theta_1 - \theta_2|.$$

Verification of Condition 2 Direct calculation shows $\tilde{\pi}_{t,N}(0) = 0$.

Verification of Condition 3 Direct calculation shows $\tilde{\pi}_{t,\infty}(\tilde{Z}_{t,\infty})$ is a linear mapping. The form of this linear mapping differs across the following three cases. We state the results of detailed calculations here providing these linear forms without including the (tedious) calculations themselves.

Case 1. $C^0_t \cup C^-_t = \varnothing$:

$$\tilde{\pi}_{t,\infty}(\tilde{Z}_{t,\infty})(s, 1) = \tilde{Z}_{t,\infty}(s), \text{ for each } s \in S.$$

Case 2. $C_t^0 \neq \varnothing$:

$\tilde{\pi}_{t,\infty}(\tilde{Z}_{t,\infty})(s,1) = \tilde{Z}_{t,\infty}(s)$, for each $s \in C_t^+$;

$\tilde{\pi}_{t,\infty}(\tilde{Z}_{t,\infty})(s,1) = -\sum_{s' \in C_t^+} \tilde{Z}_{t,\infty}(s')$, for the state $s \in C_t^0$ with highest priority-score in C_t^0;

$\tilde{\pi}_{t,\infty}(\tilde{Z}_{t,\infty})(s,1) = 0$, otherwise.

Case 3. $C_t^0 = \varnothing$, $C_t^- \neq \varnothing$:

$\tilde{\pi}_{t,\infty}(\tilde{Z}_{t,\infty})(s,1) = \tilde{Z}_{t,\infty}(s)$, for each $s \in C_t^+$;

$\tilde{\pi}_{t,\infty}(\tilde{Z}_{t,\infty})(s,1) = -\sum_{s' \in C_t^+} \tilde{Z}_{t,\infty}(s')$, for the state $s \in C_t^-$ with highest priority-score in C_t^-;

$\tilde{\pi}_{t,\infty}(\tilde{Z}_{t,\infty})(s,1) = 0$, otherwise.

To summarize, we prove the induced map of any fluid-priority policy satisfies all three conditions in Definition 2.2 and thus any fluid-priority policy is diffusion regular.

A.8 Proof of Lemma 2.6

Direct comparison of Algorithm 1 and Algorithm 2 justifies Lemma 2.6.

A.9 Proof of Lemma 2.7

Before we prove Lemma 2.7, we prove the following preliminary lemma.

Lemma A.4. *Suppose the non-degeneracy condition holds. Then there exists constants*

$\delta > 0$ and $C > 0$ s.t., $\forall \epsilon > 0, t \in [T]$, we have

$$\mathbb{P}_{\pi_R}\left[|\tilde{Z}_t^N| \geq \epsilon\sqrt{N}\right] \leq C\exp(-N\delta\epsilon^2).$$

Proof of Lemma A.4. We will show that, for $0 \leq t \leq T-1$,

$$\mathbb{P}_{\pi_R}\left[|\tilde{Z}_{t+1}^N| \geq \epsilon\sqrt{N}\right] \leq 4|S|^2\exp(-\frac{\epsilon^2 N}{8|S|^4}) + 2|S|^2\mathbb{P}_{\pi_R}\left[|\tilde{Z}_t^N| \geq \frac{\epsilon\sqrt{N}}{4|S|^2}\right]. \tag{A.16}$$

The above inequality and the observation $\tilde{Z}_1^N = 0$ would complete Lemma A.4. So in the remainder of this proof, we show inequality (A.16) holds true.

By a union bound and the fact that $|\tilde{Z}_{t+1}^N| \geq \epsilon\sqrt{N}$ implies $|\tilde{Z}_{t+1}^N(s)| \geq \epsilon\sqrt{N}/|S|$ for at least one s,

$$\mathbb{P}_{\pi_R}\left[|\tilde{Z}_{t+1}^N| \geq \epsilon\sqrt{N}\right] \leq \sum_{s\in S}\mathbb{P}_{\pi_R}\left[|\tilde{Z}_{t+1}^N(s)| \geq \frac{\epsilon\sqrt{N}}{|S|}\right].$$

Thus, we only need to show, for any $s \in S$,

$$\mathbb{P}_{\pi_R}\left[|\tilde{Z}_{t+1}^N(s)| \geq \frac{\epsilon\sqrt{N}}{|S|}\right] \leq 4|S|\exp(-\frac{\epsilon^2 N}{8|S|^4}) + 2|S|\mathbb{P}_{\pi_R}(|\tilde{Z}_t^N| \geq \frac{\epsilon\sqrt{N}}{4|S|^2}). \tag{A.17}$$

Following a similar approach to the proof of Lemma 2.4, we first write our system dynamics in a vector form:

$$Z_{t+1}^N = \sum_{s'\in S, a\in A} B_t^N(s', a), \tag{A.18}$$

where the $B_t^N(s', a)$ are conditionally independent (across s' and a) multinomial distributions with parameters $X_t^N(s', a)$ and $p(s', a) := (p(s', a, s))_{s\in S}$, i.e.,

$$B_t^N(s', a) \mid X_t^N \sim \text{Multinomial}(X_t^N(s', a), p(s', a)).$$

$B_t^N(s', a)$ is a vector counting the number of arms in each state, among those arms that were previously in state s' and for which we used action a. We use $B_t^N(s', a, s)$ to denote component s of $B_t^N(s', a)$, i.e. the number of arms that were

previously in state s', for which we used action a, and which transitioned to state s.

Recall the definition of our diffusion statistic $\tilde{Z}_{t+1}^N$ and combine it with equation (A.18),

$$\tilde{Z}_{t+1}^N(s) = \left[\sum_{s'\in S, a\in A} \frac{1}{\sqrt{N}}(B_t^N(s',a,s) - p(s',a,s)X_t^N(s',a))\right] + \left[\sum_{s'\in S,a\in A} p(s',a,s)\tilde{X}_t^N(s',a)\right].$$

So we have

$$\begin{aligned}\mathbb{P}_{\pi_R}\left[|\tilde{Z}_{t+1}^N(s)| \geq \frac{\epsilon\sqrt{N}}{|S|}\right] \leq \mathbb{P}_{\pi_R}\left[\left|\sum_{s'\in S,a\in A} \frac{1}{\sqrt{N}}(B_t^N(s',a,s) - p(s',a,s)X_t^N(s',a))\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right] \\ + \mathbb{P}_{\pi_R}\left[\left|\sum_{s'\in S,a\in A} p(s',a,s)\tilde{X}_t^N(s',a)\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right].\end{aligned}$$

Notice

$$\begin{aligned}&\mathbb{P}_{\pi_R}\left[\left|\sum_{s'\in S,a\in A} \frac{1}{\sqrt{N}}(B_t^N(s',a,s) - p(s',a,s)X_t^N(s',a))\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right] \\ &\leq \sum_{s'\in S,a\in A} \mathbb{P}_{\pi_R}\left[\frac{1}{\sqrt{N}}\left|B_t^N(s',a,s) - p(s',a,s)X_t^N(s',a)\right| \geq \frac{\epsilon\sqrt{N}}{4|S|^2}\right].\end{aligned}$$

By Hoeffding's inequality, we have

$$\mathbb{P}_{\pi_R}\left[\frac{1}{\sqrt{N}}\left|B_t^N(s',a,s) - p(s',a,s)X_t^N(s',a)\right| \geq \frac{\epsilon\sqrt{N}}{4|S|^2}\right] \leq 2\exp\left(-\frac{\epsilon^2N^2}{8|S|^4|X_t^N(s',a)|}\right) \leq 2\exp\left(-\frac{\epsilon^2N}{8|S|^4}\right).$$

Combining the above inequalities together,

$$\mathbb{P}_{\pi_R}\left[\left|\sum_{s'\in S,a\in A} \frac{1}{\sqrt{N}}(B_t^N(s',a,s) - p(s',a,s)X_t^N(s',a))\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right] \leq 4|S|\exp(-\frac{\epsilon^2N}{8|S|^4}). \tag{A.19}$$

On the other hand, combining the bound

$$\left|\sum_{s'\in S,a\in A} p(s',a,s)\tilde{X}_t^N(s',a)\right| \leq \sum_{s'\in S,a\in A} |\tilde{X}_t^N(s',a)|$$

with a union bound, we have

$$\mathbb{P}_{\pi_R}\left[\left|\sum_{s'\in S,a\in A} p(s',a,s)\tilde{X}_t^N(s',a)\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right] \leq \sum_{s'\in S,a\in A} \mathbb{P}_{\pi_R}\left[\left|\tilde{X}_t^N(s',a)\right| \geq \frac{\epsilon\sqrt{N}}{4|S|^2}\right].$$

Analysis similar to Condition 3 in §A.7 shows that $\left|\tilde{X}_t^N(s',a)\right| \leq |\tilde{Z}_t^N|$ for any $s \in S$. Thus,

$$\mathbb{P}_{\pi_R}\left[\left|\sum_{s'\in S,a\in A} p(s',a,s)\tilde{X}_t^N(s',a)\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right] \leq 2|S|\mathbb{P}_{\pi_R}\left[|\tilde{Z}_t^N| \geq \frac{\epsilon\sqrt{N}}{4|S|^2}\right]. \tag{A.20}$$

Combining inequality (A.19) and (A.20) together implies inequality (A.17), concluding the proof. □

Now we can prove Lemma 2.7.

Proof of Lemma 2.7. Let $\Omega_t := \Delta_1 \cap \Delta_2 \cap \ldots \cap \Delta_t$ and let Ω_t^c denote its complement. First we notice,

$$\begin{aligned}
\mathbb{P}_{\pi_F}(\Delta_{t+1}^c) &= \mathbb{P}_{\pi_F}(\Omega_t \cap \Delta_{t+1}^c) + \mathbb{P}_{\pi_F}(\Omega_t^c \cap \Delta_{t+1}^c) \\
&= \mathbb{P}_{\pi_R}(\Omega_t \cap \Delta_{t+1}^c) + \mathbb{P}_{\pi_F}(\Omega_t^c \cap \Delta_{t+1}^c) \\
&\leq \mathbb{P}_{\pi_R}(\Delta_{t+1}^c) + \mathbb{P}_{\pi_F}(\Omega_t^c) \\
&\leq \mathbb{P}_{\pi_R}(\Delta_{t+1}^c) + \sum_{k=1}^{t} \mathbb{P}_{\pi_F}(\Delta_k^c).
\end{aligned}$$

We will use this recursive expression show that $\mathbb{P}_{\pi_F}(\Delta_t^c) \leq L\exp(-\delta N)$ by induction on t. The base case, $t = 1$, follows immediately from $\mathbb{P}_{\pi_R}(\Delta_1^c) = 0$. Thus, it is sufficient to prove there exists constants $\delta > 0$ and L, s.t. for all t,

$$\mathbb{P}_{\pi_R}(\Delta_t^c) \leq L\exp(-\delta N). \tag{A.21}$$

We rewrite Δ_t^c in terms of $\tilde{Z}_t^N$, by first noting that there are two ways to have a budget violation event Δ_t^c. The first arises when the number of arms available to

pull in fluid-active and fluid-neutral states, $\sum_{s\in C_t^0\cup C_t^+} Z_t^N(s)$, falls below the number of arms that the optimal occupation measure plans to pull $N\sum_{s\in C_t^0\cup C_t^+} x_t(s,1)$, where we note that the optimal occupation measure never pulls arms in C_t^-. We define our diffusion statistics $\tilde{Z}_t^N$ by subtracting $(x_t(s,0)+x_t(s,1))N$ from Z_t^N and dividing the difference by $\sqrt{N}$, and so the following conditions are all equivalent:

$$\begin{aligned} N\sum_{s\in C_t^0\cup C_t^+} x_t(s,1) &> \sum_{s\in C_t^0\cup C_t^+} Z_t^N(s) \\ -N\sum_{s\in C_t^0\cup C_t^+} x_t(s,0) &> \sum_{s\in C_t^0\cup C_t^+} Z_t^N(s) - N(x_t(s,0)+x_t(s,1)), \\ -\sqrt{N}\sum_{s\in C_t^0\cup C_t^+} x_t(s,0) &> \sum_{s\in C_t^0\cup C_t^+} \tilde{Z}_t^N(s). \end{aligned}$$

Moreover, optimal occupation measures set $x_t(s,0) = 0$ for $s \in C_t^+$. Thus, the conditions above are equivalent to

$$-\sqrt{N}\sum_{s\in C_t^0} x_t(s,0) > \sum_{s\in C_t^0\cup C_t^+} \tilde{Z}_t^N(s).$$

The other way in which we can have a budget violation is to have the number of arms available to idle in fluid-inactive and fluid-neutral states fall below the number of arms that the optimal occupation measure plans to idle, $N\sum_{s\in C_t^0\cup C_t^-} x_t(s,0)$. By a similar sequence of computations, this occurs if and only if

$$\sum_{s\in C_t^+} \tilde{Z}_t^N(s) > \sqrt{N}\sum_{s\in C_t^0} x_t(s,1)$$

Thus,

$$\Delta_t^c = \Big\{ -\sqrt{N}\sum_{s\in C_t^0} x_t(s,0) > \sum_{s\in C_t^0\cup C_t^+} \tilde{Z}_t^N(s)\Big\} \bigcup \Big\{ \sum_{s\in C_t^+} \tilde{Z}_t^N(s) > \sqrt{N}\sum_{s\in C_t^0} x_t(s,1)\Big\}.$$

Thus we have,

$$\mathbb{P}_{\pi_R}(\Delta_t^c) \le \mathbb{P}_{\pi_R}\left[-\sqrt{N}\sum_{s\in C_t^0} x_t(s,0) > \sum_{s\in C_t^0\cup C_t^+}\tilde{Z}_t^N(s_t)\right] + \mathbb{P}_{\pi_R}\left[\sum_{s\in C_t^+}\tilde{Z}_t^N(s) > \sqrt{N}\sum_{s\in C_t^0} x_t(s,1)\right]$$
$$\le \sum_{s\in C_t^0\cup C_t^+}\mathbb{P}_{\pi_R}\left[|\tilde{Z}_t^N(s)| > \sqrt{N}\frac{\sum_{s\in C_t^0} x_t(s,0)}{|S|}\right] + \sum_{s\in C_t^0}\mathbb{P}_{\pi_R}\left[|\tilde{Z}_t^N(s)| > \sqrt{N}\frac{\sum_{s\in C_t^0} x_t(s,1)}{|S|}\right]$$

Using Lemma A.4, it is easy to see inequality (A.21) holds. □

A.10 Proof of Lemma 2.8

By the Fenchel Duality Theorem [47], there exists $\lambda_{1:T}^* = (\lambda_1^*, \lambda_2*, ..., \lambda_T^*)$ s.t.

$$\hat{V}_1^* = \max_{\pi} \mathbb{E}_{\pi}\sum_{t=1}^{T} r_t(s_{t,1}, a_{t,1}) + \lambda_t^*(\alpha_t - a_{t,1}) \tag{A.22}$$

where here the maximum is taken over all policies, not just those satisfying the budget constraint $\mathbb{E}_{\pi}|a_{t,1}| = \alpha_t$.

Following a dynamic programming argument, we define the value function V_t on states and the Q-factor Q_t on state-action pairs recursively as

$$Q_t(s,a) = r_t(s,a) - \lambda_t^* a + \sum_{s'\in S} p_t(s,1,s')V_{t+1}(s'),$$
$$V_t(s) = \max\{Q_t(s,0), Q_t(s,1)\}.$$

for $0 \le t \le T$ with $V_{T+1}(s) = 0$ for all $s \in S$. Thus, we can classify states into three disjoint sets

$$\text{Must-Pull}_t = \{s \in S | Q_t(s,0) < Q\},$$
$$\text{Indifferent}_t = \{s \in S | Q_t(s,0) = Q_t(s,1)\},$$
$$\text{Never-Pull}_t = \{s \in S | Q_t(s,0) > Q_t(s,1)\}.$$

A policy is optimal for (A.22) if and only if it satisfies these two conditions for each t:

- It pulls all arms whose states are in Must-Pull$_t$.
- It never pulls any arms whose states are in Never-Pull$_t$.

It can behave arbitrarily for arms whose states are in Indifferent$_t$.

Any optimal occupation measure $(x_t(s,a))_{t\in[T],s\in S,a\in A}$ achieves $\hat{V}_1^*$ and so must correspond to an optimal policy. Thus

$$\text{Must-Pull}_t \subseteq C_t^+, \ \text{Never-Pull}_t \subseteq C_t^-.$$

Any budget-relaxed fluid-priority policy π_R pulls an arm whenever its state is in C_t^+ and lets an arm idle whenever its state is in C_t^-. Thus, it is optimal for (A.22) and

$$V_N(\pi_R) + \sum_{t=1}^{T} \lambda_t^* (\alpha_t N - \mathbb{E}_{\pi_R}[|\mathbf{a}_t|]) = \hat{V}_N^*.$$

Using the fact that $|\alpha_t N - \mathbb{E}_{\pi_R}[\mathbf{a}_t]|$ is bounded above by the probability of a budget violation event times a bound N on the maximum size of a budget violation, as well as Lemma 2.7,

$$|\alpha_t N - \mathbb{E}_{\pi_R}[|\mathbf{a}_t|]| \leq N\mathbb{P}_{\pi_R}(\Delta_t^c) \leq NL\exp(-\delta N) \leq m, \ \forall t,$$

where m is a constant not depending on N.

Thus,

$$|V_N(\pi_R) - \hat{V}_N^*| = |\sum_{t=1}^{T} \lambda_t^* (\alpha_t N - \mathbb{E}_{\pi_R}[|\mathbf{a}_t|])| \leq \sum_{t=1}^{T} |\lambda_t^*||\alpha_t N - \mathbb{E}_{\pi_R}[|\mathbf{a}_t|])| \leq m \sum_{t=1}^{T} |\lambda_t^*|.$$

A.11 Proof for Proposition 2.1

For any index policy π_I, by the strong law of large numbers, $X_t^N(s,a)/N$ converges as $N \to \infty$ to a quantity that we denote $x_{I,t}(s,a)$ and refer to as the occupation measure of the index policy.

We argue that $x_{I,t}(s,a)$ has at most one state s for each t satisfying both $x_{I,t}(s,0) > 0$ and $x_{I,t}(s,1) > 0$. To see this, first recall that index policies use a strict priority order over states, pulling all arms in states higher in the priority order before pulling any arms in lower states. Then define for each state s and time t the following quantities:

- Let $P(s)$ denote the set of states that have equal or higher priority to s according to the index policy.
- Let $L_t^N(s)$ denote the number of arms whose states have equal or higher priority than s and that are not pulled.
- Let $M_t^N(s)$ denote the number of arms pulled whose states have priority strictly lower than s.

By the mechanics of an index policy's decisions, we either have $L_t^N(s) = 0$, i.e., we pull all of the arms whose states have equal or higher priority to s, or $M_t^N(s) = 0$, i.e., we pull no arms whose states have priority strictly lower than s

Then, taking the limit as $N \to \infty$ and using the strong law of large numbers, we have

$$0 = \lim_{N\to\infty} \frac{L_t^N(s) M_t^N(s)}{N^2} = \left(\sum_{s' \in P(s)} x_{I,t}(s',0) \right) \left(\sum_{s' \notin P(s)} x_{I,t}(s',1) \right),$$

This then implies that there is a unique s such that $x_{I,t}(s',0) = 1$ for $s' \in P(s) \setminus \{s\}$ and $x_{I,t}(s',1) = 0$ for all $s' \notin P(s)$. That is, states that have strictly higher priority than s are always pulled in the fluid limit, while states that have strictly lower priority than s are never pulled in this limit.

Now, since any index policy meeting the condition of the proposition has $\hat{V}_N^* - V_N(\pi_I)$ bounded above by a constant, this index policy's occupation measure $x_{I,t}$ solves Problem 2.4. We then construct a fluid-priority policy to match this index policy.

First, we note that the set of fluid-active states for the optimal occupation measure $x_{I,t}$ are those with $x_{I,t}(s,0) = 0$ and that the index policy ranks these above all other states. We take the priority score used by our fluid priority policy to rank these fluid-active states among themselves in the same way as the index policy.

Second, the set of fluid-inactive states for $x_{I,t}$ are those with $x_{I,t}(s,1) = 0$. The index policy ranks these below all other states. Again, we take the priority score used by our fluid priority policy to rank these fluid-inactive states in the same way as the index policy.

Third, the at most one state with $x_{I,t}(s,0) > 0$ and $x_{I,t}(s,1) > 0$ is a fluid-neutral state, and it is ranked by the index policy below the fluid-active states and above the fluid-inactive states.

Because our fluid-priority policy's priority score matches the index policy's prioritizations on fluid-active and fluid-inactive states, and its prioritizations also across categories (fluid-active, fluid-neutral, fluid-inactive) match those of the index policy, our fluid-priority policy is the same as the index policy.

A.12 Proof for Proposition 2.2

The proof is similar for both UCB and Thompson Sampling policy. We only show the proof for UCB here.

Under the UCB policy, there exists $z_t^{UCB}(s)$ and $x_t^{UCB}(s,a)$ which are feasible for the LP (2.4) and satisfy

$$\frac{Z_t^N(s)}{N} \to z_t^{UCB}(s), \frac{X_t^N(s,a)}{N} \to x_t^{UCB}(s,a).$$

So we have

$$\frac{R_{\pi^{UCB}}(N)}{N} \to \sum_{t=1}^{T} \sum_{s\in S} \sum_{a\in A} r_t(s,a) x_t^{UCB}(s,a).$$

UCB policy is an index policy. Thus, the occupation measure $x_t^{UCB}(s,a)$ can be calculated via forward propagation. Numerically, we can verify $x_t^{UCB}(s,a)$ is not an optimal solution for LP (2.4) under $T = 15$ and $T = 20$.

A.13 Discussion of policies in previous literature

In this section, we show the power of the techniques developed in §2.4 and 2.5 by applying them to policies proposed by previous literature to demonstrate theoretical guarantees from that literature can be seen as consequences of our results. Specifically, we observe that the Randomized Assignment Control (RAC) policy proposed by [57] is fluid-consistent, thus achieving an $o(N)$ opt gap. The policy proposed by [31] and the "optimal Lagrangian index policy" proposed by [13] are diffusion-regular, thus achieving $O(\sqrt{N})$ opt gaps.

[57] achieves $o(N)$ opt gap

This section shows the RAC policy proposed by [57] achieves an $o(N)$ opt gap. To start with, let us first describe the RAC policy. Although [57] defines RAC policy in settings more general than the binary-action bandit (referring to their more general problem setting as a "multi-action bandit"), we only focus on the binary bandit here.

Similar to our approach, [57] first solves the linear programming relaxation (2.3) and then fetch an optimal occupation measure $\{x_t(s,a)\}_{t\in[T],s\in S,a\in A}$. Then, based on the occupation measure, an activation probability is defined for each state s at period t:

$$q_t(s) = \begin{cases} \frac{x_t(s,1)}{z_t(s)}, & \text{if } z_t(s) > 0; \\ \\ 0, & \text{if } z_t(s) = 0. \end{cases}$$

Then when deciding which arm to pull at period t under the RAC policy, we first randomly choose an arm that has not been chosen in this period. If the arm's state is s, then we randomly generate a Bernoulli variable with mean $q_t(s)$. If this random realization is 0 or there is no remaining budget, idle the arm; otherwise, activate the arm. Repeat this process until no budget remains in the period.

Direct computation and the strong law of large numbers show that the RAC policy is fluid consistent. Thus, it achieves an $o(N)$ opt gap.

[31] and [13] achieve $O(\sqrt{N})$ opt gaps

The policies proposed by [31] and [13] are very similar. Thus we only discuss [13]'s policy here. The analysis for [31]'s policy can be generalized without any essential difficulty.

To start with, we first describe the "optimal Lagrangian index policy" proposed by [13]. Similar to our approach, [13] first solves the linear programming relaxation (2.3) and fetches an optimal occupation measure $\{x_t(s,a)\}_{t\in[T],s\in S,a\in A}$, which is used to do "tie-breaking" discussed later. While solving the relaxed problem using the Simplex method [43], as a byproduct, its dual problem

$$\min_{\lambda_{1:T}} \max_{\pi} \mathbb{E}_{\pi} \sum_{t=1}^{T} r_t(s_t,a_t) + \lambda_t(\alpha_t N - a_t).$$

is also solved, which yields optimal Lagrange multipliers $\{\lambda_t^*\}_{t=1}^T$.

Then, following a dynamic programming argument, the value function V_t on states and the Q-factor Q_t on state-action pairs are defined as

$$Q_t(s,a) = r_t(s,a) - \lambda_t^* a + \sum_{s'\in S} p_t(s,1,s')V_{t+1}(s'),$$

$$V_t(s) = \max\{Q_t(s,0), Q_t(s,1)\}$$

for $0 \le t \le T$ with $V_{T+1}(s) = 0$ for all $s \in S$. Finally the index of a state s at period t is defined as

$$\text{Index(s)} = Q_t(s,1) - Q_t(s,0).$$

When deciding which arm to pull, arms are activated from high index to low index until no budget remains. When there is a tie, i.e., some states share the same index value, the number of arms activated from a state is proportional to its occupation measure. More details can be found in [13] Section 4.

Now we show the optimal Lagrangian index policy is diffusion regular. First of all, we can show its associated map $\hat{\pi}_{t,N}$ is a piece-wise linear map, thus satisfying Condition 1 in Definition 2.2. Second, we can show $\hat{\pi}_{t,N}(0) = 0$, thus satisfying Condition 2. As a piece-wise linear map, we can also show that $\hat{\pi}_{t,N}$ converges as $N \to \infty$. Thus, Condition 3 is satisfied.

A.14 Choice of Occupation Measure

Multiple optimal occupation measures may exist, some degenerate and others not. A fluid-priority policy constructed from a non-degenerate optimal occupation measure is guaranteed to have an $O(1)$ OG while another constructed from a degenerate one is not. We now give a computational procedure that selects a non-degenerate optimal occupation measure, if one exists.

First, observe from (2.4) that any convex combination of optimal occupation measures is also optimal. Thus, suppose we can find a collection of optimal occupation measures, $x^{*,k}$, $k \in [K]$, such that, for each t, there is either (1) a state s that is fluid-neutral under some k, or (2) there is a state s that is fluid-active under some k and fluid-inactive under another k. Then any convex combination with strictly positive weight on each k is non-degenerate. We describe an algorithm for finding such a collection, if it exists, or establishing that it does not.

To accomplish this, first solve the LP (2.4), call the solution $x^{*,1}$, and record its optimal value for later use. Assess for each t whether there is a state s satisfying $x_t^*(s,0) > 0$ and $x_t^*(s,1) > 0$. If all t satisfy this condition, then we have found a non-degenerate optimal occupation measure.

Otherwise, we will continue iteratively in our search. In each stage k, we will maintain a collection of solutions $\{x^{*,k'} : k' = 1, \ldots, k\}$ and a set of times $A_k \subseteq [T]$. A_k contains those times for which we have not yet been able to construct a fluid-neutral state. Formally, a time t is in A_k if and only one of the following holds: (1) all states are fluid-active at t in all $x^{*,k'}$, $k' \leq k$; or (2) all states are fluid-inactive at t in all $x^{*,k'}$, $k' \leq k$. If A_k is empty, then a non-degenerate optimal occupation measure can be constructed as a convex combination of $\{x^{*,k'} : k' = 1, \ldots, k\}$ using strictly positive weights on every solution in this collection. If A_k is not empty, we will then attempt to construct an optimal occupation measure that, when added to our collection of solutions, causes A_{k+1} to be a strict subset of A_k.

Toward this goal, in stage k, choose $t_k \in A_k$. This will be the time that we seek to remove from A_k in constructing A_{k+1}. Let $C^{+,k}$ contain all of the states for which $x^{*,k'}_{t_k}(s,1) > 0$ and $x^{*,k'}_{t_k}(s,0) = 0$ for all $k' \leq k$. These are the states that are fluid-active at time t_k for all previously computed optimal occupation measures. Then solve a linear program minimizing $\sum_{s \in C^{+,k}} x_{t_k}(s,1)$ subject to all of the constraints in (2.4) and the linear constraint that the objective in (2.4) is equal to its optimal value recorded above. Call the solution $x^{*,k+1}$.

This linear program assesses whether there is an optimal occupation measure $x^{*,k+1}$ satisfying $\sum_{s \in C^{+,k}} x_{t_k}(s,1) < \alpha_k$. If no such $x^{*,k+1}$ exists, then this establishes that all optimal occupation measures are degenerate. Otherwise, if we find such a $x^{*,k+1}$, then we add it to our collection of solutions. We also construct A_{k+1} by removing the time t_k from A_k. We additionally remove any other times t for which the new solution $x^{*,k+1}$ provides a state whose category at that time t is different from those in the previous solutions $x^{*,k'}$, $k' \leq k$.

If A_{k+1} is the empty set, then this implies that there is a non-degenerate op-

timal occupation measure. We set $K = k + 1$ and construct it as described above from the collection $\{x^{*,k'} : k \leq K\}$.

APPENDIX B

APPENDIX: MULTI-ACTION MULTI-RESOURCE FINITE-HORIZON RESTLESS BANDIT

This section provides all technical proof in the main paper.

B.1 Proof for Lemma 3.1

The budget constraint is in sense of cardinality in original problem (3.1), while in the sense of expectation in relaxation problem (3.2). So a wider class of policy is feasible in the relaxation problem, which implies

$$V_N^* \leq \hat{V}_N^*. \tag{B.1}$$

To prove $\hat{V}_N^* = N\hat{V}_1^*$, we use Lagrangian Relaxation similar to [20, 26] as the key idea in the following analysis.

Through imitating straightforwardly the proof of Fenchel Duality Theorem [47],

$$\begin{aligned}
&\max_{\pi} \min_{\lambda(\boldsymbol{\omega}_t)\geq 0} \mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) + \left\langle \lambda(\boldsymbol{\omega}_t), b_t(\omega_t)N - \sum_{i=1}^{N} c_t(s_{i,t}, a_{i,t}, \omega_t) \right\rangle \\
=&\min_{\lambda(\boldsymbol{\omega}_t)\geq 0} \max_{\pi} \mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) + \left\langle \lambda(\boldsymbol{\omega}_t), b_t(\omega_t)N - \sum_{i=1}^{N} c_t(s_{i,t}, a_{i,t}, \omega_t) \right\rangle
\end{aligned} \tag{B.2}$$

where λ_t is chosen adaptively with respect the realization history $\boldsymbol{\omega}_t$. The let-hand side of equation (B.2) equals to $\hat{V}_N^*$. On the right hand side, for fixed adaptive mapping λ,

$$\begin{aligned}
&\mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) + \left\langle \lambda(\boldsymbol{\omega}_t), b_t(\omega_t)N - \sum_{i=1}^{N} c_t(s_{i,t}, a_{i,t}, \omega_t) \right\rangle \\
&\quad = \mathbb{E}_\pi \sum_{i=1}^{N}\sum_{t=1}^{T} r_t(s_{i,t}, a_{i,t}) + \left\langle \lambda(\boldsymbol{\omega}_t),\ b_t(\omega_t) - c_t(s_{i,t}, a_{i,t}, \omega_t) \right\rangle
\end{aligned}$$

Since all arms share the same transition kernel and reward function,

$$\begin{aligned}
&\mathbb{E}_\pi \sum_{i=1}^{N}\sum_{t=1}^{T} r_t(s_{i,t}, a_{i,t}) + \left\langle \lambda(\boldsymbol{\omega}_t),\ b_t(\omega_t) - c_t(s_{i,t}, a_{i,t}, \omega_t)\right\rangle \\
&= N\, \mathbb{E}_\pi \sum_{t=1}^{T} r_t(s_{1,t}, a_{1,t}) + \left\langle \lambda(\boldsymbol{\omega}_t),\ b_t(\omega_t) - c_t(s_{1,t}, a_{1,t}, \omega_t)\right\rangle.
\end{aligned}$$

So we conclude

$$\begin{aligned}
&\min_{\lambda(\boldsymbol{\omega}_t)\geq 0}\max_{\pi}\ \mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t) + \left\langle \lambda(\boldsymbol{\omega}_t), b_t(\omega_t)N - \sum_{i=1}^{N} c_t(s_{i,t}, a_{i,t}, \omega_t)\right\rangle \\
&= N \min_{\lambda(\boldsymbol{\omega}_t)\geq 0}\max_{\pi} \mathbb{E}_\pi \sum_{t=1}^{T} r_t(s_{1,t}, a_{1,t}) + \left\langle \lambda(\boldsymbol{\omega}_t),\ b_t(\omega_t) - c_t(s_{1,t}, a_{1,t}, \omega_t)\right\rangle. \qquad \text{(B.3)}
\end{aligned}$$

By using Fenchel Duality again on the one-arm problem,

$$\begin{aligned}
&\min_{\lambda(\boldsymbol{\omega}_t)\geq 0}\max_{\pi}\ \mathbb{E}_\pi \sum_{t=1}^{T} r_t(s_{1,t}, a_{1,t}) + \left\langle \lambda(\boldsymbol{\omega}_t),\ b_t(\omega_t) - c_t(s_{1,t}, a_{1,t}, \omega_t)\right\rangle \\
&= \max_{\pi}\min_{\lambda(\boldsymbol{\omega}_t)\geq 0}\ \mathbb{E}_\pi \sum_{t=1}^{T} r_t(s_{t,1}, a_{t,1}) + \left\langle \lambda(\boldsymbol{\omega}_t),\ b_t(\omega_t) - c_t(s_{1,t}, a_{1,t}, \omega_t)\right\rangle \\
&= \hat{V}_1^*. \qquad \text{(B.4)}
\end{aligned}$$

To summarize Equation (B.2), (B.3) and (B.4) together,

$$\hat{V}_N^* = N\ \hat{V}_1^*.$$

B.2 Proof of Theorem 3.1

First, we state and prove the following Lemma.

Lemma B.1. *If a policy π is fluid consistent, then*

$$\frac{Z_t^N(\boldsymbol{\omega}_t)}{N} \to z_t(\boldsymbol{\omega}_t), \frac{X_t^N(\boldsymbol{\omega}_t)}{N} \to x_t(\boldsymbol{\omega}_t)\ \textit{on}\ \boldsymbol{\omega}_t$$

for any $\boldsymbol{\omega}_t$.

Proof of Lemma B.1. We prove Lemma B.1 by induction on t.

When $t = 1$, by the initial condition all arms start at the same state $s^* \in S$,

$$\frac{Z_1^N(\omega_1)}{N} \to z_1(\omega_1) \text{ on event } \omega_1.$$

By definition of fluid consistency,

$$\frac{X_1^N(\omega_1)}{N} \to x_1(\omega_1) \text{ on event } \omega_1.$$

Thus Lemma B.1 holds true for $t = 1$.

Now assume Lemma B.1 holds for t, and we are going to prove Lemma B.1 holds for $t + 1$. By the definition of fluid consistency, we only need to prove that on event ω_{t+1},

$$\frac{Z_{t+1}^N(\omega_{t+1})}{N} \to z_{t+1}(\omega_{t+1}). \tag{B.5}$$

Recall the system dynamic equation on event $P_t(\omega_{t+1})$

$$Z_{t+1}^N(s, \omega_{t+1}) = \sum_{s',a} \sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s)$$

we only need to show that on $P_t(\omega_{t+1})$,

$$\frac{1}{N} \sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) \to x_t(s', a, P_t(\omega_{t+1}))p_t(s', a, s). \tag{B.6}$$

If $x_t(s', a, P_t(\omega_{t+1})) > 0$, then as $N \to +\infty$, on $P_t(\omega_{t+1})$

$$\begin{aligned}\frac{1}{N} X_t^N(s', a, P_t(\omega_{t+1})) &= \frac{1}{N} \sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a) \\ &\to x_t(s', a, P_t(\omega_{t+1})).\end{aligned}$$

So when N is large enough, $X_t^N(s', a, P_t(\omega_{t+1})) > 0$ and on $P_t(\omega_{t+1})$

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} I(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) \\
&= \frac{X_t^N(s', a, P_t(\omega_{t+1}))}{N} \frac{1}{X_t^N(s', a, P_t(\omega_{t+1}))} \sum_{i=1}^{N} I(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) \\
&\to x_t(s', a, P_t(\omega_{t+1})) p_t(s', a, s) \text{ as } N \to +\infty.
\end{aligned}$$

If $x_t(s', a, P_t(\omega_t + 1)) = 0$, then on $P_t(\omega_{t+1})$

$$\frac{1}{N}\sum_{i=1}^{N} I(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) \leq \frac{X_t^N(s', a, P_t(\omega_{t+1}))}{N} \to 0.$$

Combining the case of $x_t(s', a, P_t(\omega_{t+1})) > 0$ and $x_t(s', a, P_t(\omega_{t+1})) = 0$, equation (B.6) is proved.

To summarize, on $P_t(\omega_{t+1})$

$$\begin{aligned}
\frac{Z_{t+1}^N(s, \omega_{t+1})}{N} &= \sum_{s',a} \frac{1}{N} \sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) \\
&\to \sum_{s' \in S} \sum_{a \in A} x_t(s', a, P_t(\omega_{t+1})) p_t(s', a, s) \\
&= z_{t+1}(s, \omega_{t+1}).
\end{aligned}$$

□

Now we can prove Theorem 3.1 based on Lemma B.1. Because the policy π is fluid consistent, Lemma B.1 shows on ω_t

$$\frac{Z_t^N(\omega_t)}{N} \to z_t(\omega_t), \frac{X_t^N(\omega_t)}{N} \to x_t(\omega_t).$$

The total reward of the joint MDP over N

$$\begin{aligned}\frac{1}{N}\mathbb{E}_\pi \sum_{t=1}^{T} R_t(\mathbf{s}_t, \mathbf{a}_t, \omega_t) &= \frac{1}{N}\mathbb{E}_\pi \sum_{\boldsymbol{\omega}_t} 1(\boldsymbol{\omega}_t) \sum_{s,a} r_t(s,a,\omega_t) X_t^N(s,a,\boldsymbol{\omega}_t) \\ &= \mathbb{E}_\pi \sum_{\boldsymbol{\omega}_t} 1(\boldsymbol{\omega}_t) \sum_{s,a} r_t(s,a,\omega_t) \frac{X_t^N(s,a,\boldsymbol{\omega}_t)}{N} \\ &\to \sum_{\boldsymbol{\omega}_t} \mathbb{P}[\boldsymbol{\omega}_t] \sum_{s,a} r_t(s,a,\omega_t) x_t(s,a,\boldsymbol{\omega}_t) \\ &= \sum_{\boldsymbol{\omega}_t} \sum_{s,a} r_t(s,a,\omega_t) \mu_t(s,a,\boldsymbol{\omega}_t),\end{aligned}$$

Thus, we prove $V_N^* - V_N(\pi) = o(N)$.

B.3 Proof of Theorem 3.2

First we state and prove Lemma below, whose proof could be found in the next section.

Lemma B.2. *If a policy π is diffusion regular, then there exists a constant C, s.t. for all $t \in [T]$, N and $\boldsymbol{\omega}_t \in \Omega^t$,*

$$\mathbb{E}[|\tilde{Z}_t^N(\boldsymbol{\omega}_t)|^2] \le C,\ \mathbb{E}[|\tilde{X}_t^N(\boldsymbol{\omega}_t)|^2] \le C.$$

Based on Lemma B.2, we can see that the optimality gap

$$\begin{aligned}V_N^* - V_N(\pi) &\le N\,\mathbb{E}_\pi \sum_{\boldsymbol{\omega}_t} 1(\boldsymbol{\omega}_t) \sum_{s,a} r_t(s,a,\omega_t) x_t(s,a,\boldsymbol{\omega}_t) - \mathbb{E}_\pi \sum_{\boldsymbol{\omega}_t} 1(\boldsymbol{\omega}_t) \sum_{s,a} r_t(s,a,\omega_t) X_t^N(s,a,\boldsymbol{\omega}_t) \\ &= -\sqrt{N}\,\mathbb{E}_\pi \sum_{\boldsymbol{\omega}_t} 1(\boldsymbol{\omega}_t) \sum_{s,a} r_t(s,a,\omega_t) \tilde{X}_t^N(s,a,\boldsymbol{\omega}_t)\end{aligned}$$

Divide both sides by $\sqrt{N}$ and apply Lemma B.2,

$$\begin{aligned}\frac{V_N^* - V_N(\pi)}{\sqrt{N}} &\le \mathbb{E}_\pi \sum_{\boldsymbol{\omega}_t} 1(\boldsymbol{\omega}_t) \sum_{s,a} r_t(s,a,\omega_t) |\tilde{X}_t^N(s,a,\boldsymbol{\omega}_t)| \\ &\le \mathbb{E}_\pi \sum_{\boldsymbol{\omega}_t} 1(\boldsymbol{\omega}_t) \sum_{s,a} r_t(s,a,\omega_t)\,\sqrt{C}.\end{aligned}$$

Thus, we show $V_N^* - V_N(\pi) = O(\sqrt{N})$.

B.4 Proof of Lemma B.2

We only need to prove there exits a constant C s.t. $\mathbb{E}[|\tilde{Z}_t^N(\omega_t)|^2] \leq C$ on each ω_t for all N. The case of $\tilde{X}_t^N(\omega_t)$ follows directly from

$$|\tilde{X}_t^N(\omega_t)| = |\tilde{\pi}_{t,N}(\omega_t, \tilde{Z}_t^N(\omega_t))| \leq |\tilde{\pi}_{t,N}(\omega_t, \tilde{Z}_t^N) - \tilde{\pi}_{t,N}(\omega_t, 0)| + |\tilde{\pi}_{t,N}(\omega_t, 0)| \leq C_1|\tilde{Z}_t^N(\omega_t)| + C_2.$$

Recall the system dynamic equation on event $P_t(\omega_{t+1})$

$$Z_{t+1}^N(s, \omega_{t+1}) = \sum_{s',a} \sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s).$$

Recall definition of diffusion statistics and applying Lemma B.3, on event $P_t(\omega_{t+1})$

$$\tilde{Z}_{t+1}^N(\omega_{t+1}) = \frac{1}{\sqrt{N}} \sum_{s',a} C_t^N(s', a) - \frac{x_t(s', a, P_t(\omega_{t+1}))}{\sum_{s',a} x_t(s', a, P_t(\omega_{t+1}))} p(s', a)N + \frac{1}{\sqrt{N}} \sum_{s',a} \Delta_t^N(s', a),$$

where

$$C_t^N(s', a) \sim \text{Binomial}(\frac{x_t(s', a, P_t(\omega_{t+1}))}{\sum_{s',a} x_t(s', a, P_t(\omega_{t+1}))} N, p(s', a)),$$

$$\Delta_t^N(s', a) \sim \text{sgn}(\tilde{X}_t^N(s', a, P_t(\omega_{t+1}))) \text{ Binomial}(\sqrt{N}\left|\tilde{X}_t^N(s', a, P_t(\omega_{t+1}))\right|, p(s', a)).$$

Thus,

$$\begin{aligned}|\tilde{Z}_{t+1}^N(\omega_{t+1})|^2 &= \left| \frac{1}{\sqrt{N}} \sum_{s',a} C_t^N(s', a) - \frac{x_t(s', a, P_t(\omega_{t+1}))}{\sum_{s',a} x_t(s', a, P_t(\omega_{t+1}))} p(s', a)N + \frac{1}{\sqrt{N}} \sum_{s',a} \Delta_t^N(s', a) \right|^2 \\ &\leq 2 \left| \frac{1}{\sqrt{N}} \sum_{s',a} C_t^N(s', a) - \frac{x_t(s', a, P_t(\omega_{t+1}))}{\sum_{s',a} x_t(s', a, P_t(\omega_{t+1}))} p(s', a)N \right|^2 + 2 \left| \frac{1}{\sqrt{N}} \sum_{s',a} \Delta_t^N(s', a) \right|^2\end{aligned}$$

Notice

$$|\frac{1}{\sqrt{N}} \Delta_t^N(s', a)| \leq |\tilde{X}_t^N(s', a, P_t(\omega_{t+1}))|,$$

and

$$\begin{aligned}
&\mathbb{E}\left|\frac{1}{\sqrt{N}}\sum_{s',a}C_t^N(s',a)-\frac{x_t(s',a,P_t(\omega_{t+1}))}{\sum_{s',a}x_t(s',a,P_t(\omega_{t+1}))}p(s',a)N\right|^2\\
&=\frac{1}{N}\sum_{s',a}\mathbb{E}\left|C_t^N(s',a)-\frac{x_t(s',a,P_t(\omega_{t+1}))}{\sum_{s',a}x_t(s',a,P_t(\omega_{t+1}))}p(s',a)N\right|^2\\
&=\frac{1}{N}\sum_{s',a}\sum_{i=1}^{\frac{x_t(s',a,P_t(\omega_{t+1}))N}{\sum_{s',a}x_t(s',a,P_t(\omega_{t+1}))}}\mathbb{E}\left|1_i(s',a)-p(s',a)\right|^2\\
&=\sum_{s',a}\frac{x_t(s',a,P_t(\omega_{t+1}))}{\sum_{s',a}x_t(s',a,P_t(\omega_{t+1}))}\mathbb{E}\left|1(s',a)-p(s',a)\right|^2
\end{aligned}$$

Denoting $C_4 := \max_{\omega_{t+1}}\sum_{s',a}\frac{x_t(s',a,P_t(\omega_{t+1}))}{\sum_{s',a}x_t(s',a,P_t(\omega_{t+1}))}\mathbb{E}\left|1(s',a)-p(s',a)\right|^2$, we have

$$\mathbb{E}|\tilde{Z}_{t+1}^N(\omega_{t+1})|^2 \le 2C_4 + 2|S|\,\mathbb{E}|\tilde{X}_t^N(P_t(\omega_{t+1}))|^2 \le 2C_4 + 4|S|C_2^2 + 4|S||C_1|^2\mathbb{E}|\tilde{Z}_t^N(P_t(\omega_{t+1}))|^2.$$

To conclude, we prove there exits a constant C s.t. $\mathbb{E}[|\tilde{Z}_t^N(\omega_t)|^2] \le C$ for all N and ω_t. □

Lemma B.3. *Suppose random variable S is a Binomial random variable with parameter n and p, i.e., distributed as the sum of n i.i.d. Bernoulli r.v.s with mean p. Then for a given non-negative integer m, there exists random variable S_1 and S_2, s.t. $S = S_1 + S_2$, and*

$$S_1 \sim Binomial(m,p),\ S_2 \sim sgn(n-m)Binomial(|n-m|,p),$$

where $sgn(\cdot)$ is the sign function.

Proof. Proof of Lemma B.3 There exists a sequence of i.i.d random variables $X_i \sim Bin(1,p)$, s.t.

$$S = \sum_{i=1}^{n} X_i.$$

If $n > m$, taking $S_1 = \sum_{i=1}^{m} X_i, S_2 = \sum_{j=m+1}^{n} X_j$ concludes the proof. If $n \le m$, taking $S_1 = \sum_{i=1}^{m} X_i, S_2 = -\sum_{j=n+1}^{m} X_j$ concludes the proof. □

B.5 Proof of Theorem 3.3

Given a fluid-priority policy π, we directly check the induced map $\tilde{\pi}_{t,N}$ satisfies all three conditions in Definition 3.2.

Proof. Verification of Condition 1 Write the induced map in the component form $\tilde{\pi}_{t,N} = (\tilde{\pi}^1_{t,N}, ..., \tilde{\pi}^{|S|}_{t,N})$. Given $\boldsymbol{\omega}_t$, a direct calculation shows each component function $\tilde{\pi}^i_{t,N}$ $(1 \le i \le |S|)$ is continuous, piece-wise linear, and has bounded gradient when exits. Mathematically speaking, there exits a constant $\tilde{C}_1$, s.t., for any θ, $\boldsymbol{\omega}_t$ and N,

$$|\nabla \tilde{\pi}^i_{t,N}(\boldsymbol{\omega}_t, \theta)| \le \tilde{C}_1, \text{ when } \nabla \tilde{\pi}^i_{t,N}(\boldsymbol{\omega}_t, \theta) \text{ exits.}$$

For any θ_1 and θ_2, there exits a sequence $(\theta^0_{1,2}, \theta^1_{1,2}, ..., \theta^m_{1,2})$ lies on the line segment between θ_1 and θ_2, s.t.

1. $\tilde{\pi}^i_{t,N}$ restricted on line segment between $\theta^j_{1,2}$ and $\theta^{j+1}_{1,2}$ is linear for $j = 0, 1, ..., m-1$

2. $\theta^0_{1,2} = \theta_1$ and $\theta^m_{1,2} = \theta_2$.

Thus

$$|\tilde{\pi}^i_{t,N}(\boldsymbol{\omega}_t, \theta_1) - \tilde{\pi}^i_{t,N}(\boldsymbol{\omega}_t, \theta_2)| \le \sum_{j=0}^{m-1} |\tilde{\pi}^i_{t,N}(\boldsymbol{\omega}_t, \theta^j_{1,2}) - \tilde{\pi}^i_{t,N}(\boldsymbol{\omega}_t, \theta^{j+1}_{1,2})| \le \sum_{j=0}^{m-1} \tilde{C}_1 |\theta^j_{1,2} - \theta^{j+1}_{1,2}| = \tilde{C}_1 |\theta_1 - \theta_2|.$$

So by taking $C_1 = |S|\tilde{C}_1$,

$$|\tilde{\pi}_{t,N}(\boldsymbol{\omega}_t, \theta_1) - \tilde{\pi}_{t,N}(\boldsymbol{\omega}_t, \theta_2)| \le \sum_{i=1}^{|S|} |\tilde{\pi}^i_{t,N}(\boldsymbol{\omega}_t, \theta_1) - \tilde{\pi}^i_{t,N}(\boldsymbol{\omega}_t, \theta_2)| \le \sum_{i=1}^{|S|} \tilde{C}_1 |\theta_1 - \theta_2| = C_1 |\theta_1 - \theta_2|.$$

Proof. Verification of Condition 2 Direct calculation shows $\tilde{\pi}_{t,N}(0) = 0$.

Proof. Verification of Condition 3 Direct calculation shows $\tilde{\pi}_{t,\infty}(\tilde{Z}_{t,\infty})$ is a piecewise linear mapping.

B.6 Proof of Lemma 3.2

We can show that inside fluid-active category, if $Z_t^N(s,\omega_t)$ arms are available in state s under signal realization ω_t, at most $\lceil Z_t^N(s,\omega_t)\frac{x_t(s,a,\omega_t)}{\sum_a x_t(s,a,\omega_t)}\rceil$ arms are activated with action a. This can be verified by the definition of Weighted-Round-Robin procedure.

Then, directly comparing Algorithm 3 and Algorithm 5 justifies Lemma 3.2.

B.7 Proof of Lemma 3.3

Before we prove Lemma 3.3, we prove the following preliminary lemma.

Lemma B.4. *Suppose the non-degeneracy condition holds. Then there exists constants $\delta > 0$ and $C > 0$ s.t., $\forall \epsilon > 0$, we have*

$$\mathbb{P}_{\pi_R}\left[|\tilde{Z}_t^N(\omega_t)| \geq \epsilon\sqrt{N}\right] \leq C\exp(-N\delta\epsilon^2).$$

on event ω_t

Proof. Proof of Lemma B.4 We will show that, for $0 \leq t \leq T-1$,

$$\begin{aligned}&\mathbb{P}_{\pi_R}\left[|\tilde{Z}_{t+1}^N(\omega_{t+1})| \geq \epsilon\sqrt{N}\right]\\ &\leq 2|A||S|^2\exp(-\frac{\epsilon^2 N}{2|A|^2|S|^4}) + |A||S|^2\mathbb{P}_{\pi_R}\left[|\tilde{Z}_t^N(P_t(\omega_{t+1}))| \geq \frac{\epsilon\sqrt{N}}{2|A||S|^2}\right].\end{aligned} \tag{B.7}$$

The above inequality and the observation $\tilde{Z}_1^N = 0$ would complete Lemma B.4. So in the remainder of this proof, we show inequality (B.7) holds true.

Given ω_{t+1}, by a union bound and the fact that $|\tilde{Z}_{t+1}^N(\omega_{t+1})| \geq \epsilon\sqrt{N}$ implies $|\tilde{Z}_{t+1}^N(s, \omega_{t+1})| \geq \epsilon\sqrt{N}/|S|$ for at least one s,

$$\mathbb{P}_{\pi_R}\left[|\tilde{Z}_{t+1}^N(\omega_{t+1})| \geq \epsilon\sqrt{N}\right] \leq \sum_{s\in S} \mathbb{P}_{\pi_R}\left[|\tilde{Z}_{t+1}^N(s, \omega_{t+1})| \geq \frac{\epsilon\sqrt{N}}{|S|}\right].$$

Thus, we only need to show, for any $s \in S$,

$$\begin{aligned}
&\mathbb{P}_{\pi_R}\left[|\tilde{Z}_{t+1}^N(s, \omega_{t+1})| \geq \frac{\epsilon\sqrt{N}}{|S|}\right] \\
&\leq 2|A||S|\exp(-\frac{\epsilon^2 N}{2|A|^2|S|^4}) + |A||S|\mathbb{P}_{\pi_R}\left[|\tilde{Z}_t^N(P_t(\omega_{t+1}))| \geq \frac{\epsilon\sqrt{N}}{2|A||S|^2}\right] \qquad \text{(B.8)}
\end{aligned}$$

To prove B.8 above, we first write our system dynamics in a vector form: on event $P_t(\omega_{t+1})$,

$$Z_{t+1}^N(\omega_{t+1}) = \sum_{s',a} B_t^N(s', a), \qquad \text{(B.9)}$$

where the $B_t^N(s', a)$ are conditionally independent (across s' and a) multinomial distributions with parameters $X_t^N(s', a, P_t(\omega_{t+1}))$ and $p_t(s', a) := (p_t(s', a, s))_{s\in S}$, i.e.,

$$B_t^N(s', a) \mid X_t^N(s', a, P_t(\omega_{t+1})) \sim \text{Multinomial}(X_t^N(s', a, P_t(\omega_{t+1})), p(s', a)).$$

$B_t^N(s', a)$ is a vector counting the number of arms in each state, among those arms that were previously in state s' and for which we used action a. We use $B_t^N(s', a, s)$ to denote component s of $B_t^N(s', a)$, i.e. the number of arms that were previously in state s', for which we used action a, and which transitioned to state s.

Recall the definition of our diffusion statistic $\tilde{Z}_{t+1}^N$ and combine it with equa-

tion (B.9),

$$\tilde{Z}^N_{t+1}(s,\boldsymbol{\omega}_{t+1}) = \left[\sum_{s',a}\frac{1}{\sqrt{N}}(B^N_t(s',a,s) - p_t(s',a,s)X^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1})))\right] + \left[\sum_{s',a} p_t(s',a,s)\tilde{X}^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1}))\right]$$

on event $P_t(\boldsymbol{\omega}_{t+1})$.

So we have

$$\mathbb{P}_{\pi_R}\left[|\tilde{Z}^N_{t+1}(s,\boldsymbol{\omega}_{t+1})| \geq \frac{\epsilon\sqrt{N}}{|S|}\right] \leq \mathbb{P}_{\pi_R}\left[\left|\sum_{s',a}\frac{1}{\sqrt{N}}(B^N_t(s',a,s) - p_t(s',a,s)X^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1})))\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right] + \mathbb{P}_{\pi_R}\left[\left|\sum_{s',a} p_t(s',a,s)\tilde{X}^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1}))\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right].$$

Notice

$$\mathbb{P}_{\pi_R}\left[\left|\sum_{s',a}\frac{1}{\sqrt{N}}(B^N_t(s',a,s) - p_t(s',a,s)X^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1})))\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right] \leq \sum_{s',a}\mathbb{P}_{\pi_R}\left[\frac{1}{\sqrt{N}}\left|B^N_t(s',a,s) - p_t(s',a,s)X^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1}))\right| \geq \frac{\epsilon\sqrt{N}}{2|A||S|^2}\right].$$

By Hoeffding's inequality, we have

$$\mathbb{P}_{\pi_R}\left[\frac{1}{\sqrt{N}}\left|B^N_t(s',a,s) - p_t(s',a,s)X^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1}))\right| \geq \frac{\epsilon\sqrt{N}}{2|A||S|^2}\right] \leq 2\exp\left(-\frac{\epsilon^2N^2}{2|A|^2|S|^4|X^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1}))|}\right) \leq 2\exp\left(-\frac{\epsilon^2N}{2|A|^2|S|^4}\right).$$

Combining the above inequalities together,

$$\mathbb{P}_{\pi_R}\left[\left|\sum_{s',a}\frac{1}{\sqrt{N}}(B^N_t(s',a,s) - p_t(s',a,s)X^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1})))\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right] \leq 2|A||S|\exp(-\frac{\epsilon^2N}{2|A|^2|S|^4}). \tag{B.10}$$

On the other hand, combining the bound

$$\left|\sum_{s',a} p_t(s',a,s)\tilde{X}^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1}))\right| \leq \sum_{s',a}|\tilde{X}^N_t(s',a,P_t(\boldsymbol{\omega}_{t+1}))|$$

with a union bound, we have

$$\mathbb{P}_{\pi_R}\left[\left|\sum_{s',a} p_t(s',a,s)\tilde{X}_t^N(s',a,P_t(\boldsymbol{\omega}_{t+1}))\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right] \leq \sum_{s',a}\mathbb{P}_{\pi_R}\left[\left|\tilde{X}_t^N(s',a,P_t(\boldsymbol{\omega}_{t+1}))\right| \geq \frac{\epsilon\sqrt{N}}{2|A||S|^2}\right].$$

Notice $\left|\tilde{X}_t^N(s',a,P_t(\boldsymbol{\omega}_{t+1}))\right| \leq |\tilde{Z}_t^N(P_t(\boldsymbol{\omega}_{t+1}))|$ for any $s \in S$. Thus,

$$\mathbb{P}_{\pi_R}\left[\left|\sum_{s',a} p_t(s',a,s)\tilde{X}_t^N(s',a,P_t(\boldsymbol{\omega}_{t+1}))\right| \geq \frac{\epsilon\sqrt{N}}{2|S|}\right] \leq |A||S|\,\mathbb{P}_{\pi_R}\left[|\tilde{Z}_t^N(P_t(\boldsymbol{\omega}_{t+1}))| \geq \frac{\epsilon\sqrt{N}}{2|A||S|^2}\right]. \tag{B.11}$$

Combining inequality (B.10) and (B.11) together implies inequality (B.8), concluding the proof. □

Now we can prove Lemma 3.3.

Proof. Proof of Lemma 3.3 Let $\Omega(\boldsymbol{\omega}_t) := \Delta(P_1(\boldsymbol{\omega}_t)) \cap \Delta(P_2(\boldsymbol{\omega}_t)) \cap \ldots \cap \Delta(\boldsymbol{\omega}_t)$ and let $\Omega^c(\boldsymbol{\omega}_t)$ denote its complement. First we notice,

$$\begin{aligned}
\mathbb{P}_{\pi_F}(\Delta^c(\boldsymbol{\omega}_{t+1})) &= \mathbb{P}_{\pi_F}(\Omega(P_t(\boldsymbol{\omega}_{t+1})) \cap \Delta^c(\boldsymbol{\omega}_{t+1})) + \mathbb{P}_{\pi_F}(\Omega^c(P_t(\boldsymbol{\omega}_{t+1})) \cap \Delta^c(\boldsymbol{\omega}_{t+1})) \\
&= \mathbb{P}_{\pi_R}(\Omega(P_t(\boldsymbol{\omega}_{t+1})) \cap \Delta^c(\boldsymbol{\omega}_{t+1})) + \mathbb{P}_{\pi_F}(\Omega^c(P_t(\boldsymbol{\omega}_{t+1})) \cap \Delta_{t+1}^c) \\
&\leq \mathbb{P}_{\pi_R}(\Delta^c(\boldsymbol{\omega}_{t+1})) + \mathbb{P}_{\pi_F}(\Omega^c(P_t(\boldsymbol{\omega}_{t+1}))) \\
&\leq \mathbb{P}_{\pi_R}(\Delta^c(\boldsymbol{\omega}_{t+1})) + \sum_{k=1}^{t}\mathbb{P}_{\pi_F}(\Delta^c(P_k(\boldsymbol{\omega}_{t+1}))).
\end{aligned}$$

We will use this recursive expression show that $\mathbb{P}_{\pi_F}(\Delta_t^c) \leq L\exp(-\delta N)$ by induction on t. The base case, $t = 1$, follows immediately from $\Delta^c(P_1(\boldsymbol{\omega}_t)) = \varnothing$. Thus, it is sufficient to prove there exists constants $\delta > 0$ and L, s.t. for all $\boldsymbol{\omega}_t$,

$$\mathbb{P}_{\pi_R}(\Delta^c(\boldsymbol{\omega}_t)) \leq L\exp(-\delta N). \tag{B.12}$$

Rewriting $\Delta(\boldsymbol{\omega}_t)$ in terms of $\hat{Z}_t^N(\boldsymbol{\omega}_t)$ and applying Lemma B.4 concludes the proof.

B.8 Proof of Lemma 3.4

By the Fenchel Duality Theorem [47], there exists adaptive mapping λ^* s.t.

$$\lambda^* \leftarrow \arg\min_{\lambda(\omega_t)\geq 0} \max_{\pi} \mathbb{E}_{\pi} \sum_{t=1}^{T} r_t(s_{1,t}, a_{1,t}) + \left\langle \lambda(\boldsymbol{\omega}_t),\ b_t(\omega_t) - c_t(s_{1,t}, a_{1,t}, \omega_t) \right\rangle,$$

and

$$\hat{V}_1^* = \max_{\pi} \mathbb{E}_{\pi} \sum_{t=1}^{T} r_t(s_{1,t}, a_{1,t}) + \left\langle \lambda^*(\boldsymbol{\omega}_t),\ b_t(\omega_t) - c_t(s_{1,t}, a_{1,t}, \omega_t) \right\rangle. \tag{B.13}$$

Thus, we can conclude that $\lambda^*(\boldsymbol{\omega}_t)(i) = 0$ if $\mathbb{E}_{\pi}[c_t(s_{1,t}, a_{1,t}, \omega_t)(i)] = b_t(\omega_t)(i)$ on event $\boldsymbol{\omega}_t$, where i refers to the i-th resource.

Following a dynamic programming argument, we define the value function V_t on states and the Q-factor Q_t on state-action pairs recursively as

$$Q_t(s, a, \boldsymbol{\omega}_t) = r_t(s, a) - < \lambda_t^*(\boldsymbol{\omega}_t), c_t(s, a, \omega_t) > + \mathbb{E}_{\omega_{t+1}} \left[\sum_{s' \in S} p_t(s, a, s') V_{t+1}(s', (\boldsymbol{\omega}_t, \omega_{t+1})) \right],$$

$$V_t(s, \boldsymbol{\omega}_t) = \max_a \{ Q_t(s, a, \boldsymbol{\omega}_t) \}$$

for all $\boldsymbol{\omega}_{t+1}$ with $V_{T+1}(s, \boldsymbol{\omega}_{T+1}) = 0$ for all $s \in S$. Thus, we can classify states into three disjoint sets

$$\text{Must-Activate}(\boldsymbol{\omega}_t) = \{ s \in S \,|\, Q_t(s, a^*, \boldsymbol{\omega}_t) < \max_{a \neq a^*} Q_t(s, a, \boldsymbol{\omega}_t) \},$$

$$\text{Indifferent}(\boldsymbol{\omega}_t) = \{ s \in S \,|\, Q_t(s, a^*, \boldsymbol{\omega}_t) = \max_{a \neq a^*} Q_t(s, a, \boldsymbol{\omega}_t) \},$$

$$\text{Must-Idle}(\boldsymbol{\omega}_t) = \{ s \in S \,|\, Q_t(s, a^*, \boldsymbol{\omega}_t) > \max_{a \neq a^*} Q_t(s, a, \boldsymbol{\omega}_t) \}.$$

A policy is optimal for (B.13) if and only if it obeys these three criteria for each event $\boldsymbol{\omega}_t$:

- It activates all arms whose states are in Must-Activate($\boldsymbol{\omega}_t$) with proper action.

- It idles any arms whose states are in Must-Idle($\boldsymbol{\omega}_t$).
- It either idles or activates arms in Indifferent($\boldsymbol{\omega}_t$) with action a s.t. $Q_t(s, a, \boldsymbol{\omega}_t) = Q_t(s, a^*, \boldsymbol{\omega}_t)$.

Any optimal occupation measure $(x_t(s, a, \boldsymbol{\omega}_t))_{\omega_t, s, a}$ achieves $\hat{V}_1^*$ and so must correspond to an optimal policy. Thus

$$\text{Must-Activate}(\boldsymbol{\omega}_t) \subseteq C^+(\boldsymbol{\omega}_t),\ \text{Must-Idle}(\boldsymbol{\omega}_t) \subseteq C^-(\boldsymbol{\omega}_t).$$

We can check and verify that any budget-relaxed fluid-priority policy π_R obeys the above three criteria. Thus, it is optimal for (B.13) and

$$V_N(\pi_R) + \sum_{\boldsymbol{\omega}_t} \mathbb{P}[\boldsymbol{\omega}_t] \left\langle \lambda^*(\boldsymbol{\omega}_t),\ b_t(\omega_t)N - \sum_{i=1}^{N} c_t(s_{i,t}, a_{i,t}, \omega_t) \right\rangle = \hat{V}_N^*.$$

Using the fact that $\lambda^*(\boldsymbol{\omega}_t)(i) = 0$ if $\mathbb{E}_\pi[c_t(s_{1,t}, a_{1,t}, \omega_t)(i)] = b_t(\omega_t)(i)$ on event $\boldsymbol{\omega}_t$ where i refers to the i-th resource and the non-degeneracy condition, we know at each event $\boldsymbol{\omega}_t$, there exists at most 1 resource type s.t. $\lambda^*(\boldsymbol{\omega}_t)(i) > 0$. We denote it by $i^*_{\omega_t}$.

We want to show that the event

$$b_t(\omega_t)(i^*_{\omega_t})N - \sum_{i=1}^{N} c_t(s_{i,t}, a_{i,t}, \omega_t)(i^*_{\omega_t}) > 0 \tag{B.14}$$

is of negligible probability. Actually rewriting event (B.14) in term of $\hat{Z}_t^N(s, \boldsymbol{\omega}_t)$ and applying Lemma B.4 conclude the proof.

APPENDIX C

APPENDIX: BINARY-ACTION INFINITE-HORIZON RESTLESS BANDIT

This section provides all technical proof in the main paper.

C.1 Proof for Lemma 4.1

In original problem (4.1) the budget constraint is in sense of cardinality, while the expectation constraint is need for relaxation problem (4.3). So a wider class of policy is feasible in the relaxation problem, which implies

$$V_N^*(T) \leq \hat{V}_N^*(T). \tag{C.1}$$

To prove $\hat{V}_N^*(T) = N\hat{V}_1^*(T)$, we use Lagrangian Relaxation similar to [20, 26] as the key idea in the following argument.

Through imitating straightforwardly the proof of Fenchel Duality Theorem [47],

$$\max_{\pi}\min_{\lambda}\mathbb{E}_{\pi}\sum_{t=1}^{T}\gamma^t R(\mathbf{s}_t,\mathbf{a}_t)+\lambda_t(\alpha_t N-|\mathbf{a}_t|)=\min_{\lambda}\max_{\pi}\mathbb{E}_{\pi}\sum_{t=1}^{T}\gamma^t R(\mathbf{s}_t,\mathbf{a}_t)+\lambda_t(\alpha_t N-|\mathbf{a}_t|) \tag{C.2}$$

where $\lambda = (\lambda_1, \lambda_2, ..., \lambda_T)$.

The let-hand side of Equation (C.2) equals to $\hat{V}_N^*(T)$. On the right hand side, for fixed λ,

$$\mathbb{E}_{\pi}\sum_{t=1}^{T}\gamma^t R(\mathbf{s}_t,\mathbf{a}_t)+\lambda_t(\alpha_t N-|\mathbf{a}_t|)=\mathbb{E}_{\pi}\sum_{t=1}^{T}\sum_{i=1}^{N}\gamma^t r(s_{t,i},a_{t,i})+\lambda_t(\alpha_t-a_{t,i}).$$

Since all arms share the same transition kernel and reward function,

$$\mathbb{E}_{\pi}\sum_{t=1}^{T}\sum_{i=1}^{N}\gamma^t r(s_{t,i},a_{t,i})+\lambda_t(\alpha_t-a_{t,i})=N\,\mathbb{E}_{\pi}\sum_{t=1}^{T}\gamma^t r(s_{t,1},a_{t,1})+\lambda_t(\alpha_t-a_{t,1}).$$

So we conclude

$$\min_{\lambda}\max_{\pi}\mathbb{E}_{\pi}\sum_{t=1}^{T}\gamma^{t}R(\mathbf{s}_t,\mathbf{a}_t)+\lambda_t(\alpha_t N-|\mathbf{a}_t|)=N\min_{\lambda}\max_{\pi}\mathbb{E}_{\pi}\sum_{t=1}^{T}\gamma^{t}r(s_{t,1},a_{t,1})+\lambda_t(\alpha_t-a_{t,1}). \tag{C.3}$$

By using Fenchel Duality again on the one-arm problem,

$$\begin{aligned}\hat{V}_1^*(T)&=\max_{\pi}\min_{\lambda}\mathbb{E}_{\pi}\sum_{t=1}^{T}\gamma^{t}r(s_{t,1},a_{t,1})+\lambda_t(\alpha_t-a_{t,1})\\&=\min_{\lambda}\max_{\pi}\mathbb{E}_{\pi}\sum_{t=1}^{T}\gamma^{t}r(s_{t,1},a_{t,1})+\lambda_t(\alpha_t-a_{t,1}).\end{aligned} \tag{C.4}$$

To summarize Equation (C.2), (C.3) and (C.4) together,

$$\hat{V}_N^*(T)=N\ \hat{V}_1^*(T).$$

C.2 Discussion of the rounding error in budget constraints

We want to show a rounding error in the relaxation Problem (4.3) results in at most a constant difference in the optimal objective value. Mathematically speaking, denote

$$\begin{aligned}\hat{V}_N^*(T)&=\max_{\pi}\left\{\mathbb{E}_{\pi}\sum_{t=1}^{T}\gamma^{t}R(\mathbf{s}_t,\mathbf{a}_t)\middle|\mathbb{E}|\mathbf{a}_t|=\alpha_t N,\ \forall t\le T\right\},\\\hat{V}_{N,R}^*(T)&=\max_{\pi}\left\{\mathbb{E}_{\pi}\sum_{t=1}^{T}\gamma^{t}R(\mathbf{s}_t,\mathbf{a}_t)\middle|\mathbb{E}|\mathbf{a}_t|=\lfloor\alpha_t N\rfloor,\ \forall t\le T\right\}.\end{aligned}$$

Then $|\hat{V}_N^*-\hat{V}_{N,R}^*|\le c$, where c does not depend on N. Thus, all our analysis on the asymptotic regime of optimality gap holds true since the LP relaxation upper bound (in rounded version) deviates from the unrounded version at most a constant away, not affecting the asymptotic analysis.

The proof of the above statement is straight forward. As seen from Lemma 4.1, there exists a single-arm pulling strategy which pulls α_t arms per period in expectation and achieves objective value $\hat{V}_1^*$. Thus, we can pull $N-1$ arms according to this strategy and pull the only arm left with probability $\lfloor \alpha_t N \rfloor - \alpha_t(N-1)$ at period t. Thus, we show

$$\frac{N-1}{N}\hat{V}_N^*(T) - \hat{V}_{N,R}^*(T) \leq \frac{\max_{s,a} r(s,a)}{1-\gamma}.$$

Similarly, we can show

$$\frac{N-1}{N}\hat{V}_{N,R}^*(T) - \hat{V}_N^*(T) \leq \frac{\max_{s,a} r(s,a)}{1-\gamma}.$$

Combining the above two inequality concludes the statement.

C.3 Proof of Lemma 4.2

To prove Lemma 4.2, first notice

$$Z_{t+1}^N(s) = \sum_{s'\in S, a\in A} \sum_{i=1}^{N} 1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s) \tag{C.5}$$

where $1(s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s)$ is the indicator function of event $\{s_{t,i} = s', a_{t,i} = a, s_{t+1,i} = s\}$. By dynamic equation (C.5) in a vector form,

$$Z_{t+1}^N = \sum_{s'\in S, a\in A} B_t^N(s', a),$$

where $B_t^N(s', a)$ is the sum of $X_t^N(s', a)$ independent $|S|$-dimensional Bernoulli random variable with mean $(p(s', a, s))_{s\in S}$. For simplicity, we denote $p(s', a) := (p(s', a, s))_{s\in S}$ in the following proof. With this new notation,

$$B_t^N(s', a) \sim \text{Binomial}(X_t^N(s', a), p(s', a)).$$

Recall that $X_t^N(s',a)$ can be decomposed as $Nx_t(s',a) + \sqrt{N}\tilde{X}_t^N(s',a)$. According to Lemma C.1, there exists two random variables $C_t^N(s',a)$ and $\Delta_t^N(s',a)$, s.t.

$$B_t^N(s',a) = C_t^N(s',a) + \Delta_t^N(s',a), \tag{C.6}$$

and that, conditionally on $X_t^N(s',a)$, have marginal distribution:

$$\begin{aligned}
&C_t^N(s',a) \sim \text{Binomial}(\lceil Nx_t(s',a)\rceil, p(s',a)),\\
&\Delta_t^N(s',a) \sim sgn(\tilde{X}_t^N(s',a))\text{Binomial}(\lfloor \sqrt{N}\left|\tilde{X}_t^N(s',a)\right|\rfloor, p(s',a)).
\end{aligned}$$

By equation (C.6) and recall the definition of diffusion statistics,

$$\tilde{Z}_{t+1}^N = \frac{1}{\sqrt{N}}\sum_{s'\in S,a\in A} C_t^N(s',a) - x_t(s',a)p(s',a)N + \frac{1}{\sqrt{N}}\sum_{s'\in S,a\in S}\Delta_t^N(s',a).$$

Now consider the L^1-norm of $\tilde{Z}_t^N$,

$$\begin{aligned}
|\tilde{Z}_{t+1}^N| &\le \left|\frac{1}{\sqrt{N}}\sum_{s'\in S,a\in S}\Delta_t^N(s',a)\right| + \frac{1}{\sqrt{N}}\sum_{s'\in S,a\in A}\left|C_t^N(s',a) - x_t(s',a)p(s',a)N\right|\\
&\le \sum_{s\in S,a\in A}|\tilde{X}_t^N(s',a)| + \frac{1}{\sqrt{N}}\sum_{s'\in S,a\in A}\left|C_t^N(s',a) - x_t(s',a)p(s',a)N\right|\\
&\le |\tilde{Z}_t^N| + \frac{C_3|S|}{\sqrt{N}} + \frac{1}{\sqrt{N}}\sum_{s'\in S,a\in A}\left|C_t^N(s',a) - x_t(s',a)p(s',a)N\right|,
\end{aligned}$$

where the last inequality is due to diffusion regularity of the policy so that $\sum_{a\in A}|\tilde{X}_t^N(s,a)| \le |\tilde{Z}_t^N(s)| + \frac{C_3}{\sqrt{N}}$. Thus,

$$\begin{aligned}
\mathbb{E}|\tilde{Z}_{t+1}^N| &= \mathbb{E}|\tilde{Z}_t^N| + \frac{C_3|S|}{\sqrt{N}} + \frac{1}{\sqrt{N}}\sum_{s'\in S,a\in A}\mathbb{E}\left|C_t^N(s',a) - x_t(s',a)p(s',a)N\right|\\
&\le \mathbb{E}|\tilde{Z}_t^N| + \frac{C_3|S|}{\sqrt{N}} + \frac{1}{\sqrt{N}}\sum_{s'\in S,a\in A}\mathbb{E}\left|C_t^N(s',a) - x_t(s',a)p(s',a)N\right|\\
&\le \mathbb{E}|\tilde{Z}_t^N| + \frac{C_3|S|}{\sqrt{N}} + \frac{1}{\sqrt{N}}\sum_{s'\in S,a\in A,s\in S}\sqrt{\frac{1+Nx_t(s',a)}{4}},
\end{aligned}$$

where the last inequality is by Lemma C.2.

So

$$\begin{aligned}\mathbb{E}|\tilde{Z}_{t+1}^N| &\le \mathbb{E}|\tilde{Z}_t^N| + \frac{C_3|S|}{\sqrt{N}} + \frac{1}{\sqrt{N}} \sum_{s'\in S, a\in A, s\in S} \sqrt{\frac{1+Nx_t(s',a)}{4}} \\ &= \mathbb{E}|\tilde{Z}_t^N| + C_3|S| + \frac{1}{2} \sum_{s'\in S, a\in A, s\in S} \sqrt{x_t(s',a)+1} \\ &\le \mathbb{E}|\tilde{Z}_t^N| + C_3|S| + 2|S|^2.\end{aligned}$$

Thus, by induction we have

$$\mathbb{E}|\tilde{Z}_{t+1}^N| \le t(2|S|^2 + C_3|S|) + \sup_N \mathbb{E}|\tilde{Z}_1^N| \tag{C.7}$$

Taking $c_1 = \sup_N \mathbb{E}|\tilde{Z}_1^N|$ and $c_2 = 2|S|^2 + C_3|S|$ concludes the proof. □

We state and prove the following Lemma C.1 and C.2.

Lemma C.1. *Suppose random variable S is a Binomial random variable with parameter n and p, i.e., distributed as the sum of n i.i.d. Bernoulli r.v.s with mean p. Then for a given non-negative integer m, there exists random variable S_1 and S_2, s.t. $S = S_1 + S_2$, and*

$$S_1 \sim Binomial(m, p),\ S_2 \sim sgn(n-m)Binomial(|n-m|, p),$$

where $sgn(\cdot)$ is the sign function.

Proof of Lemma C.1. There exists a sequence of i.i.d random variables $X_i \sim$ Binomial$(1, p)$, s.t.

$$S = \sum_{i=1}^{n} X_i.$$

If $n > m$, taking $S_1 = \sum_{i=1}^m X_i, S_2 = \sum_{j=m+1}^m X_j$ concludes the proof. If $n \le m$, taking $S_1 = \sum_{i=1}^m X_i, S_2 = -\sum_{j=n+1}^m X_j$ concludes the proof. □

Lemma C.2. *Suppose there are n i.i.d Bernoulli random variable $X_1, X_2, ..., X_n$ with mean p. Then*

$$\mathbb{E}|\sum_{i=1}^{n} X_i - np| \leq \sqrt{\frac{n}{4}}.$$

Proof of Lemma C.2. Direct calculation

$$\mathbb{E}|\sum_{i=1}^{n} X_i - np| \leq \sqrt{\mathbb{E}|\sum_{i=1}^{n} X_i - np|^2} = \sqrt{np(1-p)} \leq \sqrt{\frac{n}{4}}$$

concludes the proof. □

C.4 Proof of Lemma 4.3

Given a fluid-balance policy π, we directly check the induced map $\tilde{\pi}_{t,N}$ satisfies all three conditions in Definition 4.1.

Verification of Condition 1 Write the induced map in the component form $\tilde{\pi}_{t,N} = (\tilde{\pi}^1_{t,N}, ..., \tilde{\pi}^{|S|}_{t,N})$, and a direct calculation shows each component function $\tilde{\pi}^i_{t,N}$ $(1 \leq i \leq |S|)$ is continuous, piece-wise linear, and has bounded gradient when exits. Mathematically speaking, there exits a constant $\tilde{C}_1$, s.t., for any θ, any t and any N,

$$|\nabla \tilde{\pi}^i_{t,N}(\theta)| \leq \tilde{C}_1, \text{ when } \nabla \tilde{\pi}^i_{t,N}(\theta) \text{ exits.}$$

For any θ_1 and θ_2, there exits a sequence $(\theta^0_{1,2}, \theta^1_{1,2}, ..., \theta^m_{1,2})$ lies on the line segment between θ_1 and θ_2, s.t.

1. $\tilde{\pi}^i_{t,N}$ restricted on line segment between $\theta^j_{1,2}$ and $\theta^{j+1}_{1,2}$ is linear for $j = 0, 1, ..., m-1$

2. $\theta_{1,2}^0 = \theta_1$ and $\theta_{1,2}^m = \theta_2$.

Thus

$$|\tilde{\pi}_{t,N}^i(\theta_1) - \tilde{\pi}_{t,N}^i(\theta_2)| \le \sum_{j=0}^{m-1} |\tilde{\pi}_{t,N}^i(\theta_{1,2}^j) - \tilde{\pi}_{t,N}^i(\theta_{1,2}^{j+1})| \le \sum_{j=0}^{m-1} \tilde{C}_1 |\theta_{1,2}^j - \theta_{1,2}^{j+1}| = \tilde{C}_1 |\theta_1 - \theta_2|.$$

So by taking $C_1 = |S|\tilde{C}_1$,

$$|\tilde{\pi}_{t,N}(\theta_1) - \tilde{\pi}_{t,N}(\theta_2)| \le \sum_{i=1}^{|S|} |\tilde{\pi}_{t,N}^i(\theta_1) - \tilde{\pi}_{t,N}^i(\theta_2)| \le \sum_{i=1}^{|S|} \tilde{C}_1 |\theta_1 - \theta_2| = C_1 |\theta_1 - \theta_2|.$$

Verification of Condition 2 Direct calculation shows $\tilde{\pi}_{t,N}(0) = 0$.

Verification of Condition 3 Direct calculation shows $\tilde{\pi}_{t,\infty}(\tilde{Z}_t^\infty)$ is a piece-wise linear map.

Verification of Condition 4 Notice

$$\min\{Z_t(s), \lceil x_t(s,1)N + \sqrt{N}|\tilde{Z}_t^N(s)|\rceil\} \ge \max\{0, \lfloor x_t(s,1)N - \sqrt{N}|\tilde{Z}_t^N(s)|\rfloor\},$$

thus showing $\tilde{X}_{t,N}(s,0)$ and $\tilde{X}_t^N(s,1)$ has the same sign with $\tilde{Z}_t^N(s)$ is equivalent to

$$\sum_s \min\{Z_t(s), \lceil x_t(s,1)N + \sqrt{N}|\tilde{Z}_t^N(s)|\rceil\} \ge \lceil \alpha_t N \rceil, \text{ and}$$
$$\sum_s \max\{0, \lfloor x_t(s,1)N - \sqrt{N}|\tilde{Z}_t^N(s)|\rfloor\} \le \lceil \alpha_t N \rceil.$$

Notice

$$\begin{aligned}\sum_s \min\{Z_t(s), \lceil x_t(s,1)N + \sqrt{N}|\tilde{Z}_t^N(s)|\rceil\} &\ge \sum_s \min\{Z_t(s), x_t(s,1)N + \sqrt{N}|\tilde{Z}_t^N(s)|\} \\ &\ge \sum_s \min\{Z_t(s), x_t(s,1)N + \sqrt{N}\tilde{Z}_t^N(s)\} \\ &= \sum_s x_t(s,1)N + \sqrt{N}\tilde{Z}_t^N(s) = \alpha_t N,\end{aligned}$$

and $\sum_s \min\{Z_t(s), \lceil x_t(s,1)N + \sqrt{N}|\tilde{Z}_t^N(s)|\rceil\}$ is an integer, we can conclude

$$\sum_s \min\{Z_t(s), \lceil x_t(s,1)N + \sqrt{N}|\tilde{Z}_t^N(s)|\rceil\} \geq \lceil \alpha_t N \rceil.$$

Similarly, since

$$\begin{aligned}\sum_s \max\{0, \lfloor x_t(s,1)N - \sqrt{N}|\tilde{Z}_t^N(s)|\rfloor\} &\leq \sum_s \min\{0, x_t(s,1)N - \sqrt{N}|\tilde{Z}_t^N(s)|\} \\ &\leq \sum_s \min\{0, x_t(s,1)N - \sqrt{N}\tilde{Z}_t^N(s)\} \\ &= \sum_s x_t(s,1)N - \sqrt{N}\tilde{Z}_t^N(s) = \alpha_t N,\end{aligned}$$

we can conclude

$$\sum_s \max\{0, \lfloor x_t(s,1)N - \sqrt{N}|\tilde{Z}_t^N(s)|\rfloor\} \leq \lceil \alpha_t N \rceil.$$

To summarize, we prove the induced map of any fluid-balance policy satisfies all four conditions in Definition 4.1. Thus, any fluid-balance policy is diffusion regular.

As for the moment bound, notice $\tilde{Z}_1^N = 0$ and $C_3 = 0$ for any fluid balance policy. Thus, direct calculation of the coefficient in (C.7) gives $\mathbb{E}_\pi[|\tilde{Z}_t^N|] \leq 2t|S|^2$ for $t \leq T$. □

C.5 Proof of Proposition 4.1

We show that

- with Lagrangian penalty $\lambda = 0$, state A is active while state B is inactive;
- with Lagrangian penalty $\lambda = \alpha$, state A is inactive while state B is active.

Thus, the problem is not indexable. Now we analyze these two cases separately.

To be consistent with [55], we denote $V_\lambda(x)$ for the reward-to-go function of initial state x with Lagrangian penalty λ.

With Lagrangian penalty $\lambda = \alpha$, we can see $V_\lambda(A') = 0, V_\lambda(C) = 0, V_\lambda(B^*) = 0$ and $V_\lambda(B') = \beta - \alpha$ via direct calculation. Thus, we have

$$V_\lambda(A) = \max\{\gamma V_\lambda(B), -\alpha\},$$
$$V_\lambda(B) = \max\{\gamma V_\lambda(A), -\alpha + \gamma(1-\epsilon)V_\lambda(B')\}.$$

Solving the above equations gives us

$$V_\lambda(A) = \gamma(\gamma^2(\beta-\alpha) - \alpha), V_\lambda(B) = \gamma^2(\beta-\alpha) - \alpha.$$

Thus, state A is inactive and state B is active.

With Lagrangian penalty $\lambda = 0$, we can see $V_\lambda(A') = \frac{\alpha}{1-\gamma}, V_\lambda(C) = 0, V_\lambda(B^*) = 0$ and $V_\lambda(B') = \beta$ via direct calculation. Thus, we have

$$V_\lambda(A) = \max\{\gamma V_\lambda(B), (1-\epsilon)\gamma V_\lambda(A')\},$$
$$V_\lambda(B) = \max\{\gamma V_\lambda(A), \gamma(1-\epsilon)V_\lambda(B')\}.$$

Solving the above equations gives us

$$V_\lambda(A) = \frac{\alpha\gamma^2}{1-\gamma}, V_\lambda(B) = \frac{\alpha\gamma^3}{1-\gamma}.$$

Thus, state A is active and state B is inactive.

C.6 Proof of Proposition 4.2

Regardless of the budget constraint,

- the maximal reward generated from an initial state A_0 is up bounded by $\frac{\gamma\alpha}{1-\gamma}$,
- the maximal reward could be generated from an initial state A is up bounded by $(1-\epsilon)\frac{\gamma\alpha}{1-\gamma}$.

Thus, the maximal reward generated from N arms is up bounded by

$$\phi_1 N(1-\epsilon)\frac{\gamma\alpha}{1-\gamma} + \phi_3 N\frac{\gamma\alpha}{1-\gamma} = \frac{\gamma\alpha}{1-\gamma}\gamma N.$$

Now we calculate the reward generated by policy π which pulls all arms in A_0 and A in the first period, and then pulls all arms in A' starting from the second period.

In the first period, zero reward is generated. In the second period, there are $\phi_3 N + \phi_1 N(1-\epsilon) = \gamma N$ arms in state A' in expectation. Starting from second period, we pull all arms in state A'. Thus, we generate $\gamma N * \frac{\alpha}{1-\gamma}\gamma = \frac{\gamma\alpha}{1-\gamma}\gamma N$ amounts of reward.

C.7 Proof of Proposition 4.3

We only need to show that the optimal policy pulls all arms in state A in the first period.

If not, then under the optimal policy π^*, there exists an arm x in state A being idled and an arm y in state A_0 or C being pulled in the first period. Now we construct a policy $\hat{\pi}$ which pulls arm x in the first period with satisfying $V_N(\hat{\pi}) \geq V_N(\pi^*)$.

The definition of $\hat{\pi}$ is pretty simple. It pull x and idles y in the first period, and starting from the second period, $\hat{\pi}$ takes the exact same action with π for every arm. Then for any trajectory ω, we can calculate the reward generated from arm x by π^* and $\hat{\pi}$ and show $V_N(\hat{\pi}) \geq V_N(\pi^*)$.

C.8 Proof of Proposition 4.4

Under fluid-balance policy π, $\hat{V}_N^* - V_N(\pi) = \frac{\gamma\alpha}{1-\gamma}\mathbb{E}[\gamma N - \min\{Z, \gamma N\}]$, where Z is number of arms in state A' starting from the second period.

Thus, we have

$$Z = (1-\gamma)N + \sum_{i=1}^{(2-\frac{1}{\gamma})N} I_i$$

where I_i is a Bernoulli r.v. with probability γ. Thus,

$$\mathbb{E}[\gamma N - \min\{Z, \gamma N\}] = \mathbb{E}\left[\sum_{i=1}^{(2-\frac{1}{\gamma})N} I_i - \gamma\right]^+.$$

By concentration inequality, we know $\mathbb{E}[\gamma N - \min\{Z, \gamma N\}] = \theta\sqrt{N} + O(1)$ where $\theta = \sqrt{\frac{(1-\gamma)(2\gamma-1)}{2\pi}}$.

BIBLIOGRAPHY

[1] Daniel Adelman and Adam J Mersereau. Relaxations of weakly coupled stochastic dynamic programs. *Operations Research*, 56(3):712–727, 2008.

[2] Rajeev Agrawal. Sample mean based index policies with o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078, 1995.

[3] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

[4] ABM Alim Al Islam, SM Iftekharul Alam, Vijay Raghunathan, and Saurabh Bagchi. Multi-armed bandit congestion control in multi-hop infrastructure wireless mesh networks. In *2012 IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, pages 31–40. IEEE, 2012.

[5] PS Ansell, Kevin D Glazebrook, José Nino-Mora, and M O'Keeffe. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research*, 57(1):21–39, 2003.

[6] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[7] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

[8] Daniel S Bernstein and Shlomo Zilberstein. Reinforcement learning for weakly-coupled mdps and an application to planetary rover control. In *Sixth European Conference on Planning*, 2014.

[9] Dimitris Bertsimas and Adam J Mersereau. A learning approach for interactive marketing to a customer segment. *Operations Research*, 55(6):1120–1135, 2007.

[10] Dimitris Bertsimas and Velibor V Mišić. Decomposable markov decision processes: A fluid optimization approach. *Operations Research*, 64(6):1537–1555, 2016.

[11] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. *arXiv preprint arXiv:2105.07965*, 2021.

[12] Robert L Bray. Markov decision processes with exogenous variables. *Management Science*, 65(10):4598–4606, 2019.

[13] David B Brown and James E Smith. Index policies and performance bounds for dynamic selection problems. *Management Science*, 2020.

[14] David B Brown and Jingwei Zhang. Dynamic programs with shared resources and signals: Dynamic fluid policies and asymptotic optimality. *Operations Research*, 2021.

[15] Felipe Caro and Jérémie Gallien. Dynamic assortment with demand learning for seasonal consumer goods. *Management science*, 53(2):276–292, 2007.

[16] Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal. Mortal

multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 273–280, 2009.

[17] Xi Chen, Qihang Lin, and Dengyong Zhou. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *International Conference on Machine Learning*, pages 64–72, 2013.

[18] Philip Cho, Vivek Farias, John Kessler, Retsef Levi, Thomas Magnanti, and Eric Zarybnisky. Maintenance and flight scheduling of low observable aircraft. *Naval Research Logistics (NRL)*, 62(1):60–80, 2015.

[19] Savas Dayanik, Warren Powell, and Kazutoshi Yamazaki. Index policies for discounted bandit problems with availability constraints. *Advances in Applied Probability*, 40(2):377–400, 2008.

[20] Vivek F Farias and Ritesh Madan. The irrevocable multiarmed bandit problem. *Operations Research*, 59(2):383–399, 2011.

[21] 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)*, pages 249–254. IEEE, 2019.

[22] John Gittins. A dynamic allocation index for the sequential design of experiments. *Progress in statistics*, pages 241–266, 1974.

[23] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[24] Kevin D Glazebrook, Diego Ruiz-Hernandez, and Christopher Kirkbride. Some indexable families of restless bandit problems. *Advances in Applied Probability*, 38(3):643–672, 2006.

[25] Sudipto Guha and Kamesh Munagala. Approximation algorithms for budgeted learning problems. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, pages 104–113. ACM, 2007.

[26] Sudipto Guha and Kamesh Munagala. Sequential design of experiments via linear programming. *arXiv:0805.2630*, 2008.

[27] Sudipto Guha, Kamesh Munagala, and Peng Shi. Approximation algorithms for restless bandit problems. *Journal of the ACM (JACM)*, 58(1):3, 2010.

[28] Neha Gupta, Ole-Christoffer Granmo, and Ashok Agrawala. Thompson sampling for dynamic multi-armed bandits. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, volume 1, pages 484–489, 2011.

[29] Jeffrey Thomas Hawkins. *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. PhD thesis, Massachusetts Institute of Technology, 2003.

[30] Alfred O Hero and Douglas Cochran. Sensor management: Past, present, and future. *IEEE Sensors Journal*, 11(12):3064–3075, 2011.

[31] Weici Hu and Peter Frazier. An asymptotically optimal index policy for finite-horizon restless bandits. *arXiv:1707.00205*, 2017.

[32] Peter Jacko and José Nino-Mora. Time-constrained restless bandits and the knapsack problem for perishable items. *Electronic Notes in Discrete Mathematics*, 28:145–152, 2007.

[33] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[34] Jerome Le Ny, Munther Dahleh, and Eric Feron. Multi-agent task assignment in the bandit framework. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 5281–5286. IEEE, 2006.

[35] Jerome Le Ny, Munther Dahleh, and Eric Feron. Multi-uav dynamic routing with partial observations using restless bandit allocation indices. In *2008 American Control Conference*, pages 4220–4225. IEEE, 2008.

[36] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[37] Keqin Liu and Qing Zhao. A restless bandit formulation of opportunistic access: Indexablity and index policy. In *2008 5th IEEE Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops*, pages 1–5. IEEE, 2008.

[38] Keqin Liu and Qing Zhao. On the myopic policy for a class of restless bandit problems with applications in dynamic multichannel access. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 3592–3597. IEEE, 2009.

[39] Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, 2010.

[40] Rahul Meshram and Kesav Kaza. Simulation based algorithms for markov decision processes and multi-action restless bandits. *arXiv preprint arXiv:2007.12933*, 2020.

[41] Sentao Miao, Stefanus Jasin, and Xiuli Chao. Asymptotically optimal lagrangian policies for one-warehouse multi-store system with lost sales. *Preprint, submitted April*, 6, 2020.

[42] Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I-Hong Hou, and Srinivas Shakkottai. Neurwin: Neural whittle index network for restless bandits via deep rl. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[43] John C Nash. The (Dantzig) simplex method for linear programming. *Computing in Science & Engineering*, 2(1):29–31, 2000.

[44] Jose Nino-Mora. Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33(1):76–98, 2001.

[45] José Niño-Mora. Dynamic priority allocation via restless bandit marginal productivity indices. *Top*, 15(2):161–198, 2007.

[46] José Niño-Mora and Sofía S Villar. Sensor scheduling for hunting elusive hiding targets via whittle's restless bandit index policy. In *International Conference on NETwork Games, Control and Optimization*, pages 1–8. IEEE, 2011.

[47] R Tyrrell Rockafellar. Convex analysis princeton university press. *Princeton, NJ*, 1970.

[48] Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32, 2019.

[49] Ilya O Ryzhov, Warren B Powell, and Peter I Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.

[50] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. *Robotica*, 17(2):229–235, 1999.

[51] Huseyin Topaloglu. Using lagrangian relaxation to compute capacity-dependent bid prices in network revenue management. *Operations Research*, 57(3):637–649, 2009.

[52] Ina Maria Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *The Annals of Applied Probability*, 26(4):1947–1995, 2016.

[53] Kai Wang, Sanket Shat, Haipeng Chen, Andrew Perrault, Finale Doshi-Velez, and Milind Tambe. Learning mdps from features: Predict-then-optimize for sequential decision problems by reinforcement learning. *arXiv preprint arXiv:2106.03279*, 2021.

[54] Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

[55] Peter Whittle. Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):143–149, 1980.

[56] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

[57] Gabriel Zayas-Caban, Stefanus Jasin, and Guihua Wang. An asymptotically optimal heuristic for general nonstationary finite-horizon rest-

less multi-armed, multi-action bandits. *Advances in Applied Probability*, 51(3):745–772, 2019.

[58] Xiangyu Zhang and Peter I Frazier. Restless bandits with many arms: Beating the central limit theorem. *arXiv preprint arXiv:2107.11911*, 2021.