Robust Planning over Restless Groups: Engagement Interventions for a Large-Scale Maternal Telehealth Program

Jackson A. Killian[*1], **Arpita Biswas**[*1], **Lily Xu**[*1], **Shresth Verma**[*2], **Vineet Nair**[2], **Aparna Taneja**[2], **Aparna Hegde**[3], **Neha Madhiwalla**[3], **Paula Rodriguez Diaz**[1], **Sonja Johnson-Yu**[1], **Milind Tambe**[1,2]

[1]Harvard University, Cambridge, United States
[2]Google Research, Bangalore, India
[3]ARMMAN, Mumbai, India
{jkillian,arpitabiswas,lily_xu}@g.harvard.edu, {vermashresth, vineetn, aparnataneja}@google.com,
{aparnahegde, neha}@armman.org, {prodriguezdiaz, sjohnsonyu}@g.harvard.edu, milindtambe@google.com

Abstract

In 2020, maternal mortality in India was estimated to be as high as 130 deaths per 100K live births, nearly twice the UN's target. To improve health outcomes, the non-profit ARMMAN sends automated voice messages to expecting and new mothers across India. However, 38% of mothers stop listening to these calls, missing critical preventative care information. To improve engagement, ARMMAN employs health workers to intervene by making service calls, but workers can only call a fraction of the 100K enrolled mothers. Partnering with ARMMAN, we model the problem of allocating limited interventions across mothers as a restless multi-armed bandit (RMAB), where the realities of *large scale* and *model uncertainty* present key new technical challenges. We address these with *GROUPS*, a double oracle–based algorithm for robust planning in RMABs with scalable *grouped arms*. Robustness over grouped arms requires several methodological advances. First, to adversarially select stochastic group dynamics, we develop a new method to optimize Whittle indices over transition probability intervals. Second, to learn group-level RMAB policy best responses to these adversarial environments, we introduce a weighted index heuristic. Third, we prove a key theoretical result that planning over grouped arms achieves the same minimax regret–optimal strategy as planning over individual arms, under a technical condition. Finally, using real-world data from ARMMAN, we show that GROUPS produces robust policies that reduce minimax regret by up to 50%, halving the number of preventable missed voice messages to connect more mothers with life-saving maternal health information.

1 Introduction

Maternal mortality, the death of a mother[1] during pregnancy or within 42 days after childbirth, is an ongoing global health crisis. In India, the maternal mortality rate is particularly stark, estimated between 99 and 130 deaths per 100K births in 2020 (Meh et al. 2021; Gates Foundation 2021), significantly higher than Sustainable Development Goal 3.1 target of 70 per 100K births (United Nations 2021). Tragically, most maternal deaths are preventable (WHO 2023), but lack of finances and awareness prevent mothers from seeking care, particularly in low-income communities (Carvalho, Salehi, and Goldie 2013).

Figure 1: Mothers enrolled with ARMMAN receive life-saving preventative care information via voice messages throughout their pregnancy, childbirth, and neonatal period. Photo courtesy of ARMMAN.

To improve maternal health outcomes, we work with ARMMAN, an India-based non-profit that provides free preventive care to millions of mothers by sending automated health voice messages, specifically targeted towards low-income communities (similar to MAMA (Johnson & Johnson 2017)). Mothers enrolled in the program receive weekly automated voice messages during pregnancy and up to one year after childbirth. Randomized control trials showed that ARMMAN's messaging program significantly improves key indicators including treatment-seeking during complications, infant breastfeeding, and post-infancy weight (Murthy et al. 2019). However, ARMMAN found that nearly 38% of mothers disengage, missing critical health information. To improve engagement, ARMMAN employs health workers to provide service calls, but there are only tens of health workers compared to hundreds of thousands of mothers in a given service area — so interventions must be carefully targeted to maximize engagement.

Working with ARMMAN, we model this resource-limited intervention planning problem as a restless multi-armed bandit (RMAB), where each mother (arm) changes their weekly engagement (state) according to a stochastic Markov

[*]These authors contributed equally.

[1]We recognize that the term "mother" is imperfect, most notably by not reflecting transgender and non-binary identities. We highlight alternative language with discussion in Appendix A.

decision process. RMABs are PSPACE-hard to solve exactly (Papadimitriou and Tsitsiklis 1999) and even the more tractable, asymptotically optimal "Whittle index policy" (Whittle 1988) is challenging to compute at scale.

To improve the scalability of real-world RMAB planning, Mate et al. (2022) proposed to organize arms into a small number of *groups*, infer transition dynamics from each group's data, then compute the Whittle index policy per group. While the scalability of their method is desirable for ARMMAN's problem setting, it ignores a key reality of *model uncertainty*: learning transition probabilities from historical data leads to imprecise and imperfect estimates which must be accounted for in planning. Computing RMAB policies that are robust to model uncertainty has only recently been studied. Existing methods achieve robustness to *interval uncertainty* over model dynamics by planning against a model-controlling "nature" adversary to yield policies that minimize max regret (Killian et al. 2022; Xu et al. 2021). Robustness is desirable for ARMMAN's setting, but these methods require training deep reinforcement learning (RL) agents for each arm, so unfortunately do not scale past hundreds of arms.

To enable large-scale, robust intervention planning for ARMMAN, we bridge the gaps in previous works by introducing *robust grouped RMAB*. Our model achieves scalability by considering a grouped-arm paradigm and optimizing for minimax regret over the uncertain model dynamics per group. Unfortunately, the grouping abstraction breaks key assumptions used in previous robust RMAB work: that (1) policies improve by collecting samples of regret by evolving a joint state of all arms, and (2) the nature adversary controls the transitions of each arm individually. We overcome (1) by *decomposing regret per arm*, freeing the planner from relying on a cumbersome joint state to enable efficient group-abstracted planning. For (2), we prove that *restricting the adversary to control dynamics only over groups does not change the equilibrium strategy*, allowing us to leverage the scalable robust grouped model to find policies over hundreds of thousands of arms without sacrificing quality.

Our contributions are as follows. *First,* we introduce *robust grouped RMABs* with a minimax regret objective and propose a solution that employs the double oracle framework (McMahan, Gordon, and Blum 2003). The approach we propose is **GROUPS**: Group RMAB Oracles for Uncertainty-robust Planning at Scale. *Second,* we develop novel methods designed for robust grouped RMABs to implement the two oracles, the planner and adversary. Planning over groups of arms allows large scale-up but presents several new algorithmic challenges as we detail above. *Third,* we prove that the minimax regret–optimal strategy is the same whether the planner and adversary play at the individual or group level. *Our proof enables massive scale-up as it is now sufficient to compute robust strategies over groups*, instead of over individual arms. *Finally,* we demonstrate empirically on real data that **GROUPS reduces worst-case regret up to 50% compared to baselines, representing potentially thousands of additional engagements with life-saving information.** We are working with ARMMAN to deploy GROUPS to positively impact maternal health.

2 Related Work

Mobile-based maternal health services are effective and affordable in low- and middle-income communities (Watterson, Walsh, and Madeka 2015; Tamrat and Kachnowski 2012). Successful programs include MatHealth in Uganda (Musiimenta et al. 2021), Aponjon in Bangladesh (Alam et al. 2017), ARMMAN in India (Murthy et al. 2019), and text4baby in the United States (Evans, Wallace, and Snider 2012). Our work is designed to support such programs.

Whittle (1988) introduced RMABs and proposed the *Whittle index policy*, which computes indices estimating each arm's "return on investment" then acts on arms with the top K. Weber and Weiss (1990) showed this policy is asymptotically optimal under a technical condition. Many RMAB studies assume known transition dynamics, although some recent works design methods to learn policies online (Wang, Huang, and Lui 2020; Nakhleh et al. 2021; Biswas et al. 2021; Killian et al. 2021; Wang et al. 2022). However, these online approaches require collecting a prohibitively large number of samples, limiting their real-world applicability in scenarios where the time horizon is short.

Most robust planning papers consider single-MDP (one arm) settings (Pinto et al. 2017; Lanctot et al. 2017; Li et al. 2019), rather than the budget-coupled N-MDP setting of RMAB. Even for single MDPs, optimizing criteria such as minimax regret (Braziunas and Boutilier 2007) requires searching massive strategy spaces; double oracle (McMahan, Gordon, and Blum 2003) is one approach to do so efficiently. Recent work combines double oracle with deep RL to solve for minimax regret–optimal robust policies for single MDPs (Xu et al. 2021). Killian et al. (2022) extended the idea to solve larger RMABs. Both Xu et al. (2021) and Killian et al. (2022) use deep RL which, if applied to a group setting, would need to explicitly account for the size of each group and state of each arm within each group, limiting their methods' ability to scale beyond hundreds of arms. For the large problem size that ARMMAN faces, our methods must scale to hundreds of thousands of arms.

Finally, robust planning for *stochastic* bandits is well studied (Maillard 2013; Huo and Fu 2017) However, stochastic bandits are stateless and lack passive rewards, and so are not expressive enough to model ARMMAN's setting.

3 Model

We consider *grouped RMABs* where N arms (enrolled mothers) comprise M groups. Each arm $n \in [N]$ follows an MDP $\langle \mathcal{S}, \mathcal{A}, P^n, r, \gamma \rangle$ where $s \in \mathcal{S} := \{0, 1\}$ is the state space indicating whether a mother is engaging ($s_n = 1$) or not engaging ($s_n = 0$) with automated voice messages; $r(s) = s$ is the reward function; $a \in \mathcal{A} := \{0, 1\}$ is the action space, i.e., {not intervene, intervene}; $P^n(s, a, s')$ is the probability that arm n transitions from state s to s' given action a; $\gamma \in [0, 1]$ is the discount factor. Let $\boldsymbol{s} \in \mathcal{S}^N$ and $\boldsymbol{a} \in \mathcal{A}^N$ be the combined state and action vectors of all arms. At each timestep t, the task is to choose K mothers to intervene on (deliver service calls to) given the state $\boldsymbol{s}_t$ at time t.

Formally, we compute RMAB policies $\pi : \mathcal{S}^N \to \mathcal{A}^N$ that respect a budget constraint $\|\pi(\boldsymbol{s}_t)\|_1 = K$ for all

t. For a given policy π and a fixed environment $P := \{P^n\}_{n\in[N]}$ representing a matrix of transition probabilities of all arms, the average discounted reward is $G(\pi, P) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(\boldsymbol{s}_t) \mid \pi, P]$. Given P, the optimal policy which maximizes reward is $\pi^\star_P := \max_\pi G(\pi, P)$. An asymptotically optimal RMAB policy is the Whittle index policy (WIP), which computes the Whittle index $W^n_P(s)$ for each arm n and state s, then intervenes on the arms with the greatest K indices. The Whittle index represents "return on investment," interpreted as a charge for acting that makes *no intervention* equally valuable as *intervention* in the long term. Let $Q^n_P(s, a, \lambda) = r(s) - \lambda a + \gamma \mathbb{E}_{s'\in\mathcal{S}}[\max_{a'\in\mathcal{A}} Q^n_P(s', a', \lambda)]$ be the long-term expected value of action a on arm n in state s. Then, for a given P, the Whittle index for arm n at state s is $W^n_P(s) = \min\{\lambda : Q^n_P(s, 1, \lambda) = Q^n_P(s, 0, \lambda)\}$.

Grouped RMAB For scalability, we organize arms into groups, extending the concept from Mate et al. (2022) to our more challenging *robust* setting, e.g., by clustering based on historical engagement patterns. We then estimate uncertainty intervals over transition probabilities per group. However, note that our robust policy computation steps in Section 4 are agnostic to the particular grouping and interval estimation methods. Let $\phi : [N] \to [M]$ be a surjective mapping of arms to groups and $\phi^{-1}(m)$ be the set of arms in group m. The uncertainty intervals are $\overline{\underline{P}}^m_{s,a,s'} := [\underline{P}^m_{s,a,s'}, \overline{P}^m_{s,a,s'}]$ for all m, s, a, s'. Then let $\overline{\underline{P}}^m := \{\overline{\underline{P}}^m_{s,a,s'}\}_{s,a,s'}$ be the interval uncertainty matrix for group m across all states and actions. Importantly, though arms in the same group have the same uncertainty intervals, they may not have the same instantiated probabilities within those intervals.

Minimax regret We define regret for grouped RMAB as:

$$R(\pi, P) := G(\pi^\star_P, P) - G(\pi, P)\ , \tag{1}$$

where P instantiates $P^m \in \overline{\underline{P}}^m$ for all groups $m \in [M]$. Our objective is to learn a policy π that minimizes max regret:

$$\min_\pi \max_P R(\pi, P)\ . \tag{2}$$

We choose minimax regret as our robust objective since it does not require probability distributions over the uncertainty intervals (Braziunas and Boutilier 2007). Such distributional information is scarce in our setting where $K \ll N$, giving us few samples of transitions for action $a = 1$.

4 Methodology

We introduce GROUPS (Group RMAB Oracles for Uncertainty-robust Planning at Scale), a four-step approach visualized end-to-end in Fig. 2. Step (3) is our key algorithmic contribution. In step (1), similar arms (mothers) are mapped into groups. In step (2), we combine data from arms in each group with historical engagement data, using bootstrapping to estimate uncertainty intervals $\overline{\underline{P}}^m$ for each group (Schomaker and Heumann 2018). In step (3), we compute a minimax regret–optimal policy over groups, where arms in a given group are treated as having the same transition probabilities, greatly improving computational efficiency. Critically, we show in Section 5 that this group-level planning is lossless — i.e., the policies we compute are the same minimax regret–optimal policies as would be computed if grouped arms were allowed *different* transition probabilities (within the same uncertainty intervals). In step (4), we map group-level policies back to individual-level policies by computing Whittle indices for each group $m \in [M]$, then assigning an index to each arm n within that group based on its current state s_n. Our policy is to intervene on mothers with the top K indices.

Double oracle In step (3), we adopt a double oracle (DO) framework (McMahan, Gordon, and Blum 2003), solving Eq. 2 by formulating the problem as a two-player zero-sum game between the RMAB *planner* and nature *adversary*, where the players aim to minimize and maximize regret respectively. The planner's *pure strategy* space is the finite set of all feasible RMAB policies π; the adversary has the continuous space of transition probabilities P within the uncertainty intervals $\overline{\underline{P}}^m$ for all $m \in [M]$. The algorithm maintains a finite pure strategy set for each player. For each iteration, we compute a mixed strategy Nash equilibrium (MSNE) on the game over the finite strategy sets. A *mixed strategy* is a probability distribution over pure strategies. In each iteration, the planner oracle computes a best response pure strategy π against the adversary's mixed strategy; π is added to the planner's finite strategy set. We follow a symmetric approach to compute a best response P for the adversary. Upon termination, we return the final planner mixed strategy, which is guaranteed, under mild conditions, to be an ϵ-optimal minimax solution (Xu et al. 2021). In practice, we terminate after T iterations (Lanctot et al. 2017). The key technical challenge of using the double oracle approach is designing *planner* and *adversary* oracles for *group* RMABs.

4.1 Planner Oracle: WI for Mixed Strategy

An adversary mixed strategy β contains tuples (P_i, β_i) where β_i is the probability of playing pure strategy P_i. Similarly, a planner mixed strategy α contains tuples (π_i, α_i) where α_i is the probability of playing pure strategy π_i.

The planner oracle must compute an intervention policy π that minimizes regret with respect to a given adversary mixed strategy β over environment settings P_i. Since β and thus all P_i are fixed, and only the second term of regret in Eq. 1 depends on π, minimizing regret is equivalent to maximizing reward, to ensure that mothers engage with as many voice messages as possible. However, existing reward-maximizing RMAB algorithms assume a *single* environment P_i, versus a mixed strategy β over multiple P_i. To address this combinatorially hard problem, we develop a new heuristic approach that computes well-performing policies π based on strategically weighted combinations of Whittle indices.

Unfortunately, optimizing *exact* regret is at least PSPACE-hard (Papadimitriou and Tsitsiklis 1999). Previous work optimized regret of the Lagrange relaxation (Killian et al. 2022), but relied on joint arm states which does not scale. *We introduce a decomposed notion of regret, allowing us to optimize regret of the full RMAB in a far more scalable*

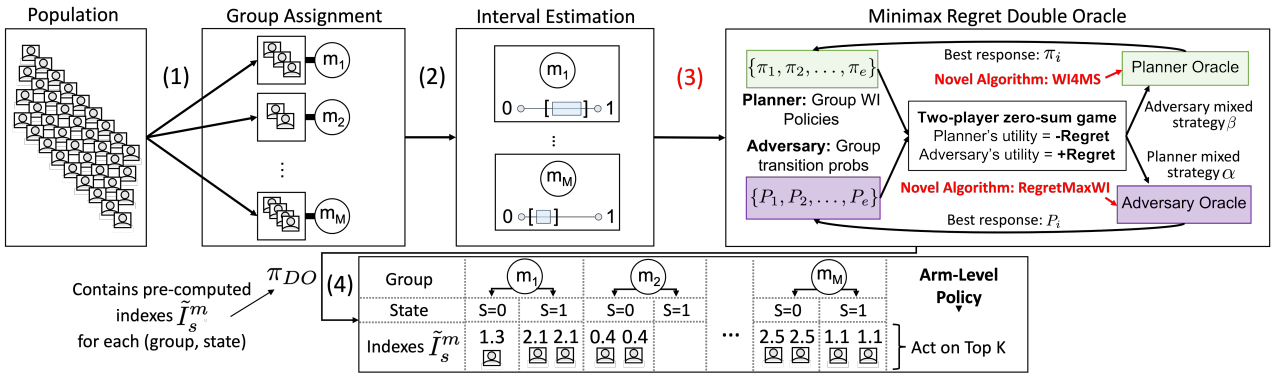

Figure 2: GROUPS pipeline for robust grouped RMABs. (1) Assign enrolled mothers (arms) to groups. (2) Estimate uncertainty intervals over transition probabilities. *(3) Novelty of this work:* Compute robust minimax regret–optimal policy via double oracle, where each oracle efficiently searches the large-scale strategy spaces by using the group abstraction. (4) To execute policies, translate group-level indices $\tilde{I}_s^m$ to arm-level intervention policy.

way. We call this *Whittle index regret*: the sum of Whittle indices played by a policy π compared to the optimal WIP. The key is that the Whittle index is a measure of "reward if played" — so agents who play arms with low Whittle indexes in lieu of arms with high Whittle indexes will incur large regret. As a further advantage, this regret notion naturally extends to groups — since the Whittle index is a function only of transition probabilities and rewards, all of which are shared in a group under P_i — improving scaling.

Given states $\boldsymbol{s}$, denote the set of arms pulled by policy π as $\Phi^\pi(\boldsymbol{s}) = \{n \in [N] : \pi_n(\boldsymbol{s}) = 1\}$ where $\pi_n(\boldsymbol{s})$ is the action on arm n. The planner's Whittle index regret $R_W^{planner}(\boldsymbol{s})$ is:

$$\sum_{(P_i,\beta_i)} \beta_i \left[\max_{\substack{\kappa \subseteq [N] \\ |\kappa|=K}} \left\{ \sum_{n\in\kappa} (W_{P_i}^n(\boldsymbol{s}^n)) \right\} - \sum_{n\in\Phi^\pi(\boldsymbol{s})} W_{P_i}^n(\boldsymbol{s}^n) \right] . \quad (3)$$

The first term in Eq. 3 corresponds to a planner's optimal mixed strategy which plays the WIP corresponding to each setting of transition probabilities P_i in β. To minimize regret $R_W^{planner}$, we seek a policy π that plays Whittle indices as close as possible to the WIPs in the first term, which equivalently maximizes the second term. How to produce a *pure* strategy π that closely follows the *mixed* WIP policies of the first term is the key challenge. *We start by making the first term more closely computable as a pure strategy* with a relaxation that leads to relaxed regret, by moving the expectation over β_i inside the max over indices:

$$\max_{\substack{\kappa \subseteq [N] \\ |\kappa|=K}} \left\{ \sum_{n\in\kappa} \sum_{(P_i,\beta_i)\in\beta} \beta_i W_{P_i}^n(\boldsymbol{s}^n) \right\} . \quad (4)$$

We replace the first term of $R_W^{planner}(\boldsymbol{s})$ (from Eq. 3) with Eq. 4 to get $\tilde{R}_W^{planner}(\boldsymbol{s})$. This illuminates a heuristic for the planner oracle. Specifically, Eq. 4 can be computed exactly by a single policy π, meaning we can make $\tilde{R}_W^{planner}(\boldsymbol{s}) = 0$ by finding a π equivalent to Eq. 4. To do so, we compute Whittle indices for each pure strategy P_i, compute the β_i–weighted average index $\tilde{I}_s^m$ for each group m and state s,

Algorithm 1: WI4MS (Planner Oracle)

Input Adversary mixed strategy β

1: **for** $(P_i, \beta_i) \in \beta$ **do** *// environment and probability i*
2: **for** $\{m = 1$ to $M\}$ and $\{s \in \mathcal{S}\}$ **do**
3: $\tilde{I}[m, s]$ += $\beta_i \times$ COMPUTEWI(m, s, P_i^m)
4: π = WIP$(\tilde{I})$ *// implements Whittle index policy*
5: **return** π *// planner pure strategy*

then follow the greedy strategy of a WIP. Since the expectation over β_i is pushed through the max (Eq. 4) we have $\tilde{R}_W^{planner}(\boldsymbol{s}) \leq R_W^{planner}$, but we show in appendix Fig. 4 that this weighted index policy performs well, despite this relaxation. We call this approach Whittle Index for Mixed Strategy (WI4MS), given in Alg. 1. Whittle indices are computed via COMPUTEWI described in Alg. 4 in the appendix.

4.2 Adversary Oracle: RegretMax Whittle Index

The adversary oracle must find one environment P that maximizes regret for the planner's current mixed strategy α over policies π_i to maximize the number of missed calls. To guide the search, we must address challenges both in maximizing regret of RMAB policies and in searching over a continuous strategy space $\overline{P}^m$. Our insight is to maximize regret by manipulating the optimal RMAB policy (a Whittle index policy) to simultaneously *minimize* the values of Whittle indices acted on by the planner and *maximize* indices that are not.

We utilize again the notion of Whittle index regret, redefined for the adversary oracle:

$$R_W^{adversary} = \mathbb{E}_{\boldsymbol{s}} \left[\sum_{n\in\Phi^{\pi^\star_P}(\boldsymbol{s})} W_P^n(\boldsymbol{s}^n)) \;\middle|\; \pi^\star_P, P \right]$$
$$- \sum_{(\pi_i,\alpha_i)\in\alpha} \alpha_i \left(\mathbb{E}_{\boldsymbol{s}} \left[\sum_{n\in\Phi^{\pi_i}(\boldsymbol{s})} W_P^n(\boldsymbol{s}^n)) \;\middle|\; \pi_i, P \right] \right) . \quad (5)$$

Algorithm 2: RegretMaxWI (Adversary Oracle)

Input: Mixed strategies (α,β), intervals $\overline{\underline{P}}^m$, group-mean budget K_M, $P = []$

$\{L_s^m\}_{s\in\mathcal{S}}^{m\in[M]} = \textsc{MonteCarlo}(\alpha, \beta)$ *// simulation*
$K_{\text{TH}} = \textsc{FindThresh}(L, K_M)$ *// returns action count of* $\lceil K_M \rceil^{th}$ *group-state*
for $\{m = 1 \text{ to } M\}$ and $\{s \in \mathcal{S}\}$ **do**
  $obj[m, s] = \min$ if $(L_s^m \geq K_{\text{TH}})$ else $\max$
for $m = 1$ to M **do**
  $P^m = \textsc{MinMaxWhittleBQP}(obj[m], \overline{\underline{P}}^m)$
return P *// Adversary pure strategy*

Given an environment, P_i, Eq. 5 captures the difference in the Whittle indices collected by the optimal policy $\pi^\star_{P_i}$ versus the Whittle indices collected by the policies of the agent mixed strategies π_i. The WIP is a proxy for finding the most effective arms on which to intervene; intuitively, this means the adversary oracle should find P_i which maximizes the Whittle indices of arms played by the optimal policy *but not* played by the planner, and simultaneously minimizes the Whittle indices of arms played *only* by the planner policies.

The first challenge is to determine which arms the planner will act on in expectation. We propose a simple but effective solution which counts the number of times the arm-state pairs are acted on during Monte Carlo simulation of the planner's mixed strategy. Since the adversary operates at the group level, we then aggregate arm-state counts into group-state counts, denoted L_s^m for each group m and state s. The next question is which group-state indices to minimize or maximize. Intuitively, if we reduced all indices an equal amount, we would reduce *reward* but not *regret* since the optimal policy, i.e., the first term of Eq. 5, would reduce the same as the second. Thus, we need to strategically minimize some indices, but *maximize* others to induce an optimal policy that plays different arms. Specifically, we choose to minimize the indices of the top $K_M = \frac{K}{N/M}$ — i.e., the budget normalized by average group size — entries of L_s^m, approximating the top K choices of the agent mixed strategy in expectation. Then we maximize the Whittle indices of all group-state pairs below that threshold.

The second challenge is to find transition probabilities P that minimize or maximize the Whittle indices of a group over its transition probability intervals. This problem has general implications, e.g., for optimistic or pessimistic search over uncertainty sets in online learning. We derive a novel binary-quadratic program that, given a group and objective for each state (min, max, or null), computes a P^m that optimizes the indices for all states simultaneously, detailed in the appendix as MinMaxWhittleBQP (Eq. 18). We give the full adversary oracle algorithm, RegretMaxWI, in Alg. 2 and empirically demonstrate its good performance in the appendix Fig. 5.

5 Theoretical Regret Guarantee

In Section 4, we proposed an approach to compute a minimax regret–optimal strategy against an adversary choosing the same transition probabilities for all arms in the same group from their corresponding intervals. However, arms within the same group may not have identical transition probabilities. Also, it is not intuitive that a minimax regret–optimal policy, when the adversary chooses the same transition probabilities for all the arms in a group, also minimizes max regret when the adversary chooses different transition probabilities for the arms in a group from their corresponding intervals. In this section, we show this is true under mild assumptions. In particular, the minimax regret–optimal strategy of the planner is the same against an adversary choosing transition probabilities at the group level as against an adversary choosing transition probabilities at the individual level.

Let Π be the planner's pure strategy space of all individual-level policies, i.e., all choices of subsets of arms with cardinality K. Then we define mixed strategy sets for the planner at *individual-level*, $\Delta_I(\Pi)$, and *group-level*, $\Delta_M(\Pi)$, where $\Delta_M(\Pi) \subseteq \Delta_I(\Pi)$ is a restricted set of mixed strategies in which the planner is indifferent between arms in the same group and state (see Appendix D.2 for definition). Next, let $\mathscr{P}$ be the adversary's pure strategy space, containing all individual-level policies, i.e., choices of transition probabilities $\{P^n\}_{n\in[N]}$ respecting the given uncertainty intervals $\overline{\underline{P}}^{\phi(n)}$. Similarly, we define mixed strategy sets for the adversary at individual-level, $\Delta_I(\mathscr{P})$, and group-level, $\Delta_M(\mathscr{P})$, where $\Delta_M(\mathscr{P}) \subseteq \Delta_I(\mathscr{P})$ is a restricted space that assigns same transition probabilities to all arms within a group.

For $X, Y \in \{I(\text{individual}), M(\text{group})\}$, the regret game with X-level planner and Y-level adversary is noted as X/Y. The X/Y regret of a planner's mixed strategy $\alpha \in \Delta_X(\Pi)$ against an adversary's mixed strategy $\beta \in \Delta_Y(\mathscr{P})$ is:

$$R(\alpha, \beta) := \sum_{i\in[|\Delta_X(\Pi)|]} \sum_{j\in[|\Delta_Y(\mathscr{P})|]} \alpha_i \beta_j R(\pi_i, P_j) \,,$$

where α_i is the i^{th} pure strategy of the X-level planner and β_j is the j^{th} pure strategy of the Y-level adversary. Let $\alpha^\star_{X,Y}$ be the planner's mixed strategy of a X/Y game, defined:

$$\min_{\alpha\in\Delta_X(\Pi)} \max_{\beta\in\Delta_Y(\mathscr{P})} R(\alpha, \beta) = \max_{\beta\in\Delta_Y(\mathscr{P})} R(\alpha^\star_{X,Y}, \beta)$$

which holds since the regret game is a two-player zero sum game, making minimax regret equal to maximin reward. We call this the worst-case regret for $\alpha^\star_{X,Y}$.

We first show in Theorem 1[2] that, when all arms within the same group have the same transition intervals, the minimax I/I regret is equal to the minimax M/I regret.

Theorem 1. *The worst-case regrets of* $\alpha^\star_{I,I}$ *and* $\alpha^\star_{M,I}$ *against an adversary operating at the individual level is equal:*

$$\max_{\beta\in\Delta_I(\mathscr{P})} R(\alpha^\star_{I,I}, \beta) = \max_{\beta\in\Delta_I(\mathscr{P})} R(\alpha^\star_{M,I}, \beta) \,.$$

Similarly, in Theorem 2, we show that, when all arms within the same group have the same transition intervals, the minimax I/M regret is equal to the minimax M/M regret.

[2]Proofs of Theorem 1, 2, and 3 are given in Appendix E.

Theorem 2. *The worst-case regrets of $\alpha^\star_{I,M}$ and $\alpha^\star_{M,M}$ against an adversary operating at the group level are equal:*

$$\max_{\beta \in \Delta_M(\mathscr{P})} R(\alpha^\star_{I,M}, \beta) = \max_{\beta \in \Delta_M(\mathscr{P})} R(\alpha^\star_{M,M}, \beta) \ .$$

Finally, we use these results to establish our main result in Theorem 3 that the worst-case regret of $\alpha^\star_{M,M}$ is equal to the worst-case regret of $\alpha^\star_{I,I}$ when (1) all arms in the same group have the same intervals and (2) there exists a surjective function ψ that maps $\Delta_I(\mathscr{P})$ to $\Delta_M(\mathscr{P})$ that preserves the regret ordering of planner and adversary strategies (formal definition and example ψ given in Appendix E.1).

Theorem 3. *If there exists an order-preserving map, then the worst-case regret of $\alpha^\star_{M,M}$ is equal to that of $\alpha^\star_{I,I}$, against an individual-level adversary, that is,*

$$\max_{\beta \in \Delta_I(\mathscr{P})} R(\alpha^\star_{M,M}, \beta) = \max_{\beta \in \Delta_I(\mathscr{P})} R(\alpha^\star_{I,I}, \beta) \ .$$

Theorems 1, 2, and 3 together establish that the minimax regret–optimal strategy is the same whether the planner and adversary play at individual or group level. In particular, this result ensures that, under some conditions, the minimax regret–optimal strategy obtained by our algorithm GROUPS, which implements group-level planner and adversary, is also minimax regret–optimal against an individual-level adversary.

6 Experiments

6.1 Experiment Setup

ARMMAN maternal health domain Every week, ARMMAN's automated system delivers prerecorded health messages to each enrolled mother with information tailored to the mother's gestational age. If mothers stop listening to the messages, healthcare workers can deliver interventions to try to improve mothers' engagement. We evaluate *the increase in number of health messages mothers listen to using GROUPS to target interventions* compared to existing baselines. To construct a simulation environment, we use a real anonymized dataset from ARMMAN's records of weekly program engagement data for 15,336 mothers (though we note that ARMMAN's larger service areas operate on the scale of hundreds of thousands). A mother is "engaged" if they listen to at least 30 seconds of a message that week. Thus, states are $\{\texttt{not engaged}, \texttt{engaged}\}$ with rewards 0 and 1, respectively. To create an arm–group mapping, we run K-means clustering on the engagement data and compute uncertainty intervals via bootstrapping followed by multiple imputation to compute standard deviations of the means (Schomaker and Heumann 2018). Statistics on the uncertainty intervals and group sizes are shown in appendix Figs. 9 and 10. For details on the dataset and consent for collection, see appendix K.

In the experiments, the default parameters match the intervention setup used by ARMMAN, i.e., budget $K = 100$, $N = 15{,}320$ mothers, and $M = 40$ groups. For sensitivity analysis, we vary the budget, horizon, and number of mothers. Additional analysis varying uncertainty interval width, number of groups, and distribution of group sizes are included in appendix Fig. 7.

Additional domains To demonstrate wider applicability, we include results from two additional domains. The **TB** domain is constructed from an anonymized dataset of daily adherence to tuberculosis medication (Killian et al. 2019). States, rewards, and groups were derived analogously to the maternal health setting; complete details are in appendix L, including group statistics in Figs. 11 and 12. In our experiments, the default setting has $N = 8{,}350$ arms, $M = 60$ groups, budget $K = N/10$, and $A_\sigma = 3$, i.e., interval width of 3 standard deviations. We vary the budget, number of groups, and A_σ. Finally, we use the **Synthetic** benchmark domain from recent robust RMAB work (Killian et al. 2022). This domain considers three "arm types" $[U, V, W]$ with different intervals, designed so that non-robust policies incur greater regret than robust ones. We augment the domain to allow homogeneous groups of each arm type, where the size and proportion of groups of each type may vary. In our experiments, the default setting has $N = 18{,}000$ arms, $M = 36$ groups, where $1/3$ of groups are composed of each of the arm types, and budget $K = 100$. We run sensitivity analysis on K, the proportion of groups made up of each arm type, and a "block group" setting which joins all arms of a given type into a single group.

Evaluation To evaluate performance, we plan at the group level but simulate individuals within groups independently, where each individual undergoes state transitions based on their own state, action, and transition probabilities. All experiments use horizon $H = 10$ and report the average of 30 seeds. We measure total reward with discount factor $\gamma = 0.9$. In Fig. 3, we evaluate each approach in terms of regret (Eq. 1), computed by simulating each planner strategy against the full set of adversary pure strategies and selecting one that maximizes regret. Note, there is no actual deployment of the proposed algorithm; all results are simulated.

Baselines First, we compare against the state-of-the-art robust RMAB method, *DDLPO*, for small settings in which DDLPO can complete (Killian et al. 2022). For larger-scale experiments with tens of thousands of arms, no other robust methods are tractable, so we compare against several scalable non-robust baselines. Mate et al.'s non-robust baseline assumes all environment parameters take the median of their uncertainty intervals then computes a reward-maximizing WIP; this strategy was employed in a recent real-world pilot (Mate et al. 2022). We consider two additional non-robust variants which assume that all parameters take the lower bound of the uncertainty interval (*pessimist*) or the upper bound (*optimist*), then compute a WIP strategy. Finally, *random* plans a WIP strategy against an environment that is uniformly randomly sampled from the uncertainty intervals.

6.2 Results

Fig. 3 shows GROUPS outperforms baselines in terms of max regret across several settings. Fig. 3(a–c) shows results for the **maternal health** setting of ARMMAN. In particular, Fig. 3(c) shows that GROUPS scales past 300,000 arms, representing more than a $1000\times$ increase over the robust state-of-the-art to meet a key need of real-world deployment settings. Moreover, across experiments, the max re-

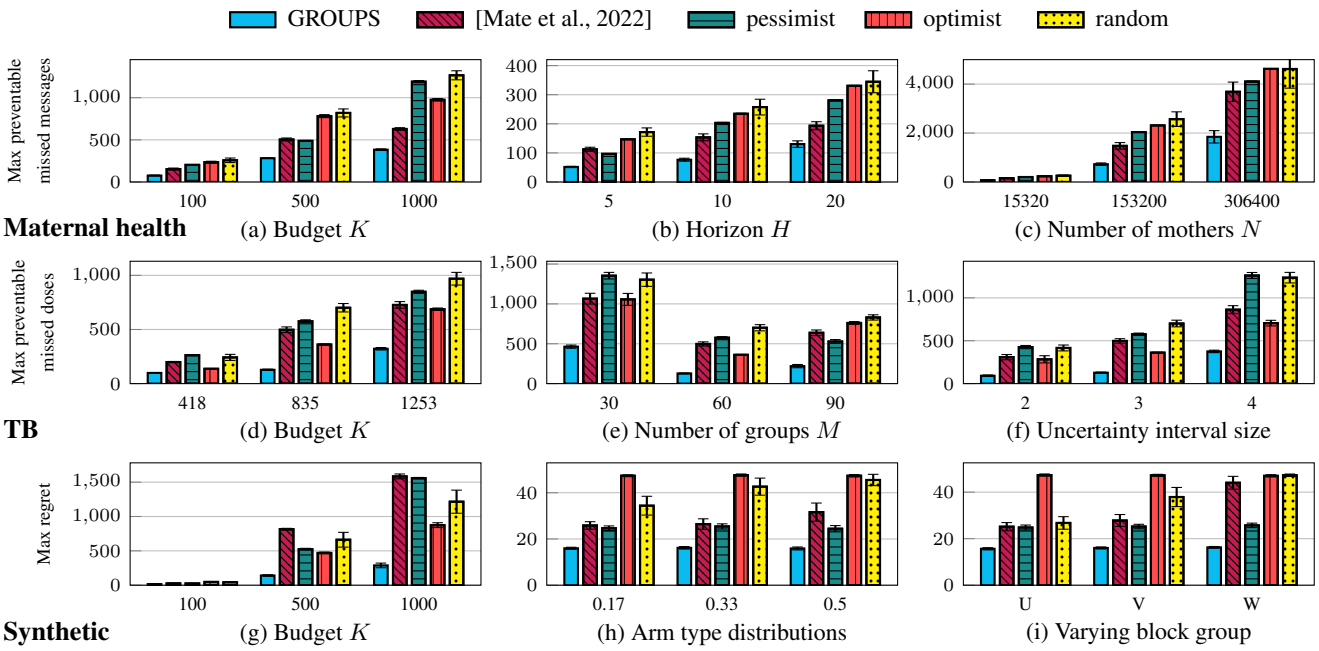

Figure 3: Max regret (lower is better) incurred by GROUPS, our robust solution approach, compared to non-robust baselines across various settings. *(a–c) Maternal health*. For (c), the number of arms is increased by multiplying each group size by a constant factor, i.e., 1, 10, and 20, but M is constant. *(d–f) TB*. For (d), budgets are 5%, 10%, and 15% of N. *(g–i) Synthetic*. For (h), the x-axis is the fraction of groups of arm type U — the fraction of type V is always 0.33, and the remaining fraction are type W. For (i) the x-axis denotes the arm type that has been combined into a single group of 6000 arms, where the other two types are split across 12 groups each of size 500. In the maternal health and TB settings, regret can be interpreted, in real-world terms, as the maximum preventable missed health messages and doses, respectively, across the uncertainty space.

gret of GROUPS is nearly half that of the non-robust strategy used in Mate et al. (2022). In other words, our simulations demonstrate that compared to the best non-robust strategy **GROUPS could prevent mothers from missing thousands of pregnancy-related health messages, each containing potentially life-saving care information.**

On the **TB** domain (Fig. 3(d–f)), we see again that GROUPS performs well across various strategies for grouping and computing uncertainty, even with very imbalanced group sizes. On the **synthetic** domain (Fig. 3(g–i)), across various budgets and grouping strategies, the non-robust baselines vary in performance and are sometimes worse than random, demonstrating the need for reliable robust policies. Moreover, Table 1 shows that GROUPS even outperforms the state-of-the-art DDLPO in terms of regret on the synthetic benchmark dataset for problems sizes small enough for DDLPO to complete (i.e., $N < 100$). The superior performance of GROUPS is due to our Whittle-based policies which specialize to two-action settings, in contrast to the more general but highly stochastic deep learning–based policies of DDLPO.

Supported by Theorem 3, GROUPS scales significantly without incurring additional regret. In Appendix I, we demonstrate the significant runtime improvement of GROUPS as M decreases, holding N constant. The scalability of our approach is critical for robust RMAB solutions to actually be deployed in real-world, low-resource settings.

	GROUPS	DDLPO
$N = 6$	0.64 ± 0.05	1.00 ± 0.06
$N = 9$	0.47 ± 0.06	0.98 ± 0.05
$N = 12$	0.45 ± 0.06	0.88 ± 0.05

Table 1: Regret of GROUPS vs. robust method DDLPO on Synthetic. We set $M = N$ and $K = 1$ to match the evaluation in Killian et al. (2022). GROUPS incurs less regret.

7 Conclusion

The GROUPS algorithm we introduce presents several key advances to make RMABs more useful in practice, enabling simultaneous *scaleup* and *robustness to uncertainty*. We are working with ARMMAN to deploy GROUPS to positively impact maternal health, demonstrating the real-world capabilities this work enables. Most notably, **our simulation experiments demonstrate that our robust planning method could help ARMMAN prevent mothers from missing thousands of health messages**, a promising result that we hope to translate into practice to help deliver life-saving health information to otherwise under-served mothers.

Acknowledgments

J.A.K. was supported by an NSF Graduate Research Fellowship under grant DGE1745303. A.B. was supported by the Harvard Center for Research on Computation and Society. L.X. was supported by a Google PhD Fellowship, and was a Student Researcher at Google for part of the project.

References

Alam, M.; D'Este, C.; Banwell, C.; and Lokuge, K. 2017. The impact of mobile phone based messages on maternal and child healthcare behaviour: a retrospective cross-sectional survey in Bangladesh. *BMC Health Serv. Res.*, 17(1).

Biswas, A.; Aggarwal, G.; Varakantham, P.; and Tambe, M. 2021. Learn to Intervene: An Adaptive Learning Policy for Restless Bandits in Application to Preventive Healthcare. In *IJCAI*.

Braziunas, D.; and Boutilier, C. 2007. Minimax regret based elicitation of generalized additive utilities. In *UAI-07*.

Carvalho, N.; Salehi, A.; and Goldie, S. 2013. National and sub-national analysis of the health benefits and cost-effectiveness of strategies to reduce maternal mortality in Afghanistan. *Health Policy Plan*, 28(1).

Evans, W. D.; Wallace, J. L.; and Snider, J. 2012. Pilot evaluation of the text4baby mobile health program. *BMC public health*, 12(1).

Gates Foundation. 2021. Global Progress and Projections for Maternal Mortality. https://www.gatesfoundation.org/goalkeepers/report/2021-report/progress-indicators/maternal-mortality/. Accessed: 2023-03-30.

Glazebrook, K. D.; Hodge, D. J.; and Kirkbride, C. 2011. General notions of indexability for queueing control and asset management. *Ann Appl Probab*, 21(3): 876–907.

Green, H.; and Riddington, A. 2020. Gender inclusive language in perinatal services: Mission statement and rationale. *Brighton, England: Brighton and Sussex University Hospitals*.

Gribble, K. D.; Bewley, S.; Bartick, M. C.; Mathisen, R.; Walker, S.; Gamble, J.; Bergman, N. J.; Gupta, A.; Hocking, J. J.; and Dahlen, H. G. 2022. Effective communication about pregnancy, birth, lactation, breastfeeding and newborn care: the importance of sexed language. *Frontiers in global women's health*, 3.

Gurobi Optimization, LLC. 2021. Gurobi Optimizer Reference Manual.

Huo, X.; and Fu, F. 2017. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society open science*, 4(11): 171377.

Johnson & Johnson. 2017. MomConnect: Connecting Women to Care, One Text at a Time. https://www.jnj.com/our-giving/momconnect-connecting-women-to-care-one-text-at-a-time. Accessed: 2023-03-30.

Killian, J. A.; Biswas, A.; Shah, S.; and Tambe, M. 2021. Q-Learning Lagrange Policies for Multi-action Restless Bandits. In *KDD*.

Killian, J. A.; Wilder, B.; Sharma, A.; Shah, D.; Choudhary, V.; Dilkina, B.; and Tambe, M. 2019. Learning to Prescribe Interventions for Tuberculosis Patients Using Digital Adherence Data. In *SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Killian, J. A.; Xu, L.; Biswas, A.; and Tambe, M. 2022. Robust Restless Bandits: Tackling Interval Uncertainty with Deep Reinforcement Learning. *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Lanctot, M.; Zambaldi, V.; Gruslys, A.; Lazaridou, A.; Tuyls, K.; Pérolat, J.; Silver, D.; and Graepel, T. 2017. A unified game-theoretic approach to multiagent reinforcement learning. *NeurIPS-17*, 30.

Li, S.; Wu, Y.; Cui, X.; et al. 2019. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *AAAI*.

Maillard, O.-A. 2013. Robust risk-averse stochastic multi-armed bandits. In *ALT*. Springer.

Mate, A.; Madaan, L.; Taneja, A.; et al. 2022. Field Study in Deploying Restless Multi-Armed Bandits: Assisting Non-Profits in Improving Maternal and Child Health. *AAAI*.

McMahan, H. B.; Gordon, G. J.; and Blum, A. 2003. Planning in the presence of cost functions controlled by an adversary. In *ICML-03*.

Meh, C.; Sharma, A.; Ram, U.; Fadel, S.; Correa, N.; Snelgrove, J. W.; Shah, P.; Begum, R.; Shah, M.; Hana, T.; et al. 2021. Trends in maternal mortality in India over two decades in nationally representative surveys. *BJOG*.

Murthy, N.; Chandrasekharan, S.; Prakash, M. P.; Kaonga, N. N.; Peter, J.; Ganju, A.; and Mechael, P. N. 2019. The impact of an mHealth voice message service (mMitra) on infant care knowledge, and practices among low-income women in India: findings from a Pseudo-Randomized controlled trial. *Maternal and child health journal*, 23(12): 1658–1669.

Musiimenta, A.; Tumuhimbise, W.; Pinkwart, N.; et al. 2021. A mobile phone-based multimedia intervention to support maternal health is acceptable and feasible among illiterate pregnant women in Uganda. *Digital Health*, 7.

Nakhleh, K.; Ganji, S.; Hsieh, P.-C.; Hou, I.; Shakkottai, S.; et al. 2021. NeurWIN: Neural Whittle Index Network For Restless Bandits Via Deep RL. *Advances in Neural Information Processing Systems*, 34.

Papadimitriou, C. H.; and Tsitsiklis, J. N. 1999. The complexity of optimal queuing network control. *Math. Oper. Res.*, 24(2).

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust adversarial reinforcement learning. In *ICML*. PMLR.

Rioux, C.; Weedon, S.; London-Nadeau, K.; Paré, A.; Juster, R.-P.; Roos, L. E.; Freeman, M.; and Tomfohr-Madsen, L. M. 2022. Gender-inclusive writing for epidemiological research on pregnancy. *J Epidemiol Community Health*, 76(9): 823–827.

Schomaker, M.; and Heumann, C. 2018. Bootstrap inference when using multiple imputation. *Stat Med*, 37(14).

Tamrat, T.; and Kachnowski, S. 2012. Special delivery: an analysis of mHealth in maternal and newborn health programs and their outcomes around the world. *Matern Child Health J.*, 16(5).

United Nations. 2021. Sustainable Development Goal 3: Ensure healthy lives and promote well-being for all at all ages. https://sdgs.un.org/goals/goal3. Accessed: 2023-04-05.

Wang, K.; Xu, L.; Taneja, A.; and Tambe, M. 2022. Optimistic Whittle Index Policy: Online Learning for Restless Bandits. *arXiv preprint arXiv:2205.15372*.

Wang, S.; Huang, L.; and Lui, J. 2020. Restless-UCB, an Efficient and Low-complexity Algorithm for Online Restless Bandits. *Advances in Neural Information Processing Systems*, 33: 11878–11889.

Watterson, J. L.; Walsh, J.; and Madeka, I. 2015. Using mHealth to improve usage of antenatal care, postnatal care, and immunization: a systematic review of the literature. *BioMed Res. Int.*

Weber, R. R.; and Weiss, G. 1990. On an index policy for restless bandits. *J. Appl. Probab.*, 27(3).

Whittle, P. 1988. Restless bandits: Activity allocation in a changing world. *J. Appl. Probab.*, 25(A).

WHO. 2023. Maternal mortality. https://www.who.int/news-room/fact-sheets/detail/maternal-mortality. Accessed: 2023-03-30.

Xu, L.; Perrault, A.; Fang, F.; Chen, H.; and Tambe, M. 2021. Robust Reinforcement Learning Under Minimax Regret for Green Security. In *UAI*.