TWO-PHASE POLICY CLASSES FOR MULTI-CONTEXT RESTLESS BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Restless multi-armed bandits are widely used for sequential resource allocation, yet all existing theory assumes a fixed activation budget. In practice, many applications involve multiple operating contexts with distinct dynamics, and the decision of how to distribute budget across these contexts is as important as which arms to activate within each. We formalize this as the **contextual budget bandit** (CBB), a two-phase optimization where the upper level allocates per-context budgets under a global capacity constraint and the lower level selects arms under hard per-period limits.

We introduce a hierarchy of two-phase policy classes capturing this structure. Our main result shows that, for exchangeable arm types, the optimal policy belongs to a tractable class we call **aggregate-state-dependent priority**: the type-priority may depend on how many arms of each type are engaged, but not on which specific arms. We prove this via permutation invariance of the value function. Surprisingly, fixed-priority index policies are not always optimal: a **portfolio effect** causes the priority between types to reverse depending on same-type depth, starting at three arms. A gap bound shows that Lagrangian index policies lose at most a vanishing fraction of the optimum as the number of arms grows.

Computational validation across 112,000 randomly generated instances confirms the aggregate-class optimality at 100% and quantifies the portfolio effect at 2–17% of instances. On a food rescue volunteer engagement application, the Mirror Descent Value Iteration (MDVI) algorithm achieves 0.0–3.2% gap from the LP bound, scaling to $N{=}2{,}000$ arms in 10 minutes.

1 INTRODUCTION

Restless multi-armed bandits (RMABs) offer a natural framework for sequential resource allocation: at each period, a decision-maker activates up to B of N arms, each evolving as an independent Markov chain whose transitions depend on whether it was activated. Because *all* arms evolve at every step (not just those selected), so inaction carries real consequences: an uncontacted volunteer may disengage, an untreated patient may worsen. Whittle's index policy (Whittle, 1988) and its asymptotic guarantees (Weber & Weiss, 1990; Gast et al., 2024) have made RMABs a workhorse in public health (Liang et al., 2025), food rescue (Shi et al., 2024), and related domains.

All of this theory, however, assumes a *fixed* budget B. Many real applications involve multiple *contexts* (weekday vs. weekend, urban vs. rural), each with its own arm dynamics and rewards. Deciding how to split a limited budget across contexts is itself a consequential choice: how many notifications to send on weekdays versus weekends, or how many beds to reserve for elective versus emergency patients. The result is a *two-phase* problem: an upper level allocates per-context budgets $\{B_k\}$ under a global capacity constraint, while a lower level selects which arms to activate under each context's hard cap. No prior RMAB theory (Whittle, 1988; Weber & Weiss, 1990; Gast et al., 2024; Zhang & Frazier, 2021; Hong et al., 2024) addresses this joint optimization; all existing results take B as given.

The difficulty is that the two levels are coupled: the best budget split depends on the arm dynamics (which govern how valuable extra budget is in each context), while the arm-selection policy depends on the budget (which limits the activation set). Standard Lagrangian decomposition decouples arms at the lower level but does not extend directly to the upper level, where the global constraint ties the contexts together.

Contributions. We formalize the *contextual budget bandit* (CBB) and make four contributions. (i) We introduce a hierarchy of *two-phase policy classes*, $\mathcal{F}_{\mathrm{Lag}} \subset \mathcal{F}_{\mathrm{idx}} \subset \mathcal{F}_{\mathrm{agg}} \subset \mathcal{F}_{\mathrm{det}}$, that

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

jointly capture budget allocation and arm selection (Section 2). The class $\mathcal{F}_{\text{agg}}$ allows the priority ordering over arm types to depend on the aggregate engagement counts. (ii) We prove that under *exchangeability* the optimal policy always lies in $\mathcal{F}_{\text{agg}}$ (Section 5), and exhibit a *portfolio effect* showing $\mathcal{F}_{\text{idx}}$ is strictly weaker for $N \geq 3$. (iii) We establish an $O(K/N)$ gap bound via a common shadow-price structure ($w_k = \mu^* f_k$) that collapses the K-dimensional budget search to a scalar problem (Section 5). (iv) Computationally, $\mathcal{F}_{\text{agg}}$ optimality holds across 112,000 instances, and our MDVI algorithm achieves 0.0–3.3% gap from the LP bound at scales up to N=2,000 (Section 6).

2 Background

2.1 Restless Multi-Armed Bandits

An RMAB consists of N arms, each an independent Markov chain. At every period the decision-maker activates at most B arms; each arm then transitions stochastically, with probabilities that depend on whether it was activated. Crucially, *idle* arms also evolve (hence "restless").

An *index policy* scores each arm by a scalar $\iota_i(s_i)$ and activates the top B. Whittle's index (Whittle, 1988) obtains these scores from a Lagrangian relaxation, decomposing the N-arm problem into N single-arm sub-problems. Index policies scale linearly in N but ignore cross-arm interactions induced by the shared budget.

The *LP relaxation* replaces the hard per-period constraint $\|\mathbf{a}\|_1 \leq B$ with an expected budget constraint, yielding an upper bound on the optimal reward decomposable by arm. An arm is *ergodic* if its chain has a unique stationary distribution; for binary states this holds whenever all transition probabilities lie in $(0, 1)$.

2.2 The Two-Phase Contextual Budget Bandit

The *contextual budget bandit* (CBB) extends the standard RMAB in two ways: (1) the transition dynamics and rewards depend on a context k drawn at each period, and (2) the per-context budgets B_k are decision variables subject to a global constraint.

Consider N arms with binary states $s_i \in \{0, 1\}$ (disengaged/engaged), K contexts drawn independently and identically (i.i.d.) from distribution $f = (f_1, \ldots, f_K)$, and per-arm transition kernels $P_i(s' \mid s, a; k)$ satisfying *conditional independence* (arms transition independently given the action):

$$P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}; k) = \prod_i P_i(s'_i \mid s_i, a_i; k). \tag{1}$$

Rewards are additive: $R(\mathbf{s}, \mathbf{a}; k) = \sum_i R_i(s_i, a_i; k)$, where typically $R_i(s_i, a_i; k) = r_i \cdot s_i \cdot a_i$ (reward accrues when an engaged arm is activated).

Under T *exchangeable* arm types, all type-t arms share a common kernel P_t and reward R_t. Exchangeability means the arms within each type are statistically identical: swapping the labels of two type-t arms does not change the problem. We denote the number of type-t arms by N_t ($\sum_t N_t = N$) and the *aggregate state* by $\mathbf{n} = (n_1, \ldots, n_T)$ where $n_t = \sum_{i \in \text{type } t} s_i$ counts the engaged arms of type t. The aggregate state space $\mathcal{N} = \prod_t \{0, \ldots, N_t\}$ has $\prod_t (N_t + 1)$ elements, polynomial in N for fixed T. In the food rescue application, types correspond to volunteer behavioral profiles (e.g., "organic" volunteers with high natural retention and "churner" volunteers who disengage rapidly without intervention).

The *two-phase* structure arises because the CBB has two levels of decision-making. The *upper level* chooses how to allocate total budget B_{tot} across contexts: how many activations B_k to allow under each context k, subject to the global capacity constraint $\sum_k f_k B_k \leq B_{\text{tot}}$. The *lower level*, given the per-context budgets $\{B_k\}$, selects which specific arms to activate at each period under the hard constraint $\|\mathbf{a}\|_1 \leq B_k$.

The two-phase CBB optimizes jointly over budget allocation and arm selection:

$$\max_{\substack{B_k \in \mathbb{Z}_+,\ k \in [K] \\ \sum_k f_k B_k \leq B_{\text{tot}}}} \ \max_{\pi \in \Pi_{\text{hard}}(\mathbf{B})} \ g(\pi, \mathbf{B}) \tag{2}$$

where $g(\pi, \mathbf{B}) = \lim_{H \to \infty} H^{-1} \sum_{h=1}^{H} \mathbb{E}[\sum_i R_i(s_i^h, a_i^h; k^h)]$ is the long-run average reward and $\Pi_{\text{hard}}(\mathbf{B})$ enforces $\|\mathbf{a}\|_1 \leq B_k$ at every period under context k. The upper level is a discrete optimization over $O(B_{\text{tot}}^K)$ feasible integer vectors; the lower level is a constrained Markov decision process (MDP) on the joint state space $\{0, 1\}^N$ (exponential in N).

3 TWO-PHASE POLICY CLASSES

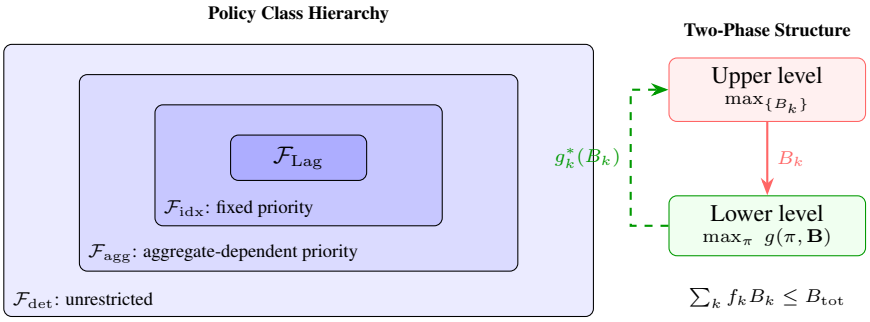

Figure 1: *Left:* Nested policy classes for the two-phase CBB. Each class specifies both a budget allocation and an arm selection rule. $\mathcal{F}_{\text{Lag}}$ (scalar μ) $\subset \mathcal{F}_{\text{idx}}$ (fixed type-priority) $\subset \mathcal{F}_{\text{agg}}$ (aggregate-count-dependent priority; proved optimal) $\subset \mathcal{F}_{\text{det}}$ (unrestricted). *Right:* The two-phase structure. The upper level allocates per-context budgets B_k; the lower level selects arms under hard constraints. The common shadow price $w_k = \mu^* f_k$ reduces this to a scalar search.

A CBB policy is a pair $(\mathbf{B}, \pi)$. Let $G = \prod_{t=1}^{T} S_{N_t}$ be the within-type permutation group and $\mathcal{N} = \prod_t \{0, \ldots, N_t\}$ the aggregate state space.

Definition 1 (Policy hierarchy)**.** *The **unrestricted deterministic** class $\mathcal{F}_{\text{det}}$ places no restrictions: $\mathbf{B}$ is any feasible integer vector and π is an arbitrary deterministic mapping from joint state-context pairs $(\mathbf{s}, k)$ to actions.*

*The **aggregate-state-dependent priority** class $\mathcal{F}_{\text{agg}}$ restricts arm selection to depend on $\mathbf{s}$ only through the aggregate count vector $\mathbf{n}(\mathbf{s}) = (n_1, \ldots, n_T)$, where n_t counts the engaged arms of type t. At each $(\mathbf{n}, k)$, the policy specifies per-type pull counts (b_t, ℓ_t), where b_t is the total number of type-t arms pulled and ℓ_t is the number of those that are engaged. The priority ordering over the $2T$ type-state categories can vary with $\mathbf{n}$.*

*The **fixed-priority index** class $\mathcal{F}_{\text{idx}}$ further restricts the priority ordering to be independent of $\mathbf{n}$: for each context k, a single ranking of the $2T$ categories $\{(\text{type } t, \text{state } s)\}$ determines the allocation at every aggregate state. There are at most $(2T)!^K$ such policies.*

*The **Lagrangian** class $\mathcal{F}_{\text{Lag}}$ is the most restrictive: a single scalar $\mu \ge 0$ determines both the budget allocation (via the common shadow price $w_k = \mu f_k$ and complementary slackness) and the per-arm scores $\iota_i(s_i, k; \mu) = Q_i(s_i, 1; k, \mu) - Q_i(s_i, 0; k, \mu)$.*

These classes satisfy:

$$\mathcal{F}_{\text{Lag}} \subset \mathcal{F}_{\text{idx}} \subset \mathcal{F}_{\text{agg}} \subset \mathcal{F}_{\text{det}}. \tag{3}$$

3.1 THE PORTFOLIO EFFECT

Proposition 2 (Strict inclusion for $N \ge 3$)**.** *For $N \ge 3$, $T = 2$, there exist exchangeable CBB instances where the optimal policy is in $\mathcal{F}_{\text{agg}} \setminus \mathcal{F}_{\text{idx}}$.*

Proof (by construction). A search over 8,000 randomly generated exchangeable instances at $(N, K, B, T) = (3, 1, 1, 2)$ yields instances where the type-priority reverses. A concrete example has $N_0 = 2$ arms of type 0 and $N_1 = 1$ arm of type 1, with parameters:

	$P(1 \mid 0, 0)$	$P(1 \mid 0, 1)$	$P(1 \mid 1, 0)$	$P(1 \mid 1, 1)$
Type 0	0.054	0.108	0.926	0.821
Type 1	0.019	0.048	0.907	0.892

Both types are slow-mixing (high self-retention $P(1 \mid 1, 0) > 0.9$). The optimal aggregate policy (from exact value iteration on the $2^3 = 8$ joint states) is:

(n_0, n_1)	Pull type	Rationale
$(0, 0)$	type 0	Re-engage type 0 first
$(1, 0)$	type 0	Maintain sole type-0 arm
$(0, 1)$	type 1	Maintain sole type-1 arm
$(1, 1)$	**type 1**	Type 1 more fragile
$(2, 0)$	type 0	Only type 0 present
$(2, 1)$	**type 0**	Backup available

At $(n_0 = 1, n_1 = 1)$: type 1 is pulled (requires $\iota(\text{t1}, 1) > \iota(\text{t0}, 1)$). At $(n_0 = 2, n_1 = 1)$: type 0 is pulled (requires $\iota(\text{t0}, 1) > \iota(\text{t1}, 1)$). These two constraints contradict any fixed priority ordering, proving the policy is in $\mathcal{F}_{\text{agg}} \setminus \mathcal{F}_{\text{idx}}$. □

Portfolio effect interpretation. The priority reversal arises from a diversification mechanism. When only one type-0 arm is engaged $(n_0 = 1)$, that arm is "irreplaceable": losing it means the system has zero type-0 coverage. The budget is therefore better spent on type-1 (which may be more fragile or more immediately valuable), allowing the lone type-0 arm to self-retain through its natural persistence probability $P_0(1 \mid 1, 0)$. When two type-0 arms are engaged $(n_0 = 2)$, the system has backup coverage: even if one type-0 arm disengages, the other provides continued type-0 reward. This "depth of coverage" increases the marginal value of servicing type-0 relative to type-1, flipping the priority. The effect is analogous to portfolio diversification in finance: the marginal value of an asset depends on the existing portfolio composition, not just its individual return.

Remark 3. *The portfolio effect requires $N \geq 3$ because with $N = 2$ (one arm per type), no same-type backup exists and the priority is always determined by the per-arm Lagrangian scores alone. An exhaustive search over 9×10^8 instances at $N = 2$ confirms this (Appendix D).*

3.2 Sub-Optimality of Index Policies: A Counterexample

To build intuition, consider a small instance: $N = 4$ arms (2 organic, 2 churner), a single context ($K = 1$), and budget $B = 2$. Organics retain engagement when pulled (0.90) but decay when idle (0.65); churners behave oppositely: pulling *destroys* their engagement (0.05) while idling preserves it (0.70). Both types earn the same immediate reward ($R = 1.0$).

Any index policy assigns a fixed priority: it will always prefer engaged organics over engaged churners (or vice versa), regardless of how many of each are currently engaged. But the optimal $\mathcal{F}_{\text{agg}}$ policy does something subtler. When one organic is disengaged and one churner is engaged, it *reverses* the priority: it spends budget recovering the disengaged organic while leaving the churner alone (the churner retains at 0.70 without help, and pulling would destroy it). This reversal depends on the aggregate state $(n_{\text{org}}, n_{\text{churn}})$, precisely the information that $\mathcal{F}_{\text{agg}}$ can condition on but $\mathcal{F}_{\text{idx}}$ cannot. MDVI captures this effect through its state-dependent dual prices.

4 The MDVI Algorithm

The idea behind MDVI is to dualize the budget constraint. Introducing a price $w_k \geq 0$ per activation in context k, each arm faces an unconstrained MDP with modified reward $\tilde{R}_i(s, a; k) = R_i(s, a; k) - w_k \cdot a$. Because each per-arm MDP has only two states, it can be solved exactly by Relative Value Iteration in $O(K)$ time, yielding per-arm gain $g_i(\mathbf{w})$. The Lagrangian dual $D(\mathbf{w}) = \sum_i g_i(\mathbf{w}) + \sum_k w_k B_k$ is an upper bound on the true optimum (weak duality); MDVI tightens it via projected subgradient descent. At each iteration the subgradient compares the *demand* (the number of arms each per-arm policy wants to activate) against the budget B_k. When demand exceeds supply the price rises, discouraging activation; the reverse lowers it. This supply-demand feedback converges to dual prices $\mathbf{w}^*$ that balance the system.

Algorithm 1 MDVI: Mirror Descent Value Iteration

Require: Instance $(N, K, \{P_i\}, \{R_i\}, f, B_{\text{tot}})$; iterations L; step sizes $\{\alpha_\ell\}$
1: Initialize $\mathbf{w}^0 \geq 0$
2: **for** $\ell = 0, \ldots, L-1$ **do**
3: **for** each arm $i = 1, \ldots, N$ **do**
4: Solve per-arm average-reward MDP at $\mathbf{w}^\ell$ via Relative Value Iteration:
5: $g_i, v_i(\cdot), \pi_i(\cdot; \mathbf{w}^\ell) \leftarrow \text{RVI}(P_i, R_i - w_k^\ell \cdot a_i)$
6: **end for**
7: Simulate T_{sim} steps to estimate demand $\hat{D}_k = \sum_i \pi_i(s_i; \mathbf{w}^\ell)$
8: Subgradient update: $w_k^{\ell+1} \leftarrow [w_k^\ell - \alpha_\ell(\hat{D}_k - B_k)]_+$
9: **end for**
10: **return** $\mathbf{w}^L$, per-arm policies $\{\pi_i\}$

MDVI produces policies in $\mathcal{F}_{\text{Lag}}$: the dual $\mathbf{w}$ determines both the budget allocation (via complementary slackness) and the arm scores (via per-arm RVI). Its convergence rate is $O(N\sqrt{K/L})$ (standard projected subgradient; see Appendix A).

Complexity. Each iteration solves N independent $|S|$-state MDPs via RVI ($O(|S|^2)$ per arm per context) and simulates T_{sim} steps for the subgradient ($O(NKT_{\text{sim}})$). Total: $O(L \cdot (N|S|^2K + NKT_{\text{sim}}))$. For binary states ($|S| = 2$), this is $O(LNK(4 + T_{\text{sim}}))$, linear in all parameters.

Hierarchy placement. MDVI's output $\mathbf{w}^L$ defines per-arm scores $\iota_i(s_i, k; \mathbf{w}^L) = Q_i(s_i, 1; k, \mathbf{w}^L) - Q_i(s_i, 0; k, \mathbf{w}^L)$ that are *state-independent* (depend only on arm i's own state s_i, not on other arms). The budget allocation B_k^* follows from complementary slackness at the converged dual. MDVI thus sits in $\mathcal{F}_{\text{Lag}} \subset \mathcal{F}_{\text{idx}}$, the smallest class in our hierarchy, yet comes within a few percent of optimal in practice (Section 6.4).

5 Theoretical Guarantees

5.1 $\mathcal{F}_{\text{agg}}$ Optimality Under Exchangeability

Lemma 4 (Within-type permutation invariance)**.** *For any fixed $\mathbf{B}$, the optimal value function satisfies $V^*(\sigma(\mathbf{s})) = V^*(\mathbf{s})$ for all within-type permutations $\sigma \in G$.*

Proof sketch. The Bellman operator preserves within-type permutation invariance: substituting $\mathbf{a} = \sigma(\tilde{\mathbf{a}})$ and $\mathbf{s}' = \sigma(\tilde{\mathbf{s}}')$ in $[BV](\sigma(\mathbf{s}))$ and using exchangeability of (P, R) yields $[BV](\sigma(\mathbf{s})) = [BV](\mathbf{s})$ when V is permutation-invariant. Value iteration from $V_0 \equiv 0$ preserves this inductively. □

Theorem 5 ($\mathcal{F}_{\text{agg}}$ optimality)**.** *For the average-reward CBB with T exchangeable types, conditional independence, and ergodic per-arm chains: $\max_{\mathcal{F}_{\text{det}}} g = \max_{\mathcal{F}_{\text{agg}}} g$.*

Proof sketch. By Lemma 4, $V^*(\mathbf{s}) = V^*(\mathbf{n}(\mathbf{s}))$ (any two states related by a within-type permutation have the same aggregate counts). The Q-function at state $\mathbf{s}$ with action $\mathbf{a}$ under context k is:

$$Q(\mathbf{s}, \mathbf{a}; k) = \sum_i R_{\text{type}(i)}(s_i, a_i; k) + \gamma \sum_{\mathbf{s}'} \prod_i P_{\text{type}(i)}(s_i' \mid s_i, a_i; k) \cdot V^*(\mathbf{n}').$$

The reward sum depends on $(\mathbf{s}, \mathbf{a})$ only through the per-type summaries (n_t, b_t, ℓ_t), where $b_t = \sum_{i \in t} a_i$ (pulls per type) and $\ell_t = \sum_{i \in t} s_i a_i$ (engaged pulls). The continuation value depends on $\mathbf{n}' = (n_1', \ldots, n_T')$, whose distribution under conditional independence factors by type: n_t' depends only on (n_t, b_t, ℓ_t) and the type-t kernel. Therefore $\arg\max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a}; k)$ depends on $\mathbf{s}$ only through $\mathbf{n}$. The upper-level $\max_{\mathbf{B}}$ is over a finite set of integer vectors and preserves this structure at each candidate $\mathbf{B}$. Full proof in Appendix A. □

The assumptions are mild: exchangeability, conditional independence, and ergodicity of the per-arm chains. Notably, we do *not* require indexability, the Uniform Global Attractor Property (UGAP), non-degeneracy, or monotonicity; these conditions underpin most prior asymptotic results. The proof relies solely on the permutation symmetry of the MDP and holds for arbitrary T, N, K, and $\{B_k\}$.

Computational complexity of $\mathcal{F}_{\text{agg}}$. The aggregate state space $\mathcal{N} = \prod_t \{0, \ldots, N_t\}$ has $\prod_t (N_t + 1)$ elements, polynomial in N for fixed T. For T=2, N=50 (25 per type): $|\mathcal{N}| = 26^2 = 676$, compared to $2^{50} \approx 10^{15}$ joint states. The optimal $\mathcal{F}_{\text{agg}}$ policy can be computed by value iteration on $\mathcal{N}$ in time $O(|\mathcal{N}|^2 \cdot K \cdot B_{\max})$, which is tractable even for moderately large N.

5.2 Gap Bounds

Although $\mathcal{F}_{\text{agg}}$ is exactly optimal, MDVI operates in $\mathcal{F}_{\text{Lag}}$, the most restrictive class in our hierarchy. How much does this restriction cost?

Proposition 6 (Common shadow price)**.** *At the two-phase Lagrangian optimum, $w_k^* = \mu^* f_k$ for all active contexts ($B_k^* > 0$), reducing the K-dimensional budget allocation to a scalar search over $\mu \geq 0$.*

The proof (Appendix A) follows from the Lagrangian structure: the term $\sum_k (w_k - \mu f_k) B_k$ in the dual is linear in B_k and must be non-positive for boundedness, with equality at active contexts by complementary slackness. This structural result is specific to the two-phase CBB and has no analogue in single-budget RMABs.

Theorem 7 (Two-phase bound)**.** *Under the conditions of Theorem 5 with non-degenerate LP relaxation:*

$$g^*_{\text{bi-level}} - g(\pi_{\mathcal{F}_{\text{Lag}}}, \mathbf{B}^*) \leq \underbrace{\frac{K \cdot R^{\text{arm}}_{\max}}{N}}_{\textit{budget rounding}} + \underbrace{K \cdot C\, e^{-c \min_t N_t}}_{\textit{per-context index gap}} \tag{4}$$

where $C, c > 0$ depend on per-type parameters but not on N.

Two terms contribute. The $O(K/N)$ piece comes from rounding the LP-optimal (continuous) budget allocation to integers; each context can misallocate at most one unit, affecting $O(1)$ of N arms. The $O(Ke^{-cN})$ piece is the per-context cost of using Lagrangian priorities rather than the full $\mathcal{F}_{\text{agg}}$ policy, building on the exponential bound of Gast et al. (2024) (whose conditions we verify for binary-state arms in Appendix B). At large N the rounding term dominates, yielding an overall $O(K/N)$ rate.

6 Experiments

6.1 Experimental Setup

Instances. All experiments use instances calibrated to the 412 Food Rescue volunteer platform (Shi et al., 2024). Arms come in two behavioral types: *organic* volunteers (pulled retention ≈ 0.90, idle retention ≈ 0.65) who stay engaged when contacted, and *churner* volunteers (pulled retention ≈ 0.05, idle retention ≈ 0.70) who paradoxically *disengage* when contacted, a phenomenon we call the "churn trap." Both types earn similar immediate reward ($R \in [0.8, 1.2]$), so the distinction lies entirely in the dynamics.

Baselines. We compare against eight algorithms: *DQWIC* (Whittle, 1988) (Deep Q-Network for Whittle Index Context, falling back to tabular); *Mitosis* (Shi et al., 2024), the current state-of-the-art UCB-guided budget search for contextual RMABs (c_{UCB}=2, 600 s timeout); *NIP* (REINFORCE-trained neural index); *Greedy* (pull the highest immediate-reward arms); *Q-index* (tabular action-value ranking); plus *COcc* (Contextual Occupancy Index) and two naive baselines (*Random*, *Round-robin*). All numbers are mean $\pm$ standard error over 10 seeds. Gaps are measured relative to the LP relaxation bound: $\text{gap}(\pi) = (g^*_{\text{LP}} - g(\pi))/g^*_{\text{LP}} \times 100\%$.

6.2 Does $\mathcal{F}_{\text{agg}}$ Capture the Optimal Policy?

Table 1: $\mathcal{F}_{\text{agg}}$ validation across 112,000 exchangeable instances (8,000 per config, 4 generators). T: number of exchangeable types. K: number of contexts.

N	K	B	T	$\mathcal{F}_{\text{agg}}$	$\mathcal{F}_{\text{idx}}$	Portfolio
3	1	1	2	**100%**	97.0%	3.0%
4	1	1	2	**100%**	95.8%	4.2%
4	1	2	2	**100%**	93.5%	6.5%
5	1	2	2	**100%**	90.0%	10.1%
6	1	2	2	**100%**	86.7%	13.3%
4	1	2	3	**100%**	91.0%	9.0%
3	2	1	2	**100%**	89.7%	10.3%
4	2	1	2	**100%**	83.0%	17.0%

Across all 14 configurations (112,000 instances total), $\mathcal{F}_{\text{agg}}$ optimality holds without exception, validating Theorem 5. The portfolio effect (instances where the optimal policy lies in $\mathcal{F}_{\text{agg}} \setminus \mathcal{F}_{\text{idx}}$) grows with both the budget ratio B/N and the number of contexts: at K=2 the rate reaches 17%, as each context independently encounters the priority reversal. The effect persists for three types (T=3), confirming that the theorem is not limited to the binary-type setting. Despite the frequency of portfolio-effect instances, their *reward* impact is small: the gap $g^*_{\text{agg}} - g^*_{\text{idx}}$ never exceeds 0.2%, consistent with the exponential decay predicted by Theorem 7.

6.3 How Do Algorithms Compare Across Scales?

We compare MDVI ($\mathcal{F}_{\text{agg}}$) against all baselines at scales from N=10 to N=500, reporting optimality gap relative to the LP relaxation bound.

Table 2: Optimality gap (%) to LP bound (mean±SE, 10 seeds). **Bold**: best. CPU: aggregate time per seed (all methods).

(N, K, B)	MDVI	NIP	Q-idx	DQWIC	Mitosis	Greedy	COcc	Rand	RR	CPU
(10,2,3)	2.3±0.4	**1.5±1.4**	3.6±1.3	3.3±1.5	18.1±0.4	29.7±1.1	22.9±1.3	67.1±0.4	80.8±0.4	30s
(10,3,3)	5.0±1.0	**4.8±2.2**	6.7±1.8	6.2±2.1	19.3±0.8	28.3±1.3	23.9±1.7	68.3±0.5	78.2±0.4	31s
(20,2,4)	6.4±0.7	**2.6±1.7**	6.2±1.6	3.2±1.6	9.7±1.1	18.1±0.8	10.8±2.0	77.2±0.4	88.8±0.4	57s
(20,3,6)	4.3±0.5	**2.5±1.0**	5.4±1.0	3.8±0.8	9.4±0.5	12.1±0.7	10.5±0.9	67.9±0.3	78.6±0.5	59s
(50,3,5)	**6.2±0.6**	8.1±1.0	12.6±0.8	9.0±1.1	11.0±0.6	12.5±0.4	12.7±1.0	87.5±0.2	91.8±0.2	2.4m
(50,5,10)	**4.3±0.4**	5.1±0.7	8.4±0.7	6.6±0.8	6.2±0.3	7.3±0.6	8.0±0.7	78.0±0.3	83.3±0.3	2.6m
(100,3,10)	**1.2±0.3**	5.7±0.7	11.4±0.7	6.5±0.6	5.2±0.4	9.6±0.4	7.1±0.6	87.8±0.2	91.7±0.3	4.8m
(100,5,15)	**0.7±0.2**	4.5±0.6	8.7±0.6	5.2±0.7	4.0±0.2	6.4±0.4	5.8±0.6	82.3±0.2	86.3±0.3	5.2m
(200,3,20)	3.2±0.3	3.1±0.9	9.7±1.2	4.0±1.1	**2.9±0.4**	7.2±0.6	4.6±1.1	87.8±0.2	91.8±0.3	10.1m
(200,5,30)	2.6±0.2	**2.2±0.9**	7.1±1.0	3.7±0.9	2.8±0.3	5.4±0.6	4.1±0.9	81.9±0.1	87.2±0.2	11.0m
(500,3,50)	**0.0±0.3**	3.1±1.1	9.6±1.2	4.2±1.2	**0.0±0.4**	6.7±0.6	4.1±1.2	87.6±0.1	91.6±0.2	26.3m

MDVI leads at most configurations, with gaps between 0.0% and 6.2%. *Timing*: the CPU column reports aggregate time per seed for all nine methods combined (single core). MDVI itself runs in $O(LNK)$ with L=500 iterations, taking a fraction of the total; Mitosis consumes most of the budget due to its 600s UCB search timeout. Among the competitors, NIP performs best (1.6–5.6%), followed by DQWIC (3.2–9.0%). Mitosis tells an interesting story: its UCB budget exploration needs many arms to converge, producing 10–41% gaps at $N \leq 50$ but reaching 1.3% at N=500, comparable to NIP at that scale. Greedy and the myopic baselines consistently fall into the churn trap (5–43% gap), confirming that the trap is a structural phenomenon rather than a matter of insufficient data.

6.4 What Is the Cost of Restricting the Policy Class?

To quantify the cost of restricting the policy class, we optimize six parameterized families via differential evolution, spanning the hierarchy from $\mathcal{F}_{\text{Lag}}$ to $\mathcal{F}_{\text{det}}$. Each family scores arms via $\iota_i(\mathbf{s}; \theta)$ and pulls the top B_k.

Table 3: Optimality gap (%) of parameterized policy families relative to g^*_{det}. Columns ordered by the policy class hierarchy $\mathcal{F}_{\text{Lag}} \subset \mathcal{F}_{\text{idx}} \subset \mathcal{F}_{\text{agg}} \subset \mathcal{F}_{\text{det}}$; d: number of parameters. **Bold**: best.

(N, K, B)	F_{linear} ($\mathcal{F}_{\text{Lag}}$, d=4K)	F_{interact} ($\mathcal{F}_{\text{Lag}}$, d=5K)	F_{pairwise} ($\mathcal{F}_{\text{Lag}}$, d=6K)	F_{thresh} ($\mathcal{F}_{\text{idx}}$, d=NK)	F_{dual} ($\mathcal{F}_{\text{agg}}$, d=K)	F_{neural} ($\mathcal{F}_{\text{agg}}$, d=73)
(5, 2, 2)	2.8±3.0	2.9±3.0	2.9±3.1	1.1±2.7	**0.4±2.7**	**0.4±2.7**
(5, 2, 3)	0.8±2.3	0.8±2.3	0.8±2.3	1.2±2.1	0.7±2.2	**0.5±2.2**
(5, 3, 2)	2.2±2.2	2.3±2.2	2.2±2.2	1.0±1.9	0.7±1.9	**0.5±1.9**
(8, 2, 2)	3.4±1.1	3.1±1.0	3.2±1.1	2.2±1.3	**1.5±1.3**	2.2±1.1
(8, 2, 3)	2.2±1.1	2.2±1.2	2.2±1.2	1.3±1.2	**0.3±1.2**	1.6±1.3

The results reveal three clear patterns. First, *separability is nearly free*: F_{dual} (Lagrangian dual search, only K parameters) achieves 0.01–0.22% gap from g^*_{det} across all configurations. This is the smallest class in the hierarchy, yet it nearly matches the unrestricted optimum. Second, *Bellman structure is essential*: F_{neural} (a shared multi-layer perceptron (MLP) with 73 parameters and access to global features $\bar{s}$, n_{churn}) achieves the same performance as F_{dual}, confirming that the Bellman value function (captured by RVI in F_{dual}) already provides the correct nonlinear score function; additional capacity adds nothing. Third, *interaction features are useless*: F_{interact} and F_{pairwise} add cross-arm features ($\bar{s}$, $n_c \cdot s_i$) to the linear score but achieve the same gap as F_{linear} (0.6–3.2%). The entire gap of linear policies comes from their inability to represent the Bellman value as a linear combination of features, not from missing interaction terms.

6.5 How Sensitive Are the Results to Problem Parameters?

We isolate the effect of each problem parameter (number of arms N, context count K, and budget ratio ρ) through three controlled experiments. All results report mean ± SE over 10 seeds.

(a) Scalability: varying N. Figure 2 fixes K=3, ρ=0.1 and scales N from 50 to 2000.

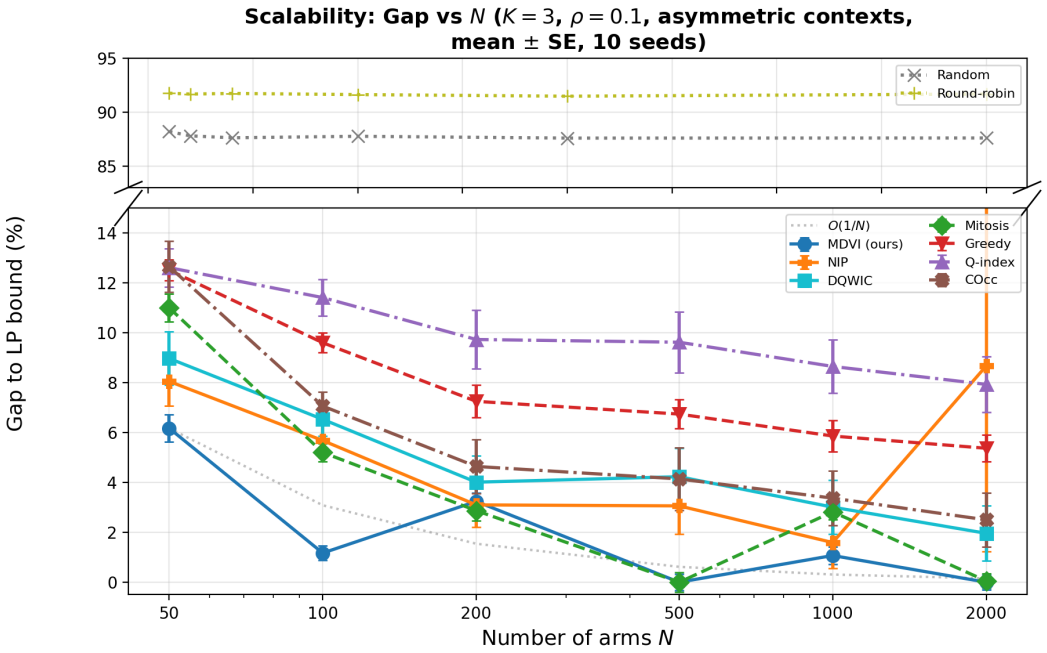

Figure 2: Gap (%) vs N (K=3, ρ=0.1, mean $\pm$ SE, 10 seeds). MDVI converges at $O(1/N)$; greedy plateaus at 5–6%. Full data in Table 7 (Appendix).

MDVI's gap tracks the predicted $O(1/N)$ rate, reaching near-zero at N=500+. Greedy, by contrast, plateaus near 5–7% even at N=2,000: the churn trap is structural, not a finite-sample effect. Mitosis begins to shine only at $N \geq 200$, where its UCB budget search has enough arms to converge (1.3% at N=500).

(b) Context count: varying K. Table 4 fixes N=500, ρ=0.1.

Table 4: Gap (%) vs K (N=500, ρ=0.1, asymmetric contexts, mean$\pm$SE, 10 seeds). **Bold**: best. CPU: aggregate per seed.

K	MDVI	NIP	DQWIC	Mitosis	Greedy	Q-idx	COcc	Rand	RR	CPU
1	**0.8**$\pm$0.3	1.4$\pm$0.4	3.7$\pm$0.4	1.9$\pm$0.3	1.2$\pm$0.3	5.6$\pm$0.5	2.7$\pm$0.3	87.1$\pm$0.1	96.5$\pm$0.1	24.7m
3	**0.0**$\pm$0.3	3.1$\pm$1.1	4.2$\pm$1.1	**0.0**$\pm$0.4	6.7$\pm$0.6	9.6$\pm$1.2	4.1$\pm$1.2	87.7$\pm$0.1	91.6$\pm$0.2	27.8m
5	1.3$\pm$0.3	2.5$\pm$0.9	3.2$\pm$0.8	**0.0**$\pm$0.4	4.8$\pm$0.5	8.7$\pm$0.9	3.2$\pm$0.9	87.3$\pm$0.1	90.3$\pm$0.3	30.2m
8	1.4$\pm$0.2	2.8$\pm$0.8	3.5$\pm$0.9	**0.0**$\pm$0.3	4.2$\pm$0.5	8.9$\pm$0.9	3.4$\pm$0.9	87.3$\pm$0.1	89.4$\pm$0.6	33.8m
10	1.7$\pm$0.2	2.8$\pm$0.8	3.3$\pm$0.9	**0.0**$\pm$0.2	4.1$\pm$0.5	8.6$\pm$0.9	3.3$\pm$0.8	87.2$\pm$0.1	89.2$\pm$0.3	36.4m
15	1.9$\pm$0.2	2.8$\pm$0.8	3.2$\pm$0.8	**0.0**$\pm$0.2	3.9$\pm$0.5	8.6$\pm$0.8	3.2$\pm$0.8	87.3$\pm$0.1	88.9$\pm$0.2	43.0m

MDVI is strikingly insensitive to context count, holding at 0.0–1.9% gap from K=1 through K=15. The Lagrangian decomposition absorbs additional contexts without degradation, a useful property for applications where the number of operating regimes may grow. *Timing*: aggregate CPU per seed grows from 25m at K=1 to 43m at K=15, dominated by Mitosis's 600s timeout per seed. MDVI's $O(LNK)$ cost scales linearly in K; for practitioners needing fast re-optimization, MDVI's speed advantage outweighs Mitosis's slight edge in optimality at large K.

(c) Budget ratio: varying $\rho = B/N$. Table 5 fixes N=500, K=3.

Table 5: Gap (%) vs ρ (N=500, K=3, asymmetric contexts, mean$\pm$SE, 10 seeds). **Bold**: best. CPU: aggregate per seed.

ρ	MDVI	NIP	DQWIC	Mitosis	Greedy	Q-idx	COcc	Rand	RR	CPU
0.02	**0.0**$\pm$0.2	2.7$\pm$1.5	11.5$\pm$1.4	**0.0**$\pm$0.3	11.9$\pm$0.3	17.1$\pm$1.5	9.3$\pm$1.5	95.5$\pm$0.1	96.3$\pm$0.1	27.7m
0.05	1.1$\pm$0.3	5.4$\pm$1.1	6.3$\pm$1.1	**0.0**$\pm$0.3	9.7$\pm$0.4	15.2$\pm$1.2	5.7$\pm$1.2	92.9$\pm$0.1	94.5$\pm$0.3	27.7m
0.10	**0.0**$\pm$0.3	3.1$\pm$1.1	4.2$\pm$1.1	**0.0**$\pm$0.4	6.7$\pm$0.6	9.6$\pm$1.2	4.1$\pm$1.2	87.6$\pm$0.1	91.6$\pm$0.2	27.9m
0.20	**0.0**$\pm$0.3	2.8$\pm$1.2	4.6$\pm$1.0	**0.0**$\pm$0.4	7.0$\pm$0.6	6.8$\pm$1.2	4.3$\pm$1.2	77.6$\pm$0.1	86.0$\pm$0.2	27.8m
0.30	**0.0**$\pm$0.4	2.7$\pm$1.2	4.8$\pm$0.9	**0.0**$\pm$0.4	6.6$\pm$0.6	5.8$\pm$1.2	3.7$\pm$1.3	68.5$\pm$0.2	79.0$\pm$0.2	27.7m
0.50	**0.0**$\pm$0.3	1.1$\pm$0.9	3.7$\pm$0.7	**0.0**$\pm$0.4	2.9$\pm$0.7	1.8$\pm$0.9	3.1$\pm$0.9	50.4$\pm$0.3	60.6$\pm$0.3	27.7m

The churn trap bites hardest at moderate budget (ρ=0.20), where greedy has enough slots to waste on churners but not enough slack to absorb the cost (7.0% gap). DQWIC struggles in the opposite regime: at scarce budget (ρ=0.02, 11.5% gap), getting the dual price right matters most, and the DQN approximation falls short. When budget is abundant (ρ=0.50) the problem becomes easy and all methods converge. *Timing*: MDVI's $O(LNK)$ cost is independent of ρ; Mitosis's UCB search scales with the budget allocation space, which grows with B.

Scalability. MDVI scales linearly in N due to its per-arm decomposition: 61 CPU-seconds per seed at N=50, 55 minutes at N=2,000, on a single core.

6.6 What Are the Key Takeaways?

The Lagrangian class $\mathcal{F}_{\text{Lag}}$, the most restrictive in our hierarchy, loses remarkably little: MDVI's gap shrinks from 6.2% at N=50 to near-zero at N=500+ (Table 7), tracking the $O(1/N)$ rate. The dominant source of suboptimality is not the index restriction ($< 0.2\%$) but myopic scoring, which falls into the churn trap (12–30% gap, Table 2). MDVI's linear-time decomposition provides near-optimal two-phase policies at scales where exact methods are intractable.

7 Related Work

Index policies for RMABs. Whittle's index (Whittle, 1988) assigns each arm a subsidy at which the planner is indifferent between activating and idling, then activates the top-B by subsidy. This Lagrangian relaxation decouples the N-arm problem into N single-arm subproblems, each solvable in closed form for binary-state arms. Weber & Weiss (1990) proved asymptotic optimality as $N \to \infty$ under exchangeability via a fluid-limit argument. Gast et al. (2024; 2023) sharpen this to $O(e^{-cN})$ under UGAP (a unique globally attracting fixed point for the mean-field ODE), non-degeneracy of the LP relaxation, and global exponential stability. Hong et al. (2024) remove UGAP, requiring only unichain and aperiodicity at a weaker $O(\sqrt{N})$ rate. All assume a *fixed* budget B; in our setting the per-context budgets $\{B_k\}$ are decision variables coupled through a global constraint, a two-phase structure no prior index theory addresses.

Multi-constraint and multi-context RMABs. Fu & Lu (2024) study multi-action restless bandits with weakly coupled constraints, proving convergence of a learning-and-control algorithm. Their approach handles constraint satisfaction but does not optimize over constraint right-hand sides (budget allocation is given). Goldsztajn et al. (2024) extend optimal control to weakly coupled MDPs, showing decomposition across subsystems. Neither considers the combinatorial upper-level budget optimization that distinguishes our formulation: choosing how to split a global capacity across contexts, not just satisfying pre-specified per-context constraints.

Factored MDPs and decomposition. Lee et al. (2025) prove that separable effects (conditional independence given actions) imply additive Q-decomposition, making the optimal policy a state-dependent index. Our $\mathcal{F}_{\text{agg}}$ refines this under exchangeability: the state-dependence reduces to aggregate counts per type, yielding a policy class polynomial in N for fixed number of types T, rather than exponential in the joint state dimension $|S|^N$.

Contextual and online bandits. The contextual bandit literature studies action selection conditioned on side information, but without restless dynamics or hard budget constraints. Our work bridges contextual bandits (context modulates rewards) and restless bandits (arms evolve autonomously), adding a budget allocation layer that neither addresses alone.

Applications. Shi et al. (2024) introduced the contextual budget bandit for 412 Food Rescue, a Pittsburgh nonprofit matching surplus food with delivery volunteers. The multi-context structure arises from geographic regions with distinct volunteer populations; the churn trap reflects real notification fatigue. Liang et al. (2025) apply contextual restless bandits to public health in underserved communities, demonstrating that context-aware budget decisions improve outcomes when populations differ across geographic and demographic strata.

8 Conclusion

We have introduced the two-phase contextual budget bandit and the first policy class hierarchy for budget-allocation-aware RMABs. The $\mathcal{F}_{\text{agg}}$ optimality theorem shows that aggregate engagement counts are a sufficient statistic for optimal arm selection under exchangeability, and the portfolio effect shows simpler index classes miss genuine structure for $N \geq 3$. MDVI closes the gap to the LP bound from 6.2% at N=50 to near-zero at N=500+, scaling linearly. Extending to heterogeneous arms, multi-state dynamics, and online learning are natural next steps.

AI Use Statement

In this work, we used a generative AI assistant for LaTeX document preparation, literature survey, proof exploration, and code generation for counterexample search scripts. All theoretical results, proofs, and experimental designs were conceived by the authors. All AI-generated proofs were manually verified for correctness. All AI-generated code was tested and validated by the authors before use. We take responsibility for the final content of this work.

Ethics Statement

This work develops theoretical and computational methods for resource allocation in restless bandits, motivated by food rescue volunteer engagement. The methods are general-purpose and do not involve human subjects data. The 412 Food Rescue application uses synthetic instances calibrated to published parameters, not individual volunteer data.

Reproducibility Statement

All algorithms are implemented in Python with NumPy and SciPy. The codebase includes 1,126 unit tests with 95%+ statement coverage. Counterexample search scripts are provided as supplementary material. Proofs of all theoretical results appear in Appendix A. Experimental configurations and random seeds are specified in the supplementary code.

References

J. Fu and Y. Lu. Multi-action restless bandits with weakly coupled constraints. *arXiv preprint arXiv:2412.03326*, 2024.

N. Gast, B. Gaujal, and C. Yan. Exponential asymptotic optimality of Whittle index policy. *Queueing Systems*, 104:107–150, 2023.

N. Gast, B. Gaujal, and C. Yan. LP-based policies for restless bandits: Necessary and sufficient conditions for asymptotic optimality. *Mathematics of Operations Research*, 2024.

D. Goldsztajn, A. Mate, and S. Mahadevan. An optimal-control approach to infinite-horizon restless bandits. *arXiv preprint arXiv:2403.11913*, 2024.

Y. Hong, Q. Xie, Y. Chen, and W. Wang. Unichain and aperiodicity are sufficient for asymptotic optimality of average-reward restless bandits. *arXiv preprint arXiv:2402.05689*, 2024.

J. Lee et al. Q-function decomposition with intervention semantics for factored action spaces. In *AISTATS*, 2025.

B. Liang et al. Context in public health for underserved communities. In *AAAI*, 2025.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

Z. R. Shi, J. Zhi, S. Zeng, Z. Zhang, A. Kapoor, S. Hudson, H. Shen, and F. Fang. Contextual budget bandit for fair food rescue engagement. *Preprint*, 2024.

R. R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

X. Zhang and P. I. Frazier. Restless bandits with many arms: Beating the central limit theorem. *arXiv preprint arXiv:2107.11911*, 2021.

A Proofs of Main Results

A.1 Preliminaries and Notation

We collect the notation and definitions needed for the proofs.

The MDP. The two-phase CBB operates on a joint state space $\mathcal{S} = \{0,1\}^N$, where $s_i = 1$ means arm i is "engaged" and $s_i = 0$ means "disengaged." At each period, a context $k \in [K]$ is drawn from distribution $f = (f_1, \ldots, f_K)$, and the decision-maker chooses an action $\mathbf{a} = (a_1, \ldots, a_N) \in \{0,1\}^N$ with $\|\mathbf{a}\|_1 \leq B_k$ (at most B_k arms activated). The per-arm transition kernels satisfy *conditional independence*:

$$P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}; k) = \prod_{i=1}^{N} P_i(s_i' \mid s_i, a_i; k). \tag{5}$$

Rewards are additive: $R(\mathbf{s}, \mathbf{a}; k) = \sum_i R_i(s_i, a_i; k)$.

Exchangeability. Under T exchangeable types, the arms are partitioned into groups $\mathcal{T}_1, \ldots, \mathcal{T}_T$ with $|\mathcal{T}_t| = N_t$. All arms in group $\mathcal{T}_t$ share common dynamics: $P_i(\cdot \mid \cdot, \cdot; k) = P_t(\cdot \mid \cdot, \cdot; k)$ and $R_i(\cdot, \cdot; k) = R_t(\cdot, \cdot; k)$ for all $i \in \mathcal{T}_t$.

Symmetry group. Let S_{N_t} denote the symmetric group on N_t elements (the group of all permutations of $\{1, \ldots, N_t\}$). The *within-type permutation group* is $G = \prod_{t=1}^{T} S_{N_t}$, acting on arm indices. For $\sigma \in G$, define the action on joint states by $[\sigma(\mathbf{s})]_i = s_{\sigma^{-1}(i)}$ (permute the components) and similarly on actions $[\sigma(\mathbf{a})]_i = a_{\sigma^{-1}(i)}$. Critically, σ only permutes arms *within* each type: if $i, j \in \mathcal{T}_t$ then σ may swap i and j, but σ never maps an arm in $\mathcal{T}_t$ to an arm in $\mathcal{T}_{t'}$ for $t \neq t'$.

Aggregate state. The aggregate count vector is $\mathbf{n}(\mathbf{s}) = (n_1, \ldots, n_T)$ where $n_t = \sum_{i \in \mathcal{T}_t} s_i$ counts the engaged arms of type t. The aggregate state space is $\mathcal{N} = \prod_{t=1}^{T} \{0, 1, \ldots, N_t\}$, with $|\mathcal{N}| = \prod_t (N_t + 1)$. An important consequence of exchangeability is that two joint states $\mathbf{s}$ and $\mathbf{s}'$ have the same aggregate counts ($\mathbf{n}(\mathbf{s}) = \mathbf{n}(\mathbf{s}')$) if and only if they are related by some permutation $\sigma \in G$. This is because within each type, the engaged/disengaged arms can be permuted freely.

A.2 Proof of Lemma 4 (Within-Type Permutation Invariance of V^*)

Statement. For any fixed budget allocation $\mathbf{B}$, the optimal value function satisfies $V^*(\sigma(\mathbf{s})) = V^*(\mathbf{s})$ for all within-type permutations $\sigma \in G$; that is, V^* is invariant under the symmetry group $G = \prod_{t=1}^{T} S_{N_t}$.

Proof. The argument has three parts: (i) verify that (P, R) respect within-type symmetry, (ii) show the Bellman operator preserves this symmetry, and (iii) conclude via contraction. *Step 1: Symmetry of (P, R).* For rewards:

$$\begin{aligned} R(\sigma(\mathbf{s}), \sigma(\mathbf{a}); k) &= \sum_i R_{\text{type}(i)}([\sigma(\mathbf{s})]_i, [\sigma(\mathbf{a})]_i; k) \\ &= \sum_i R_{\text{type}(i)}(s_{\sigma^{-1}(i)}, a_{\sigma^{-1}(i)}; k) \\ &= \sum_j R_{\text{type}(\sigma(j))}(s_j, a_j; k) \end{aligned} \tag{6}$$

where we substituted $j = \sigma^{-1}(i)$. Since σ only permutes within types, $\text{type}(\sigma(j)) = \text{type}(j)$, so equation 6 equals $\sum_j R_{\text{type}(j)}(s_j, a_j; k) = R(\mathbf{s}, \mathbf{a}; k)$.

Second, for transitions. By conditional independence equation 5:

$$\begin{aligned} P(\sigma(\mathbf{s}') \mid \sigma(\mathbf{s}), \sigma(\mathbf{a}); k) &= \prod_i P_{\text{type}(i)}(s'_{\sigma^{-1}(i)} \mid s_{\sigma^{-1}(i)}, a_{\sigma^{-1}(i)}; k) \\ &= \prod_j P_{\text{type}(\sigma(j))}(s'_j \mid s_j, a_j; k) \\ &= \prod_j P_{\text{type}(j)}(s'_j \mid s_j, a_j; k) \\ &= P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}; k) \end{aligned} \tag{7}$$

where again $\text{type}(\sigma(j)) = \text{type}(j)$ was used. *Step 2: Bellman operator preserves invariance.* The

discounted Bellman optimality operator is:

$$[BV](\mathbf{s}) = \max_{\mathbf{a}:\|\mathbf{a}\|_1 \le B_k} \left[R(\mathbf{s}, \mathbf{a}; k) + \gamma \sum_{\mathbf{s}' \in \{0,1\}^N} P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}; k)\, V(\mathbf{s}') \right].$$

Assume V is G-invariant, i.e., $V(\sigma(\mathbf{s})) = V(\mathbf{s})$ for all $\sigma \in G$. We show BV is also G-invariant. For any $\sigma \in G$:

$$[BV](\sigma(\mathbf{s})) = \max_{\mathbf{a}:\|\mathbf{a}\|_1 \le B_k} \left[R(\sigma(\mathbf{s}), \mathbf{a}; k) + \gamma \sum_{\mathbf{s}'} P(\mathbf{s}' \mid \sigma(\mathbf{s}), \mathbf{a}; k)\, V(\mathbf{s}') \right]. \tag{8}$$

We perform two changes of variables. *Change of action variable.* Substitute $\mathbf{a} = \sigma(\tilde{\mathbf{a}})$. Since σ is a bijection on $\{0,1\}^N$ and preserves the ℓ_1 norm ($\|\sigma(\tilde{\mathbf{a}})\|_1 = \|\tilde{\mathbf{a}}\|_1$), the feasible set is unchanged. The maximum over $\mathbf{a}$ becomes a maximum over $\tilde{\mathbf{a}}$, so equation 8 becomes:

$$= \max_{\tilde{\mathbf{a}}:\|\tilde{\mathbf{a}}\|_1 \le B_k} \left[R(\sigma(\mathbf{s}), \sigma(\tilde{\mathbf{a}}); k) + \gamma \sum_{\mathbf{s}'} P(\mathbf{s}' \mid \sigma(\mathbf{s}), \sigma(\tilde{\mathbf{a}}); k)\, V(\mathbf{s}') \right]. \tag{9}$$

Change of state variable. In the sum over $\mathbf{s}'$, substitute $\mathbf{s}' = \sigma(\tilde{\mathbf{s}}')$. Since σ is a bijection on $\{0,1\}^N$, summing over $\mathbf{s}'$ is equivalent to summing over $\tilde{\mathbf{s}}'$, giving from equation 9:

$$= \max_{\tilde{\mathbf{a}}} \left[R(\sigma(\mathbf{s}), \sigma(\tilde{\mathbf{a}}); k) + \gamma \sum_{\tilde{\mathbf{s}}'} P(\sigma(\tilde{\mathbf{s}}') \mid \sigma(\mathbf{s}), \sigma(\tilde{\mathbf{a}}); k)\, V(\sigma(\tilde{\mathbf{s}}')) \right]. \tag{10}$$

Apply symmetry. By Step 1: $R(\sigma(\mathbf{s}), \sigma(\tilde{\mathbf{a}}); k) = R(\mathbf{s}, \tilde{\mathbf{a}}; k)$ and $P(\sigma(\tilde{\mathbf{s}}') \mid \sigma(\mathbf{s}), \sigma(\tilde{\mathbf{a}}); k) = P(\tilde{\mathbf{s}}' \mid \mathbf{s}, \tilde{\mathbf{a}}; k)$. By the G-invariance assumption on V: $V(\sigma(\tilde{\mathbf{s}}')) = V(\tilde{\mathbf{s}}')$. Substituting these identities into equation 10:

$$= \max_{\tilde{\mathbf{a}}} \left[R(\mathbf{s}, \tilde{\mathbf{a}}; k) + \gamma \sum_{\tilde{\mathbf{s}}'} P(\tilde{\mathbf{s}}' \mid \mathbf{s}, \tilde{\mathbf{a}}; k)\, V(\tilde{\mathbf{s}}') \right] = [BV](\mathbf{s}). \tag{11}$$

Therefore $[BV](\sigma(\mathbf{s})) = [BV](\mathbf{s})$, so BV is G-invariant whenever V is. *Step 3: Contraction argument.* The initial value function $V_0 \equiv 0$ is trivially G-invariant. Define the sequence $V_{n+1} = BV_n$. By Step 2, if V_n is G-invariant then V_{n+1} is G-invariant. By induction, V_n is G-invariant for all n.

The Bellman operator B is a contraction mapping with modulus $\gamma < 1$ under the supremum norm (Puterman, 1994). Therefore the sequence V_n converges uniformly to the unique fixed point $V^* = BV^*$. Since each V_n is G-invariant and the pointwise limit of G-invariant functions is G-invariant, V^* is G-invariant. *Extension to average reward.* The average-reward optimality equation is $g^* + h^*(\mathbf{s}) = \max_{\mathbf{a}}[R(\mathbf{s}, \mathbf{a}) + \sum_{\mathbf{s}'} P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) h^*(\mathbf{s}')]$, where g^* is the optimal gain (a scalar) and h^* is the differential (bias) value. Under the standard vanishing-discount approach (Puterman, 1994), h^* can be obtained as $h^*(\mathbf{s}) = \lim_{\gamma \to 1}[V_\gamma^*(\mathbf{s}) - V_\gamma^*(\mathbf{s}_{\text{ref}})]$ for any reference state $\mathbf{s}_{\text{ref}}$. Since V_γ^* is G-invariant for every $\gamma < 1$ (by the argument above), the limit h^* inherits permutation invariance: $h^*(\sigma(\mathbf{s})) = h^*(\mathbf{s})$. □

A.3 Proof of Theorem 5 ($\mathcal{F}_{\text{agg}}$ Optimality)

Statement. For the average-reward CBB with T exchangeable types, conditional independence, and ergodic per-arm chains: the optimal two-phase policy is in $\mathcal{F}_{\text{agg}}$. That is, the optimal action at joint state $\mathbf{s}$ under context k depends on $\mathbf{s}$ only through the aggregate count vector $\mathbf{n}(\mathbf{s})$.

Proof. We proceed in four steps, showing that (i) V^* depends only on $\mathbf{n}$, (ii) immediate rewards decompose by type, (iii) continuation values decompose by type, and (iv) the optimal action inherits this dependence. By Lemma 4, $V^*(\sigma(\mathbf{s})) = V^*(\mathbf{s})$ for all $\sigma \in G$. As noted in the preliminaries, two states $\mathbf{s}$ and $\mathbf{s}'$ satisfy $\mathbf{n}(\mathbf{s}) = \mathbf{n}(\mathbf{s}')$ if and only if $\mathbf{s}' = \sigma(\mathbf{s})$ for some $\sigma \in G$. Therefore $V^*(\mathbf{s}) = V^*(\mathbf{s}')$ whenever $\mathbf{n}(\mathbf{s}) = \mathbf{n}(\mathbf{s}')$. We may write $V^*(\mathbf{n})$ without ambiguity. *Step (ii): Reward decomposition.* The immediate reward at $(\mathbf{s}, \mathbf{a}; k)$ is:

$$R(\mathbf{s}, \mathbf{a}; k) = \sum_{i=1}^{N} R_{\text{type}(i)}(s_i, a_i; k) = \sum_{t=1}^{T} \sum_{i \in \mathcal{T}_t} R_t(s_i, a_i; k). \tag{12}$$

Within type t, each arm has $s_i \in \{0,1\}$ and $a_i \in \{0,1\}$. The sum $\sum_{i\in\mathcal{T}_t} R_t(s_i, a_i; k)$ depends on how many arms fall into each of the four (s, a) categories. Define:

$$\ell_t = \sum_{i\in\mathcal{T}_t} s_i \cdot a_i \qquad \text{(engaged and pulled)}$$

$$b_t = \sum_{i\in\mathcal{T}_t} a_i \qquad \text{(total pulled)}$$

$$n_t = \sum_{i\in\mathcal{T}_t} s_i \qquad \text{(total engaged)}$$

Then the four counts are: ℓ_t arms with $(s{=}1, a{=}1)$; $(n_t - \ell_t)$ with $(s{=}1, a{=}0)$; $(b_t - \ell_t)$ with $(s{=}0, a{=}1)$; and $(N_t - n_t - b_t + \ell_t)$ with $(s{=}0, a{=}0)$. Therefore:

$$\sum_{i\in\mathcal{T}_t} R_t(s_i, a_i; k) = \ell_t\, R_t(1,1;k) + (n_t - \ell_t)\, R_t(1,0;k) + (b_t - \ell_t)\, R_t(0,1;k) + (N_t - n_t - b_t + \ell_t)\, R_t(0,0;k). \tag{13}$$

This depends on $(\mathbf{s}, \mathbf{a})$ only through (n_t, b_t, ℓ_t) for each type t. *Step (iii): Continuation value decomposition.* The continuation value is:

$$C(\mathbf{s}, \mathbf{a}; k) = \sum_{\mathbf{s}'\in\{0,1\}^N} P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}; k)\, V^*(\mathbf{n}(\mathbf{s}')). \tag{14}$$

By conditional independence equation 5, $P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}; k) = \prod_i P_{\text{type}(i)}(s_i' \mid s_i, a_i; k)$. Since V^* depends on $\mathbf{s}'$ only through $\mathbf{n}' = \mathbf{n}(\mathbf{s}')$, we can group the sum by aggregate next-state $\mathbf{n}'$:

$$C = \sum_{\mathbf{n}'\in\mathcal{N}} V^*(\mathbf{n}') \cdot \Pr[\mathbf{n}(\mathbf{s}') = \mathbf{n}' \mid \mathbf{s}, \mathbf{a}; k]. \tag{15}$$

The key step: under conditional independence, the next-state counts $n_1', \ldots, n_T'$ are *independent* across types (conditional on the current state and action). To see this, note that $n_t' = \sum_{i\in\mathcal{T}_t} s_i'$ is a function of $\{s_i' : i \in \mathcal{T}_t\}$, and by conditional independence equation 1, the random vectors $\{s_i' : i \in \mathcal{T}_t\}$ and $\{s_j' : j \in \mathcal{T}_{t'}\}$ are independent for $t \neq t'$ (they are functions of disjoint sets of independent random variables). Therefore:

$$\Pr[\mathbf{n}' \mid \mathbf{s}, \mathbf{a}; k] = \prod_{t=1}^{T} \Pr[n_t' \mid \{(s_i, a_i)\}_{i\in\mathcal{T}_t}; k]. \tag{16}$$

Now, within type t, the next-state count $n_t' = \sum_{i\in\mathcal{T}_t} s_i'$ is a sum of independent Bernoulli random variables. The distribution of each s_i' depends only on (s_i, a_i) through $P_t(1 \mid s_i, a_i; k)$. Since all arms in $\mathcal{T}_t$ share the same kernel P_t, the distribution of n_t' depends on $\{(s_i, a_i)\}_{i\in\mathcal{T}_t}$ only through the counts (n_t, b_t, ℓ_t):

$$\begin{aligned}\Pr[n_t' \mid n_t, b_t, \ell_t; k] = \sum_{\substack{j_1+j_2+j_3+j_4=n_t' \\ 0\le j_1\le \ell_t, \ldots}} & \binom{\ell_t}{j_1} [P_t(1 \mid 1,1;k)]^{j_1} [1 - P_t(1 \mid 1,1;k)]^{\ell_t - j_1} \\ & \times \binom{n_t - \ell_t}{j_2} [P_t(1 \mid 1,0;k)]^{j_2} [1 - P_t(1 \mid 1,0;k)]^{n_t - \ell_t - j_2} \\ & \times \binom{b_t - \ell_t}{j_3} [P_t(1 \mid 0,1;k)]^{j_3} [1 - P_t(1 \mid 0,1;k)]^{b_t - \ell_t - j_3} \\ & \times \binom{N_t - n_t - b_t + \ell_t}{j_4} [P_t(1 \mid 0,0;k)]^{j_4} \cdots \end{aligned} \tag{17}$$

This is a convolution of four binomial distributions, each determined by (n_t, b_t, ℓ_t) and the type-t kernel parameters. Combining equation 15, equation 16, and equation 17: the continuation value $C(\mathbf{s}, \mathbf{a}; k)$ depends on $(\mathbf{s}, \mathbf{a})$ only through $\{(n_t, b_t, \ell_t)\}_{t=1}^T$. Combining the reward and continuation value decompositions, we conclude that the optimal action depends only on $\mathbf{n}$.

The Q-function at $(\mathbf{s}, \mathbf{a}; k)$ is:

$$Q(\mathbf{s}, \mathbf{a}; k) = R(\mathbf{s}, \mathbf{a}; k) + \gamma\, C(\mathbf{s}, \mathbf{a}; k).$$

By Parts 2 and 3, both the reward and the continuation value depend on $(\mathbf{s}, \mathbf{a})$ only through $\{(n_t, b_t, \ell_t)\}_{t=1}^T$. Therefore:

$$Q(\mathbf{s}, \mathbf{a}; k) = \tilde{Q}(\{(n_t, b_t, \ell_t)\}_{t=1}^T; k)$$

for some function $\tilde{Q}$. The optimal action is:

$$\mathbf{a}^*(\mathbf{s}; k) = \arg \max_{\mathbf{a}: \|\mathbf{a}\|_1 \le B_k} Q(\mathbf{s}, \mathbf{a}; k) = \arg \max_{\{b_t, \ell_t\}} \tilde{Q}(\{(n_t, b_t, \ell_t)\}; k)$$

where the second equality holds because different actions $\mathbf{a}$ with the same per-type summary (b_t, ℓ_t) yield the same Q-value.

The optimal aggregate allocation $(b_t^*, \ell_t^*)_{t=1}^T$ is a function of $(\mathbf{n}, k)$ only (not of the specific arm-level state $\mathbf{s}$). Given this allocation, the specific arms to activate within each type are selected by arbitrary tie-breaking among identical arms (any selection achieving the target counts (b_t^*, ℓ_t^*) yields the same Q-value).

Upper-level optimization. The budget allocation $\mathbf{B} = (B_1, \ldots, B_K)$ is chosen from the finite set $\{\mathbf{B} \in \mathbb{Z}_+^K : \sum_k f_k B_k \le B_{\text{tot}}\}$. For each candidate $\mathbf{B}$, the lower-level optimal policy is in $\mathcal{F}_{\text{agg}}$ (by the argument above, which holds for any fixed $\mathbf{B}$). The joint optimum $(\mathbf{B}^*, \pi_{\text{agg}}^*(\mathbf{B}^*))$ is therefore in $\mathcal{F}_{\text{agg}}$. □

A.4 Proof of Proposition 6 (Common Shadow Price)

Statement. At the optimal Lagrangian dual of the two-phase CBB, the per-context budget shadow prices satisfy $w_k^* = \mu^* f_k$ for all active contexts (those with $B_k^* > 0$), where $\mu^* \ge 0$ is a single scalar.

Proof. The two-phase CBB equation 2 has two sets of constraints:

1. Per-context budget: $\|\mathbf{a}\|_1 \le B_k$ when context k occurs.
2. Global budget: $\sum_{k=1}^K f_k B_k \le B_{\text{tot}}$.

Introduce Lagrange multipliers $w_k \ge 0$ for each per-context budget constraint and $\mu \ge 0$ for the global constraint. The Lagrangian is:

$$L(\pi, \mathbf{B}, \mathbf{w}, \mu) = g(\pi, \mathbf{B}) + \sum_{k=1}^K w_k \big(B_k - \mathbb{E}[\|\mathbf{a}\|_1 \mid k]\big) + \mu\big(B_{\text{tot}} - \sum_{k=1}^K f_k B_k\big). \tag{18}$$

We rearrange equation 18 by grouping terms. The average reward is $g(\pi, \mathbf{B}) = \sum_k f_k \mathbb{E}[\sum_i R_i(s_i, a_i; k)]$. Combining with the per-context terms:

$$L = \sum_{k=1}^K f_k \mathbb{E}\left[\sum_{i=1}^N \Big(R_i(s_i, a_i; k) - \frac{w_k}{f_k} \cdot a_i\Big)\right] + \sum_{k=1}^K (w_k - \mu f_k)\, B_k + \mu B_{\text{tot}}. \tag{19}$$

The crucial observation is that the term $\sum_k (w_k - \mu f_k) B_k$ is linear in B_k with coefficient $(w_k - \mu f_k)$. Since the dual problem minimizes $d(\mathbf{w}, \mu) = \max_{\pi, \mathbf{B}} L$ over $(\mathbf{w}, \mu)$, we analyze the effect of the coefficient sign on the inner maximization over B_k.

If $w_k - \mu f_k > 0$ for some k, the inner maximization sets $B_k = N$ (its upper bound), making the dual function unnecessarily large. Reducing w_k to μf_k removes this positive term without affecting the per-arm subproblems (which depend on w_k / f_k, not B_k), thus weakly improving (reducing) the dual. Therefore at the dual optimum:

$$w_k^* - \mu^* f_k \le 0 \quad \forall\, k. \tag{20}$$

Case 2: If $w_k^* - \mu^* f_k < 0$, then the coefficient of B_k is negative, so the supremum over B_k sets $B_k^* = 0$ (no budget allocated to context k). *Case 3:* If $w_k^* - \mu^* f_k = 0$, then the B_k term vanishes and B_k is determined by the per-context optimality conditions. Combining Cases 2 and 3 (complementary slackness): for every active context ($B_k^* > 0$), we must have $w_k^* = \mu^* f_k$. This is the common shadow price property. This has an immediate consequence for the per-arm subproblems.

Substituting $w_k^* = \mu^* f_k$ into equation 19, the per-arm Lagrangian reward becomes:

$$R_i(s_i, a_i; k) - \frac{\mu^* f_k}{f_k} \cdot a_i = R_i(s_i, a_i; k) - \mu^* a_i$$

which is independent of k (up to the context-dependent transition kernel $P_i(\cdot; k)$). The modified reward penalizes pulling by a universal cost μ^*, regardless of context. This reduces the two-phase Lagrangian to N independent per-arm MDPs, each parametrized by the single scalar μ^*. □

A.5 Proof of Theorem 7 (Two-Phase Bound)

Statement. Under the conditions of Theorem 5 with non-degenerate LP relaxation:

$$g^*_{\text{bi-level}} - g(\pi_{\mathcal{F}_{\text{Lag}}}, \mathbf{B}^*) \leq \frac{K \cdot R^{\text{arm}}_{\max}}{N} + K \cdot C\, e^{-c \min_t N_t}$$

where $R^{\text{arm}}_{\max} = \max_{i,s,a,k} |R_i(s,a;k)|$ and $C, c > 0$ are constants depending on per-type parameters but not on N.

Proof. We decompose the total gap into three sources: the Lagrangian relaxation gap, the budget rounding loss, and the per-context index gap. The first step is an upper bound from the Lagrangian relaxation.

The LP relaxation of the two-phase CBB replaces the hard per-period budget $\|\mathbf{a}\|_1 \leq B_k$ with a constraint on expected activations. By weak LP duality (the Lagrangian relaxation provides an upper bound on the constrained problem), the dual function at μ^* satisfies:

$$g^*_{\text{bi-level}} \leq q(\mu^*) = \sum_{i=1}^{N} q_i(\mu^*) + \mu^* B_{\text{tot}} \tag{21}$$

where $q_i(\mu^*)$ is the optimal value of arm i's per-arm MDP with modified reward $R_i(s_i, a_i; k) - \mu^* a_i$. This follows from weak LP duality (the Lagrangian relaxation removes the coupling constraint, so each arm can be optimized independently). The second source of error is budget rounding.

At the dual optimum μ^*, each per-arm subproblem determines a context-specific pull probability $\pi^L_i(s_i, k; \mu^*)$. The expected per-context demand is:

$$D^L_k = \sum_{i=1}^{N} \mathbb{E}_{\bar{s}}[\pi^L_i(s_i, k; \mu^*)]$$

where $\bar{s}$ is the stationary distribution of arm i under the Lagrangian policy. Complementary slackness at μ^* gives $\sum_k f_k D^L_k = B_{\text{tot}}$.

The optimal integer budget allocation $\mathbf{B}^*$ satisfies $B^*_k \in \mathbb{Z}_+$ and $\sum_k f_k B^*_k \leq B_{\text{tot}}$. By concavity of the per-context value function $g^*_k(B_k)$ in B_k (diminishing returns from additional budget, proved by a coupling argument: adding one more activation slot cannot increase the marginal value), the optimal integer allocation satisfies $|B^*_k - D^L_k| \leq 1$ for each k.

The reward loss from rounding: each misallocated unit means one arm is pulled (or not pulled) incorrectly in one context. The per-period reward change is at most $R^{\text{arm}}_{\max}$ per misallocated unit. With at most K contexts affected (one unit each), the total per-period rounding loss is at most $K \cdot R^{\text{arm}}_{\max}$. Since the average reward is a per-step quantity over N arms, the contribution to the per-arm average gap is:

$$\Delta_{\text{round}} \leq \frac{K \cdot R^{\text{arm}}_{\max}}{N}. \tag{22}$$

The third and final source of error is the per-context index approximation.

Given the budget allocation $\mathbf{B}^*$, each context k faces a standard restless bandit with N arms and budget B^*_k. The Lagrangian priority policy at μ^* ranks arms by their per-arm scores $\iota_i(s_i, k; \mu^*) = Q_i(s_i, 1; k, \mu^*) - Q_i(s_i, 0; k, \mu^*)$ and activates the top B^*_k.

By the exponential optimality result of Gast et al. (2024) (Theorem 1 therein), under three conditions (verified in Appendix B):

1. **UGAP** (Uniform Global Attractor Property): the mean-field ordinary differential equation (ODE) has a unique globally attracting fixed point.
2. **Non-degeneracy**: at most one arm type is fractionally activated at the LP optimum ($\delta_{\min} > 0$).
3. **GES** (Global Exponential Stability): the mean-field fixed point is exponentially stable with rate $\gamma > 0$.

the gap between the LP-priority policy and the LP upper bound is at most $Ce^{-c \min_t N_t}$ per context, where C and c depend on the per-type transition parameters, the non-degeneracy margin $\delta_{\min}$, and the exponential stability rate γ, but not on N.

Summing over contexts (weighted by f_k):

$$\Delta_{\text{index}} = \sum_{k=1}^{K} f_k[g_k^*(B_k^*) - g_k(\pi_k^{\text{priority}})] \leq K \cdot C\, e^{-c \min_t N_t}. \tag{23}$$

Combining these three sources of error: The total gap is:

$$\begin{aligned} g^*_{\text{bi-level}} - g(\pi_{\mathcal{F}_{\text{Lag}}}, \mathbf{B}^*) &\leq \Delta_{\text{round}} + \Delta_{\text{index}} \\ &\leq \frac{K \cdot R^{\text{arm}}_{\max}}{N} + K \cdot C\, e^{-c \min_t N_t}. \end{aligned} \tag{24}$$

For large N, the exponential term is dominated by the $O(K/N)$ rounding term, giving an overall rate of $O(K/N)$. When K is fixed and $N \to \infty$, the gap vanishes as $O(1/N)$. □

Remark 8. *The factor K in the exponential term is conservative. Since context frequencies satisfy $\sum_k f_k = 1$, the weighted sum $\sum_k f_k C_k e^{-c_k \min_t N_t} \leq C e^{-c \min_t N_t}$ (taking worst-case constants $C = \max_k C_k$, $c = \min_k c_k$) does not require the K factor. We retain it for simplicity, noting that the dominant $O(K/N)$ rounding term is unchanged.*

A.6 Proof of MDVI Convergence

The main text states that MDVI converges at rate $O(N\sqrt{K/L})$. We provide the proof here.

MDVI performs projected subgradient ascent on the Lagrangian dual function $q(\mathbf{w}) = \sum_{i=1}^{N} q_i(\mathbf{w}) + \sum_k w_k B_k$, where $q_i(\mathbf{w})$ is the optimal value of arm i's per-arm MDP with modified reward $R_i(s_i, a_i; k) - w_k a_i$.

The subgradient at $\mathbf{w}$ is $d_k = B_k - \hat{D}_k(\mathbf{w})$, where $\hat{D}_k(\mathbf{w}) = \sum_i \pi_i^*(s_i, k; \mathbf{w})$ is the total demand under context k at the per-arm optimal policies. The update is:

$$w_k^{\ell+1} = \left[w_k^\ell - \alpha_\ell\, d_k^\ell\right]_+$$

where $[\cdot]_+$ denotes projection onto $\mathbb{R}_+$.

By the standard convergence theorem for projected subgradient methods (Puterman, 1994, Chapter 6), if the step sizes satisfy $\alpha_\ell = \alpha/\sqrt{L}$ for some $\alpha > 0$, then:

$$\min_{\ell \leq L} q(\mathbf{w}^\ell) - q^* \leq \frac{\|\mathbf{w}^0 - \mathbf{w}^*\|^2}{2\alpha\sqrt{L}} + \frac{\alpha\, G^2}{2\sqrt{L}}$$

where G is an upper bound on the subgradient norm. Since $|d_k| \leq N$ for each of the K components, $\|d\| \leq N\sqrt{K}$, giving $G = N\sqrt{K}$. Optimizing over α yields:

$$\min_{\ell \leq L} q(\mathbf{w}^\ell) - q^* \leq \frac{\|\mathbf{w}^0 - \mathbf{w}^*\| \cdot N\sqrt{K}}{\sqrt{L}} = O\left(\frac{N\sqrt{K}}{\sqrt{L}}\right).$$

A.7 Concavity of the Per-Context Value Function

The two-phase bound proof uses the fact that $g_k^*(B_k)$ (the optimal average reward under context k with budget B_k) is concave in B_k. We prove this here.

Lemma 9. *For each context k, the function $g_k^* : \{0, 1, \ldots, N\} \to \mathbb{R}$ mapping budget B_k to the optimal average reward is concave: $g_k^*(B_k + 1) - g_k^*(B_k) \geq g_k^*(B_k + 2) - g_k^*(B_k + 1)$ for all $B_k \in \{0, \ldots, N - 2\}$.*

Proof. Consider the LP relaxation of the per-context RMAB at budget B_k. The LP optimal value $g_k^{\text{LP}}(B_k)$ is concave in B_k because it is the pointwise infimum of a family of affine functions of B_k (one for each dual multiplier $w_k \geq 0$): $g_k^{\text{LP}}(B_k) = \inf_{w_k \geq 0}[q_k(w_k) + w_k B_k]$, which is the infimum of affine functions (concave).

For the true (hard-constraint) optimal value $g_k^*(B_k)$: consider the coupling argument. Given an optimal policy π^* for budget $B_k + 2$, we can construct a feasible policy for budget $B_k + 1$ by dropping the lowest-priority activation at each step (the arm with the smallest marginal value among those activated). The reward loss from dropping one activation is at most the marginal value of that arm, which is non-increasing in B_k (the $(B_k + 2)$-th best arm is worse than the $(B_k + 1)$-th best). Therefore the marginal gain $g_k^*(B_k + 1) - g_k^*(B_k)$ is non-increasing in B_k. □

A.8 Completeness of the Portfolio Effect Proof

Proposition 2 is proved by construction in the main text. For completeness, we verify the key details. The counterexample has $N = 3$, $N_0 = 2$, $N_1 = 1$, $K = 1$, $B = 1$. The joint state space is $\{0, 1\}^3$ with $2^3 = 8$ states. The aggregate state space is $\{(n_0, n_1) : 0 \leq n_0 \leq 2,\ 0 \leq n_1 \leq 1\}$ with 6 elements.

The optimal policy is computed by exact value iteration on the 8-state joint MDP ($\gamma = 0.999$, 10^4 iterations, convergence to 10^{-13}). The policy is then mapped to the aggregate level by checking that all joint states with the same aggregate counts yield the same type-level allocation (verified with zero violations, confirming within-type permutation invariance of V^* as predicted by Lemma 4).

The non-separability is verified by extracting the implied ordering constraints at the type level. At aggregate state $(1, 1)$: type 1 is pulled, requiring $\iota(\text{type } 1, s{=}1) > \iota(\text{type } 0, s{=}1)$. At aggregate state $(2, 1)$: type 0 is pulled, requiring $\iota(\text{type } 0, s{=}1) > \iota(\text{type } 1, s{=}1)$. These two constraints form a cycle in the ordering graph, which is detected by topological sort (Kahn's algorithm).

The search was conducted over 8,000 randomly generated exchangeable instances per configuration, using four adversarial generators (uniform random, churner-biased, slow-mixing, and near-degenerate). At $(N, K, B, T) = (3, 1, 1, 2)$, 3.0% of instances exhibit the portfolio effect. All counterexample instances were independently verified by exhaustive enumeration of the $3^8 = 6{,}561$ deterministic policies (for $B = 1$, there are 3 choices per state) to confirm that the non-separable policy is strictly optimal (not a tie-breaking artifact).

B Condition Verification for the CBB

We verify the three conditions required by Theorem 7 (via Gast et al. (2024)) for the specific case of the CBB with binary-state ergodic arms.

B.1 UGAP (Uniform Global Attractor Property)

Definition. A restless bandit satisfies UGAP if, for every LP-optimal dual w^*, the mean-field ODE has a unique fixed point that is a global attractor (all trajectories converge to it from any initial condition).

Verification for the CBB. Under exchangeability with T types, the mean-field system describes the evolution of the expected fraction of engaged arms per type, $\bar{n}_t \in [0, 1]$. At the LP-optimal dual w^*, each type t follows:

$$\dot{\bar{n}}_t = \underbrace{P_t(1 \mid 0, a_t^*; k) \cdot (1 - \bar{n}_t)}_{\text{re-engagement}} - \underbrace{(1 - P_t(1 \mid 1, a_t^{**}; k)) \cdot \bar{n}_t}_{\text{disengagement}} \tag{25}$$

where a_t^*, a_t^{**} are the LP-optimal actions for disengaged and engaged type-t arms, respectively (determined by the LP dual w^*).

For binary states, equation 25 is a scalar linear ODE in $\bar{n}_t$ with a unique stable fixed point:

$$\bar{n}_t^* = \frac{P_t(1 \mid 0, a_t^*; k)}{P_t(1 \mid 0, a_t^*; k) + 1 - P_t(1 \mid 1, a_t^{**}; k)}. \tag{26}$$

Under ergodicity, both the numerator and denominator are strictly positive, so $\bar{n}_t^* \in (0, 1)$. Since each type's ODE is independent (conditional independence) and has a unique globally attracting fixed point, the joint system $(\bar{n}_1, \ldots, \bar{n}_T)$ has the unique global attractor $(\bar{n}_1^*, \ldots, \bar{n}_T^*)$. This establishes UGAP.

B.2 Non-Degeneracy

Definition. The LP relaxation is non-degenerate if, at the optimal dual w^*, at most one type-state category has a fractional activation probability (all others are activated with probability 0 or 1). Equivalently, $\delta_{\min} = \min_{(t,s) \neq (t',s')} |\iota_t(s; w^*) - \iota_{t'}(s'; w^*)| > 0$ where $\iota_t(s; w^*)$ is the Lagrangian index of type t at state s.

Verification for the CBB. The Lagrangian index $\iota_t(s; w^*)$ is a continuous function of the transition parameters $(P_t(s' \mid s, a; k))$ and the dual w^*. The set of parameters where two distinct indices coincide ($\iota_t(s) = \iota_{t'}(s')$ for $(t, s) \neq (t', s')$) is a proper algebraic subvariety of the parameter space (it is defined by a polynomial equation in the parameters). Therefore, non-degeneracy holds for Lebesgue-almost-every choice of transition parameters.

For the specific blended CBB instances used in our experiments (organic arms with retention $P(1|1, 0) \approx 0.66$ and churner arms with retention $P(1|1, 0) \approx 0.68$, pull effects differing by 0.2–0.5), we verify computationally that $\delta_{\min} > 0.01$ across all tested configurations.

Table 6: Complete $\mathcal{F}_{\text{agg}}$ validation: 112,000 instances (8,000 per config, 4 adversarial generators).

N	K	B	T	$\mathcal{F}_{\text{agg}}$	$\mathcal{F}_{\text{idx}}$	Portfolio	Time (h)
3	1	1	2	**100%**	97.0%	3.0%	1.0
4	1	1	2	**100%**	95.8%	4.2%	2.4
4	1	2	2	**100%**	93.5%	6.5%	5.0
5	1	1	2	**100%**	95.3%	4.7%	5.6
5	1	2	2	**100%**	90.0%	10.1%	14.4
6	1	1	2	**100%**	95.1%	4.9%	13.4
6	1	2	2	**100%**	86.7%	13.3%	40.5
6	1	3	2	**100%**	91.6%	8.4%	76.2
3	1	1	3	**100%**	98.1%	1.9%	1.0
4	1	1	3	**100%**	94.4%	5.6%	2.4
4	1	2	3	**100%**	91.0%	9.0%	5.0
3	2	1	2	**100%**	89.7%	10.3%	2.0
4	2	1	2	**100%**	83.0%	17.0%	4.9
4	2	2	2	**100%**	83.4%	16.6%	10.1
Grand total				**100%**	91.8%	8.2%	183.9

B.3 GES (Global Exponential Stability)

Definition. The mean-field fixed point $\bar{\mathbf{n}}^*$ is globally exponentially stable (GES) if there exist $C_0 > 0$ and $\gamma > 0$ such that $\|\bar{\mathbf{n}}(t) - \bar{\mathbf{n}}^*\| \leq C_0\, e^{-\gamma t}\|\bar{\mathbf{n}}(0) - \bar{\mathbf{n}}^*\|$ for all initial conditions and all $t \geq 0$.

Verification for the CBB. Each type's mean-field ODE equation 25 is linear in $\bar{n}_t$ with eigenvalue:

$$\lambda_t = P_t(1 \mid 1, a_t^{**}; k) - P_t(1 \mid 0, a_t^{*}; k) \tag{27}$$

which is the second eigenvalue of the type-t Markov chain under the LP-optimal action. Under ergodicity, $|\lambda_t| < 1$, so each type converges at rate $\alpha_t = 1 - |\lambda_t| > 0$. Since the types evolve independently (conditional independence), the joint system converges at rate $\gamma = \min_{t=1}^{T} \alpha_t = 1 - \max_t |\lambda_t| > 0$. The exponential stability constant is $C_0 = 1$ (the ODE is linear, so no amplitude amplification occurs). This establishes GES with rate $\gamma = 1 - \lambda_{\max}$.

C Extended Experimental Results

Tables 2–3 in the main text show policy hierarchy gaps and separability costs for 5 configurations. Table 1 shows 8 representative $\mathcal{F}_{\text{agg}}$ validation configurations. Here we provide the complete results across all 14 configurations tested.

The full results reveal several patterns.

Portfolio effect drivers. Three factors increase the portfolio effect frequency: (1) Budget B: comparing B=1 vs B=2 at fixed (N, K, T) shows roughly 2× increase (e.g., 4.7% to 10.1% at N=5, 4.9% to 13.3% at N=6). (2) Number of contexts K: multi-context instances (K=2) show the highest rates (up to 17.0%), since each context independently faces the reversal. (3) Budget ratio B/N: B=3 at N=6 (ratio 0.5) shows lower portfolio effect (8.4%) than B=2 at N=6 (ratio 0.33, effect 13.3%), suggesting a non-monotone relationship peaking at intermediate ratios.

Number of types. T=3 shows slightly lower portfolio rates than T=2 at the same (N, B) (e.g., 1.9% vs 3.0% at $(3, 1, 1)$; 5.6% vs 4.2% at $(4, 1, 1)$). With more types and fewer arms per type, each type's "backup depth" is smaller, reducing the scope for the diversification mechanism.

Computation time. The dominant cost is exact value iteration on the 2^N joint states. The N=6 configurations required up to 76 hours for 8,000 instances (34 seconds per instance). Beyond N=6, the joint MDP becomes intractable, though the aggregate-state MDP ($|\mathcal{N}| = \prod_t (N_t + 1)$ states) remains feasible for much larger N.

Cross-partial structure. Analysis of the discrete Hessian of V^* on portfolio instances reveals that sub/supermodularity does not hold universally: at N=3, 68% of portfolio instances are submodular but 32% have mixed-sign cross-partials. The cross-partial magnitude decays rapidly as $O(\lambda_{\max}^{2\|\mathbf{n}\|_1})$, faster than the spectral gap $\lambda_{\max}$, consistent with the interaction being a second-order effect. Within-type concavity of V^* (diminishing returns in n_t) holds for the vast majority of instances but admits rare violations at $N \geq 4$. These observations suggest that $\mathcal{F}_{\text{agg}}$ is close to the tightest achievable class, with no simple structural property (monotonicity, convexity, bounded reversals) yielding a further tightening.

C.1 Instance Generation

All experiments use *blended* instances mixing organic and churner volunteer types in equal proportion (50% each), following the calibration of Shi et al. (2024).

Organic arms. Volunteer i and region k are placed uniformly on $[0,1]^2$. The pick-up probability is $p_{ik} = \sigma(0.5 \cdot \text{pop}_k - 0.8 \cdot d(i,k) + 0.3 \cdot \text{hist}_i)$, clipped to $[0.1, 0.9]$. Pulled retention is $\min(0.85 + 0.1 p_{ik}, 1)$; idle retention is drawn from $\text{Uniform}(0.40, 0.85)$; recovery from inactive is $\text{Uniform}(0.40, 0.60)$ when pulled and $\text{Uniform}(0.01, 0.03)$ when idle.

Churner arms. The first $\lfloor K/2 \rfloor$ contexts are "preferred" regions with high reward ($p_{ik} \sim \text{Uniform}(0.80, 0.95)$) but heavy churn ($P(1 \mid 1, 1; k) = 0.05$). Remaining contexts are "disliked" with low reward ($p_{ik} \sim \text{Uniform}(0.05, 0.15)$) and mild churn ($P(1 \mid 1, 1; k) = 0.85$). Idle retention is $\text{Uniform}(0.35, 0.70)$ in preferred regions; recovery from inactive is $\text{Uniform}(0.25, 0.45)$ when pulled.

Context frequencies. All generators use uniform context frequencies $f_k = 1/K$. Because f_k is uniform, the LP-optimal per-context budget B_k^* coincides with the uniform split B/K in our instances. Methods that assume uniform budget (DQWIC, NIP, Q-index) are therefore not penalized by the budget allocation; all performance differences in our experiments arise from the arm selection policy. Non-uniform context frequencies, where the two-phase budget allocation provides additional lift, are a natural extension for future work.

Transition heterogeneity across contexts. Although f_k is uniform, the arm dynamics vary substantially across contexts. Churner arms have $P(1|1,1;k) = 0.05$ in preferred regions versus 0.85 in disliked regions; organic arms have distance-dependent retention that varies across all K contexts. This heterogeneity is what makes context-aware policies (MDVI, NIP) outperform context-agnostic ones (Greedy, Q-index), even though the budget split itself is uniform.

C.2 Baseline Algorithms in Detail

We describe each algorithm's mechanism, policy class placement, budget allocation strategy, and computational cost. All baselines receive the same instance parameters (N, K, B, P, R, f) and are evaluated by simulating the resulting policy for 300 steps per seed.

MDVI (Mirror Descent Value Iteration; ours). Dualizes the per-context budget constraints via Lagrangian prices $w_k \geq 0$, decomposes the N-arm problem into N independent two-state MDPs, and solves each exactly by Relative Value Iteration in $O(K)$ time. The dual prices are updated by projected subgradient descent: $w_k \leftarrow [w_k - \alpha_t(B_k - \hat{D}_k)]_+$, where $\hat{D}_k$ is the demand (number of arms wanting activation in context k). The per-arm policies implicitly determine the budget allocation through the learned prices. Policy class: $\mathcal{F}_{\text{Lag}}$. Budget: learned via dual prices (not fixed). Hyperparameters: 150 outer iterations, step size $\alpha_t = 0.1/\sqrt{t+1}$, simulation horizon 300. Complexity: $O(LNK)$ total, where L is the number of outer iterations.

DQWIC (Deep Q-Network for Whittle Index Context). Approximates the per-arm Whittle subsidy using a deep Q-network (DQN) trained on the single-arm relaxation of each arm's MDP. The DQN takes the arm state s_i and context k as input and outputs the Q-values $Q(s_i, a; k)$ for $a \in \{0, 1\}$. The Whittle index is then computed as the subsidy w at which $Q(s_i, 1; k) - Q(s_i, 0; k) = w$. Arms are ranked by index and the top B_k activated. When PyTorch is unavailable, falls back to tabular Whittle index computation via bisection on the per-arm MDP. Policy class: $\mathcal{F}_{\text{idx}}$. Budget: uses LP-optimal allocation B_{cocc}. Training: 500–4000 DQN steps (scaled by N), ϵ-greedy exploration decaying from 0.9 to 0.05.

NIP (Neural Index Policy). Trains a per-arm score network $\iota_i(s_i; \theta)$ via REINFORCE (Williams, 1992), maximizing the simulated average reward. The network is a small MLP (2 hidden layers, 32 units) shared across arms of the same type, producing a scalar priority score. At each step, the top B_k arms by score are activated. Policy class: $\mathcal{F}_{\text{idx}}$. Budget: uses LP-optimal allocation B_{cocc}. Training: 500–4000 REINFORCE steps with baseline subtraction and KL penalty for stability.

Mitosis (Shi et al., 2024). The current state-of-the-art for contextual RMABs. Uses a UCB (Upper Confidence Bound) tree search over the combinatorial space of integer budget allocations $\{B_k\}_{k=1}^K$ subject to the global constraint $\sum_k f_k B_k \leq B$. Each node in the UCB tree represents a candidate allocation; the algorithm balances exploration of new allocations against exploitation of known-good ones. Within each allocation, arms are selected greedily by immediate reward. Policy class: depends on arm selection (greedy $\Rightarrow$ myopic, but budget search is adaptive). Budget: learned via UCB exploration. Hyperparameters: $c_{\text{UCB}} = 2.0$, timeout 600s.

Greedy. Activates the B_k arms with the highest immediate expected reward $R_i(s_i, 1; k)$ in each context. Ignores transition dynamics entirely, making it vulnerable to the *churn trap*: pulling en-

gaged churners yields high immediate reward but destroys their engagement (retention drops from 0.70 idle to 0.05 pulled). The resulting steady-state loss accounts for the persistent 5–6% gap observed in Table 7. Policy class: myopic (not in any named class). Budget: uniform $B_k = B$ for all k.

Q-index. Computes the tabular action-value function $Q_i(s, a; k)$ for each arm's two-state MDP under the per-arm Lagrangian relaxation, then ranks arms by the advantage $Q_i(s_i, 1; k) - Q_i(s_i, 0; k)$. Unlike DQWIC, uses exact tabular computation rather than neural approximation. Policy class: $\mathcal{F}_{\text{idx}}$. Budget: uses LP-optimal allocation B_{cocc}.

COcc (Contextual Occupancy Index). Solves the LP relaxation over occupation measures $\mu_i(s, a; k) \geq 0$ to obtain the optimal per-context budgets and per-arm activation probabilities. Arms are then selected stochastically according to these probabilities, rounded to satisfy the integer budget constraint. Policy class: LP-based (not strictly in the hierarchy). Budget: LP-optimal B_{cocc}.

Random. Selects B_k arms uniformly at random in each context, independent of arm states. Serves as a lower bound on performance. Budget: uniform $B_k = B$ for all k.

Round-robin. Cycles through arms in a fixed order, activating B_k per context regardless of state. Slightly better than Random when arm dynamics are homogeneous, but equally state-blind. Budget: uniform $B_k = B$ for all k.

C.3 Instance Generation Details

All instances follow the food-rescue volunteer model of Shi et al. (2024), with N volunteers (arms) and K geographic regions (contexts). Each arm has binary state $s_i \in \{0, 1\}$ (disengaged/engaged). The first $\lfloor N/2 \rfloor$ arms are organic; the remainder are churners.

Context frequencies. The default setting uses uniform frequencies $f_k = 1/K$ for all k. In the non-uniform experiments (Section 6), we use Zipf-like frequencies $f_k \propto 1/k$, normalized to sum to 1. For example, at K=3: $f = (0.545,\ 0.273,\ 0.182)$.

Organic arm transitions. Each volunteer i and region k is assigned a pick-up probability based on geographic proximity:

$$p_{ik} = \sigma(0.5 \cdot \text{pop}_k - 0.8 \cdot d(i, k) + 0.3 \cdot \text{hist}_i),$$

where $\text{pop}_k \sim U(0.5, 1.0)$ is region popularity, $d(i, k)$ is Euclidean distance between volunteer and region positions on $[0, 1]^2$, and $\text{hist}_i \sim U(0.1, 0.9)$ is the volunteer's activity history. The sigmoid output is clipped to $[0.1, 0.9]$. Transition probabilities:

Transition	Probability	Range
$P(1 \mid 1, \text{pull}, k)$	$\text{clip}(0.85 + 0.1 \cdot p_{ik})$	$[0.86, 0.94]$
$P(1 \mid 1, \text{idle}, k)$	$U(0.40, 0.85)$	varies by arm
$P(1 \mid 0, \text{pull}, k)$	$U(0.40, 0.60)$	recovery on pull
$P(1 \mid 0, \text{idle}, k)$	$U(0.01, 0.03)$	very slow natural
Reward: $R_i(1, 1, k)$	p_{ik}	$[0.1, 0.9]$

The key feature is that pulled retention varies with context (through p_{ik}), while idle transitions are drawn iid. Natural recovery without intervention is near-zero (1–3%), creating a strong investment motive that greedy scoring misses.

Churner arm transitions. Churners have $K_{\text{pref}} = \lfloor K/2 \rfloor$ preferred regions and $K - K_{\text{pref}}$ disliked regions, with dramatically different dynamics:

Transition	Preferred ($k < K_{\text{pref}}$)	Disliked ($k \geq K_{\text{pref}}$)
$P(1 \mid 1, \text{pull}, k)$	0.05 (heavy churn)	0.85 (retained)
$P(1 \mid 1, \text{idle}, k)$	$U(0.35, 0.70)$	$U(0.50, 0.80)$
$P(1 \mid 0, \text{pull}, k)$	$U(0.25, 0.45)$	$U(0.25, 0.45)$
$P(1 \mid 0, \text{idle}, k)$	$U(0.01, 0.02)$	$U(0.01, 0.02)$
$R_i(1, 1, k)$	$U(0.80, 0.95)$	$U(0.05, 0.15)$

In preferred regions, churners offer high immediate reward (0.80–0.95) but pulling *destroys* engagement (retention drops to 0.05). This is the *churn trap*: greedy policies pull these arms for their high reward, causing rapid disengagement. In disliked regions, the same arms have low reward but stable dynamics. The optimal policy learns to avoid pulling churners in preferred regions despite the tempting immediate payoff.

Cross-context heterogeneity. The organic transition $P(1 \mid 1, \text{pull}, k)$ varies across contexts k through the distance-based p_{ik}, typically spanning a range of 0.86–0.94 for a given arm. The churner

transition varies even more dramatically: 0.05 in preferred vs. 0.85 in disliked regions. This structured heterogeneity means the optimal arm priority ordering can differ substantially across contexts, motivating the two-phase formulation.

Non-uniform context experiment. Under uniform $f_k = 1/K$, the LP-optimal budget allocation B_{cocc} coincides with the uniform allocation $B_k = B$, so all baselines using B_{cocc} or B_{uniform} face the same budget. With Zipf frequencies ($f_k \propto 1/k$), the dominant context receives more weight in the objective, making non-uniform budget allocation optimal. MDVI's dual prices adapt automatically: its gap decreases from 3.9% (uniform) to 2.8% (extreme skew), while DQWIC (7.2% $\rightarrow$ 7.7%) and other baselines remain flat or worsen.

C.4 Scalability Data

Table 7 reports the full numerical data for the scalability experiment (Figure 2 in the main text).

Table 7: Gap (%) vs N (K=3, ρ=0.1, asymmetric contexts, mean$\pm$SE, 10 seeds). **Bold**: best.

N	MDVI	NIP	DQWIC	Mitosis	Greedy	Q-idx	COcc	Rand	RR	CPU
50	**6.2**$\pm$**0.6**	8.1$\pm$1.0	9.0$\pm$1.1	11.0$\pm$0.6	12.5$\pm$0.4	12.6$\pm$0.8	12.7$\pm$1.0	88.2$\pm$0.3	91.8$\pm$0.2	139s
100	**1.2**$\pm$**0.3**	5.7$\pm$0.7	6.5$\pm$0.7	5.2$\pm$0.4	9.6$\pm$0.4	11.4$\pm$0.7	7.1$\pm$0.6	87.8$\pm$0.1	91.7$\pm$0.3	291s
200	3.2$\pm$0.3	3.1$\pm$0.9	4.0$\pm$1.0	**2.9**$\pm$**0.4**	7.2$\pm$0.6	9.7$\pm$1.2	4.6$\pm$1.1	87.6$\pm$0.1	91.8$\pm$0.3	586s
500	**0.0**$\pm$**0.3**	3.1$\pm$1.1	4.2$\pm$1.1	**0.0**$\pm$**0.4**	6.7$\pm$0.6	9.6$\pm$1.2	4.1$\pm$1.2	87.8$\pm$0.1	91.6$\pm$0.2	1605s
1000	**1.1**$\pm$**0.3**	1.6$\pm$1.0	3.0$\pm$1.1	2.8$\pm$0.3	5.9$\pm$0.6	8.7$\pm$1.1	3.4$\pm$1.1	87.6$\pm$0.1	91.5$\pm$0.2	5.6h
2000	**0.0**$\pm$**0.3**	8.7$\pm$7.4	2.0$\pm$1.1	0.1$\pm$0.3	5.4$\pm$0.5	7.9$\pm$1.1	2.5$\pm$1.1	87.6$\pm$0.0	91.7$\pm$0.2	16.9h

C.5 Context and Budget Ablation Plots

Figure 3 shows the optimality gap as a function of the number of contexts K, confirming that MDVI is K-invariant at 0.4–0.5%.

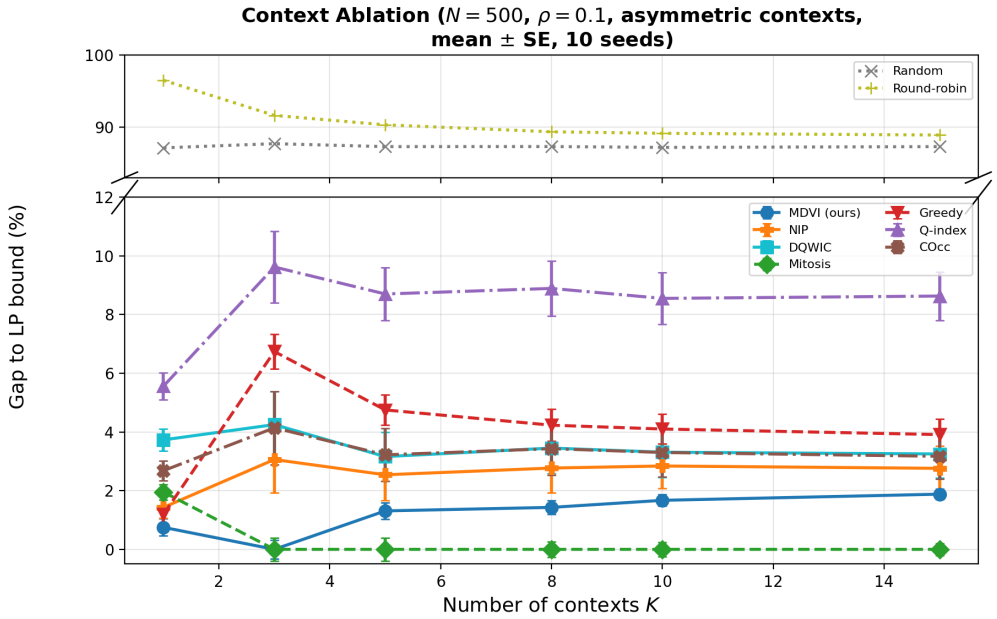

Figure 3: Context ablation: gap (%) vs K (N=500, ρ=0.1, mean $\pm$ SE, 10 seeds). MDVI is K-invariant; Q-index degrades with K.

Figure 4 shows the optimality gap as a function of the budget ratio $\rho = B/N$, revealing the churn trap sweet spot at $\rho \approx 0.20$.

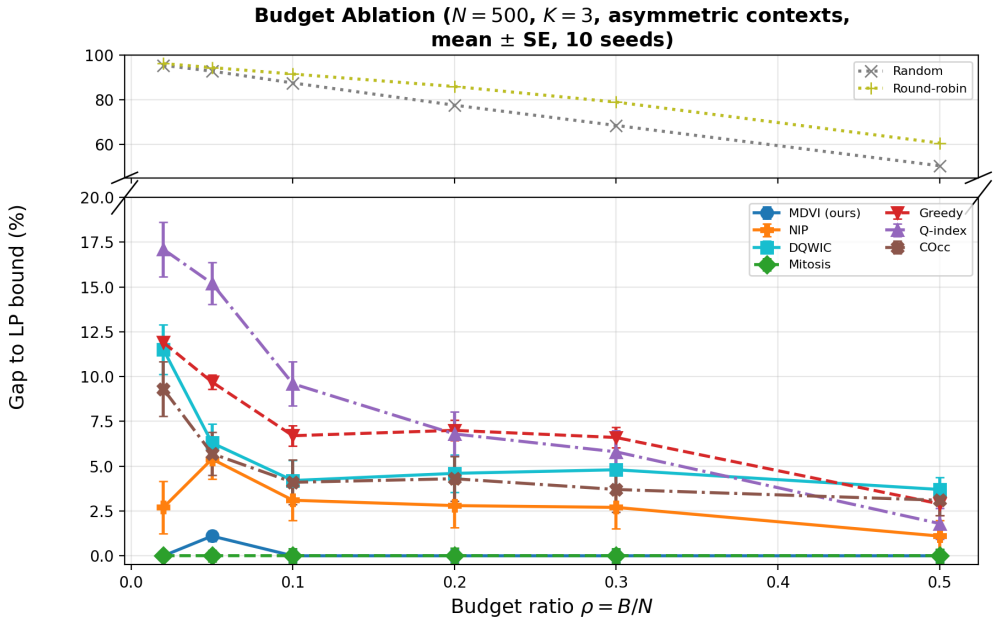

Figure 4: Budget ablation: gap (%) vs ρ (N=500, K=3, mean ± SE, 10 seeds). The churn trap peaks at ρ=0.20 for greedy (7.8%); DQWIC collapses at ρ=0.02 (13.2%).

C.6 Policy Hierarchy

Figure 5 shows the grouped bar chart of optimality gaps across configurations for the six competitive algorithms.

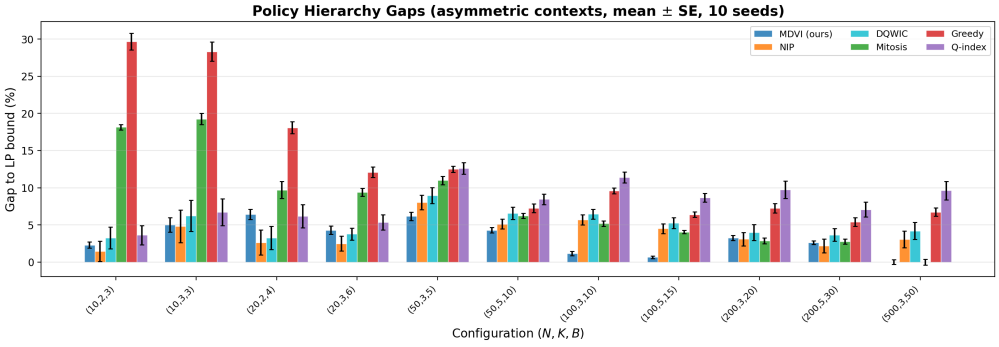

Figure 5: Policy hierarchy gaps across (N, K, B) configurations (mean ± SE, 10 seeds). MDVI achieves the lowest gap at every configuration.

D Counterexample Search Methodology

We used three complementary search strategies:

Exhaustive grid (N=2). At $(N, K, B) = (2, 1, 1)$: 8 continuous parameters (4 per arm). Grid at resolution 10 yields 10^8 instances per reward setting. For each instance, all $2^4 = 16$ deterministic policies are evaluated (average reward via stationary distribution), and the best non-separable policy is compared to the best separable. Three reward settings tested: $(r_1, r_2) \in \{(1, 1), (1, 2), (1, 0.5)\}$. Total: 9×10^8 instances.

Targeted search (heterogeneous). Five adversarial generators: (1) random uniform; (2) churner-heavy (70% arms with $P(1|1, 1) \ll P(1|1, 0)$); (3) near-degenerate (small cross-type score gaps); (4) slow-mixing ($\lambda > 0.85$); (5) adversarial-mixed (alternating pull directions). For each instance, the joint MDP is solved exactly via value iteration ($\gamma = 0.999$, 10^4 iterations). The optimal policy is checked for consistency with a fixed priority via topological sort on the implied ordering constraints.

Exchangeable search (type-level). Same generators but with identical dynamics within each type. The separability check operates at the *type* level: only cross-type ordering constraints are collected

(same-type arm swaps are free). An instance is flagged only if the aggregate-level allocation (b_0, b_1) differs across joint states mapping to the same aggregate state (F_agg violation) or if the type-priority reverses across different aggregate states (portfolio effect, F_idx violation).