Cosmic Bandits: Exploration versus Exploitation in CMB B-Mode Experiments

Ely D. Kovetz\textsuperscript{1} and Marc Kamionkowski\textsuperscript{2}

\textsuperscript{1}Theory Group, Department of Physics and Texas Cosmology Center, The University of Texas at Austin, TX 78712, USA and
\textsuperscript{2}Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA

A preferred method to detect the curl-component, or B-mode, signature of inflationary gravitational waves (IGWs) in the cosmic microwave background (CMB) polarization, in the absence of foregrounds and lensing, is a prolonged integration over a single patch of sky of a few square degrees. In practice, however, foregrounds abound and the sensitivity to B modes can be improved considerably by finding the region of sky cleanest of foregrounds. The best strategy to detect B modes thus involves a tradeoff between exploration (to find lower-foreground patches) and exploitation (through prolonged integration). This problem is akin to the multi-armed bandit (MAB) problem in probability theory, wherein a bandit faces a series of slot machines with unknown winning odds and must develop a strategy to maximize his/her winnings with some finite number of pulls. While the optimal MAB strategy remains to be determined, a number of algorithms have been developed in an effort to maximize the winnings. Here, we formulate the search for IGW B modes in the presence of spatially-varying foregrounds as an MAB problem and develop adaptive survey strategies to optimize the sensitivity to IGW B modes. We demonstrate, using realistic foreground models and taking lensing-induced B modes into account, that adaptive experiments can substantially improve the upper bound on the tensor-to-scalar ratio (by factors of 2–3 in single frequency experiments, and possibly even more). Similar techniques can be applied to other surveys, including 21-cm measurements of signatures of the epoch of reionization, searches for a stochastic primordial gravitational wave background, deep-field imaging by the James Webb Space Telescope or various radio interferometers, and transient follow-up searches.

I. INTRODUCTION

Cosmology has become a science of surveys. Ever larger surveys are used to seek ever-more-subtle correlations to shed light on novel early-Universe phenomena or the physics of galaxy formation. The separation of the signals of interest from similar ones due to astrophysical foregrounds requires more sensitive measurements and clever algorithms. The issue of foregrounds can also be dealt with by restricting the survey to “clean” regions, where the foregrounds are absent or at least smaller in amplitude. But finding these clean regions requires a search which may then take time away from integration on a single patch of sky. Optimization of the sensitivity to a given signal may thus involve a tradeoff between exploration of several patches of sky, to find the cleanest one, and exploitation, deep integration on a single patch. What is the best strategy, under these circumstances, to optimize the sensitivity to the signal?

This question is analogous to the Multi-Armed Bandit (MAB) problem, a well-known problem from probability theory and machine learning in computer science\textsuperscript{1}\textsuperscript{1}. In this problem, a gambler is faced with a set of slot machines with different reward probability distributions and has to maximize the total reward in a given number of plays, or actions. This is a classic learning problem, as repeated plays allow the gambler to learn the distributions of the different machines, with a tradeoff between exploration and exploitation governed by the total number of allowed plays. A popular manifestation of this problem, which has garnered growing attention in recent decades, is clinical trials\textsuperscript{3}, where rewards—in the form of survival/fatality—are of particular importance. Theoretical study of the MAB problem has led to several theorems regarding the ultimate prospects of solution methods in the asymptotic limit of infinite number of plays\textsuperscript{5}. In realistic scenarios, however, with only a finite number of plays, one must resort to heuristic approaches, and over the years several classes of these have been suggested in the literature and compared empirically to some extent\textsuperscript{6,7,8}.

In this paper we focus on the search for the curl, or B-mode, signature of inflationary gravitational waves (IGWs) in the cosmic microwave background (CMB) polarization\textsuperscript{9,11}. These B modes are the target of a number of ongoing and forthcoming CMB-polarization experiments\textsuperscript{12,24}. The strategy of many of these experiments is to integrate deeply on a small patch of sky, as this optimizes the sensitivity to IGW B modes in an experiment with fixed detector sensitivity, or noise-equivalent temperature (NET), and duration\textsuperscript{27,28}. Realistically, though, these experiments will have to contend with foreground emission from Galactic dust and synchrotron radiation\textsuperscript{28,33}. Since the amplitudes of these foregrounds may vary considerably from one region of the sky to another\textsuperscript{31,33}, the sensitivity to IGWs may be improved considerably by integrating on the cleanest patch. While measurements (mostly unpolarized) at other frequencies can be used to steer the experimentalist toward a clean region of the sky, the polarized foregrounds...
in the electromagnetic and spatial frequencies of interest have never been measured. One can thus do an initial exploration of a broad region to find clean patches, but that then takes time away from exploiting any particular region. The challenge is thus to balance the tradeoff between exploration and exploitation in an optimal way, given the limits set by instrumental properties (including the total observation time of the experiment) and the expected distribution of foreground noise on the sky.

The purpose of this paper is to present a method inspired by heuristic solutions to the MAB problem to optimally perform the integration over sky patches so that noise from polarized foregrounds is minimized and the strongest possible upper bound can be placed on the amplitude of IGW B modes. We consider several fiducial experiments with instrumental properties representative of current and next-generation experiments, all operating at a single frequency of 150 GHz (a value common to many of the leading B-mode experiments) and focus on the dominant foreground source at this frequency, which is polarized emission from dust (PED) in the galaxy.

In order to forecast the variation of this foreground source across the sky, we use the FGPol/\cite{32} foreground templates for PED. We perform simulations of different survey (bandit) strategies on patches of sky within a low-noise region accessible from the South Pole, for which PED amplitudes are randomly drawn from the FGPol template, and calculate the improvement (or degradation) in the upper bound on the tensor-to-scalar ratio $r$. While our analysis makes a number of simplifications (although we do include lensing-induced B modes \cite{25} \cite{26}, an essential ingredient), our results demonstrate that the adaptive survey strategies we consider provide considerable advantage over prolonged integration on naively-chosen patches.

While our focus here is on CMB polarization, the methods described in this work can also be applied to other observations in cosmology and astrophysics, such as 21-cm measurements \cite{44} \cite{46}, searches for a stochastic primordial gravitational wave background \cite{37} \cite{38}, deep-field telescope imaging \cite{39} \cite{40}, and transient searches \cite{41}. We discuss potential issues pertaining to such applications, but leave their full study to future work.

The plan of the paper is as follows: In Section II we describe the PED templates used in our analysis, discuss the instrumental noise of our fiducial experiments and present the statistical tools for estimating the errors in measurements of the relevant power spectra. In Section III, we describe the MAB problem, discuss different heuristic solutions and explain our prescription for simulating adaptive B-mode experiments. We present our results in Section IV and discuss several assumptions and possible additional implementations in Section V. We conclude in Section VI.

## II. B-MODE MEASUREMENTS

### A. PED Foreground

In order to remove the different foreground contributions, most experiments operate at several frequencies and use component separation \cite{39} \cite{41} or template-based techniques \cite{42} to extract as clean a signal as possible. In any such process, residuals remain at some level and will hinder the ability to detect the desired signal.

The major contributions of polarized foreground noise in the relevant frequency range ($\sim 20-300$ GHz) of CMB experiments are sourced by PED in the Galaxy and by synchrotron radiation. Synchrotron is more dominant at lower frequencies ($\lesssim 100$ GHz), while PED overwhelms the CMB at higher frequencies ($\gtrsim 100$ GHz). The well-known CMB foreground “sweet spot” is around 90 GHz, where the noise sources are similar in amplitude and both are comparably low.

For simplicity, we shall address fiducial experiments operating at a single frequency of 150 GHz, which is adopted by many of the sub-orbital polarization experiments. Therefore, PED would be the major source for concern in terms of foregrounds and its subtraction would be difficult. To estimate the sky variation of the PED power spectrum, we use the FGPol templates \cite{32} \cite{33}.

These templates are based on a three-dimensional bi-symmetric spiral \cite{32} model of the Galactic magnetic field (including the turbulent component) and are normalized according to the results of WMAP \cite{29} so that the average dust polarization fraction outside the WMAP P06 polarization mask \cite{43} is 3.6%. Until polarization results from the Planck experiment are released (we do not use the pre-launch Planck sky model \cite{44} in this work), the best constraint on the dust polarization fraction at higher frequencies comes from the partial sky ($f_{\text{sky}} = 17\%$) measurement at 353 GHz of the ARCHEOPS balloon-borne experiment \cite{15}, which detected a polarization fraction around 4–5% in the Galactic plane and a maximum of 10–20% in some localized clouds. We shall therefore consider 3.6% and 10% normalizations as conservative and worst-case scenarios, respectively.

In Fig. 1 we plot the high-resolution FGPol template for a 45°-radius low-noise region centered around $(b,l) = (-69.5,241.5)$, which is accessible by ground-based experiments such as BICEP, POLARBEAR, the KECK Array, QUAD, QUIET and SPTPOL. We also plot the average between the variance in the Q and U polarization components in non-overlapping patches of $15^\circ \times 15^\circ$ inside this region.

In order to estimate the PED amplitude in a given patch, we follow the prescription of Ref. \cite{33}. The PED amplitude

\[^2\] http://www3.imperial.ac.uk/people/c.contaldi/fgpol
angular power spectrum is assumed to obey a power law,

$$C^D_{\ell} = A\ell^m,$$

where the power-law index $m$ is fixed to its full-sky best-fit value $m = -0.22$ in this template, and $A$ is allowed to vary between different patches. To estimate $A_p$ for a given patch $p$, we calculate the variances $\sigma_Q^2 = \langle Q^2 \rangle_p - \langle Q \rangle_p^2$ and $\sigma_U^2 = \langle U^2 \rangle_p - \langle U \rangle_p^2$ of the polarization in both the $Q$ and $U$ components of the patch, and infer a power-spectrum amplitude from each, using the relation

$$\sigma^2 = \frac{1}{4\pi} \sum_{\ell=2}^{\ell_{\max}} (2\ell + 1) C^D_{\ell} B^2_{\ell}(\theta_s). \tag{2}$$

We then take the average of the resulting amplitudes to be the PED B-mode power-spectrum amplitude in the desired patch, under the assumption that power is equally distributed between the E and B modes.

Finally, to remain with a realistic set of possible patches, we immediately apply a cutoff to remove the noisiest patches from our sample. The choice of cutoff represents prior knowledge regarding the expected PED amplitudes in this region given the templates and results of previous surveys. We choose a cutoff of $33\%$ throughout. A histogram of PED amplitudes $A_p$ in the remaining 19 patches within the region shown in Fig. 1 is plotted in Fig. 2, with a normalization of $3.6\%$. We see that the amplitudes in this sample still vary over more than an order of magnitude.

### B. Instrumental Noise

The instrumental noise in a CMB-polarization experiment is determined by the detector-array sensitivity (or noise-equivalent temperature NET) $s$, its angular resolution $\theta_{\text{rebm}}$, the sky coverage $f_{\text{sky}}$, and the total observation time $T$ (which is reduced in practice by the optical efficiency). We consider three types of fiducial experiments whose parameters are given in Table I. Experiment 1 is inspired by the POLARBEAR experiment and observes a single $15^\circ \times 15^\circ$ patch. Experiment 2 has similar properties, but a smaller sky coverage, optimized for detecting the peak of the primordial B-mode power spectrum at $\ell \sim 10^2$ (without de-lensing, the optimal size would be a bit higher). Experiment 3 is a fiducial lower-cost, lower-resolution experiment targeted at the large scale primordial B-mode signal (hence the larger sky coverage). We assume an optical efficiency of $20\%$ for all three experiments.

| Experiment | $\theta_{\text{rebm}}$ [arcmin] | $f_{\text{sky}}$ [%] | $T$ [years] | $s=\text{NET}_{\text{array}}$ [$\mu K \sqrt{\text{s}}$] |
|------------|---------------------------------|----------------------|-------------|----------------------------------|
| 1          | 3.5                             | 0.55                 | 2           | $480\sqrt{\ell}/\sqrt{1274} = 19$ |
| 2          | 5                               | 0.06                 | 4           | 15                                |
| 3          | 30                              | 1.32                 | 6           | 25                                |

TABLE I: Parameters for three CMB polarization experiments we use for our analysis. Here, $s$ is the single-detector polarization sensitivity divided by the square root of the number of detectors.

The pixel noise $\sigma_{\text{pix}} = s/\sqrt{w_{\text{pix}}}$ is determined by the detector sensitivity $s$ and the observation time $t_{\text{pix}} = T/N_{\text{pix}}$ dedicated to each pixel. Defining the inverse weight per solid angle, $w^{-1} = 4\pi s^2/T$, the angular power spectrum of the instrumental noise, assuming the experimental beam is approximately Gaussian in shape, is given by

$$C^N_{\ell} = \frac{\Omega \sigma_{\text{pix}}^2 e^{\ell^2 \sigma_{\ell}^2}}{N_{\text{pix}}} = \frac{\Omega s^2 e^{\ell^2 \sigma_{\ell}^2}}{T} e^{\ell^2 \sigma_{\ell}^2} = f_{\text{sky}} w^{-1} e^{\ell^2 \sigma_{\ell}^2}, \tag{3}$$

where $\Omega = 4\pi f_{\text{sky}}$ and $\sigma_{\ell}^2 = \theta_{\text{rebm}}^2/(8 \ln 2)$.  

\[\text{FIG. 1: Left: The polarization amplitude } \sqrt{Q^2 + U^2} \text{ in a } 45^\circ\text{-radius region centered around } (b, l) = (-69.5, 241.5) \text{, taken from the FGPol template} \] 

\[\text{FIG. 2: A histogram of dust amplitudes } A_s \text{ in 19 patches of } 15^\circ \times 15^\circ \text{ within the region plotted in Fig. 1, calculated using Eq. (2), after a cutoff at the 67th-percentile to remove the noisiest patches from our sample altogether.}\]
In order to choose between exploration and exploitation on the fly, we need to estimate the PED amplitude \( \hat{A}_p \) in a targeted patch within a given observation time. With partial sky coverage, one has to be mindful of the tradeoff between increasing sample variance and reducing the noise as the patch size is decreased. For a given sky coverage \( f_{\text{sky}} \), the “1-sigma” error for an individual \( \ell \) in the estimated value \( \hat{A}_p \) is

\[
\sigma_{\ell}^2 = \frac{2}{f_{\text{sky}}(2\ell + 1)} \left( \hat{A}_p C^D_\ell + \alpha C^D_\ell + f_{\text{sky}} w^{-1} e^{i2\sigma_\ell^2} \right),
\]

where \( C^D_\ell = C^D_\ell / A_p = 2\pi \ell^{-m}/[\ell(\ell + 1)] \) encodes the \( \ell \) dependence of the PED power spectrum and \( C^D_\ell \) the lensing B-mode contribution. The quantity \( 1 - \alpha \) parametrizes the level of de-lensing [51, 52] that was applied to the data. When comparing forecasts for upper bounds on the tensor-to-scalar ratio, we will consider \( \alpha = 0.2 \) and \( \alpha = 1 \) as best and worst-case scenarios, respectively. The total “1-sigma” error in the measurement of \( \hat{A}_p \) over a time \( t_p \) is thus

\[
\sigma^2 = \left[ \frac{f_{\text{sky}}}{2} \frac{\ell_{\text{max}}}{\ell_{\text{min}}} \frac{(2\ell+1)(\hat{C}^D_\ell)^2}{A_p C^D_\ell + \alpha C^D_\ell + f_{\text{sky}} w^{-1}(t_p)e^{i2\sigma_\ell^2}} \right]^{-\frac{1}{2}},
\]

where \( \ell_{\text{min}} = 180/\theta \) is the largest scale accessible by an experiment with sky coverage \( f_{\text{sky}} = \theta^2 \) and \( w^{-1}(t_p) \) is the inverse weight per solid angle for the patch \( p \).

In Fig. 4, we plot the normalized (Gaussian) likelihood curves for the measured amplitudes of the patches with maximum, median, and minimum PED amplitudes in the FGPol template (see Fig. 3). The goal of the adaptive survey (bandit) strategies will be to distinguish between the means of these distributions in order to spend more time observing lower-noise patches. Clearly, this task is harder when the similarity between the distributions is larger (we will quantify this in the next Section).

### C. Statistical Estimators on Partial Sky

The measurements we deal with are those of power spectra. In a maximum-likelihood analysis, the Fisher forecast for the error in the measurement of the amplitude \( A \) of a power spectrum \( C_\ell \) is given by [27],

\[
\frac{1}{\sigma_A^2} = \sum \ell \left( \frac{\partial C_\ell}{\partial A} \right)^2 \frac{1}{\sigma_{\ell}^2}.
\]

We will work under the assumption that the likelihood function is Gaussian in the vicinity of its maximum [40]. To emphasize this approximation we use quotation marks around “1-sigma” when referring to the error \( \sigma_A \).
Our figure of merit when comparing the prospects of different survey (bandit) strategies will be the smallest amplitude of primordial B modes that can be distinguished from the null hypothesis at “1-sigma” confidence after a fixed total observation time $T$. Under the approximation that the PED power spectrum $A_p \hat{C}_l^P$ acts as a source of extra Gaussian noise, and denoting the $\ell$ dependence of the IGW B-mode power spectrum by $\hat{C}_\ell^B$, the smallest amplitude in a patch $p$ detectable at “1-sigma,” according to Eq. (4), is

\[
\sigma_p^\tau = \left[ \frac{f_{\text{sky}}}{2} \sum_{\ell_{\text{min}}}^{\ell_{\text{max}}} \left( \frac{\sqrt{(2\ell+1)\hat{C}_\ell^B}}{A_p \hat{C}_\ell^P + \alpha \hat{C}_\ell^I + f_{\text{sky}} \omega(t_p)^{-1} e^{r/\sigma_\ell^B}} \right)^2 \right]^{-\frac{1}{2}}. 
\]

Notice that we set the sample variance of the primordial signal to zero in this expression (since we are comparing to the null hypothesis).

Finally, if an adaptive survey ends up spending an observing time $t_p$ (where $\sum_{p=1}^{n_p} t_p = T$) on each of $n_p$ patches with known dust amplitudes $A_p = A_1, \ldots, A_{n_p}$, then the smallest detectable primordial B-mode amplitude will be

\[
\sigma^\tau = \left[ \frac{f_{\text{sky}}}{2} \sum_{p=1}^{n_p} \sum_{\ell_{\text{min}}}^{\ell_{\text{max}}} \left( \frac{\sqrt{(2\ell+1)\hat{C}_\ell^B}}{A_p \hat{C}_\ell^P + \alpha \hat{C}_\ell^I + f_{\text{sky}} \omega(t_p)^{-1} e^{r/\sigma_\ell^B}} \right)^2 \right]^{-\frac{1}{2}}. 
\]

III. EXPLORATION VS. EXPLOITATION

In this Section we describe the methodology for testing MAB strategies, present a series of heuristic solutions and explain how we apply them to B-mode experiments.

A. Constructing and Evaluating MAB Strategies

A strategy to solve an MAB problem uses a set of estimates for the expected rewards of each arm in order to decide which one to select at each play. The expected (or mean) reward of each arm $a$ is called its action-value, and we denote it by $\mu^*(a)$.

A natural way to estimate the action-value of an arm after $t$ plays is through sample-averaging of its previous rewards. If at play $t$ an arm $a$ has been chosen $N_t(a)$ times and has yielded a set of rewards $r_1, r_2, \ldots, r_{N_t(a)}$, then its action-value estimate will be given by

\[
\mu_t(a) = \frac{r_1 + r_2 + \cdots + r_{N_t(a)}}{N_t(a)}, 
\]

where the values $\mu_0(a)$ are defined by some chosen method of initialization. The law of large numbers then guarantees that $\mu_t(a) \to \mu^*(a)$ as $N_t(a) \to \infty$.

The standard figure of merit for the success of a proposed algorithm to solve the MAB problem is its total regret. If $V^* = \max_a \mu^*(a)$ is the expected reward of the optimal arm $a^*$, then after $t$ plays the total regret is given by

\[
R_t = \sum_{a=1}^{A} (V^* - \mu_t(a)) = \sum_a (N_t(a)) \Delta_a, 
\]

where $\Delta_a = V^* - \mu_t(a)$ is the gap between the optimal reward and the expected reward of arm $a$. Therefore, a good strategy ensures smaller counts for larger gaps.

The performance of any strategy depends on the similarity between the optimal arm and the other arms. The hardest problems will have similar-appearing arms with slightly different means. Lai and Robbins [5] have derived a lower bound for the asymptotic total regret in the limit of unlimited number of plays. According to their theorem,

\[
\lim_{t \to \infty} R_t \geq \log t \sum_{a|\Delta_a > 0} \frac{\Delta_a}{KL(\mathcal{P}(a)||\mathcal{P}(a^*))}, 
\]

where $KL(\mathcal{P}(a)||\mathcal{P}(a^*))$ is the Kullback-Leibler (KL) distance between the probability distributions of an arm $a$ and the optimal arm $a^*$. If the rewards have normal distributions, $N(\mu^*(a), \sigma_a)$ and $N(\mu^*(a^*), \sigma_a^*)$, the KL distance between two arms is given by

\[
KL(\mathcal{P}(a)||\mathcal{P}(a^*)) = \ln \frac{\sigma_a^*}{\sigma_a} + \frac{\sigma_a^*^2 + (\mu^*(a^*) - \mu^*(a))^2}{2\sigma_a^*^2} - \frac{1}{2}. 
\]

In practice, achieving logarithmic total regret is a challenge, requiring an ideal tradeoff between exploration and exploitation. It is easy to see from Eq. (10) that an algorithm that never explores will have linear total regret and the same is true for an algorithm that explores forever.

B. Families of Heuristic Solution Methods

We briefly describe a set of algorithms that represent some of the prevailing heuristics in the literature.

- Greedy

The most naive approach to the MAB problem is to simply select at each play $t$ the arm with the highest estimated action-value

\[
a_t = \arg\max_a \mu_t(a). 
\]

This method merely exploits past knowledge in order to maximize the immediate reward (hence greedy [53]) and results in linear total regret. In the absence of any initial knowledge regarding the expected rewards (setting all $\mu_0(a) = -\infty$), this method chooses one of the arms randomly on the first play and then continuously exploits it, without any further exploration.
• $\epsilon$-greedy algorithms

A simple enhancement to the regular greedy method is to occasionally force the selection of an arm at random, regardless of the action-value estimates. In the $\epsilon$-greedy method, the greedy arm is chosen at probability $1 - \epsilon$, while at probability $\epsilon$ a different arm is chosen uniformly at random

$$p_t(a) = \begin{cases} 1 - \epsilon & \text{if } a = \text{argmax } \mu_t(a') \\ \epsilon & \text{otherwise} \end{cases}. \quad (14)$$

In the limit $\epsilon \to 0$, this reduces to the regular greedy method. When $\epsilon = 1$, this method forever exploits all arms uniformly.

While perpetual exploration ensures the convergence of all the action-value estimators $\mu_t(a) \to \mu^*(a)$ (since $N_t(a) \to \infty$ for all arms), it also means that the total regret will be linear, as exploration continues even after the optimal arm is identified with great (and ever-growing) confidence.

To solve this problem, a further enhancement to this method is to decrease the exploration probability $\epsilon$ with time, thus asymptotically limiting the regret. Decaying $\epsilon$-greedy methods have logarithmic asymptotic total regret, but in practice, especially when the number of plays is finite, it is hard to choose a single decay strategy for $\epsilon$ that works well under different circumstances (different reward distributions).

• Optimistic initialization

We implement the absence of prior knowledge by taking the limit $\mu_0(a) \to -\infty$. Any other choice for the initial action-value estimates would introduce a bias in all the methods described so far, but an effective choice for this bias can be used as a means of forcing an early stage of increased exploration. For example, by choosing optimistically-high initial values in the greedy method, we delay the convergence onto a single arm until the initial values no longer dominate the sample average. Until that happens, some knowledge will have been gathered about all arms, ensuring a more informed greedy choice from then on, at the expense of this initial stage of exploration.

• Probability matching (Boltzmann)

An obvious path towards more sophisticated algorithms is to vary the probability with which inferior arms are explored, instead of sampling them uniformly as in $\epsilon$-greedy. After all, $\epsilon$-greedy is likely to keep exploring extremely unfavorable arms even after they are known to be vastly inferior to others. The straightforward solution is to set the probability to explore each arm according to its estimated value. In the Boltzmann method (or Thompson sampling [56]), for example, the probability to choose arm $a$ at play $t$ is given by

$$p_t(a) = \frac{e^{\mu_t(a)/\tau}}{\sum_{a' \neq a} e^{\mu_t(a')/\tau}}, \quad (15)$$

where $\tau$ is a positive parameter, often referred to as the temperature. In the high-temperature limit, this reduces to the $\epsilon = 1$ uniform method, while in the $\tau \to 0$ limit we retrieve the regular greedy method with absence of initial knowledge.

Although some quantitative studies have been performed to compare the Boltzmann and $\epsilon$-greedy methods, no consistent conclusions as to which is superior have been reached [6,7]. Both methods depend on the tuning of their respective free parameter and one performs better than the other under different circumstances. We shall use the Boltzmann method with optimistic initialization, as our experience deems this favorable.

• Upper confidence bound

The upper confidence bound (UCB) method [57] tackles the exploration-versus-exploitation tradeoff according to the principle that the more uncertain we are about an arm’s action-value, the more important it is to explore it further (since it could turn out to be the optimal one). In this method, we estimate an upper confidence $U_t(a)$ for each action value such that $\mu(a) \leq \mu_t(a) + U_t(a)$ with high probability, and then select the action that maximizes

$$a_t = \text{argmax}_a \{\mu_t(a) + U_t(a)\}. \quad (16)$$

The confidence bound depends on the number of times $a$ has been selected (as increased selection reduces the uncertainty bound). When the rewards are Gaussian, a useful upper bound is given by $c \sigma_a / \sqrt{N_t(a)}$, where $c > 0$ is some constant. Naturally, there is no need for optimistic initialization when using the UCB method, as it is always optimistic about unexplored arms.

Before moving on to the implementation of these bandit strategies in B-mode experiments, we demonstrate their performance for a simple MAB problem with 10 arms with reward distributions $\mathcal{N}(\mu^*_a, 1)$, where the mean rewards $\mu^*_a$ are themselves drawn from a normal distribution, $\mathcal{N}(0, 1)$. In Fig. 5, we plot the behavior of each of the methods with time. Most methods converge onto the optimal arm in greater percentages and decrease their daily regret as time goes by. The UCB method seems more successful eventually than others, while the second-best Boltzmann method is the most efficient after a very limited number of plays. In Fig. 6, we compare the total regret for three algorithms: greedy, $\epsilon$-greedy with $\epsilon = 1$, and UCB. We plot the average in solid lines and the...
full range of performance in the corresponding shaded regions. The greedy method will always have the widest range of performance, as it chooses and stays with the optimal arm or worst arm at 10% probability. The UCB method has logarithmic total regret, much like the Lai and Robbins asymptotic bound, also plotted.

C. Simulating a B-Mode bandit strategy

In order to simulate a B-mode bandit strategy, we do the following:

- Use the fiducial experimental parameters in Table I to set the parameters $w^{-1}, f_{\text{sky}},$ and $\sigma_b$.
- Apply an upper cutoff in amplitude to the desired-size patches in the South Pole region FGPol template to remain with a certain percentage of lowest-noise patches, representing our prior knowledge regarding the target area, either from the noise templates themselves or from other surveys. As mentioned above, for all experiments we limit our sample to patches with PED amplitudes below the 67th percentile.
- Randomly draw a set of $n_p$ dust amplitudes from the remaining patches.
- Choose the time-step size, typically a few observation days.
- Use Eq. (6) to calculate the “1-sigma” error $\sigma_p$ in our estimate of the amplitude in each patch within a single time step (corresponding to the width of the likelihood functions in Fig. 4).
- Retrieve the reward $V$ from the patch $p$ chosen at a given time step through a random draw from a normal distribution $V = -\min(0, \mathcal{N}(A_p, \sigma_p))$. Notice that in order for the methods to converge onto lower-noise patches, we multiply the measured PED amplitude by $-1$ to get the reward.
- Update our action-value estimate of the chosen patch $p$ according to the reward.
- Calculate the figure-of-merit (the smallest detectable tensor-to-scalar ratio at “1-sigma”) at each time $T^i$ by plugging in the observation times spent so far on each patch $t^i_p$ (where $\sum_{p=1}^{n} t^i_p = T^i$) and the (known) dust amplitudes $A_p$ into Eq. (8).
- Calculate the regret, measured by the difference in the noise amplitude between the chosen patch and the lowest-noise patch and update the total regret.
- Decide which patch to observe at the next time step according to the strategy and keep going until the total observation time is exhausted.

Our choice of parameters for the different methods:
- Absence of initial knowledge: $\mu_0(p) = -5(\sigma_p)$.
- For $\epsilon$-greedy, we choose $\epsilon = 0.1$.
- Our $\epsilon$-decay strategy is $\epsilon = \min(1, n_p/t^{3/2})$.
- Optimistic initialization: $\mu_0(p) = 5(\sigma_p)$.
- For the Boltzmann method, we choose $\tau = (\langle \sigma_p \rangle)/10$.
- For UCB we choose $c = 1$: $U_t(p) = (\sigma_p)/\sqrt{N_t(p)}$.

We then repeat the full list of steps above 500 times for each proposed method in order to acquire an ensemble average for comparison. In the next Section we study the performance of each of the strategies above for our fiducial experiments (see Table I) and for the best-, conservative- and worst-case scenarios regarding the PED normalization.

IV. RESULTS

We are now ready to test the MAB strategies on simulations of B-mode experiments. We first address the extreme cases (see Fig. 3) of 10% average dust polarization fraction with no-de-lensing (a pessimistic scenario) and 3.6% with an efficient de-lensing process, leaving only a 20% residual (an optimistic scenario). Focusing on Experiment 1, which has properties similar to the POLARBEAR experiment [12], we compare the performance of the different strategies in Fig. 7 using our simulations. We see that the UCB method fares better than other methods in identifying the optimal patch, although it does not achieve the same rate of success as in our simple Gaussian test case above. With a distribution of foreground amplitudes taken from the FGPol templates in the region of Fig. 1 after a cutoff at the 67th percentile, the UCB method manages to converge onto the optimal patch after 2 years of observation in only 80% of the simulations in the pessimistic scenario and in less than half of the simulations in the optimistic one. Nevertheless, we see that the total regret in the optimal methods (Boltzmann, UCB) is much lower than with naive methods. The total regret in the greedy method, which is the default version of a POLARBEAR-type experiment, is roughly 2–3 times higher when comparing the average or worst-case performance to those of Boltzmann or UCB. Greedy with $\epsilon = 1$, which corresponds to a BICEP-type experiment, where a larger region is uniformly observed (or several patches uniformly integrated over), is better in the worst case than greedy, but also never reaches below the average of the Boltzmann method in the pessimistic scenario (and barely below the UCB method in the optimistic scenario), even in its best-case performance.

As we explained above, our figure-of-merit in comparing survey (bandit) strategies is the smallest detectable tensor-to-scalar ratio $r$ at “1-sigma” confidence. In
FIG. 5: Different algorithms balance exploration and exploitation differently. Left: The percentage of simulations in which the optimal arm was chosen in each play. The UCB method reaches a higher percentage than all other methods as the number of plays is increased. Greedy and $\epsilon$-greedy with $\epsilon = 1$ choose the optimal arm 10% of the time, which is expected with a total of 10 arms. Right: The average regret at each play. Notice that the Boltzmann method is inferior to UCB in the long run, but it reduces the daily regret faster than UCB at the beginning.

FIG. 6: Left: The average and ranges of total regret for greedy (range: shaded gray, average: solid black), $\epsilon$-greedy with $\epsilon = 1$ (red) and UCB (blue). Right: Zooming in → the UCB method assumes logarithmic behavior, similar to that of the Lai and Robbins asymptotic bound (in dashed-black), after a relatively small number of plays.

We see that the improvement on average when comparing to the greedy method ranges from $\sim 25\%$ to $\sim 70\%$ (and it goes without saying that it will only improve further if the observation is prolonged). Thus, adaptive survey strategies have great potential to improve the sensitivity in IGW searches.

V. DISCUSSION

A. Simplifying assumptions and caveats

It is worthwhile to discuss several simplifications we have made throughout the analysis above. First, we have made exclusive use of the FGPol templates [32] both to motivate the need for adaptive survey strategies and in order to estimate the expected distribution of foreground levels across the sky. The use of other templates that have been developed in the literature [31, 44, 58] (or in

Fig. 5, we plot our results for the three experiments in Table I for three scenarios: a pessimistic scenario with 10% dust polarization fraction and no de-lensing, a conservative scenario with 3.6% normalization and no de-lensing and an optimistic scenario with 3.6% normalization and a 20% lensing residual. In the pessimistic case, the Boltzmann method achieves the best results for all experiments (though it is very close to the UCB method). In the two other scenarios, UCB is the best performing method. In the case of experiment 2 in the conservative scenario, even the best-case performance of the greedy method is inferior to the averages of the UCB method (and also $\epsilon$-greedy with $\epsilon = 1$). This is because with this low-noise experiment, prolonging the integration, even over the cleanest patch, only goes so far and it is often preferable to spend equal time on the two (or more) cleanest patches.

Then, we can conclude that the trend of the results will only improve further if the observation is prolonged. Thus, adaptive survey strategies have great potential to improve the sensitivity in IGW searches.
FIG. 7: The percentage of 2000 simulations in which the optimal patch was chosen per day (left), the average daily regret (center) and the average total regret (right) for Experiment 1. Top Row: A pessimistic scenario of no-delensing and 10% average dust polarization fraction. Bottom Row: The optimistic case of 3.6% normalization and 80% de-lensing. We see that the UCB method ultimately chooses the optimal patch in larger percentages compared to other methods. In the pessimistic scenario, the Boltzmann method has the lowest average total regret. In the optimistic case, The UCB method achieves the lowest total regret on average, and even in its worst case (top of shaded-blue region), it is relatively close to the averages of greedy (black line, overlapping the red line, with gray shading) and $\epsilon$-greedy with $\epsilon = 1$ (red), whose own worst cases are 4 and 10 times the UCB average.

The near future, the results of actual measurements by experiments such as Planck may lead to somewhat different results, but most likely will not alter our conclusions.

Secondly, we have limited our analysis to the case of experiments operating at a single frequency. This allowed us to focus on a single source of foregrounds and work under the assumption that it could not be efficiently removed through component separation or other methods. Most B-mode experiments will eventually gather measurements from multiple frequencies in order to enable foreground subtraction. In this case, the foreground amplitudes we have considered here will be replaced by the amplitude of the remaining residuals. In a very optimistic scenario, where these would be smaller than the IGW B-mode signal, the advantage of our methods would be restricted.

Additional assumptions include the neglect of E and B mode mixing, which may pose a problem in partial sky measurements [59–61] (although optimized estimators greatly reduce this complication [62]) and the approximation that the likelihood curves for the power
spectra estimators are Gaussian (in our B-mode simulations we simply drew a measured amplitude from a normalized distribution matching the likelihood mean and width, as opposed to drawing the estimated value for each multipole from a corresponding $\chi^2$ distribution). We do not expect these assumptions to have a meaningful impact on our results.

Finally, we have limited our analysis to just a handful of scenarios, ranging from pessimistic to optimistic, in an attempt to provide a complete picture regarding the prospects of the methods proposed. We have verified that changing some of the parameters chosen here, such as the power-law index for the PED power spectrum, the level of dust polarization fraction outside the Galactic plane and the efficiency of the de-lensing process, do not substantially affect our conclusions. However, a more detailed study should be conducted when implementing these methods to a specific B-mode experiment. One enhancement upon our analysis is to calculate the full foreground power spectrum directly from the template patch, instead of relying on a best-fit power law. Another important ingredient to include is the cost of moving between different sky patches when calculating the total regret of different observing strategies. While this may degrade the improvement in sensitivity somewhat, it will be offset by the continued gain from identifying the optimal patches for observation as the experiments progress to their next stages, which we have not taken into account here.

B. MAB strategies elsewhere

Analogous adaptive survey strategies might also be used to help detect 21-cm intensity fluctuations from the EoR [34, 35], where the signal must be distinguished, using angular and/or frequency information, from a variety of Galactic and extragalactic foregrounds [63–69] that are expected to be larger by several orders of magnitude. The desired signal is expected to extend from several arcminutes to several degrees. What is currently envisioned, for example, for the Low Frequency Array [70] and the Square Kilometer Array [71], is a deep integration on a $\sim 5^\circ \times 5^\circ$ region of the sky. As is well known by now, the amplitude of Galactic foregrounds may vary considerably from one such region on the sky to another. While existing Galactic dust and synchrotron maps can be used to identify regions of the sky that are likely to be clean, again, detailed maps at the several-hundred MHz frequencies of interest, with the angular resolutions of interest, do not exist. There are also likely to be foregrounds in frequency space contributed by radio recombination lines [65], and little is known about their distribution at the high Galactic latitudes that present the likeliest targets for these deep EoR integrations. The distribution of the extragalactic-foreground intensity across the sky is expected to be smoother than the Galactic foregrounds. Even so, clustering of highly-biased radio sources at high redshift may give rise to significant variation in foreground intensities from one patch of the sky to another. The aim of an adaptive bandit strategy would in this case be to identify the cleanest patch in a fixed survey time.

Another example of a foreground-limited stochastic measurement is that of a gravitational wave background (GWB) [37, 38], which may be pursued by terrestrial experiments such as The Laser Interferometer Gravitational Wave Observatory [74] and several Pulsar Timing Array consortia [75]. In the relevant frequency bands, the stochastic GWB signal may be obscured by confusion foreground stemming from bright sources, including supernovae, black hole binaries [76], neutron star binaries [77] and Galactic white dwarf binaries [78]. In an effort to minimize the effect of foregrounds, bandit strategies may be used to improve the sensitivity of the relevant ground-based experiments.

Likewise, similar strategies should be explored for a JWST deep field aimed to detect the first stars/galaxies [79, 80]. CMB measurements of the reionization optical depth suggest that the first stars formed at redshifts $\sim 10$, and these stars should emit radiation that today falls in the wavelengths probed by JWST [81, 82]. Still, the abundance of these objects, their luminosity function, etc., are largely a matter of guesswork. It may well be that the population of the first stars/galaxies will fill the entire image, or that there may be a vast population of low-flux sources from lower redshifts that may obscure the EoR objects of interest, as suggested in Ref. [83]. The very faint EoR sources, in this case, will be easiest identified in the clearest field on the sky. Again, an adaptive survey strategy to identify the clearest of several candidate patches may well be warranted.

In this regard, one should be mindful of an inherent difference between the application of adaptive survey strategies to stochastic measurements and deep-field imaging. While a choice of an optimal sky patch for deep-field imaging may considerably improve the sensitivity of a given experiment, the initial stage of exploration is mostly wasted. Therefore, the figure-of-merit when comparing different strategies should be the speed of convergence onto the optimal arm instead of the total regret, and the best-performing algorithms may be different. Nevertheless, an initial stage of exploration may still prove invaluable in reducing the error in the integrated signal from the chosen patch to image.

A somewhat different version of the MAB problem in the context of astronomical observations shows up in the case of follow-up observations of transients [41], for example. In this case, the exploration versus exploitation tradeoff is manifested in the challenge of allocating...

---

3 Very recently, the performance of two non-adaptive modes of observation for the Murchinson Widefield Array [72] — corresponding to the greedy and $\epsilon$-greedy with $\epsilon = 1$ methods described above — was investigated in detail [73].
ing the wide array of resources to the task, each with very different characteristics and cost functions. Hence, the figure-of-merit for comparing various algorithms may once again be different than considered here.

We leave the detailed study of the prospects of MAB strategies in these setups to future work.

VI. CONCLUSION

The era of CMB B-mode measurements has only just begun [25, 20]. The race for the detection of IGW B modes is picking up and the competing efforts will be mainly characterized by their ability to tighten the upper bound on the tensor-to-scalar ratio. This endeavor is extremely difficult and will grow more so as the upper bound is decreased, particularly due to the major obstacles which need to be efficiently removed — foreground contamination and lensed E modes — whose amplitudes are independent of the targeted IGW B-modes. Therefore, any novelty in the approach to conducting these measurements which may mitigate these problems is both timely and important.

In this paper we have proposed exactly such a novelty, in the form of adaptive survey (bandit) strategies, inspired by heuristic solutions to the MAB problem. Focusing on polarized foreground components, whose amplitudes are predicted to vary considerably across the sky, our target was to balance the time spent on exploration to find lower-noise patches of sky with that devoted to exploitation of the optimal patches through deep integration. By formulating this tradeoff as an MAB problem, we were able to adopt machine learning algorithms developed in the general context of the MAB challenge and implement them in the settings of fiducial B-mode experiments.

Attempting to go beyond a mere proof-of-concept, we have incorporated a number of crucial ingredients into our analysis, such as prior knowledge from existing surveys, several fiducial sets of instrumental properties, different possible forecasts regarding the average amplitude of foregrounds, and realistic prospects for the level of lensing residuals (including pessimistic cases of no de-lensing). Relying on advanced templates for polarized emission from dust in the Galaxy, we demonstrated that single frequency experiments could improve their upper bounds on the tensor-to-scalar ratio by factors of a few on average and even higher when comparing worst-case performances of standard methods versus the ones proposed here.

The assumptions made in this work were discussed in the previous section, along with some possible caveats in the implementation of the bandit survey strategies proposed here in B-mode experiments. While the precise improvement these methods may enable remains to be seen, we feel that the case has been made for their consideration in any ground-based experimental setup focusing on CMB B-mode detection with partial sky coverage.

Lastly, as the tradeoff between exploration and exploitation shows up in other realms of observational cosmology as well, this work opens the door for additional applications of bandit survey strategies. These may include other stochastic measurements, such as the power spectrum of 21-cm fluctuations [33–36] or a primordial GWB [37, 38], deep-field imaging by radio interferometers [70, 71] or optical telescopes [39, 40], and the allocation of limited resources in follow-ups of identified astrophysical transients [41], to name a few.

Acknowledgments

EDK was supported by the National Science Foundation under Grant Number PHY-0969020 and by the Texas Cosmology Center. This work was supported at Johns Hopkins by DoE SC-0008108, NASA NNX12AE86G, and NSF 0244990. EDK thanks the New College Oxford - Johns Hopkins Centre for Cosmological Studies for hospitality at Johns Hopkins University, where most of this work was carried out.

[1] H. Robbins, Bull. Am. Math. Soc. 58, 527535 (1952)
[2] D. A. Berry and B. Fristedt, Bandit Problems: Sequential Allocation of Experiments (Chapman & Hall, London, 1985).
[3] W. H. Press, talk at “Astrophysics At The Extreme, The Hebrew University of Jerusalem, Israel, December 2009,
http://tsvi.phys.huji.ac.il [4] W. H. Press, Proc. Nat. Acad. Sci., 106 52, 22387 (2009). [5] T. L. Lai and H. Robbins, Advances in Applied Mathematics 6,4 (1985).
[6] R. S. Sutton and A. G. Barton, Reinforcement Learning: An Introduction (MIT Press, Cambridge, 1998).
[7] V. Kuleshov and D. Precup, sub. to J. of Mach. Learn. Res.,
http://www.cs.mcgill.ca/~vkules/bandits.pdf [8] J. Gittins, K. Glazebrook and R. Weber, Multi-Armed Bandit Allocation Indices (Wiley, New York City, 2011).
[9] M. Kamionkowski, A. Kosowsky and A. Stebbins, Phys. Rev. Lett. 78, 2058 (1997) [arXiv:astro-ph/9609132].
[10] U. Seljak and M. Zaldarriaga, Phys. Rev. Lett. 78, 2054 (1997) [arXiv:astro-ph/9609169].
[11] M. Kamionkowski and A. Kosowsky, Ann. Rev. Nucl. Part. Sci. 49, 77 (1999) [arXiv:astro-ph/9904108].
[12] Z. Kermish, P. Ade, A. Anthony, K. Arnold, K. Arnold, D. Barron, D. Boettger and J. Borrill et al.,
arXiv:1210.7768 [astro-ph.IM].
[13] B. G. Keating, P. A. R. Ade. J. J. Bock, E. Hivon, W. L. Holzapfel et al., Proc. SPIE 4843, Polarimetry in Astronomy, 284 (2003); [doi:10.1117/12.459274].
[14] C. Bischoff et al. [QUIET Collaboration],
575 (2005) [arXiv:astro-ph/0408515].

[67] J. D. Bowman, M. F. Morales and J. N. Hewitt, Astrophys. J. 695, 183 (2009) [arXiv:0807.3956 [astro-ph]].

[68] A. Liu and M. Tegmark, Phys. Rev. D 83, 103006 (2011) [arXiv:1103.0281 [astro-ph.CO]].

[69] L. Gleser, A. Nusser and A. J. Benson, arXiv:0712.0497 [astro-ph].

[70] LOFAR website: www.lofar.org

[71] SKA website: http://www.skatelescope.org

[72] MWA website: http://www.mwatelescope.org/

[73] N. Thyagarajan, N. U. Shankar, R. Subrahmanyan, W. Arcus, G. Bernardi, J. D. Bowman, F. Briggs and J. D. Bunton et al., arXiv:1308.0565 [astro-ph.CO].

[74] LIGO website http://www.ligo.caltech.edu

[75] IPTA website http://www.ipta4gw.org

[76] F. A. Jenet, G. B. Hobbs, W. van Straten, R. N. Manchester, M. Bailes, J. P. W. Verbiest, R. T. Edwards and A. W. Hotan et al., Astrophys. J. 653, 1571 (2006) [astro-ph/0609013].

[77] T. Regimbau and J. A. de Freitas Pacheco, Astron. Astrophys. 447, 1 (2006) [astro-ph/0509880].

[78] C. R. Evans, I. Iben, and L. Smarr, Astrophys. J. 323, 129 (1987).

[79] V. Bromm and R. B. Larson, Ann. Rev. Astron. Astrophys. 42, 79 (2004) [astro-ph/0311019].

[80] R. Barkana and A. Loeb, Phys. Rept. 349, 125 (2001) [astro-ph/0010468].

[81] M. R. Santos, V. Bromm and M. Kamionkowski, Mon. Not. Roy. Astron. Soc. 336, 1082 (2002) [astro-ph/0111467].

[82] M. Trenti and M. Stiavelli, Astrophys. J. 694, 879 (2009) [arXiv:0901.0711 [astro-ph.CO]].

[83] R. Windhorst et al., Bull. Amer. Astron. Soc. 39, 977 (2007).
FIG. 8: The smallest detectable tensor-to-scalar ratio with different survey (bandit) strategies. The greedy method is drawn in black, with its best and worst performances bracketing the area shaded in gray. Greedy with $\epsilon = 1$ is drawn in red and its performance interval is also shaded in red. The best-performing method in each scenario is shaded in its corresponding color (see Fig. 5). We also note the best improvement achieved with respect to greedy. Column 1 is for the worst-case 10% normalization scenario, Column 2 is for the conservative scenario of 3.6% normalization and Column 3 is for the optimistic case of 3.6% normalization and 80% de-lensing ($\alpha = 0.2$ in Eq. 8). Top Row: Experiment 1, with $n_p = 10$ patches and step-size of one observation day. Middle Row: Experiment 2, with $n_p = 15$ patches and step-size of four observation days. Bottom Row: Experiment 3, $n_p = 5$ and the step-size is two days.