Scalable Decision-Focused Learning in Restless Multi-Armed Bandits with Application to Maternal and Child Health

Kai Wang[*†1], Shresth Verma[*2], Aditya Mate[†1], Sanket Shah[1], Aparna Taneja[2], Neha Madhiwalla[3], Aparna Hegde[3], Milind Tambe[1,2]

[1]Harvard University
[2]Google Research
[3]ARMMAN
{kaiwang, aditya_mate, sanketshah}@g.harvard.edu, {aparnataneja, milindtambe}@google.com,
{neha, aparnahegde}@armman.org

Abstract

This paper studies restless multi-armed bandit (RMAB) problems with unknown arm transition dynamics but with known correlated arm features. The goal is to learn a model to predict transition dynamics given features, where the Whittle index policy solves the RMAB problems using predicted transitions. However, prior works often learn the model by maximizing the predictive accuracy instead of final RMAB solution quality, causing a mismatch between training and evaluation objectives. To address this shortcoming, we propose a novel approach for decision-focused learning in RMAB that directly trains the predictive model to maximize the Whittle index solution quality. We present three key contributions: (i) we establish differentiability of the Whittle index policy to support decision-focused learning; (ii) we significantly improve the scalability of decision-focused learning approaches in sequential problems, specifically RMAB problems; (iii) we apply our algorithm to a previously collected dataset of maternal and child health to demonstrate its performance. Indeed, our algorithm is the first for decision-focused learning in RMAB that scales to real-world problem sizes.

Introduction

Restless multi-armed bandits (RMABs) (Weber and Weiss 1990; Tekin and Liu 2012) are composed of a set of heterogeneous arms and a planner who can pull multiple arms under budget constraint at each time step to collect rewards. Different from the classic stochastic multi-armed bandits (Gittins, Glazebrook, and Weber 2011; Bubeck and Cesa-Bianchi 2012), the state of each arm in an RMAB can change even when the arm is not pulled, where each arm follows a Markovian process to transition between different states with transition probabilities dependent on arms and the pulling decision. Rewards are associated with different arm states, where the planner's goal is to plan a sequential pulling policy to maximize the total reward received from all arms. RMABs are commonly used to model sequential scheduling problems where limited resources must be strategically assigned to different tasks sequentially to maximize performance. Examples include machine maintenance (Glazebrook, Ruiz-Hernandez, and Kirkbride 2006), cognitive radio sensing problem (Bagheri and Scaglione 2015), and healthcare (Mate et al. 2022).

In this paper, we study offline RMAB problems with unknown transition dynamics but with given arm features. The goal is to learn a mapping from arm features to transition dynamics, which can be used to infer the dynamics of unseen RMAB problems to plan accordingly. Prior works (Mate et al. 2022; Sun et al. 2018) often learn the transition dynamics from the historical pulling data by *maximizing the predictive accuracy*. However, RMAB performance is evaluated *by its solution quality* derived from the predicted transition dynamics, which leads to a mismatch in the training objective and the evaluation objective. Previously, decision-focused learning (Wilder, Dilkina, and Tambe 2019) has been proposed to directly optimize the solution quality rather than predictive accuracy, by integrating the one-shot optimization problem (Donti, Amos, and Kolter 2017; Perrault et al. 2020) or sequential problems (Wang et al. 2021; Futoma, Hughes, and Doshi-Velez 2020) as a differentiable layer in the training pipeline. Unfortunately, while decision-focused learning can successfully optimize the evaluation objective, it is computationally extremely expensive due to the presence of the optimization problems in the training process. Specifically, for RMAB problems, the computation cost of decision-focused learning arises from the complexity of the sequential problems formulated as Markov decision processes (MDPs), which limits the applicability to RMAB problems due to the PSPACE hardness of finding the optimal solution (Papadimitriou and Tsitsiklis 1994).

Our main contribution is a novel and scalable approach for decision-focused learning in RMAB problems using Whittle index policy, a commonly used approximate solution in RMABs. Our three key contributions are (i) we establish the differentiability of Whittle index policy to support decision-focused learning to directly optimize the RMAB solution quality; (ii) we show that our approach of differen-

[*]These authors contributed equally.
[†]Work done during an internship at Google Research.

tiating through Whittle index policy improves the scalability of decision-focused learning in RMAB; (iii) we apply our algorithm to an anonymized maternal and child health RMAB dataset previously collected by ARMMAN (2022) to evaluate the performance of our algorithm in simulation.

We establish the differentiability of Whittle index by showing that Whittle index can be expressed as a solution to a full-rank linear system reduced from Bellman equations with transition dynamics as entries, which allows us to compute the derivative of Whittle index with respect to transition dynamics. On the other hand, to execute Whittle index policy, the standard selection process of choosing arms with top-k Whittle indices to pull is non-differentiable. We relax this non-differentiable process by using a differentiable soft top-k selection to establish differentiability. Our differentiable Whittle index policy enables decision-focused learning in RMAB problems to backpropagate from final policy performance to the predictive model. We significantly improve the scalability of decision-focused learning, where the computation cost of our algorithm $O(NM^{\omega+1})$ scales linearly in the number of arms N and polynomially in the number of states M with $\omega \approx 2.373$, while previous work scales exponentially $O(M^{\omega N})$. This significant reduction in computation cost is crucial for extending decision-focused learning to RMAB problems with large number of arms.

In our experiments, we apply decision-focused learning to RMAB problems to optimize importance sampling-based evaluation on synthetic datasets as well as an anonymized RMAB dataset about a maternal and child health program previously collected by (ARMMAN 2022) – these datasets are the basis of comparing different methods in simulation. We compare decision-focused learning with the two-stage method that trains to minimize the predictive loss. The two-stage method achieves the best predictive loss but significantly degraded solution quality. In contrast, decision-focused learning reaches a slightly worse predictive loss but with a much better importance sampling-based solution quality evaluation and the improvement generalizes to the simulation-based evaluation that is built from the data. Lastly, the scalability improvement is the crux of applying decision-focused learning to real-world RMAB problems: our algorithm can run decision-focused learning on the maternal and child health dataset with hundreds of arms, whereas state of the art is a 100-fold slower even with 20 arms and grows exponentially worse.

Related Work

Restless multi-armed bandits with given transition dynamics This line of research primarily focuses on solving RMAB problems to get a sequential policy. The complexity of solving RMAB problems optimally is known to be PSPACE hard (Papadimitriou and Tsitsiklis 1994). One approximate solution is proposed by Whittle (1988), where they use Lagrangian relaxation to decompose arms and compute the associated Whittle indices to define a policy. Specifically, the indexability condition (Akbarzadeh and Mahajan 2019; Wang et al. 2019) guarantees this Whittle index policy to be asymptotically optimal (Weber and Weiss 1990). In practice, Whittle index policy usually provides a near-optimal solution to RMAB problems.

Restless multi-armed bandits with missing transition dynamics When the transition dynamics are unknown in RMAB problems but an interactive environment is available, prior works (Tekin and Liu 2012; Liu, Liu, and Zhao 2012; Oksanen and Koivunen 2015; Dai et al. 2011) consider this as an online learning problem that aims to maximize the expected reward. However, these approaches become infeasible when interacting with the environment is expensive, e.g., healthcare problems (Mate et al. 2022). In this work, we consider the offline RMAB problem, and each arm comes with an arm feature that is correlated to the transition dynamics and can be learned from the past data.

Decision-focused learning The predict-then-optimize framework (Elmachtoub and Grigas 2021) is composed of a predictive problem that makes predictions on the parameters of the later optimization problem, and an optimization problem that uses the predicted parameters to come up with a solution, where the overall objective is the solution quality of the proposed solution. Standard two-stage learning method solves the predictive and optimization problems separately, leading to a mismatch of the predictive loss and the evaluation metric (Huang et al. 2019; Lambert et al. 2020; Johnson and Khoshgoftaar 2019). In contrast, decision-focused learning (Wilder, Dilkina, and Tambe 2019; Mandi et al. 2020; Elmachtoub, Liang, and McNellis 2020) learns the predictive model to directly optimize the solution quality by integrating the optimization problem as a differentiable layer (Amos and Kolter 2017; Agrawal et al. 2019) in the training pipeline. Our offline RMAB problem is a predict-then-optimize problem, where we first (offline) learn a mapping from arm features to transition dynamics from the historical data (Mate et al. 2022; Sun et al. 2018), and the RMAB problem is solved using the predicted transition dynamics accordingly. Prior work (Mate et al. 2022) is limited to using two-stage learning to solve the offline RMAB problems. While decision-focused learning in sequential problems were primarily studied in the context of MDPs (Wang et al. 2021; Futoma, Hughes, and Doshi-Velez 2020) they come with an expensive computation cost that immediately becomes infeasible in large RMAB problems.

Model: Restless Multi-armed Bandit

An instance of the restless multi-armed bandit (RMAB) problem is composed of a set of N arms, each is modeled as an independent Markov decision process (MDP). The i-th arm in a RMAB problem is defined by a tuple $(\mathcal{S}, \mathcal{A}, R_i, P_i)$. $\mathcal{S}$ and $\mathcal{A}$ are the identical state and action spaces across all arms. $R_i, P_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ are the reward and transition functions associated to arm i. We consider finite state space with $|\mathcal{S}| = M$ fully observable states and action set $\mathcal{A} = \{0, 1\}$ corresponding to not pulling or pulling the arm, respectively. For each arm i, the reward is denoted by $R_i(s_i, a_i, s'_i) = R(s_i)$, i.e., the reward $R(s_i)$ only depends on the current state s_i, where $R : \mathcal{S} \to \mathbb{R}$ is a vector of size M. Given the state s_i and action a_i,

$P_i(s_i, a_i) = [P_i(s_i, a_i, s'_i)]_{s'_i \in \mathcal{S}}$ defines the probability distribution of transitioning to all possible next states $s'_i \in \mathcal{S}$.

In a RMAB problem, at each time step $t \in [T]$, the learner observes $\boldsymbol{s}_t = [s_{t,i}]_{i \in [N]} \in \mathcal{S}^N$, the states of all arms. The learner then chooses action $\boldsymbol{a}_t = [a_{t,i}]_{i \in [N]} \in \mathcal{A}^N$ denoting the pulling actions of all arms, which has to satisfy a budget constraint $\sum_{i \in [N]} a_{t,i} \leq K$, i.e., the learner can pull at most K arms at each time step. Once the action is chosen, arms receive action $\boldsymbol{a}_t$ and transitions under P with rewards $\boldsymbol{r}_t = [r_{t,i}]_{i \in [N]}$ accordingly. We denote a full trajectory by $\tau = (\boldsymbol{s}_1, \boldsymbol{a}_1, \boldsymbol{r}_1, \cdots, \boldsymbol{s}_T, \boldsymbol{a}_T, \boldsymbol{r}_T)$. The total reward is defined by the summation of the discounted reward across T time steps and N arms, i.e., $\sum_{t=1}^{T} \gamma^{t-1} \sum_{i \in [N]} r_{t,i}$, where $0 < \gamma \leq 1$ is the discount factor.

A policy is denoted by π, where $\pi(\boldsymbol{a} \mid \boldsymbol{s})$ is the probability of choosing action $\boldsymbol{a}$ given state $\boldsymbol{s}$. Additionally, we define $\pi(a_i = 1 \mid \boldsymbol{s})$ to be the marginal probability of pulling arm i given state $\boldsymbol{s}$, where $\pi(\boldsymbol{s}) = [\pi(a_i = 1 \mid \boldsymbol{s})]_{i \in [N]}$ is a vector of arm pulling probabilities. Specifically, we use π^* to denote the optimal policy that optimizes the cumulative reward, while π^{solver} to denote a near-optimal policy solver.

Problem Statement

This paper studies the RMAB problem where we do not know the transition probabilities $P = \{P_i\}_{i \in [N]}$ in advance. Instead, we are given a set of features $\boldsymbol{x} = \{x_i \in \mathcal{X}\}_{i \in [N]}$, each corresponding to one arm. The goal is to learn a mapping $f_w : \mathcal{X} \to \mathcal{P}$, parameterized by weights w, to make predictions on the transition probabilities $P = f_w(\boldsymbol{x}) := \{f_w(x_i)\}_{i \in [N]}$. The predicted transition probabilities are later used to solve the RMAB problem to derive a policy $\pi = \pi^{\text{solver}}(f_w(\boldsymbol{x}))$. The performance of the model f is evaluated by the performance of the proposed policy π.

Training and Testing Datasets

To learn the mapping f_w, we are given a set of RMAB instances as training examples $\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}, \mathcal{T})\}$, where each instance is composed of a RMAB problem with feature $\boldsymbol{x}$ that is correlated to the unknown transition probabilities P, and a set of realized trajectories $\mathcal{T} = \{\tau^{(j)}\}_{j \in J}$ generated from a given behavior policy π_{beh} that determined how to pull arms in the past. The testing set $\mathcal{D}_{\text{test}}$ is defined similarly but hidden at training time.

Evaluation Metrics

Predictive loss To measure the correctness of transition probabilities $P = \{P_i\}_{i \in [N]}$, we define the predictive loss as the average negative log-likelihood of seeing the given trajectories $\mathcal{T}$, i.e., $\mathcal{L}(P, \mathcal{T}) := -\log \Pr(\mathcal{T} \mid P) = -\mathbb{E}_{\tau \sim \mathcal{T}} \sum_{t \in [T]} \log P(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$. Therefore, we can define the predictive loss of a model f_w on dataset $\mathcal{D}$ by:

$$\mathbb{E}_{(\boldsymbol{x}, \mathcal{T}) \sim \mathcal{D}} \, \mathcal{L}(f_w(\boldsymbol{x}), \mathcal{T}) \tag{1}$$

Policy evaluation On the other hand, given transition probabilities P, we can solve the RMAB problem to derive a policy $\pi^{\text{solver}}(P)$. We can use the historical trajectories $\mathcal{T}$ to evaluate how good the policy performs, denoted by $\text{Eval}(\pi^{\text{solver}}(P), \mathcal{T})$. Given dataset $\mathcal{D}$, we can evaluate the predictive model f_w on dataset $\mathcal{D}$ by:

$$\mathbb{E}_{(\boldsymbol{x}, \mathcal{T}) \sim \mathcal{D}} \, \text{Eval}(\pi^{\text{solver}}(f_w(\boldsymbol{x})), \mathcal{T}) \tag{2}$$

Two common types of policy evaluation are importance sampling-based off-policy policy evaluation and simulation-based evaluation, which will be discussed in Section .

Learning Methods

Two-stage learning To learn the predictive model f_w, we can minimize Equation 1 by computing gradient $\frac{d\mathcal{L}(f_w(\boldsymbol{x}), \mathcal{T})}{dw}$ to run gradient descent. However, this training objective (Equation 1) differs from the evaluation objective (Equation 2), which often leads to suboptimal performance.

Decision-focused learning In contrast, we can directly run gradient ascent to maximize Equation 2 by computing the gradient $\frac{d\text{Eval}(\pi^{\text{solver}}(f_w(\boldsymbol{x})), \mathcal{T})}{dw}$. However, in order to compute the gradient, we need to differentiate through the policy solver π^{solver} and the corresponding optimal solution. Unfortunately, finding the optimal policy in RMABs is expensive and the policy is high-dimensional. Both of these challenges prevent us from computing the gradient to achieve decision-focused learning.

Decision-focused Learning in RMABs

In this paper, instead of grappling with the optimal policy, we consider the Whittle index policy (Whittle 1988) – the dominant solution paradigm used to solve the RMAB problem. Whittle index policy is easier to compute and has been shown to perform well in practice. In this section we establish that it is also possible to backpropagate through the Whittle index policy. This differentiability of Whittle index policy allows us to run decision-focused learning to directly maximize the performance in the RMAB problem.

Whittle Index and Whittle Index Policy

Informally, the Whittle index of an arm captures the added value derived from pulling that arm. The key idea is to determine the Whittle indices of all arms and to pull the arms with the highest values of the index.

To evaluate the value of pulling an arm i, we consider the notion of 'passive subsidy', which is a hypothetical exogenous compensation m rewarded for not pulling the arm (i.e. for choosing action $a = 0$). Whittle index is defined as the smallest subsidy necessary to make pulling as rewarding as not pulling, assuming indexability (Liu and Zhao 2010):

Definition 0.1 (Whittle index)**.** Given state $u \in \mathcal{S}$, we define the Whittle index associated to state u by:

$$W_i(u) := \inf\nolimits_m \{Q_i^m(u; a = 0) = Q_i^m(u; a = 1)\} \tag{3}$$

where the value functions are defined by the following Bellman equations, augmented with subsidy m for action $a = 0$.

$$V_i^m(s) = \max\nolimits_a Q_i^m(s; a) \tag{4}$$

$$Q_i^m(s; a) = m\mathbf{1}_{a=0} + R(s) + \gamma \sum\nolimits_{s'} P_i(s, a, s') V_i^m(s') \tag{5}$$

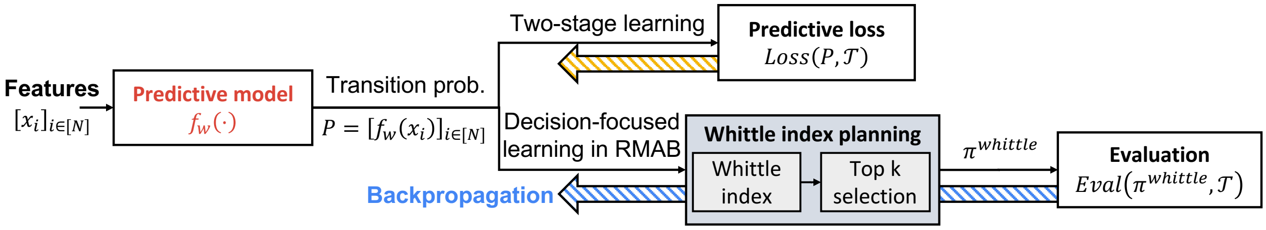


Figure 1: This flowchart visualizes different methods of learning the predictive model. Two-stage learning directly compares the predicted transition probabilities with the given data to define a predictive loss to run gradient descent. Decision-focused learning instead goes through a policy solver using Whittle index policy to estimate the final evaluation and run gradient ascent.

Given the Whittle indices of all arms and all states $W = [W_i(u)]_{i\in[N],u\in\mathcal{S}}$, the Whittle index policy is denoted by $\pi^{\text{whittle}} : \mathcal{S}^N \longrightarrow [0,1]^N$, which takes the states of all arms as input to compute their Whittle indices and output the probabilities of pulling arms. This policy repeats for every time step to pull arms based on the index values.

Decision-focused Learning Using Whittle Index Policy

Instead of using the optimal policy π^* to run decision-focused learning with expensive computation cost, we use Whittle index policy π^{whittle} to determine how to pull arms as an approximate solution. In this case, in order to run decision-focused learning, we need to compute the derivative of the evaluation metric by chain rule:

$$\frac{d\text{Eval}(\pi^{\text{whittle}}, \mathcal{T})}{dw} = \frac{d\text{Eval}(\pi^{\text{whittle}}, \mathcal{T})}{d\pi^{\text{whittle}}} \frac{d\pi^{\text{whittle}}}{dW} \frac{dW}{dP} \frac{dP}{dw} \tag{6}$$

where W is the Whittle indices of all states under the predicted transition probabilities P. The policy π^{whittle} is the Whittle index policy induced by W. The flowchart is illustrated in Figure 1.

The term $\frac{d\text{Eval}(\pi^{\text{whittle}}, \mathcal{T})}{d\pi^{\text{whittle}}}$ can be computed via policy gradient theorem (Sutton, Barto et al. 1998), and the term $\frac{dP}{dw}$ can be computed using auto-differentiation. However, there are still two challenges remaining: (i) how to differentiate through Whittle index policy to get $\frac{d\pi^{\text{whittle}}}{dW}$ (ii) how to differentiate through Whittle index computation to derive $\frac{dW}{dP}$.

Differentiability of Whittle Index Policy

A common choice of Whittle index policy is defined by:

Definition 0.2 (Strict Whittle index policy)**.**

$$\pi_W^{\text{strict}}(\boldsymbol{s}) = \mathbf{1}_{\text{top-k}([W_i(s_i)]_{i\in[N]})} \in \{0,1\}^N \tag{7}$$

which selects arms with the top-k Whittle indices to pull.

However, the strict top-k operation in the strict Whittle index policy is non-differentiable, which prevents us from computing a meaningful estimate of $\frac{d\pi^{\text{whittle}}}{dW}$ in Equation 6. We circumvent this issue by relaxing the top-k selection to a soft-top-k selection (Xie et al. 2020), which can be expressed as an optimal transport problem with regularization, making it differentiable. We apply soft-top-k to define a new differentiable soft Whittle index policy:

Definition 0.3 (Soft Whittle index policy)**.**

$$\pi_W^{\text{soft}}(\boldsymbol{s}) = \text{soft-top-k}([W_j(s_i)]_{i\in[N]}) \in [0,1]^N \tag{8}$$

Using the soft Whittle index policy, the policy becomes differentiable and we can compute $\frac{d\pi^{\text{whittle}}}{dW}$.

Differentiability of Whittle Index

The second challenge is the differentiability of Whittle index. Whittle indices are often computed using value iteration and binary search (Qian et al. 2016; Mate et al. 2020) or mixed integer linear program. However, these operations are not differentiable and we cannot compute the derivative $\frac{dW}{dP}$ in Equation 6 directly.

Main idea After computing the Whittle indices and the value functions of each arm i, the key idea is to construct linear equations that link the Whittle index with the transition matrix P_i. Specifically, we achieve this by resolving the $\max$ operator in Equation 4 of Definition 0.1 by determining the optimal actions a from the pre-computed value functions. Plugging back in Equation 5 and manipulating as shown below yields linear equations in the Whittle index $W_i(u)$ and transition matrix P_i, which can be expressed as a full-rank linear system in P_i, with the Whittle index as a solution. This makes the Whittle index differentiable in P_i.

Selecting Bellman equation Let u and arm i be the target state and target arm to compute the Whittle index. Assume we have precomputed the Whittle index $m = W_i(u)$ for state u and the corresponding value functions $[V_i^m(s)]_{s\in\mathcal{S}}$ for all states under the same passive subsidy $m = W_i(u)$. Equation 5 can be combined with Equation 4 to get:

$$V_i^m(s) \geq \left\{ \begin{array}{l} m + R(s) + \gamma \sum_{s'\in\mathcal{S}} P_i(s, a=0, s') V_i^m(s') \\ R(s) + \gamma \sum_{s'\in\mathcal{S}} P_i(s, a=1, s') V_i^m(s') \end{array} \right. \tag{9}$$

where $m = W_i(u)$.

For each $s \in S$, at least one of the equalities in Equation 9 holds because one of the actions must be optimal and match the state value function $V_i^m(s)$. We can identify which equality holds by simply plugging in values of precomputed value functions $[V_i^m(s)]_{s\in\mathcal{S}}$. Furthermore, for the target state u, both equalities must hold because by the definition of Whittle index, the passive subsidy $m = W_i(u)$ makes both actions equally optimal, i.e. in Equation 3, $V_i^m(u) = Q_i^m(u, a=0) = Q_i^m(u, a=1)$ for $m = W_i(u)$.

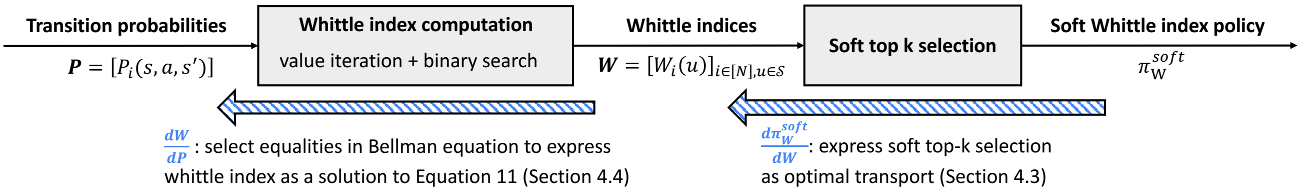

Transition probabilities
$\boldsymbol{P} = [P_i(s,a,s')]$
Whittle index computation
value iteration + binary search
Whittle indices
$\boldsymbol{W} = [W_i(u)]_{i\in[N],u\in\mathcal{S}}$
Soft top k selection
Soft Whittle index policy
π_{W}^{soft}
$\frac{dW}{dP}$: select equalities in Bellman equation to express whittle index as a solution to Equation 11 (Section 4.4)
$\frac{d\pi_W^{soft}}{dW}$: express soft top-k selection as optimal transport (Section 4.3)

Figure 2: We establish the differentiability of Whittle index policy using a soft top-k selection to construct a soft Whittle index policy, and the differentiability of Whittle index by expressing Whittle index as a solution to a linear system in Equation 11.

Thus Equation 9 can be written in matrix form:

$$\begin{bmatrix} \boldsymbol{V}_i^m \\ \boldsymbol{V}_i^m \end{bmatrix} \geq \begin{bmatrix} \boldsymbol{1}_M & \gamma \boldsymbol{P}_i(\mathcal{S}, a{=}0, \mathcal{S}) \\ \boldsymbol{0}_M & \gamma \boldsymbol{P}_i(\mathcal{S}, a{=}1, \mathcal{S}) \end{bmatrix} \begin{bmatrix} m \\ \boldsymbol{V}_i^m \end{bmatrix} + \begin{bmatrix} \boldsymbol{R}(S) \\ \boldsymbol{R}(S) \end{bmatrix} \quad (10)$$

where $\boldsymbol{V}_i^m := [V_i^m(s)]_{s\in\mathcal{S}}$, $\boldsymbol{R}(\mathcal{S}) = [R(s)]_{s\in\mathcal{S}}$, and $\boldsymbol{P}_i(\mathcal{S}, a, \mathcal{S}) := [P_i(s,a,s')]_{s,s'\in\mathcal{S}} \in \mathbb{R}^{M\times M}$.

By the aforementioned discussion, we know that there are at least $M+1$ equalities in Equation 10 while there are also only $M+1$ variables ($m \in \mathbb{R}$ and $\boldsymbol{V}_i^m \in \mathbb{R}^M$). Therefore, we rearrange Equation 10 and pick only the rows where equalities hold to get:

$$A \begin{bmatrix} \boldsymbol{1}_M & \gamma \boldsymbol{P}_i(\mathcal{S}, a = 0, \mathcal{S}) - I_M \\ \boldsymbol{0}_M & \gamma \boldsymbol{P}_i(\mathcal{S}, a = 1, \mathcal{S}) - I_M \end{bmatrix} \begin{bmatrix} m \\ \boldsymbol{V}_i^m \end{bmatrix} = A \begin{bmatrix} -\boldsymbol{R}(S) \\ -\boldsymbol{R}(S) \end{bmatrix} \quad (11)$$

where we use a binary matrix $A \in \{0,1\}^{(M+1)\times 2M}$ with a single 1 per row to extract the equality. For example, we can set $A_{ij} = 1$ if the j-th row in Equation 10 corresponds to the equality in Equation 9 with the i-th state in the state space S for $i \in [M]$, and the last row $A_{(M+1),j} = 1$ to mark the additional equality matched by the Whittle index definition (see Appendix for more details). Matrix A picks $M+1$ equalities out from Equation 10 to form Equation 11.

Equation 11 is a full-rank linear system with $m = W_i(u)$ as a solution. This expresses $W_i(u)$ as an implicit function of $\boldsymbol{P}$, allowing for computation of $\frac{dW_i(u)}{d\boldsymbol{P}}$ via autodifferentiation, thus achieving differentiability of the Whittle index. We repeat this process for every arm $i \in [N]$ and every state u. Figure 2 summarizes the differentiable Whittle index policy and the algorithm is shown in Algorithm 1.

Computation Cost and Backpropagation

It is well studied that Whittle index policy can be computed more efficiently than solving the RMAB problem as a large MDP problem. Here, we show that the use of Whittle index policy also demonstrates a large speed up in terms of backpropagating the gradient in decision-focused learning.

In order to use Equation 11 to compute the gradient of Whittle indices, we need to invert the left-hand-side of Equation 11 with dimensionality $M+1$, which takes $O(M^\omega)$ where $\omega \approx 2.373$ (Alman and Williams 2021) is the best known matrix inversion constant. Therefore, the overall computation of all N arms and M states is $O(NM^{\omega+1})$ per gradient step.

In contrast, the standard decision-focused learning differentiates through the optimal policy using the full Bellman equation with $O(M^N)$ variables, where inverting the large Bellman equation requires $O(M^{\omega N})$ cost per gradient step. Thus, our algorithm significantly reduces the computation cost to a linear dependency on the number of arms N. This significantly improves the scalability of decision-focused learning.

Extension to Partially Observable RMAB

For partially observable RMAB problem, we focus on a subclass of RMAB problem known as collapsing bandits (Mate et al. 2020). In collapsing bandits, belief states (Monahan 1982) are used to represent the posterior belief of the unobservable states. Specifically, for each arm i, we use $b_i \in \mathcal{B} = \Delta(\mathcal{S}) \subset [0,1]^M$ to denote the posterior belief of an arm, where each entry $b_i(s_i)$ denotes the probability that the true state is $s_i \in \mathcal{S}$. When arm i is pulled, the current true state $s_i \sim b_i$ is revealed and drawn from the posterior belief with expected reward $b_i^\top R$, where we can define the transition probability on the belief states. This process reduces partially observable states to fully observable belief states with in total MT states since the maximal horizon is T. Therefore, we can use the same technique to differentiate through Whittle indices of partially observable states.

Policy Evaluation Metrics

In this paper, we use two different variants of evaluation metric: importance sampling-based evaluation (Sutton, Barto et al. 1998) and simulation-based (model-based) evaluation.

Importance sampling-based Evaluation We adopt Consistent Weighted Per-Decision Importance Sampling (CWPDIS) (Thomas 2015) as our importance sampling-based evaluation. Given target policy π and a trajectory $\tau = \{s_1, a_1, r_1, \cdots, s_T, a_T, r_T\}$ executed by the behavior policy π_{beh}, the importance sampling weight is defined by $\rho_{ti} = \prod_{t'=1}^{t} \frac{\pi(a_{t',i}|s_{t'})}{\pi_{\text{beh}}(a_{t',i}|s_{t'})}$. We evaluate the policy π by:

$$\text{Eval}_{\text{IS}}(\pi, \mathcal{T}) = \sum\nolimits_{t\in[T], i\in[N]} \gamma^{t-1} \frac{\mathbb{E}_{\tau\sim\mathcal{T}}\left[r_{t,i}\rho_{ti}(\tau)\right]}{\mathbb{E}_{\tau\sim\mathcal{T}}\left[\rho_{ti}(\tau)\right]} \quad (12)$$

Importance sampling-based evaluations are often unbiased but with a larger variance due to the unstable importance sampling weights. CWPDIS normalizes the importance sampling weights to achieve a consistent estimate.

Simulation-based Evaluation An alternative way is to use the given trajectories to construct an empirical transition probability $\bar{P}$ to build a simulator and evaluate the target policy π. The variance of simulation-based evaluation is

Algorithm 1: Decision-focused Learning in RMAB

```
**Input:** training set $\mathcal{D}_{\text{train}}$, learning rate $r$, model $f_w$
**for** epoch $= 1, 2, \cdots$ and $(x, \mathcal{T}) \in \mathcal{D}_{\text{train}}$ **do**
    Predict $P = f_w(x)$ and compute Whittle indices $W(P)$.
    Let $\pi^{\text{whittle}} = \pi_W^{\text{soft}}$ and compute $\text{Eval}(\pi^{\text{whittle}}, \mathcal{T})$.
    Update $w = w + r \frac{d\text{Eval}(\pi^{\text{whittle}}, \mathcal{T})}{d\pi^{\text{whittle}}} \frac{d\pi^{\text{whittle}}}{dW} \frac{dW}{dP} \frac{dP}{dw}$, where $\frac{dW}{dP}$ is computed from Equation 11.
**end for**
**Return:** predictive model $f_w$
```

small, but it may require additional assumptions on the missing transition when the empirical transition $\bar{P}$ is not fully reconstructed.

Experiments

We compare two-stage learning (**TS**) with our decision-focused learning (**DF-Whittle**) that optimizes importance sampling-based evaluation directly. We consider three different evaluation metrics including predictive loss, importance sampling evaluation, and simulation-based evaluation to evaluate all learning methods. We perform experiments on three synthetic datasets including 2-state fully observable, 5-state fully observable, and 2-state partially observable RMAB problems. We also perform experiments on a real dataset on maternal and child health problem modelled as a 2-state fully observable RMAB problem with real features and historical trajectories. For each dataset, we use 70%, 10%, 20% of the RMAB problems as the training, validation, and testing sets, respectively. All experiments are averaged over 50 independent runs.

Synthetic datasets We consider RMAB problems composed of $N = 100$ arms, M states, budget $K = 20$, and time horizon $T = 10$ with a discount rate of $\gamma = 0.99$. The reward function is given by $R = [\frac{i-1}{M-1}]_{i \in [M]}$, while the transition probabilities are generated uniformly at random but with a constraint that pulling the arm ($a = 1$) is strictly better than not pulling the arm ($a = 0$) to ensure the benefit of pulling. To generate the arm features, we feed the transition probability of each arm to a randomly initialized neural network to generate fixed-length correlated features with size 16 per arm. The historical trajectories $\mathcal{T}$ with $|\mathcal{T}| = 10$ are produced by running a random behavior policy π_{beh}. The goal is to predict transition probabilities from the arm features and the training trajectories.

Real dataset The Maternal and Child Healthcare Mobile Health program operated by ARMMAN (2022) aims to improve dissemination of health information to pregnant women and mothers with an aim to reduce maternal, neonatal and child mortality and morbidity. ARMMAN serves expectant/new mothers in disadvantaged communities with *median daily family income of $3.22 per day* which is seen to be below the world bank poverty line (World Bank 2020). The program is composed of multiple enrolled beneficiaries and a planner who schedules service calls to improve the overall engagement of beneficiaries; engagement is measured in terms of total number of automated voice (health related) messages that the beneficiary engaged with. More precisely, this problem is modelled as a $M = 2$-state fully observable RMAB problem where each beneficiary's behavior is governed by an MDP with two states - Engaging and Non-Engaging state; engagement is determined by whether the beneficiary listens to an automated voice message (average length 115 seconds) for more than 30 seconds. The planner's task is to recommend a subset of beneficiaries every week to receive service calls from health workers to further improve their engagement behavior. We do not know the transition dynamics, but we are given beneficiaries' socio-demographic features to predict transition dynamics.

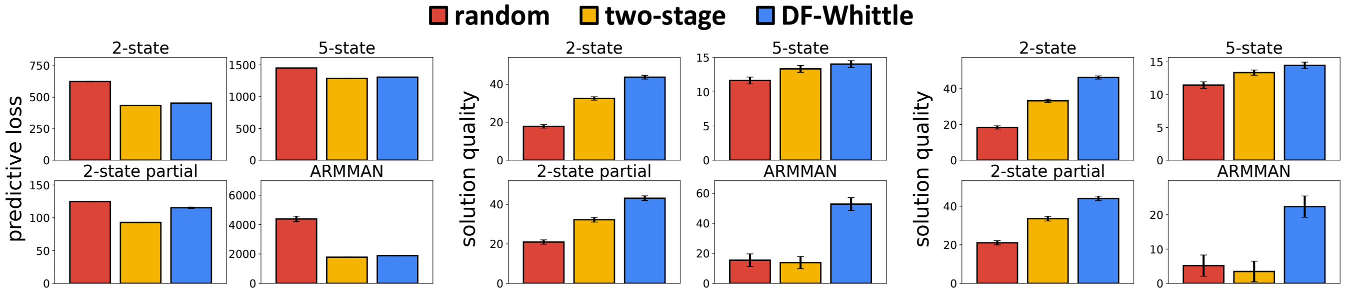
We use a subset of data from the large-scale anonymized quality improvement study performed by ARMMAN for $T = 7$ weeks, obtained from Mate et al. (2022), with beneficiary consent. In the study, a cohort of beneficiaries received Round-Robin policy, scheduling service calls in a fixed order, with a single trajectory $|\mathcal{T}| = 1$ per beneficiary that documents the calling decisions and the engagement behavior in the past. We randomly split the cohort into 8 training groups, 1 validation group, and 3 testing groups each with $N = 639$ beneficiaries and $K = 18$ budget formulated as an RMAB problem. The demographic features of beneficiaries are used to infer the missing transition dynamics.

Data usage All the datasets are anonymized. The experiments are secondary analysis using different evaluation metrics with approval from the ARMMAN ethics board. There is no actual deployment of the proposed algorithm at ARMMAN. For more details about the dataset, consent of data collection, please refer to Appendix and .

Experimental Results

Performance improvement and justification of objective mismatch In Figure 3, we show the performance of random policy, two-stage, and decision-focused learning (DF-Whittle) on three evaluation metrics - predictive loss, importance sampling-based evaluation and simulation-based evaluation for all domains. For the evaluation metrics, we plot the improvement against the no-action baseline that does not pull any arms throughout the entire RMAB problem. We observe that two-stage learning consistently converges to a smaller predictive loss, while DF-Whittle outperforms two-stage on all solution quality evaluation metrics significantly (p-value < 0.05) by alleviating the objective mismatch issue. This result also provides evidence of aforementioned objective mismatch, where the advantage of two-stage in the predictive loss does not translate to solution quality.

Significance in maternal and child care domain In the ARMMAN data in Figure 3, we assume limited resources that we can only select 18 out of 638 beneficiaries to make service call per week. Both random and two-stage method lead to around 15 more (IS-based evaluation) listening to automated voice messages among all beneficiaries throughout the 7-week program by $18 \times 7 = 126$ service calls, when compared to not scheduling any service call; this low



a Predictive loss

b IS-based evaluation

c Simulation-based evaluation

Figure 3: Comparison of predictive loss, importance sampling-based evaluation, and simulation-based evaluation on all synthetic domains and the real ARMMAN dataset. For the evaluation metrics, we plot the improvement against the no-action baseline that does not pull any arm. Although two-stage method achieves the smallest predictive loss, decision-focused learning consistently outperforms two-stage method in both *solution quality* evaluation metrics across all domains.

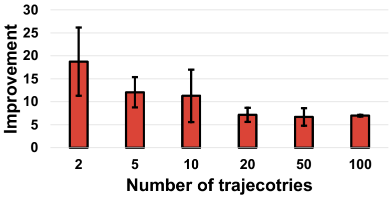
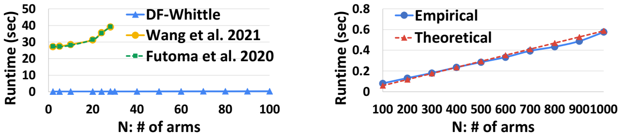


Figure 4: Performance improvement of decision-focused v.s. two-stage method with varying number of trajectories.



a Comparing out algorithm to decision-focused baselines.

b Computation cost with varying number of arms N.

Figure 5: We compare the computation cost of our decision-focused learning with other baselines and the theoretical complexity $O(NM^{\omega+1})$ with varying number of arms N.

improvement also reflects the hardness of maximizing the effectiveness of service calls. In contrast, decision-focused learning achieves an increase of beneficiaries listening to 50 more voice messages overall; DF-whittle achieves a much higher increase by strategically assigning the limited service calls using the right objective in the learning method. The improvement is statistically significant (p-value < 0.05).

In the testing set, we examine the difference between those selected for service call in two-stage and DF-Whittle. We observe that there are some interesting differences. For example, DF-Whittle chooses to do service calls to expectant mothers earlier in gestational age (22% vs 37%), and to a lower proportion of those who have already given birth (2.8% vs 13%) compared to two-stage. In terms of the income level, there is no statistic significance between two-stage and DFL (p-value = 0.20 see Appendix). In particular, 94% of the mothers selected by both methods are below the poverty line (World Bank 2020).

Impact of Limited Data Figure 4 shows the improvement between decision-focused learning and two-stage method with varying number of trajectories given to evaluate the impact of limited data. We notice that a larger improvement between decision-focused and two-stage learning is observed when fewer trajectories are available. We hypothesize that less samples implies larger predictive error and more discrepancy between the loss metric and the evaluation metric.

Computation cost comparison Figure 5a, compares the computation cost per gradient step of our Whittle index-based decision-focused learning and other baselines in decision-focused learning (Wang et al. 2021; Futoma, Hughes, and Doshi-Velez 2020) by changing N (the number of arms) in $M = 2$-state RMAB problem. The other baselines fail to run with $N = 30$ arms and do not scale to larger problems like maternal and child care with more than 600 people enrolled, while our approach is 100x faster than the baselines as shown in Figure 5a and with a linear dependency on the number of arms N.

In Figure 5b, we compare the empirical computation cost of our algorithm with the theoretical computation complexity $O(NM^{\omega+1})$ in N arms and M states RMAB problems. The empirical computation cost matches with the linear trend in N. Our computation cost significantly improves the computation cost $O(M^{\omega N})$ of previous work as discussed in Section .

Conclusion

This paper presents the first decision-focused learning in RMAB problems that is scalable for large real-world datasets. We establish the differentiability of Whittle index policy in RMAB by providing new method to differentiate through Whittle index and using soft-top-k to relax the arm selection process. Our algorithm significantly improves the performance and scalability of decision-focused learning, and is scalable to real-world RMAB problem sizes.

Acknowledgments

Aditya Mate was supported by the ARO and was accomplished under Grant Number W911NF-17-1-0370. Sanket Shah and Kai Wang were also supported by W911NF-17-1-0370 and ARO Grant Number W911NF-18-1-0208. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARO or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Kai Wang was additionally supported by Siebel Scholars.

References

Agrawal, A.; Amos, B.; Barratt, S.; Boyd, S.; Diamond, S.; and Kolter, Z. 2019. Differentiable convex optimization layers. *arXiv preprint arXiv:1910.12430*.

Akbarzadeh, N.; and Mahajan, A. 2019. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 7294–7300. IEEE.

Alman, J.; and Williams, V. V. 2021. A refined laser method and faster matrix multiplication. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 522–539. SIAM.

Amos, B.; and Kolter, J. Z. 2017. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, 136–145. PMLR.

ARMMAN. 2022. ARMMAN Helping Mothers and Children. https://armman.org/. Accessed: 2022-05-19.

Bagheri, S.; and Scaglione, A. 2015. The restless multi-armed bandit formulation of the cognitive compressive sensing problem. *IEEE Transactions on Signal Processing*, 63(5): 1183–1198.

Bai, S.; Kolter, J. Z.; and Koltun, V. 2019. Deep equilibrium models. *arXiv preprint arXiv:1909.01377*.

Benamou, J.-D.; Carlier, G.; Cuturi, M.; Nenna, L.; and Peyré, G. 2015. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2): A1111–A1138.

Bubeck, S.; and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Dai, W.; Gai, Y.; Krishnamachari, B.; and Zhao, Q. 2011. The non-Bayesian restless multi-armed bandit: A case of near-logarithmic regret. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2940–2943. IEEE.

Donti, P. L.; Amos, B.; and Kolter, J. Z. 2017. Task-based end-to-end model learning in stochastic optimization. *arXiv preprint arXiv:1703.04529*.

Elmachtoub, A.; Liang, J. C. N.; and McNellis, R. 2020. Decision trees for decision-making under the predict-then-optimize framework. In *International Conference on Machine Learning*, 2858–2867. PMLR.

Elmachtoub, A. N.; and Grigas, P. 2021. Smart "predict, then optimize". *Management Science*.

Futoma, J.; Hughes, M. C.; and Doshi-Velez, F. 2020. Popcorn: Partially observed prediction constrained reinforcement learning. *arXiv preprint arXiv:2001.04032*.

Gittins, J.; Glazebrook, K.; and Weber, R. 2011. *Multi-armed bandit allocation indices*. John Wiley & Sons.

Glazebrook, K. D.; Ruiz-Hernandez, D.; and Kirkbride, C. 2006. Some indexable families of restless bandit problems. *Advances in Applied Probability*, 38(3): 643–672.

Huang, C.; Zhai, S.; Talbott, W.; Martin, M. B.; Sun, S.-Y.; Guestrin, C.; and Susskind, J. 2019. Addressing the loss-metric mismatch with adaptive loss alignment. In *International Conference on Machine Learning*, 2891–2900. PMLR.

Jiang, S.; Song, Z.; Weinstein, O.; and Zhang, H. 2020. Faster dynamic matrix inverse for faster lps. *arXiv preprint arXiv:2004.07470*.

Johnson, J. M.; and Khoshgoftaar, T. M. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1): 1–54.

Krishnan, S.; Garg, A.; Patil, S.; Lea, C.; Hager, G.; Abbeel, P.; and Goldberg, K. 2017. Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning. *The International Journal of Robotics Research*, 36(13-14): 1595–1618.

Lambert, N.; Amos, B.; Yadan, O.; and Calandra, R. 2020. Objective mismatch in model-based reinforcement learning. *arXiv preprint arXiv:2002.04523*.

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

Liu, K.; and Zhao, Q. 2010. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11): 5547–5567.

Mandi, J.; Stuckey, P. J.; Guns, T.; et al. 2020. Smart predict-and-optimize for hard combinatorial optimization problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1603–1610.

Mate, A.; Killian, J. A.; Xu, H.; Perrault, A.; and Tambe, M. 2020. Collapsing Bandits and Their Application to Public Health Intervention. In *NeurIPS*.

Mate, A.; Madaan, L.; Taneja, A.; Madhiwalla, N.; Verma, S.; Singh, G.; Hegde, A.; Varakantham, P.; and Tambe, M. 2022. Field Study in Deploying Restless Multi-Armed Bandits: Assisting Non-Profits in Improving Maternal and Child Health. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Monahan, G. E. 1982. State of the art—a survey of partially observable Markov decision processes: theory, models, and algorithms. *Management science*, 28(1): 1–16.

Oksanen, J.; and Koivunen, V. 2015. An order optimal policy for exploiting idle spectrum in cognitive radio networks. *IEEE Transactions on Signal Processing*, 63(5): 1214–1227.

Papadimitriou, C. H.; and Tsitsiklis, J. N. 1994. The complexity of optimal queueing network control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*, 318–322. IEEE.

Perrault, A.; Wilder, B.; Ewing, E.; Mate, A.; Dilkina, B.; and Tambe, M. 2020. End-to-end game-focused learning of adversary behavior in security games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1378–1386.

Qian, Y.; Zhang, C.; Krishnamachari, B.; and Tambe, M. 2016. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 123–131.

Ranchod, P.; Rosman, B.; and Konidaris, G. 2015. Nonparametric bayesian reward segmentation for skill discovery using inverse reinforcement learning. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 471–477. IEEE.

Sun, Y.; Feng, G.; Qin, S.; and Sun, S. 2018. Cell association with user behavior awareness in heterogeneous cellular networks. *IEEE Transactions on Vehicular Technology*, 67(5): 4589–4601.

Sutton, R. S.; Barto, A. G.; et al. 1998. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

Tekin, C.; and Liu, M. 2012. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8): 5588–5611.

Thomas, P. S. 2015. *Safe reinforcement learning*. Ph.D. thesis, University of Massachusetts Libraries.

Wang, K.; Shah, S.; Chen, H.; Perrault, A.; Doshi-Velez, F.; and Tambe, M. 2021. Learning MDPs from Features: Predict-Then-Optimize for Sequential Decision Making by Reinforcement Learning. *Advances in Neural Information Processing Systems*, 34.

Wang, K.; Yu, J.; Chen, L.; Zhou, P.; Ge, X.; and Win, M. Z. 2019. Opportunistic scheduling revisited using restless bandits: Indexability and index policy. *IEEE Transactions on Wireless Communications*, 18(10): 4997–5010.

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

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

Wilder, B.; Dilkina, B.; and Tambe, M. 2019. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1658–1665.

World Bank, . 2020. *Poverty and shared prosperity 2020: Reversals of fortune*. The World Bank.

Xie, Y.; Dai, H.; Chen, M.; Dai, B.; Zhao, T.; Zha, H.; Wei, W.; and Pfister, T. 2020. Differentiable top-k operator with optimal transport. *arXiv preprint arXiv:2002.06504*.