Editors' Suggestion

Sampling-based inference of the primordial CMB and gravitational lensing

Marius Millea[*,†]
Berkeley Center for Cosmological Physics and Department of Physics, University of California, Berkeley, California 94720, USA

Ethan Anderes[†]
Department of Statistics, University of California, Davis, California 95616, USA

Benjamin D. Wandelt
Sorbonne Université, CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 bis bd Arago, 75014 Paris, France,
Sorbonne Université, Institut Lagrange de Paris (ILP), 98 bis bd Arago, 75014 Paris, France, and Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, 10010 New York, New York, USA

(Received 10 February 2020; accepted 24 September 2020; published 28 December 2020)

The search for primordial gravitational waves in the cosmic microwave background (CMB) will soon be limited by our ability to remove the lensing contamination to B-mode polarization. The often-used quadratic estimator for lensing is known to be suboptimal for surveys that are currently operating and will continue to become less and less efficient as instrumental noise decreases. While foregrounds can, in principle, be mitigated by observing in more frequency bands, progress in delensing hinges entirely on algorithmic advances. We demonstrate here a new inference method that solves this problem by sampling the exact Bayesian posterior of any desired cosmological parameters, of the gravitational lensing potential, and of the delensed CMB maps, given lensed temperature and polarization data. We validate the method using simulated CMB data with nonwhite noise and masking on up to 650 deg^2 patches of sky. A unique strength of this approach is the ability to perform joint inference of cosmological parameters, which control both the primordial CMB and the lensing potential, which we demonstrate here for the first time by sampling both the tensor-to-scalar ratio, r, and the amplitude of the lensing potential, A_ϕ. The method allows us to perform the most precise check to-date of several important approximations underlying CMB-S4 r forecasting, and we confirm these yield the correct expected uncertainty on r to better than 10%.

DOI: 10.1103/PhysRevD.102.123542

I. INTRODUCTION

The gravitational lensing of the cosmic microwave background (CMB) is a key cosmological observable. Current and next generation CMB probes are all targeting significant improvements in sensitivity to the lensing effect [1–8]. These will correspond to large improvements in the precision with which we can reconstruct the gravitational lensing potential, ϕ, and with which we can "delense" the CMB to reveal the unaltered primordial signal. The inferred maps of ϕ encode a wealth of information about the late-time structure and geometry of the Universe, both by themselves and in cross-correlation with other tracers of matter. Delensing, which can remove the spurious foreground B-mode polarization generated by lensing, will be crucial in searching for the hypothesized primordial B-mode signal sourced by inflationary gravitational waves. Despite the importance of the lensing effect, however, it is still an open question how in practice to optimally extract cosmological information from the very low-noise observations of the lensed CMB achievable in the near future.

Up until very recently, all CMB lensing analyses have used a quadratic estimator (QE) [9,10] to produce a point estimate of ϕ. Obtaining cosmological constraints then proceeds by either 1) taking the auto power spectrum of this reconstructed ϕ, debiasing the spectrum, and computing error bars with a combination of analytic calculations and Monte Carlo simulations, then comparing to model $C_\ell^{\phi\phi}$ power spectra, or 2) cross-correlating ϕ with other low-redshift probes of structure, and similarly, computing the expected response with various semianalytic techniques. This is the approach taken in the first detection of the lensing effect in the CMB from cross-correlating WMAP with NVSS galaxies [11], the first CMB-only detection by

[*]mariusmillea@gmail.com
[†]These authors contributed equally to this work.

the Atacama Cosmology Telescope [12], the first detection of lensing in the B-mode polarization by the South Pole Telescope [13], the Planck lensing results [14–18], as well as in the large body of other work steadily improving the fidelity of the lensing measurements [19–24]. Delensing can be implemented by using the estimate of ϕ to undo the lensing deflection in the data maps or by creating a B mode template that can be subtracted. Again, this requires using Monte Carlo simulations to quantify the resulting bias and uncertainties in the power spectra of the delensed maps. The first CMB-only delensing analysis used the QE to estimate ϕ maps from *Planck* temperature data and then inverted the lensing deflection [25].

As successful as the QE has been, however, it will soon become obsolete because it becomes statistically suboptimal as instrumental noise levels dip below ~ 5 μK-arcmin [26–28]. This threshold is being crossed with currently available data sets.

Several methods have been proposed to improve upon aspects of the standard QE procedure. Mirmelstein *et al.* [29] derive a more optimal spatial weighting of the quadratically estimated ϕ before taking its power spectrum, although do not improve the ϕ estimate itself. Horowitz *et al.* [28] and Hadzhiyska *et al.* [30] work in the small-scale limit ($\ell \gtrsim 5000$), where a lower variance ϕ estimator can be analytically derived, but which is not optimal on all scales, in particular, not on the intermediate and large scales which are relevant for r estimation. Caldeira *et al.* [31] train a neural network to extract a ϕ map from noisy lensed CMB data, finding near optimality on relevant scales, but it is not straightforward how one would quantify uncertainties on ϕ in such an analysis. Finally, there are a class of near-optimal maximum *a posteriori* (MAP) estimators of ϕ generated by maximizing the Bayesian posterior $\mathcal{P}(\phi|d,\theta)$, where d is the data and θ represents cosmological parameters or directly the theoretical bandpowers (we will refer to this as the "marginal posterior" and the associated "marginal MAP" for reasons that will be clear in a moment). Hirata and Seljak [32,33] were the first to explore such an approach and to develop an approximate maximization technique, while Carron and Lewis [34] recently made the maxmization procedure exact.

A major challenge associated with any new point estimate of ϕ is the quantification and propagation of uncertainty when trying to estimate cosmological parameters from the estimated ϕ or from data delensed by the estimate. Although Monte Carlo simulations can help, these will generally depend on the same cosmological parameters one is trying to estimate in the first place. As an example, consider attempting to use the marginal MAP ϕ to infer the theoretical ϕ bandpowers (in our notation, the case, where $\theta \equiv \{C_\ell^{\phi\phi}\}$). Since $\mathcal{P}(\phi|d,\theta)$ depends on $C_\ell^{\phi\phi}$, the resulting estimate inherits a Wiener-filter-like multiplicative bias, which depends explicitly—but not analytically—on $C_\ell^{\phi\phi}$ itself. This circularity seriously complicates any attempt to debias and/or probe properties of $C_\ell^{\phi\phi}$ in this way.

Despite these challenges, some progress has been made using these new ϕ estimates. Adachi *et al.* [35] were recently the first to apply a non-QE method to actual CMB data, demonstrating that delensing data from the POLARBEAR telecope with the algorithm from [34] yielded a 22% reduction in lensing B modes, compared to only 14% when delensing with the QE. The circularity problem is partially ameliorated by a procedure they develop termed "overlapping B-mode deprojection," wherein for each bandpower that is delensed, a ϕ estimate is constructed only from modes outside of that multipole range. This reduces the size of the bias and its dependence on the theoretical spectra themselves but at the price of a 5%–35% reduction in the delensing efficiency depending on the multipole range considered. Skipping ahead slightly, we remark that the new methodology introduced in this paper would fully remove this delensing efficiency penalty, as well allowing inference of other cosmological parameters governing $C_\ell^{\phi\phi}$ or the delensed bandpowers themselves.

In parallel, there have also been attempts to unify near-optimal estimation of ϕ with simultaneous inference of cosmological parameters. The main approach has been to extend the marginal posterior from $\mathcal{P}(\phi|d,\theta)$ to include the θ as free parameters rather than fixing them, then marginalize out ϕ to arrive at constraints on θ given by $\mathcal{P}(\theta|d) = \int \mathrm{d}\phi \mathcal{P}(\phi,\theta|d)$. Hirata and Seljak [32,33] consider the case of $\theta \equiv \{C_\ell^{\phi\phi}\}$, use the Laplace approximation to perform the integral over ϕ, then compute a maximum likelihood estimator with Gaussian error bars for the resulting $\mathcal{P}(C_\ell^{\phi\phi}|d)$. Carron [36] developed a similar method for $\theta \equiv r$ which does not assume Gaussian error bars on r but still uses an underlying Laplace approximation. Both are useful forecasting methods, but the former has never been checked in the presence of required analysis complexities such as pixel masking, and the brute-force integration employed by the latter does not scale computationally to these cases.

In this paper, we develop a complete Bayesian solution which unifies optimal inference of ϕ along with delensing and cosmological parameter inference. This is achieved by further extending $\mathcal{P}(\phi,\theta|d)$ to include the unlensed CMB fields, hereafter f, rather that analytically marginalizing over them as was implicit in the marginal posterior (hence, the name). The resulting "joint posterior," $\mathcal{P}(f,\phi,\theta|d)$, theoretically extracts all of the information in d for (f,ϕ,θ) and completely summarizes the uncertainty on all of these quantities. As we will demonstrate, it also allows us to perform parameter inference by using Monte Carlo sampling to compute the integral in $\mathcal{P}(\theta|d) = \int \mathrm{d}f\mathrm{d}\phi \mathcal{P}(f,\phi,\theta|d)$. This avoids use of the Laplace approximation, whose accuracy is difficult to check and may be poor due to the nonlinearity of the lensing problem.

The challenge is that this is a very high-dimensional and non-Gaussian posterior, with around $\sim 10^6$ dimensions for the cases considered in this work. Previous attempts at sampling in this space have been blocked by the extreme degeneracies generated by parameter expansion—from ϕ to (f, ϕ, θ)—resulting in more parameter degrees of freedom than data. These nonlinear degeneracies render the exploration of the joint posterior surface extremely difficult. To make progress, one has to find a way to condition the posterior into a more manageable form. We do so here by finding a reparametrization of the posterior from variables (f, ϕ, θ) to new variables (f', ϕ', θ), which have a posterior distribution, which we are then able to sample efficiently with the combination of a Gibbs block sampler and Hamiltonian Monte Carlo (HMC) [37]. The resulting fast-mixing chain yields samples of (f', ϕ', θ), which can be easily converted to samples of (f, ϕ, θ) in postprocessing.

The final piece of the procedure is LenseFlow, which is a numerical algorithm for lensing a map [38]. LenseFlow reformulates lensing into solving an ordinary differential equation (ODE) and makes it possible to compute the gradients and determinants that arise in the reparametrization.

We use our method to compute, for the first time, the exact Bayesian posterior, $\mathcal{P}(r|d)$, in the presence of realistic analysis complexities, notably pixel masking. Doing so, we can check existing forecasts for r similar to those performed for CMB-S4, South Pole Observatory, or Simons Observatory [5,8,39,40]. These rely on approximations which, among other things, ignore masking [27]. Pixel masking couples modes together and leaks E into B mode polarization exactly like lensing, so it is particularly worrisome that it might impact delensing in some unexpected way. We present these results in Sec. V D.

The power of the methodology developed here is not just that it works for forecasting but that it is ready to be applied to analysis of real data, including the many extra complexities which arise. We demonstrate this with simulations, which include the effect of beams, nonwhite noise, and Fourier and pixel masking. We work in the flat-sky approximation and consider patches of sky as large as 512×512 pixels or ~ 650 deg^2. We focus on the specific problem of delensing and inference on the tensor-to-scalar, r, and the amplitude of the lensing potential, A_ϕ. The procedure is conceptually straightforward to generalize to sampling other cosmological parameters (or to sample bandpowers directly), to the curved sky and larger sky area, and to include foreground components. An accompanying software package, CMBLensing.jl (see Sec. VI A), is available online.[1]

[1]See Ref. [41].

II. THE DATA MODEL AND PRIOR ASSUMPTIONS

The Bayesian posterior for the lensing problem is completely specified by a data model and a set of priors. The data model we use, which is flexible enough to handle real experiments, is

$$d = \mathbb{A}\mathbb{L}(\phi) f + n, \tag{1}$$

where d is the data, f are the unlensed CMB fields, and n is the instrumental noise. In this paper, we will work with only polarization data since they give the tightest constraints for low noise levels, although the equations (and our code) are generic to temperature, polarization, or temperature and polarization data. The term $\mathbb{L}(\phi)$ encodes the lensing displacement operation, which can be written for f in the $T/Q/U$ basis as a function of 2D position on the sky x,

$$(\mathbb{L}(\phi) f)(\mathbf{x}) = f(\mathbf{x} + \nabla\phi(\mathbf{x})). \tag{2}$$

Note that $\mathbb{L}(\phi)$ is a linear operator acting on f but has a nonlinear dependence on ϕ. We use LenseFlow [38] to implement $\mathbb{L}(\phi)$ numerically. This is a necessity for our application because no other known numerical approximation allows practical calculation of determinants or of gradients of inverse lensing[2] with respect to ϕ, both of which are needed by the reparametrization which we will describe in Sec. III. Another advantage of LenseFlow is that it allows us easily to apply the full lensing displacement, rather than, e.g., having to rely on a truncated Taylor approximation. We will assume the lensing Born approximation, although it would be straightforward to include a curl potential to the deflection field to model these effects. We omit a detailed treatment of post-Born effects because their importance in the context of this paper will be marginal for current and upcoming surveys [42–45].

Instrumental transfer functions and user-chosen masking are encoded in the operator,

$$\mathbb{A} \equiv \mathbb{K}\mathbb{M}\mathbb{B}, \tag{3}$$

which is the product of a Fourier mask $\mathbb{K}$, a pixel mask $\mathbb{M}$, and a beam $\mathbb{B}$. In general, $\mathbb{M}$ can be chosen to mask the boundaries of the field and any foreground contaminated areas (such the as the areas around detected discrete sources), and $\mathbb{K}$ can be chosen to restrict the analysis to only certain modes in the 2D Fourier plane. Typical choices we use for these operators as well as data simulated according to Eq. (1) are shown in Fig. 1.

[2]Gradients of forward lensing are simple for many algorithms, but easy gradients of inverse lensing appear unique to LenseFlow.

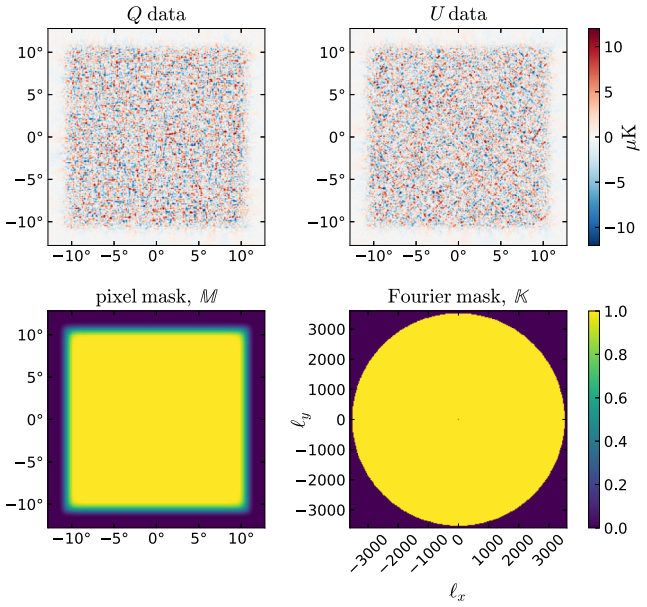

FIG. 1. Typical simulated data and mask choices used in this work. Specifically, these correspond to the configuration BIG (see Table II) with a true value of $r = 0.01$. Reconstructed maps from this exact data are shown in Fig. 6. We note that an apodized pixel mask and an isotropic Fourier mask are not algorithm requirements, rather arbitrary choices we made for this example.

We take Gaussian priors on the fields f, ϕ, and n,

$$f \sim \mathcal{N}(0, \mathbb{C}_f(r)) \tag{4}$$

$$\phi \sim \mathcal{N}(0, \mathbb{C}_\phi(A_\phi)) \tag{5}$$

$$n \sim \mathcal{N}(0, \mathbb{C}_n), \tag{6}$$

where $\mathbb{C}_n$, $\mathbb{C}_f(r)$, and $\mathbb{C}_\phi(A_\phi)$ denote the covariance operators for the experimental noise, unlensed CMB polarization, and lensing potential. The latter two have explicit dependence on the scalar-to-tenser ratio, r, and a lensing spectral amplitude parameter, A_ϕ, given by

$$\mathbb{C}_f(r) = \mathbb{C}^*_{sf} + (r/r^*)\mathbb{C}^*_{tf} \tag{7}$$

$$\mathbb{C}_\phi(A_\phi) = A_\phi \mathbb{C}^*_\phi, \tag{8}$$

where $\mathbb{C}^*_{sf}$, $\mathbb{C}^*_{tf}$, and $\mathbb{C}^*_\phi$ are covariance operators for CMB scalar perturbations, tensor perturbations, and the lensing potential field, computed at fiducial ΛCDM parameters.[3]

Finally, we chose the following weakly informative priors for r and A_ϕ [47]:

$$\pi(r) \propto r^{-1/2}, \qquad \pi(A_\phi) \propto A_\phi^{-1/2}. \tag{9}$$

We find little impact on our sampling algorithm for different priors, and different choices that can be of importance sampled into the final chains if desired.

With this final ingredient specified, the posterior distribution is now fully defined and given by Eq. (10),

$$\mathcal{P}(f, \phi, r, A_\phi | d) \propto \frac{\exp\left\{\frac{(d - \mathbb{A}\mathbb{L}(\phi) f)^2}{2\mathbb{C}_n}\right\}}{\det \mathbb{C}_n^{1/2}} \frac{\exp\left\{-\frac{f^2}{2\mathbb{C}_f(r)}\right\}}{\det \mathbb{C}_f(r)^{1/2}} \frac{\exp\left\{-\frac{\phi^2}{2\mathbb{C}_\phi(A_\phi)}\right\}}{\det \mathbb{C}_\phi(A_\phi)^{1/2}} \frac{1}{(rA_\phi)^{1/2}} \tag{10}$$

$$\mathcal{P}(\phi, r, A_\phi | d) \propto \frac{\exp\left\{-\frac{d^2}{2\mathbf{\Sigma}_d}\right\}}{\det \mathbf{\Sigma}_d^{1/2}} \frac{\exp\left\{-\frac{\phi^2}{2\mathbb{C}_\phi(A_\phi)}\right\}}{\det \mathbb{C}_\phi(A_\phi)^{1/2}} \frac{1}{(rA_\phi)^{1/2}}, \tag{11}$$

where $\mathbf{\Sigma}_d \equiv \mathbb{C}_n + \mathbb{A}\mathbb{L}(\phi)\mathbb{C}_f(r)\mathbb{L}(\phi)^\dagger \mathbb{A}^\dagger$, and we use the shorthand $x^2/\mathbb{N} \equiv x^\dagger \mathbb{N}^{-1} x$.

Note that the conditional distribution $\mathcal{P}(f|\phi, r, A_\phi, d)$ is Gaussian in f (although all the other conditionals are non-Gaussian). Because of this, it is possible analytically to marginalize over f,

$$\mathcal{P}(\phi, r, A_\phi | d) = \int \mathrm{d}f \mathcal{P}(f, \phi, r, A_\phi | d), \tag{12}$$

to arrive at Eq. (11), which, as previously mentioned, we refer to as the marginal posterior.

As discussed in [38], the joint and marginal posteriors have a crucial distinction. All of the operators whose determinants and inverses appear in the joint posterior are sparse in simple bases; e.g., $\mathbb{C}_n$ is sparse in pixel space for typical instrumental noise, and $\mathbb{C}_f$ and $\mathbb{C}_\phi$ are diagonal (and even isotropic) in Fourier space. The action of these operators can thus be evaluated in $\mathcal{O}(N_{\mathrm{pix}} \log N_{\mathrm{pix}})$, where N_{pix} is the number of pixels in the maps, as the limiting step is an FFT to transform into the sparse bases. However, $\mathbf{\Sigma}_d$, which is introduced in the marginal posterior, is not sparse in any simple basis.

This would limit us in several ways if we were attempting to use the marginal posterior for sampling.

[3]Spectra are computed using CAMB (see Ref. [46]) with fiducial settings $k^* = 0.002$, $r^* = 0.1$, $A^*_\phi = 1$, $\omega_\mathrm{b} = 0.0224567$, $\omega_\mathrm{c} = 0.118489$, $\tau = 0.055$, $\theta\mathrm{s} = 0.0104098$, $\log A = 3.043$, $n_\mathrm{s} = 0.968602$, and $n_\mathrm{t} = -r^*/8$. Note that for simplicity, in Eq. (7), we are implicitly fixing n_t rather than enforcing the single field consistency relation.

Evaluating gradients of $\det \mathbf{\Sigma}_d$ with respect to ϕ, which would be needed by the HMC sampler (see Sec. IV), would now have to be done through a costly Monte Carlo procedure [34]. This procedure involves solving $N_{\rm MC} \sim 500$ conjugate gradient problems, each of which require $N_{\rm CG} \sim 100$ conjugate gradient iterations, with each iteration having similar computation cost as a single joint posterior gradient. Hence, marginal posterior gradients are slower than joint posterior gradients by a factor of an order $N_{\rm MC} N_{\rm CG}$, which can in practice be a very large number. Even if this were overcome (if the total CPU cost was not prohibitive, the $N_{\rm MC}$ steps can at least be done in parallel), there is another even more serious limitation. No algorithm we are aware of can robustly evaluate $\det \mathbf{\Sigma}_d$ itself faster than $\mathcal{O}(N_{\rm pix}^3)$, which in practice makes this impossible for maps larger than about 32×32 pixels. Without an ability to evaluate this determinant and hence, the value of our log posterior, the accept/reject step of the HMC is impossible. For these reasons, we find that sampling the joint posterior is the more promising path, and the one which we take.

In summary, we choose to work with the higher dimensional joint posterior because it has a structure that allows the use of powerful Markov chain Monte Carlo (MCMC) sampling techniques such as HMC. This approach is typical for the implementation of high-dimensional Bayesian hierarchical models, starting with their first application in cosmology [48] which applied Gibbs sampling to CMB power spectrum inference, or the more recent application to nonlinear large scale structure reconstruction and inference in the Bayesian origin reconstruction from galaxies (BORG) sampler [49–51].

III. REPARAMETRIZING THE POSTERIOR

The joint posterior, parametrized as in Eq. (10) by the unlensed CMB fields and the lensing potential, is nearly unusable in practice due to the presence of large non-Gaussianities and degeneracies. These issues already appeared in a milder form in the temperature-only CMB lensing posterior [52], where the solution was to change from the unlensed to the lensed (or from a "sufficient" to an "ancillary") parametrization. The situation is more challenging for the polarized CMB lensing/delensing problem we treat in this paper, and the solution in [52] is not powerful enough. In the context of polarization, Millea *et al.* [38] encountered the same underlying problem when maximizing $\mathcal{P}(f, \phi | d, \theta)$, but the "cooling scheme" solution presented there does not have an obvious analog for sampling. Additionally, here we have the complexity of degeneracies in the full (f, ϕ, θ) space, which must be dealt with.

A key aspect of this work is that we develop a physically motivated reparametrization, which works for polarization and yields a posterior, which is significantly less degenerate and more Gaussian than the original $\mathcal{P}(f, \phi, r, A_\phi | d)$. The reparametrization is fully invertible and consequently, does not introduce any approximations to the inference; it only serves to increase the efficiency of sampling or maximization. We first describe the reparametrization (which we also refer to as "mixing," since it mixes the various parameters) and afterwards explain the motivation behind it.

We perform a change of variables from (f, ϕ) to new variables, which we call (f', ϕ'), which are defined by

$$\phi' \equiv \mathbb{G}(A_\phi)\phi \tag{13}$$

$$f' \equiv \mathbb{L}(\phi)\mathbb{D}(r) f. \tag{14}$$

The operator $\mathbb{D}(r)$ is defined to be diagonal in the E, B Fourier domain, and $\mathbb{G}(A_\phi)$ is diagonal in the Fourier domain, with

$$\mathbb{D}(r) \equiv \left[\frac{\tilde{\mathbb{C}}_f(r) + 2\mathbb{N}_f}{\tilde{\mathbb{C}}_f(r)}\right]^{1/2} \left[\frac{\tilde{\mathbb{C}}_f(r)}{\mathbb{C}_f(r)}\right]^{1/2} \tag{15}$$

$$\mathbb{G}(A_\phi) \equiv \left[\frac{\mathbb{C}_\phi(A_\phi) + 2\mathbb{N}_\phi}{\mathbb{C}_\phi(A_\phi)}\right]^{1/2}, \tag{16}$$

where $\tilde{\mathbb{C}}_f(r) = \mathbb{C}_f(r) + \mathbb{N}_{\rm len}$ and $\mathbb{N}_{\rm len}$ denotes the effective power contribution of lensing to the CMB polarization, which we set equal to 5 μK-arcmin white noise (this seems to work better than using the actual lensing contribution that rolls off at higher ℓ). The operators $\mathbb{N}_f$ and $\mathbb{N}_\phi$ are taken to be diagonal in the Fourier domain and are intended to represent the effective noise for f and ϕ in the data. Even if the noise covariance is not actually diagonal in Fourier space, the requirement is only that it needs to be approximated sufficiently well by a Fourier diagonal approximation. Since we explicitly take the instrumental noise in our simulations to be diagonal in Fourier space, we use directly $\mathbb{N}_f = \mathbb{C}n$. For $\mathbb{N}_\phi$, we compute an iterated "N_0" noise as described in Smith *et al.* [27].

The reparametrized posterior needs the determinant of the Jacobian of the transformation, where the Jacobian is

$$\frac{\partial(f', \phi')}{\partial(f, \phi)} = \begin{bmatrix} \mathbb{L}(\phi)\mathbb{D}(r) & \frac{\partial}{\partial\phi}\mathbb{L}(\phi)\mathbb{D}(r) f \\ 0 & \mathbb{G}(A_\phi) \end{bmatrix}. \tag{17}$$

We have intentionally chosen the reparametrization such that the Jacobian is upper triangular, since in this case, the determinant does not involve the complicated off diagonal term. Additionally, because we model $\mathbb{L}(\phi)$ with LenseFlow, we have $\det \mathbb{L}(\phi) = 1$, independent of ϕ [38]. Also, since $\mathbb{D}(r)$ and $\mathbb{G}(A_\phi)$ are diagonal in Fourier space, their determinants are easy to compute. This gives a final tractable reparametrized posterior, which is given by

$$\begin{aligned} &\log \mathcal{P}(f', \phi', r, A_\phi | d) \\ &= \log \mathcal{P}(f(f', \phi', r, A_\phi), \phi(\phi', A_\phi), r, A_\phi | d) \\ &\quad - \log\det \mathbb{G}(A_\phi) - \log\det \mathbb{D}(r). \end{aligned} \tag{18}$$

Note that, by design, the new determinant terms are independent of f and ϕ. This means that the best-fit (f, ϕ) at fixed (r, A_ϕ) can be computed by running the maximization in the mixed parametrization, then taking the best-fit (f', ϕ') and unmixing them. The maximization is much easier in the mixed parametrization, and can be done with coordinate descent similarly as in [38], but with the cooling scheme no longer needed.

Gradients of the mixed posterior can be computed from gradients of the original posterior with an application of the chain rule using the Jacobian in Eq. (17). Both evaluating the value of and gradients of the reparametrized posterior are only about twice the computational cost of the original posterior, stemming from the presence of a second lensing operation $\mathbb{L}(\phi)$, which appears in Eq. (14).

The choice of $\mathbb{D}(r)$ and $\mathbb{G}(A_\phi)$ can be motivated as follows. Consider the toy statistical problem of obtaining constraints on a scalar parameter, θ, given data, d, where

$$d = s + n, \qquad s \sim \mathcal{N}(0, \mathbb{S}(\theta)), \qquad n \sim \mathcal{N}(0, \mathbb{N}).$$

The field n represents noise and s the signal field, with a covariance operator $\mathbb{S}(\theta)$ depending on the unknown parameter. The goal in this toy example is to find an invertible reparametrization $s \to s'$ of the form $s' = \mathbb{G}(\theta)s$, which minimizes the dependence between θ and s' given d. In the ideal case, such a choice of $\mathbb{G}(\theta)$ would have the property that $\mathcal{P}(\theta|s', d) \approx \mathcal{P}(\theta|d)$, meaning s' provides minimal additional information for θ beyond what is already contained in d. Such a property would imply that a single iteration of a Gibbs sampling algorithm for (θ, s') would return an approximate marginal draw from $\mathcal{P}(\theta|d)$.

Another way of phrasing this goal is to choose $\mathbb{G}(\theta)$ such that the information content in (d, s') for θ is minimized. Note that the marginal information in d for θ is fixed regardless of $\mathbb{G}(\theta)$ since we are simply considering reparametrization of the same data model. So by minimizing the joint information in (d, s') for θ, we are implicitly minimizing the additional information in s' for θ beyond that given by d.

A way to describe this mathematically is to start by letting $\mathcal{F}(\theta; \mathbb{G})$ denote the Fisher information for θ given (d, s'); in particular,

$$\mathcal{F}(\theta; \mathbb{G}) = \left\langle -\frac{\partial^2}{\partial\theta^2} \log \mathcal{P}(d, s'|\theta) \right\rangle_{d,s' \sim \mathcal{P}(d,s'|\theta)}, \tag{19}$$

where the dependence on $\mathbb{G}(\theta)$ is implicit in the reparametrized density $\mathcal{P}(d, s'|\theta)$. Then, $\mathcal{F}(\theta; \mathbb{G})$ can be explicitly computed using standard matrix algebra/calculus to arrive at

$$\begin{aligned}\mathcal{F}(\theta; \mathbb{G}) = \mathrm{tr}[&\mathbb{S}(\mathbb{G}^{-1}\dot{\mathbb{G}})^\dagger(\mathbb{N}^{-1} + \mathbb{S}^{-1})(\mathbb{G}^{-1}\dot{\mathbb{G}}) \\ &+ (\mathbb{G}^{-1}\dot{\mathbb{G}})^2 + 2\dot{\mathbb{S}}\mathbb{S}^{-1}(\mathbb{G}^{-1}\dot{\mathbb{G}}) + \frac{1}{2}(\mathbb{S}^{-1}\dot{\mathbb{S}})^2],\end{aligned} \tag{20}$$

where the overdots refer to derivative with respect to the scalar θ. Finally, we seek to minimize the Fisher information, and rather than doing so at any fixed θ, we integrate over the prior for θ, which can be any arbitrary probability function, $\mathcal{P}(\theta)$. Thus, we seek $\mathbb{G}(\theta)$, which is a minimizer in

$$\arg\min_{\mathbb{G}} \int \mathrm{d}\theta \mathcal{F}(\theta; \mathbb{G})\mathcal{P}(\theta). \tag{21}$$

In the case that $\mathbb{G}(\theta)$, $\mathbb{S}(\theta)$, and $\mathbb{N}$ are diagonal with positive entries, we can define $\mathbb{H}(\theta) = \log \mathbb{G}(\theta)$ such that $\dot{\mathbb{H}} = \mathbb{G}^{-1}\dot{\mathbb{G}}$. Now $\mathcal{F}(\theta; \mathbb{G})\mathcal{P}(\theta) = \mathcal{L}(\theta, \dot{\mathbb{H}}(\theta))$ for a Lagrangian $\mathcal{L}$, which yields N Euler-Lagrange equations (corresponding to N diagonal entries of $\mathbb{G}$) that characterize the stationary points of (21) given by

$$(\mathbb{N}^{-1}\mathbb{S} + 2\mathbb{I})\frac{d}{d\theta}\log \mathbb{G}(\theta) + \frac{d}{d\theta}\log \mathbb{S}(\theta) = 0, \tag{22}$$

where we have applied a boundary condition such that Eq. (22) is invariant to the choice of prior. An explicit solution is then given by

$$\mathbb{G}(\theta) \propto \left[\frac{\mathbb{S}(\theta) + 2\mathbb{N}}{\mathbb{S}(\theta)}\right]^{1/2}, \tag{23}$$

which is additionally invariant to multiplication by any diagonal matrix that does not depend on θ.

Notice that for coordinates which are noise dominated, we have $\mathbb{G}(\theta) \propto \mathbb{S}(\theta)^{-1/2}$. It is simple to see analytically that in this limit, the posterior becomes exactly separable given this choice of $\mathbb{G}(\theta)$. This also conforms to the expectation from Jewell *et al.* [53], who derived the same result in this limit. For coordinates which are signal dominated, we instead have $\mathbb{G}(\theta) \propto \mathbb{1}$, which also matches intuition since in this limit, the data determine s perfectly. Equation (23) is thus in some sense an optimal way to connect these two limits. One can additionally regard this result as an extension of Racine *et al.* [54], who derived a modified Gibbs proposal step, which also works in both limits. The advantage of our result is that it is generic and not limited to Gibbs sampling, and that it does not affect the detailed balance of the Monte Carlo chains or force us to include any extra efficiency-reducing accept/reject steps.

This toy example directly explains the mixing matrix $\mathbb{G}(A_\phi)$ given in Eq. (16); it is just Eq. (23) with $\mathbb{S} = \mathbb{C}_\phi$ and $\mathbb{N}$ chosen as previously described. Although the (ϕ, A_ϕ) block of the lensing posterior is not exactly the same as

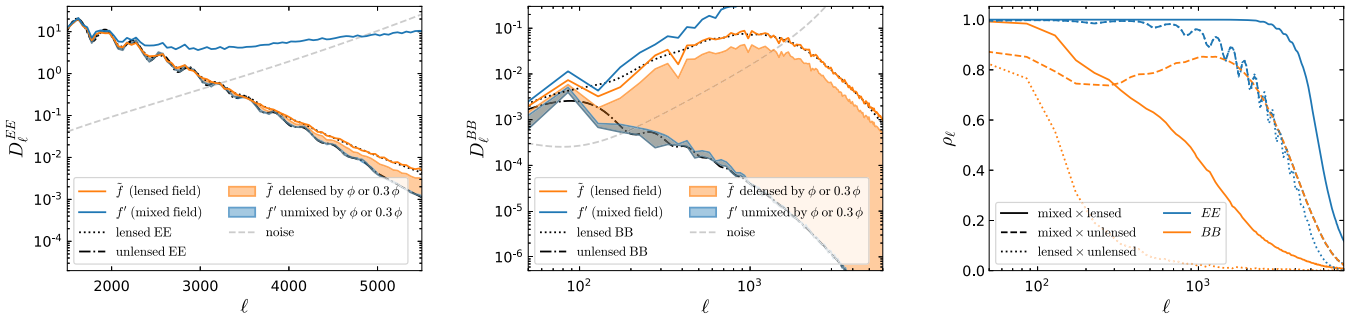

FIG. 2. Three figures which are helpful in understanding the benefit of the reparametrization (described in Sec. III), which makes sampling possible. Data configuration `2PARAM` (see Table II) is assumed for these figures. Left two panels: The reparamaterization includes switching from sampling the unlensed CMB fields, f, to sampling the "mixed" fields, f'. The left two panels show that the power variation in a typical f' unmixed by various ϕ is very small, an indication that large moves are allowed in Gibbs samples from the conditional $\mathcal{P}(\phi'|f', d)$. For comparison, the much larger variation in a typical $\tilde{f}$ when delensed by various ϕ is shown, indicating that the lensed parametrization performs very poorly for polarization as is the case here. Right panel: Empirically, one finds that mixed E is mostly lensed E at all scales, while mixed B is lensed B at large scales but unlensed B at small scales. We demonstrate this here by cross-correlating the mixed with the lensed or unlensed fields. This qualitatively conforms to the expectations of what should give a parametrization that is minimally degenerate (see discussion in Sec. III).

(s, θ) in the toy example, in particular the $\mathcal{P}(s|\theta, d)$ conditional is Gaussian in the toy example whereas as the corresponding ϕ conditional is not, the problems are sufficiently similar that this works very well.

The mixing matrix $\mathbb{D}(r)$ given in Eq. (14) is also similar; the first term indeed is just Eq. (23) with $\mathbb{S} = \tilde{\mathbb{C}}_f$ and $\mathbb{N} = \mathbb{C}_n$. There is, however, an additional prefactor of $(\tilde{\mathbb{C}}_f/\mathbb{C}_f)^{1/2}$ present and also a lensing operation in Eq. (14) before arriving at the final mixed field, f'. The motivation for this can be understood by applying a similar argument as in our toy example. Suppose we wish to make f' and ϕ' more independent and increase the width of the conditional $\mathcal{P}(\phi'|f', \theta, d)$ so that it is on the order of the marginal distribution $\mathcal{P}(\phi'|\theta, d)$. We have argued that a way to do this is to decrease the information content for ϕ' in (d, f'). One way to do so is to prevent the power in f' from being informative about ϕ'. The prefactor in Eq. (14) serves exactly this purpose, since it boosts power in f' to look like lensed power, independent of whether ϕ' causes a large or small lensing. This shifts the information in $\mathcal{P}(\phi'|f', \theta, d)$ from the lensed B-mode power to the less informative lensed B-mode phase coupling. Typical power spectra of f' are shown in the left two panels of Fig. 2, as well as an illustration of how the power in f' is less informative than, e.g., the power in the lensed field, $\tilde{f}$, explaining why $\tilde{f}$ does not work well as a parameter when considering polarization data.

Another way to understand why the mixed parametrization works well is to ask what choice of variables render the posterior distribution in Eq. (10) explicitly independent between f' and ϕ'. In the limit of low signal-to-noise where only the prior terms matter, an independent choice of variables is trivially (f, ϕ) since the prior is explicitly separable between them. As we move away from this limit, the data likelihood begins to couple f and ϕ, so it is clear the right choice will be some combination of them. The mixing indeed has exactly this behavior in these limits, as demonstrated in the right panel of Fig. 2. Here, we plot the cross-correlation coefficient at different scales between the mixed maps and either lensed or unlensed ones. For scales where signal-to-noise is low (like medium and small scales in B), the mixed field f' looks like the unlensed field. In the high signal-to-noise limit (such as in E, or at very large scales in B), f' becomes a mixture of f and ϕ. In particular, we find it tracks the lensed field.

The end result of all of this is a dramatically better conditioned posterior, resulting in large Gibbs moves and much faster chain mixing for the sampling procedure we describe in the next section. The improvement is not limited to our particular Gibbs sampler, however, and we expect that any sampling algorithm applied to this problem would benefit drastically from this reparametrization. Finally, we note that although the reparametrization in our toy example is optimal in the sense that it can be rigorously and analytically derived, the full mixing in Eqs. (13) and (14) is almost certainly not optimal. Instead, it is based on physical intuition and simple analogy to the covariance estimation problem, and it would be worthwhile to investigate even better choices.

Algorithm 1. $\mathcal{P}(f', \phi', r, A_\phi|d)$ sampler. The Gibbs sampling algorithm,

1: Initialize $A_{\phi,0}$ and r_0 anywhere within the prior range.
2: Initialize fields f'_0 and ϕ'_0 with quasisamples.
3: **for** $i = 1 \ldots n$ **do**
4: $f'_i \sim \mathcal{P}(f'|\phi'_{i-1}, A_{\phi,i-1}, r_{i-1}, d)$ ▷CG
5: $\phi'_i \sim \mathcal{P}(\phi'|A_{\phi,i-1}, r_{i-1}, f'_i, d)$ ▷HMC
6: $A_{\phi,i} \sim \mathcal{P}(A_\phi|r_{i-1}, f'_i, \phi'_i, d)$ ▷Slice
7: $r_i \sim \mathcal{P}(r|f'_i, \phi'_i, A_{\phi,i}, d)$ ▷Slice
8: **end for**

TABLE I. List of tuning parameters used in Gibbs algorithm 1.

$\mathbb{N}_f$	Effective noise used in $\mathbb{D}(r)$	See Sec. III
$\mathbb{N}_\phi$	Effective noise used in $\mathbb{G}(A_\phi)$	See Sec. III
$\tilde{\mathbf{\Lambda}}_f(r)$, $n_{\rm cg}$	Parameters for conjugate gradient sample of f'	See Sec. IV B
$\epsilon_{\rm h}$, $n_{\rm h}$, $\mathbf{\Lambda}_{\phi'}(A_\phi)$	HMC leapfrog and momentum parameters for ϕ'	See Sec. IV C
K	Number of over-relaxation samples for r and A_ϕ	See Sec. IV D

IV. THE GIBBS CHAIN

Next, we outline the details of our Gibbs chain for sampling $\mathcal{P}(f', \phi', r, A_\phi|d)$. The procedure itself is summarized in algorithm 1 and is a standard block Gibbs sampler with each of f', ϕ', r, and A_ϕ sampled on separate passes. A list of all of the tuning parameters that will be needed are also summarized in Table I.

There is a fair amount of freedom in setting up the sampler; our motivation comes from two considerations. First, the conditional distribution of f' is Gaussian; hence, it is advantageous to split this piece off into its own Gibbs pass and use a sampling technique specifically tailored for this situation. Second, the r and A_ϕ slices are qualitatively quite different from the other parameters since they are "global" parameters that are correlated at a small level with everything else, making it more difficult to simply include them in a joint HMC pass. We therefore split these off as well, and since they are one dimensional, it is easy to use slice sampling. This also has the advantage of letting us build up a Blackwell-Rao posterior for these parameters.

We now describe the different passes in more detail.

A. Initializing f'_0 and ϕ'_0 with quasisamples

The choice of initialization can shorten the "burn-in time," that is, the number of samples required for the Markov chain to equilibrate. Although initialization is less critical for our case since our reparametrization results in good mixing properties of the chains, the method described here is so simple it is worth utilizing. First, we note that while we do have easy access to the best fit of the distribution, which would seem like reasonable starting point, in very high-dimensional spaces, the best fit is often extremely far from the bulk of the posterior mass (e.g., for an n-dimensional standard normal distribution, the probability mass associated with the interior of the unit sphere centered on the origin goes to 0 as $n \to \infty$). Instead, we use the following cheap way to generate a point which more closely resembles a true sample and should reside closer to the bulk of the posterior.

First, we randomly sample $A_{\phi,0}$ and r_0 from their priors to generate their starting values in the chain. We then initialize f'_0 and ϕ'_0 to zero and iterate the following two steps:

$$f'_0 \sim \mathcal{P}(f'|\phi'_0, A_{\phi,0}, r_0, d) \tag{24}$$

$$\phi'_0 = \phi'_0 + \alpha \mathbf{\Lambda}^{-1}_{\phi'} \nabla_{\phi'} \log \mathcal{P}(\phi'|f'_0, A_{\phi,0}, r_0, d)|_{\phi'_0}. \tag{25}$$

The first step [Eq. (24)] is a draw from the conditional distribution of f', which, as we will describe below, can be done with one run of a conjugate gradient solver. The second step [Eq. (25)] is a quasi Newton-Raphson iteration where α is a step size, which we compute via line search to maximize the resulting $\log \mathcal{P}$ at each iteration, and $\mathbf{\Lambda}_{\phi'}$ is an approximate negative Hessian of $\log \mathcal{P}$ with respect to ϕ', which we take as

$$\mathbf{\Lambda}_{\phi'}(A_\phi) = \mathbb{G}[A_\phi]^{-2}[\mathbb{N}_\phi^{-1} + \mathbb{C}_\phi(A_\phi)^{-1}], \tag{26}$$

where $\mathbb{N}_\phi$ is the same approximate noise covariance appearing in Eq. (16).

Note that if we replaced Eq. (25) with a conditional sample of ϕ', we would recover exactly our sampling algorithm given in algorithm 1 with a fixed $A_{\phi,0}$ and r_0. Hence, we call the point generated by this procedure a "quasisample," since it involves sampling in the f' direction but maximization in the ϕ' direction. In practice, an important aspect of quasisamples is that they do not contain the mean-field feature, which would otherwise exist in the joint best fit, $\hat{\phi}_J$, [38] and which would slow the initial convergence of our chains. We find 20 iterations of Eqs. (24)–(25) are sufficient.

B. The f' Gibbs pass

The first step of each full chain iteration is to draw a conditional sample of f'. We can do so by solving one conjugate gradient problem [48]. This is because the conditional f posterior is Gaussian,

$$\begin{aligned}&\mathcal{P}(f|\phi, A_\phi, r, d)\\ &= \mathcal{N}(\mathbf{\Lambda}_f(r,\phi)^{-1}\mathbb{L}(\phi)^\dagger\mathbb{A}^\dagger\mathbb{C}_n^{-1}d, \mathbf{\Lambda}_f(r,\phi)^{-1}),\end{aligned} \tag{27}$$

where the inverse covariance $\mathbf{\Lambda}_f(r,\phi)$ is given by

$$\mathbf{\Lambda}_f(r,\phi) = \mathbb{L}(\phi)^\dagger\mathbb{A}^\dagger\mathbb{C}_n^{-1}\mathbb{A}\mathbb{L}(\phi) + \mathbb{C}f(r)^{-1}. \tag{28}$$

A sample, f_i, is then drawn by computing,

$$\begin{aligned}f_i = \mathbf{\Lambda}_f(r,\phi)^{-1} \times [&\mathbb{L}(\phi)^\dagger\mathbb{A}^\dagger\mathbb{C}_n^{-1}d\\ &+ \mathbb{L}(\phi)^\dagger\mathbb{A}^\dagger\mathbb{C}_n^{-1/2}\xi_1 + \mathbb{C}_f(r)^{-1/2}\xi_2],\end{aligned} \tag{29}$$

where ξ_1 and ξ_2 are independent unit normal random fields, resampled at each iteration, and the inversion of $\mathbf{\Lambda}_f(r,\phi)$ is done via conjugate gradient. Finally, because the mixing is a linear function of f, a sample of the mixed field, f'_i, is simply given by $f'_i = \mathbb{L}(\phi_i)\mathbb{D}(r_i)f_i$. Note that conjugate gradient is, by design, tailored to exploit the positive definiteness of $\mathbf{\Lambda}_f$, or equivalently, the convexity of the f conditional. This is why it is advantageous to split f' into its own Gibbs step, rather than, e.g., including it in a larger HMC pass which would not be exploiting the convexity and hence, be much less efficient.

For the conjugate gradient solver, we use a simple diagonal preconditioner, $\tilde{\mathbf{\Lambda}}_f(r)$, given by

$$\tilde{\mathbf{\Lambda}}_f(r) = \mathbb{B}^\dagger\mathbb{K}^\dagger\mathbb{C}_n^{-1}\mathbb{K}\mathbb{B} + \mathbb{C}_f(r)^{-1}. \tag{30}$$

Although we find this is sufficient for the simulated data considered here, this step does account for roughly half of the total run time of the entire sampling algorithm and is thus worth improving further. A promising avenue we expect to try in the future is to use the neural network-based Wiener filter given by Münchmeyer and Smith [55]; this assumes $\phi = 0$ but could potentially be a powerful preconditioner. Other techniques developed for Wiener filtering without preconditioner could be adapted to the lensing problem, possibly in combination with a neural preconditioner [56,57]. We also note that one could absorb the final mixing step into the quantity in brackets in Eq. (29), although in practice we do not do so and instead solve Eq. (29) exactly as written, which we find to be more numerically stable.

C. The ϕ' Gibbs pass

The next step of the sampling algorithm is to draw a conditional sample of ϕ'. Because this conditional distribution is not Gaussian, no specialized tricks like in the previous subsection exist, and we instead use a single HMC pass [37] to draw a sample.

There are only two tunable inputs to the HMC algorithm: 1) a mass matrix, which should approximate the Hessian of the distribution to give the most efficient sampling, and 2) a prescription for the length of each Hamiltonian trajectory. For the mass matrix, we again use the Hessian approximation, $\mathbf{\Lambda}_{\phi'}$, given in Eq. (26). For the trajectories, we perform a leapfrog symplectic integration with $n_h = 25$ steps of size $\epsilon_h = 0.02$. This choice is hand tuned to work well for a range of configurations similar to the main ones we consider in this work but may need to be retuned for sufficiently different analyses.

Fortunately, it is fairly straightforward to perform this tuning. To begin with, the choice of ϵ_h is set uniquely by the need to limit symplectic integration error. This error comes from two sources: 1) errors in the posterior gradient itself, and 2) errors due to the finite step size, ϵ_h. Before choosing ϵ_h, we first make sure the contribution from (1) is subdominant. For this, the number of LenseFlow ODE steps is relevant because we compute gradients of the lensing operator by running a separate ODE for the gradient, rather than by backpropagating a gradient through the original ODE [see Sec. IV of [38]]. The gradient generated by the gradient ODE will differ from the true gradient due to ODE integration error. In practice, we find we need a fourth order Runge-Kutta integration with ten steps before the LenseFlow gradient error is a subdominant contribution to the symplectic integration error. Another source of error in the posterior gradient is floating point truncation. We find the dominant source comes from the sums involved in the inner products in the posterior in Eq. (10) and that these errors can be significantly reduced with Kahan summation [58]. With this, we are able to run the entire analysis with 32-bit instead of 64-bit floating point numbers, which doubles performance on most CPUs and gives potentially much more drastic speed improvements on GPUs, depending on hardware (fast 64-bit support on GPUs is limited to high-end models). Once this and the number of LenseFlow ODE steps are set, ϵ_h is then simply tuned to give small enough integration errors such that the HMC acceptance is near 80%.

Given ϵ_h, the choice of n_h comes from integrating long enough to meet the "no U-turn criteria" [59]. We have checked the integration length on representative data configurations and multiple random starting points, and find $n_h = 25$ is adequate. We note that we do not adaptively change either n_h or ϵ_h throughout our chains (the full "no U-turn sampler" of [59] usually refers to an algorithm where the integration length is adaptively chosen at each step). We do this for simplicity and since we have not found very obvious regions of parameter space which appear to need significantly different values. The reparametrization of Sec. III in particular helps us avoid the "funnel problem," [60] which might otherwise cause such a need. Nevertheless, it is worth exploring more sophisticated HMC sampling techniques in the future, since, as we will discuss in Sec. V C, our chains have autocorrelation lengths which could be even further improved.

D. The A_ϕ and r Gibbs passes

Finally, we sample the conditional distribution of each of A_ϕ and r on separate Gibbs passes. Because these are one-dimensional probability distributions, we can directly probe these functions on a grid and use inverse transform sampling (often called "slice sampling") to draw a sample. Moreover, we find that the log conditional densities are typically quite smooth and close to quadratic, so we can compute a very accurate interpolation of the log probability. For the simulations given in this paper, we use 200 grid points over the intervals $A_\phi \in [0.75, 1.25]$ and $r \in [10^{-6}, 0.1]$, respectively, with the r grid points quadratically spaced to ensure sufficient resolution near $r = 0$.

There are two additional tricks, which come at no extra computational cost, which we utilize to reduce the number of samples required for convergence. First, we use MCMC over-relaxation [61]; instead of drawing a single sample from the discretized density, K samples are drawn independently, one of which is chosen depending on the rank (among the K draws) of the parameter value from the previous Gibbs iteration parameter. In the simulations below, we set $K = 15$, and we find that this can sometimes reduce the chain autocorrelation time by 10%–20%. Second, we save the interpolated conditional densities at each step and use these to construct Rao-Blackwell estimates of the marginal posterior densities,

$$\mathcal{P}(r|d) \approx \frac{1}{n}\sum_{i=1}^{n} \mathcal{P}(r|f'_i, \phi'_i, A_{\phi,i}, d) \tag{31}$$

$$\mathcal{P}(A_\phi|d) \approx \frac{1}{n}\sum_{i=1}^{n} \mathcal{P}(A_\phi|r_{i-1}, f'_i, \phi'_i, d). \tag{32}$$

This helps reduce the variance of the estimated posteriors slightly faster than just building up a histogram of the Monte Carlo samples, particularly deep in the tails of distribution.

V. SIMULATION RESULTS

A. Description of runs

With the details of our posterior and the sampling algorithm specified, we now turn to actually running chains and interpreting results. We have picked three different configurations of simulated data, the details summarized in Table. II, which are meant to resemble possible CMB-S4 resolutions and noise levels, but slightly smaller sky area. We will describe these runs first, then come back to a more quantitative discussion of chain convergence as well as some scientific conclusions that can be extracted from these results.

All of our configurations include a Gaussian beam with a 2–3 arcmin full width half max (FWHM). We take isotropic Gaussian 1 μK-arcmin polarization noise with a power spectrum, which includes a contribution from a $1/f$ knee, modeled via $\ell_{\rm knee}$ and $\alpha_{\rm knee}$ parameters [62]. The runs are all in the flat-sky approximation and include a border mask, $\mathbb{M}$, of various widths. Although the runs we have chosen here use an apodized border mask, we find that unapodized masks work just as well. This is helpful if, for example, there are so many point sources that apodizing them all would discard too much data. In addition to a pixel mask, we also apply an isotropic low-pass mask in Fourier space, $\mathbb{K}$, generally near the Nyquist frequency. Although we do not do so here, it is completely straightforward to use an anisotropic Fourier mask instead, which can be useful in limiting systematics by masking scan-parallel and scan-perpendicular directions differently. Finally, we use grid sizes between 256×256 and 512×512 pixels. The latter is around the limit of what is currently computationally possible on performance hardware and covers about 650 deg^2, with an effective unmasked region of around 450 deg^2. This is about a third to a fifth of the planned CMB-S4 deep field, where our procedure is most applicable, with several years remaining to scale up to the full patch or beyond.

The first run we describe uses data simulated in configuration `2PARAM`. In this configuration, we sample both r and A_ϕ. We show a trace of the sampled values for these two parameters in Fig. 3. We will asses convergence and correctness of the chains in the next subsection, but for now, one can at least see by eye the stationarity of the samples and that they cover the true input values, as expected. For this case, we have also run an identical copy of the chain, including identical starting random seed, but which uses $\mathbb{G}(A_\phi) = \mathbb{1}$ instead of the fiducial choice, which we described in Sec. III. The impact of not using the fiducial $\mathbb{G}(A_\phi)$ is shown in orange. There is a dramatic reduction in the convergence of the A_ϕ samples (the

TABLE II. Parameters for the different configurations of simulated data used in this work.

	Configuration `2PARAM`	Configuration `MANY`	Configuration `BIG`
Map size	256×256	256×256	512×512
Pixel width	2 arcmin	3 arcmin	3 arcmin
Total area	73 deg^2	160 deg^2	650 deg^2
White noise level in P	1 μK-arcmin	1 μK-arcmin	1 μK-arcmin
$(\ell_{\rm knee}, \alpha_{\rm knee})$	(100,3)	(100,3)	(100,3)
Beam FWHM	2 arcmin	3 arcmin	3 arcmin
Fourier masking ($\mathbb{K}$)	$2 < \ell < 5000$	$2 < \ell < 3500$	$2 < \ell < 3500$
Pixel masking ($\mathbb{M}$)	$0.4°$ border + $0.6°$ apod	$0.6°$ border + $0.9°$ apod	$1.2°$ border + $1.8°$ apod
Sampled parameters (θ)	r, A_ϕ	r	r
Fiducial r	$r = 0.04$	$r = \{0.04, 0.02, 0\}$	$r = \{0.02, 0.01, 0\}$
Chain iterations	10000	5000	4000
Autocorrelation length for θ	22	5–33	12
Wall time (one GPU)	48 h	19 h	50 h

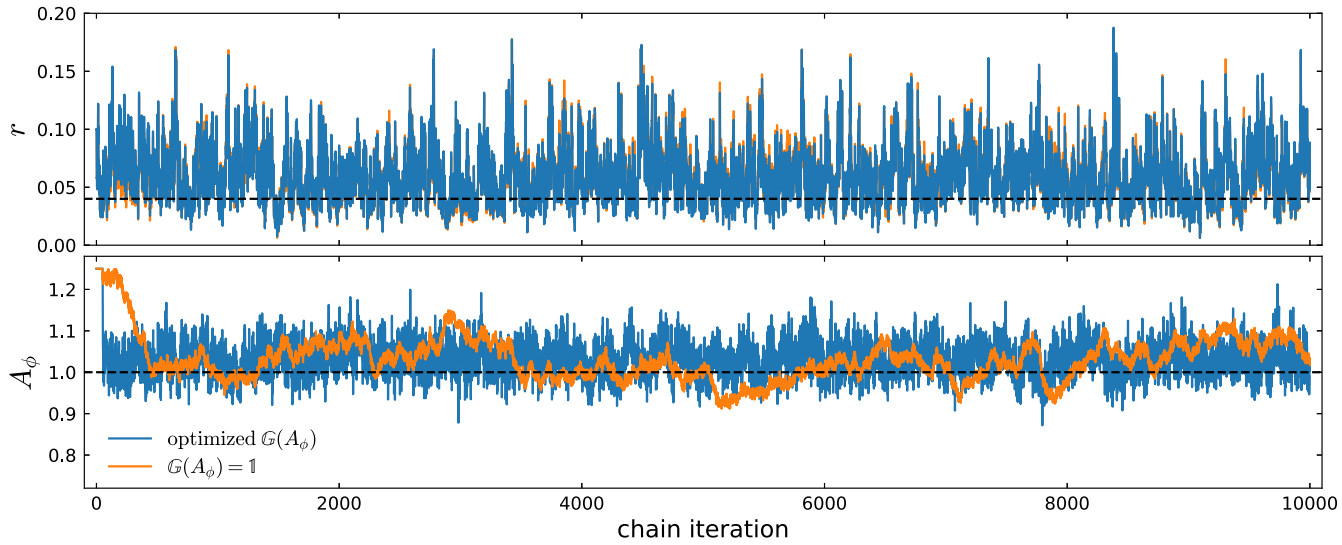

FIG. 3. Samples of r and A_ϕ at each iteration for two chains with the same data and starting random seed, but different choices for parametrizing the posterior (see Sec. III). The blue line corresponds to using our optimized $\mathbb{G}(A_\phi)$ reparametrization, whereas the orange line shows the highly suboptimal choice of $\mathbb{G}(A_\phi) = \mathbb{1}$. No burn in is removed in either case. The simulated data here are generated according to configuration `2PARAM` (see Table II). The run time for a chain of this length is 48 hours on one GPU.

autocorrelation length is ∼25 times larger), highlighting the importance of our reparametrization. We do not show a case where we set the other mixing matrix, $\mathbb{D}(r)$, to the identity matrix; in that case, the impact would be so drastic that it would be impossible to even run a chain at all.

In Fig. 4, we show the posterior distribution for r and A_ϕ computed from these samples, for demonstration plotted using the `getdist` [63] package instead of our Blackwell-Rao estimate. This ability to compute joint constraints on parameters which control both the unlensed CMB fields and lensing potential, with the Bayesian procedure having implicitly performed an optimal lensing reconstruction and delensing, is a unique strength of our procedure and a key result of this work. Note the very small correlation between r and A_ϕ ($\rho = 0.10$); this is evidence that estimates of r are not strongly limited by knowledge of the theoretical lensing spectrum amplitude, or conversely, that lensing reconstruction and hence, delensing efficiency is not strongly limited by the true value of r. This was expected from the intuition that the lensing reconstruction is mostly dominated by small scales whereas r is mainly estimated from large scales, but our procedure allows us to quantify this explicitly.

Next, we describe a set of simulations in the configuration `BIG`. Since we have ascertained that there is little dependence on A_ϕ for r estimation, in these runs, we fix $A_\phi = 1$. We also increase the grid size to 512×512 and the pixel size to 3 arcmin pixels, giving a total sky area of $\sim 650\text{deg}^2$, which is the largest sky area we analyze in this work. We note that although 3 arcmin pixels may seem large compared to ∼1 arcmin typical lensing deflections, LenseFlow is able to lense maps accurately up to scales very

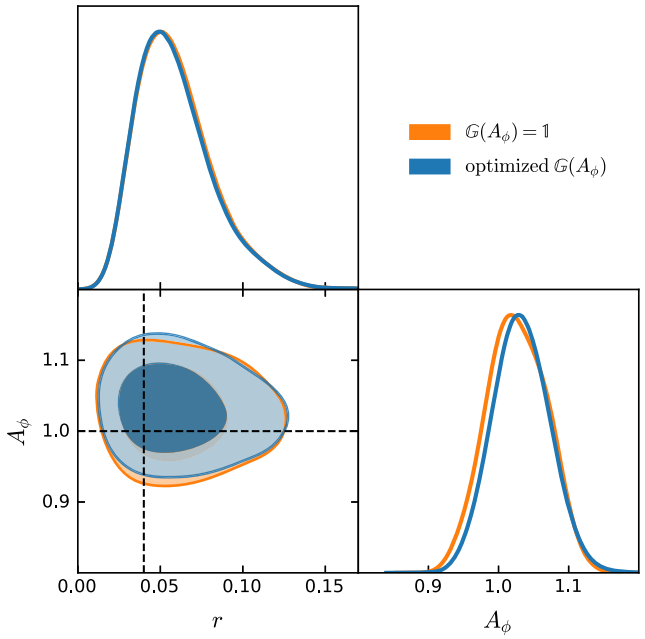

FIG. 4. Posterior distribution for r and A_ϕ from a chain on simulated data in configuration `2PARAM` (see Table II). The samples that comprise this plot are shown in Fig. 3. For demonstration, here we use the `getdist` [63] plotting package rather than our Blackwell-Rao posterior density estimate. The ability to examine joint constraints on these parameters while performing optimal delensing for a realistic data set with masking is a unique strength of our approach. Here, we find these two parameters are highly uncorrelated, providing evidence that A_ϕ can be fixed without impacting r estimation. The orange curve shows a suboptimal choice of the $\mathbb{G}$ matrix, which causes that chain to converge more slowly.

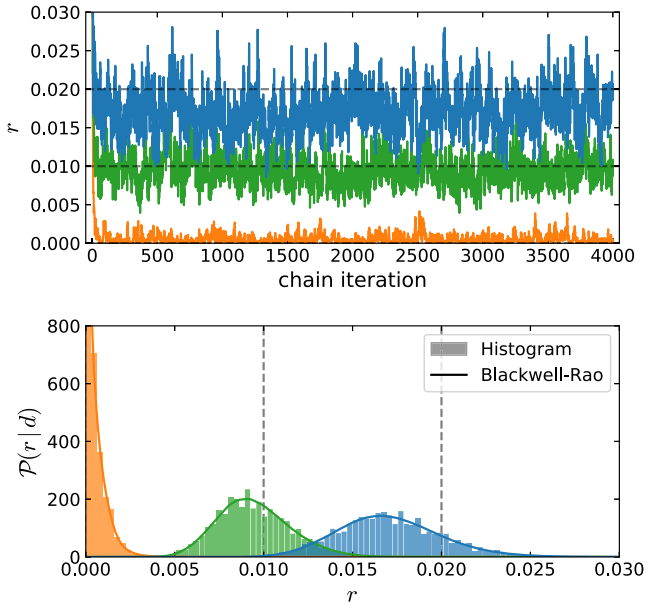

FIG. 5. Top panel: The trace of r samples from chains in configuration BIG (see Table II). Three different fiducial values of r are explored, with the true value given by the black dashed line and each chain in a different color. No burn in is removed. Bottom panel: The same samples binned into histograms, as well Blackwell-Rao estimates of the posterior density, as described in Sec. IV D. These estimates recover very smooth distributions, even in the case where the true r is zero and the constraint is just an upper bound.

close to the Nyquist frequency [38], which here is $\ell = 3400$ and contains nearly all of the available information given our choice of beam for this configuration. For these runs, we use simulated data with three different fiducial values for the tensor-to-scalar ratio, $r = \{0, 0.01, 0.02\}$. Posterior distributions for r are shown in Fig. 5, this time, using the Blackwell-Rao estimate. We can see that each case covers the truth, and that in the $r = 0$ case, the chain samples of r oscillate against zero, as expected.

Of course, the chains contain not just samples of the parameters θ, but also samples of f and ϕ at each iteration. In Fig. 6, we compare the posterior mean of ϕ and some quantities derived from f against the simulation truth for configuration BIG. In the first column, we show the posterior mean reconstructed ϕ, multiplied in Fourier space by the wave number ℓ to make smaller scale structure more easily visible. The posterior mean can be regarded as the "optimal" point estimate of ϕ in the sense that it minimizes the posterior expected squared error against the truth. This estimate is slightly lower variance than the marginal MAP estimate given by [34] [the two differ only due to the non-Gaussianity of $\mathcal{P}(\phi|d)$], although we leave to a future work determining whether there is a meaningful difference. The remaining two columns of Fig. 6 show the posterior mean "E-lensed-into-B" maps (the average over all chain samples of unlensed E and zero B, lensed by ϕ), as well as the posterior mean of the unlensed B map. These latter two quantities are useful data products from the chains, as we will describe in the next section.

B. What can the f and ϕ samples be used for?

Despite the seemingly valuable information contained in the samples of full maps or their associated posterior mean, it is worth asking "what explicitly can these actually be used for?" In terms of a principled statistical analysis for parameter inference within a standard cosmological sky model with Gaussian initial conditions, the answer is actually "not much"; the map samples are just a by-product of the Monte Carlo marginalization, which we used to obtain constraints on the cosmological quantities which we were really after, here r and A_ϕ. Indeed, we cannot readily use the samples of f and ϕ to estimate any other cosmological parameters that were not jointly sampled in the first place.

The real situation is somewhat less pessimistic, however. For example, if we have a physical reason to believe that having jointly sampled extra parameters would not actually impact the lensing reconstruction and delensing, then it may be still be a valid approximation to derive further constraints from the samples. One such case is the search for primordial scalar non-Gaussianity, where constraints on the local type non-Gaussianity become limited at small scales by lensing-induced variance and could be significantly improved by delensing [64]. Although part of the locally non-Gaussian primordial signal would affect the reconstruction, Coulton *et al.* [64] demonstrated this effect is small and quantifiable, meaning our posterior mean delensed maps would be excellent candidates to be used in these searches. Furthermore, our posterior delensed B maps could be used in the search for primordial tensor non-Gaussianity as well [65], with near-optimal results as long as any potential non-Gaussianity is perturbatively small.

The outlook on samples is even better when we consider what can be done in cross-correlation with other probes. Take, for example, the posterior unlensed B map. We could cross-correlate this map with some tracer of foreground B contamination from the Milky Way; if any correlation were detected, it would indicate that whatever foreground cleaning had been performed was insufficient, and we would deduce that our corresponding r samples could be biased. Similarly, the sampled maps, their mean, or even the mean power spectrum of the maps, could be inspected for anything that correlates with an instrumental effect as a way to search for systematics.

From searching for contaminants, it is only a small step to using our posterior samples to check all aspects of the data model (containing the cosmological model, the lensed sky signal, noise, etc.) itself. It is worth recalling the well-known quote by George Box that "all models are wrong but some are useful." [66] This quote applies to CMB data just as much as to any other data set. One way to check if the

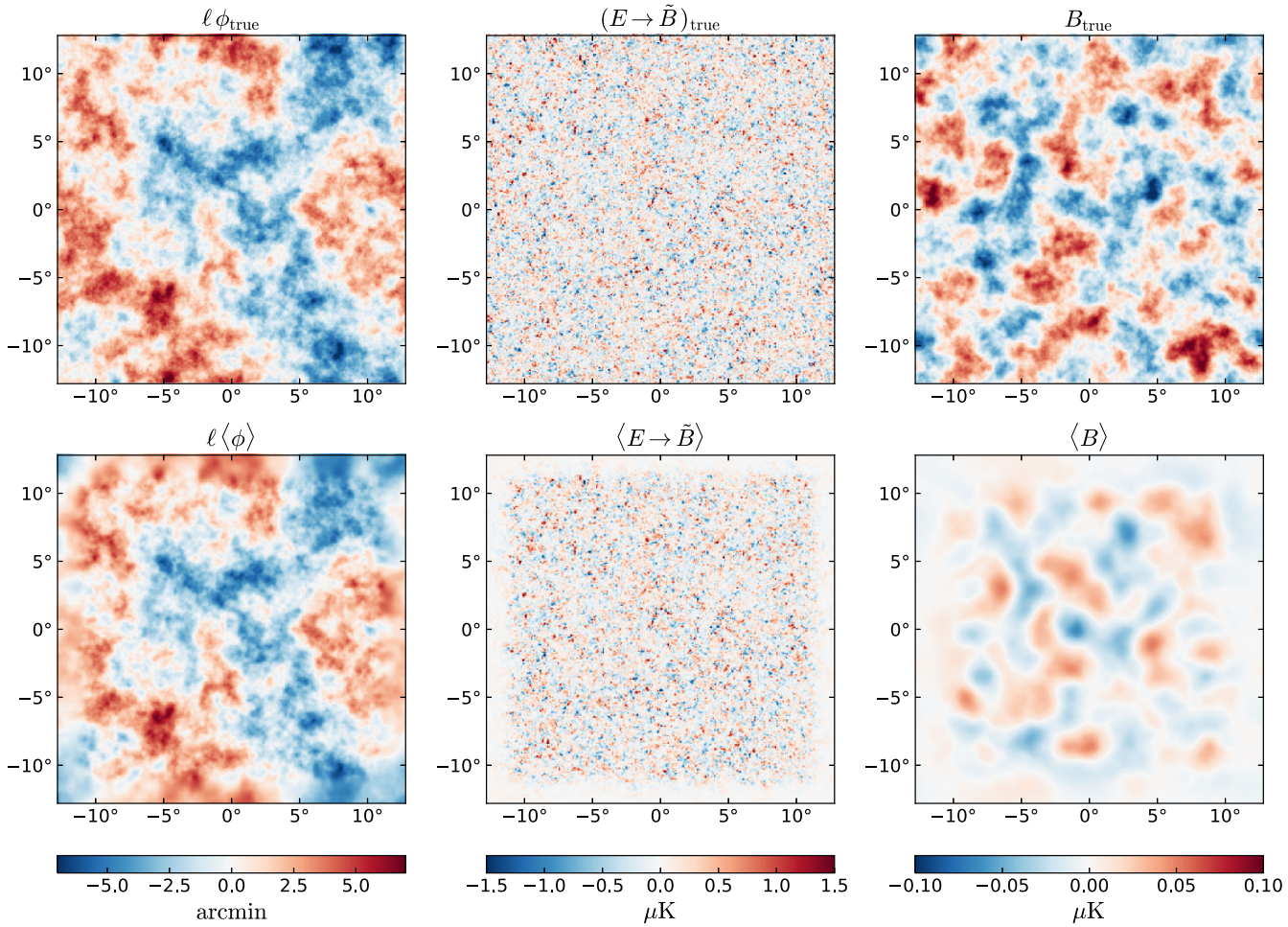

FIG. 6. True input maps (top row) as compared to posterior mean maps (bottom row) computed by averaging over chain samples. This chain uses configuration BIG (see Table II) with a true value of $r = 0.01$. The data for this chain are shown in Fig. 1. The first column shows the ϕ map multiplied in Fourier space by ℓ to visually enhance smaller scales, the middle column shows E modes that have been lensed into B, and the final column shows the reconstructed primordial B modes.

standard model of lensed CMB data is useful is to use it to simulate data starting from the posterior samples and then to check whether this replicated data reproduce the salient features of the actual data. This technique for model evaluation is called "posterior predictive checks" (PPCs) and was introduced in a Bayesian context in [67]; see [68] for a recent application in cosmology. In the literature, PPCs are typically based on the parameters θ; using the samples of the latent fields f, and ϕ would allow defining much more fine-grained PPCs of the model.

The samples can also be used in other more quantitative ways. Consider, for example, a cross-correlation analysis between the CMB and another low-redshift probe of matter fluctuations. One can generally write down the likelihood, $\mathcal{L}(d_{\text{low-z}}|\phi,\theta)$, where $d_{\text{low-z}}$ is the low-redshift data. The full posterior given both data sets is

$$\mathcal{P}(f,\phi,\theta|d,d_{\text{low-z}}) = \mathcal{P}(f,\phi,\theta|d)\mathcal{L}(d_{\text{low-z}}|\phi,\theta). \quad (33)$$

If the low-redshift data are sufficiently less constraining on ϕ than the CMB data, then the importance sampling of the CMB chain is an easy and efficient way of obtaining a Monte Carlo representation of the new posterior for both data sets.

Another analysis which could use the samples would be to split delensing into two steps: 1) obtain E-lensed-into-B samples from small scale CMB data, then 2) use these samples to delense large-scale CMB data and search for nonzero r. Delensing via the samples rather than via a single point estimate of ϕ is a convenient way to propagate the (fully non-Gaussian) delensing uncertainty into the large-scale analysis. A practical reason for doing such a split analysis instead of simply jointly estimating r from the entire CMB data set might be that large-scale foregrounds and systematics are easier to deal with outside of the Bayesian framework.

We leave further development of any of these ideas to future work. Regardless of how these samples may be used,

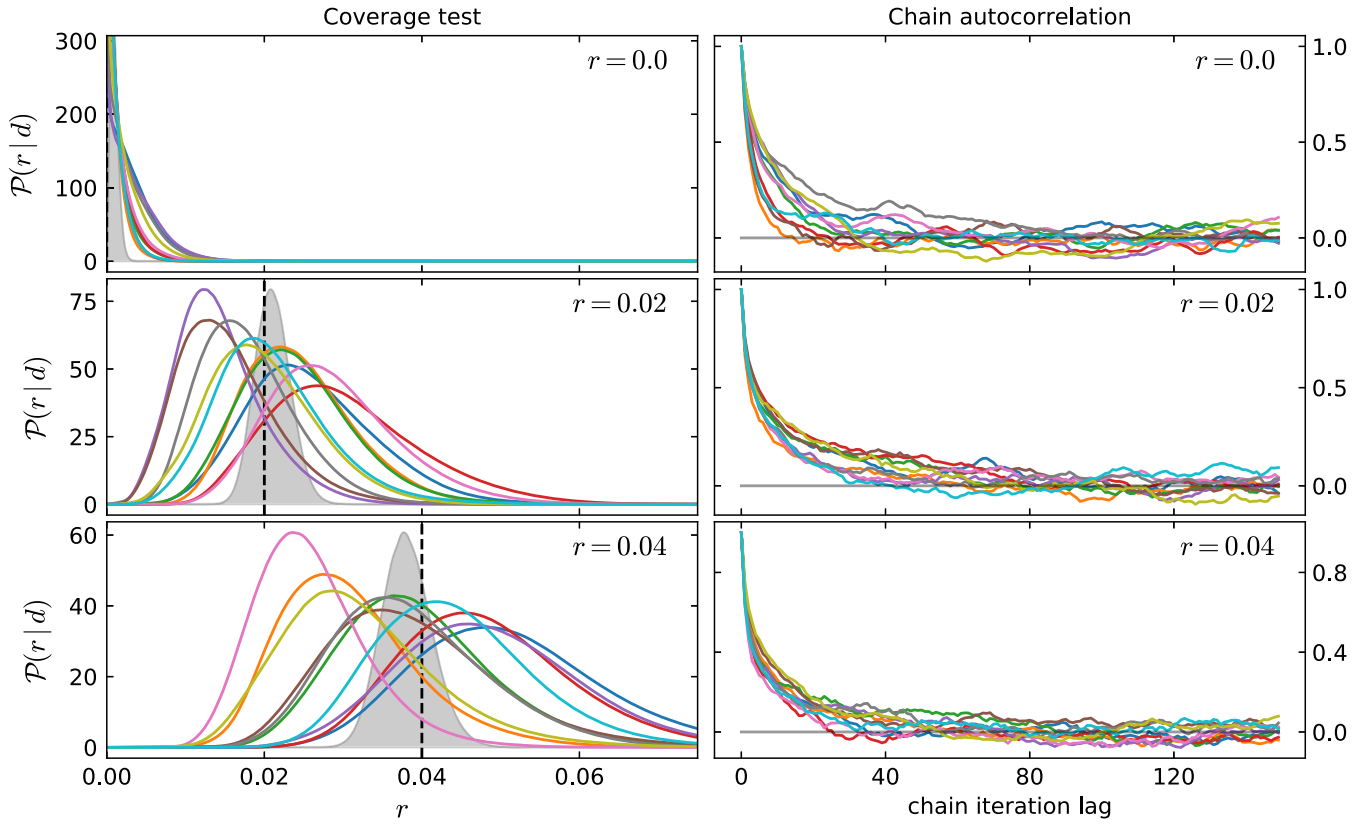

FIG. 7. Left column: Blackwell-Rao posteriors from each of ten chains on different simulated data for three different true values of the tensor-to-scalar ratio indicated in each row. The gray band is the product of the posteriors for each case, with the prior on r importance sampled to be uniform (and with arbitrary normalization constant so as to fit on these axes). We expect that the gray band covers the fiducial value of r to within its own width, as is indeed the case. This is a test of the coverage of our $\mathcal{P}(r|d)$ posteriors and hence, a test of the correctness of our procedure. Right column: The chain autocorrelation function for each of the chains in the left panel. The integrated autocorrelation time for these chains ranges from 5–33.

the key point is that they are a useful way to capture the entire information content in the CMB data that generate them, and they fully represent the uncertainty in the reconstruction due to noise, modeled systematics, and incomplete knowledge of the cosmological parameters, which were free parameters in the posterior.

C. Convergence diagnostics

Having described some of the results from the chains, we now turn to more quantitatively assessing chain convergence. We begin using a final set of chains with data simulated from configuration MANY. These chains only sample r and have been reduced to 256×256 pixels; however, we run ten chains on different simulated data for each of three fiducial values, $r = \{0, 0.02, 0.04\}$.

The posteriors from each of these chains are shown in Fig. 7. It is worth noting the scatter in the mean and width of the different data realizations (here, σ_r can vary by almost a factor of 2) as a reminder that any one experiment can be lucky or unlucky depending on the particular patch of sky observed. It would be interesting to determine how much of the contribution to this scatter comes from the non-Gaussian uncertainty in the lensing reconstruction as opposed to Gaussian sample variance, although that is beyond our scope here.

One way to check the correctness and convergence of these chains is to multiply the ten posteriors together. We expect that the resulting distribution should tighten around the true of r, with scatter such that roughly ∼68% of the time the truth will be covered by the 1σ contours. This is indeed what we see in Fig. 7 for all values of r. Formally, with only ten chains, we can only check for the presence of biases in our posteriors at the $\sigma/\sqrt{10} \approx 30\%\,\sigma$ level; however in the absence of a coding error, there is no reason to believe these contours would not continue to shrink further around the truth.

Another way to check the convergence of our chains is by computing the integrated autocorrelation time and the accompanying effective sample size [69]. The right-hand panel of Fig. 7 shows the autocorrelation function for the r samples from each of these chains. In all cases, it takes about ∼40 iterations of our sampler before the autocorrelation drops to near-zero, and we obtain an independent sample. More exactly, the integrated autocorrelation time is

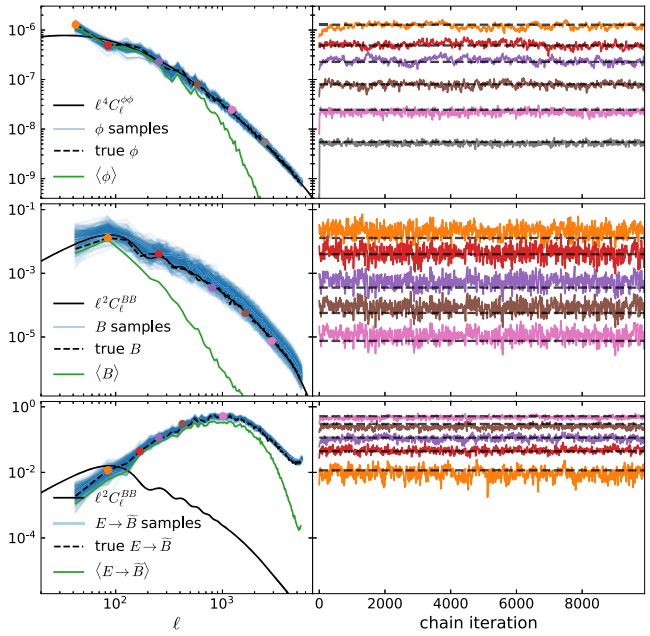

FIG. 8. Left column: In blue, we overlay the power spectra of chain samples of ϕ and of two quantities derived from f. The black dashed line gives the power spectrum of the truth, and the green line is the power spectrum of the posterior mean map. The three rows correspond to ϕ, unlensed B, and E-lensed-into-B. The posterior mean maps exhibit Wiener-filter like suppression, as expected, while the samples scatter around the true spectrum and quantify uncertainty. Right column: The same power spectra that are overlayed on the left but picking some specific multipoles and plotting the trace of their value throughout the chain. Visually one can see the acceptable correlation length of the power spectrum samples, as well as that they cluster around the input truth, confirmation that the likelihood is dominating over the prior that would otherwise pull these quantities towards zero. This is the same chain in configuration `2PARAM` (see Table II) used in Figs. 3 and 4.

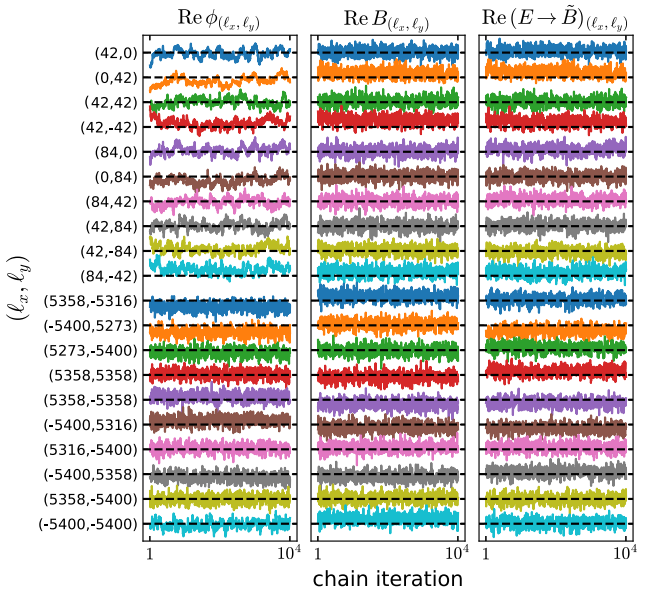

FIG. 9. Chain samples of the real part of the ten largest-scale and ten smallest-scale Fourier modes of the posterior ϕ, B, and E-lensed-into-B maps. Each set of samples is normalized to unit variance, but the relative distance to the truth (shown in the black dashed line) is preserved. This chain uses configuration `2PARAM` (see Table II). Visually, we achieve great convergence even at the individual mode level. Out of the $\sim$200,000 modes which are sampled, the only exceptions are perhaps the two largest scale ϕ modes, which could benefit from a slightly longer chain. However, these two modes are not informative for A_ϕ, which remains very well converged (Fig. 3).

in the range of 5–33, corresponding to an effective sample size of 150–1000 given the 5000 total iterations in each chain (autocorrelation lengths for all configurations are listed in Table II). In turn, this means we should expect a Monte Carlo error on the posterior mean of r on the order of 3%–10% of σ_r.

This is consistent with another estimate of the error which we can get by splitting our chains into multiple pieces or running multiple chains, and computing the mean from each. We have performed this test for the chain in configuration `2PARAM` by splitting the 10000 samples into two halves and checking the difference in the resulting posterior mean for both A_ϕ and for r. We find that the mean agrees to within 5% of σ_{A_ϕ} and 8% of σ_r, respectively.

The posterior distribution of any quantity derived from (f,ϕ,θ) can be explored by postprocessing the Monte Carlo chain, and its convergence can be tested. Bandpowers are one such quantity, and these have a very direct relation to the convergence of r and A_ϕ. In particular, only the bandpowers of f and ϕ enter the $\mathcal{P}(r,A_\phi|f,\phi,d)$ conditional distribution. In Fig. 8, we show the trace of various bandpowers of ϕ, B, and E-lensed-into-B. Visually, we see these samples are still consistent with being drawn from a stationary distribution.

Delving deeper into the $\sim$200,000 parameters, which are sampled in this configuration, we plot in Fig. 9 the trace of the real part of individual Fourier modes of ϕ, B, and E-lensed-into-B. The choice of plotting the real part is arbitrary as it has identical statistical properties to the imaginary part under the assumption of isotropy (nevertheless, we have checked that the imaginary part does behave similarly). Even here, we mostly see very good convergence of the samples. For an internal CMB analysis, the convergence of these individual modes is not particularly important, since, as previously stated, what really matters is the convergence of the θ parameters. However, for a cross-correlation analysis such as the ones described in the previous subsection, the individual modes (and hence, the full maps themselves) must be adequately converged. Figure 9 is evidence that this is indeed the case.

We do note that ϕ modes at the largest scales converge slightly slower than others, as can be seen in Figs. 8 and 9. We believe this is related to the mean field, which also arises in both quadratic or MAP estimation [34]. At these

large scales where the mean field is very big, frequentist analyses require a large number of Monte Carlo simulations to estimate the mean field precisely enough so that the error on the mean-field determination is subdominant to sample variance. In our Bayesian analysis, this challenge is not solved "for free", rather it manifests as a need for longer chains to overcome the larger correlation length at these same scales. Evidence that this is the case comes from the fact that removing the mask and hence, reducing the mean-field yields more rapid relative convergence at these large scales. We do stress, however, that because the majority of information on A_ϕ is not sourced by these handful of largest scale modes, their slower convergence does not significantly impact the very good convergence of A_ϕ that we see in Fig. 3.

The results in this section demonstrate that the θ, the bandpowers, and even individual Fourier modes are well converged in these chains. However, it is not implausible that one could find pathological combinations of parameters for which this is not the case. We caution users of these chains to first verify convergence of arbitrary derived quantities that they may need. This can be done using tests similar to the ones described in this section.

D. Fisher information on r and S4 forecasting

The chains give us the ability to check existing forecasts for, e.g., CMB-S4, South Pole Observatory, or Simons Observatory to a precision which has not been possible before. We will refer to these as CMB-S4-like forecasts since the methodology we are testing is the same between all of them. The approach is to use chains on simulated data to compute exactly (up to Monte Carlo errors) the Fisher information on r contained in lensed CMB data. This can be done even in the presence of real instrumental complexities such as the pixel masking we apply here. We will use chains in configuration `MANY` for this test. Although this is a smaller patch of sky than the planned CMB-S4 observations, the noise levels are similar, and this lets us validate the forecasting procedure itself.

To begin, consider the Fisher information,

$$\mathcal{F}_{rr}(r_{\rm fid}) = -\left\langle \frac{d^2}{dr^2} \log \mathcal{L}(d|r)\Big|_{r_{\rm fid}} \right\rangle_{d\sim\mathcal{L}(d|r_{\rm fid})}. \quad (34)$$

It is an average over data, d, of the Hessian of the log-likelihood of r for each of these data, evaluated at $r = r_{\rm fid}$, and where the data are themselves simulated given $r = r_{\rm fid}$. If we run our chains with a uniform prior on r (or importance sample it to be uniform after the fact), then we have $\mathcal{L}(d|r) = \mathcal{P}(r|d)$. Thus, we can take the log of the posterior $\mathcal{P}(r|d)$ estimated from the chains, numerically compute the second derivative, and explicitly perform the average in Eq. (34) over several chains with different simulated data. Alternatively, we can swap the order of the derivative and expectation value in Eq. (34) and take the geometric mean of the chain posteriors first. The second derivative of the log of this function at $r_{\rm fid}$ is then again the Fisher information, but instead of looking just at one value, we can simply plot the entire function. Loosely speaking, this maps out something like the "typical posterior" that one might expect given possible data, which is also a useful forecasting quantity, particularly for $r_{\rm fid} = 0$, where Monte Carlo noise prevents us from computing a stable numerical derivative. For configuration `MANY`, these functions, as well as Gaussians with standard deviations given by $1/\sqrt{\mathcal{F}_{rr}}$ are shown in Fig. 10.

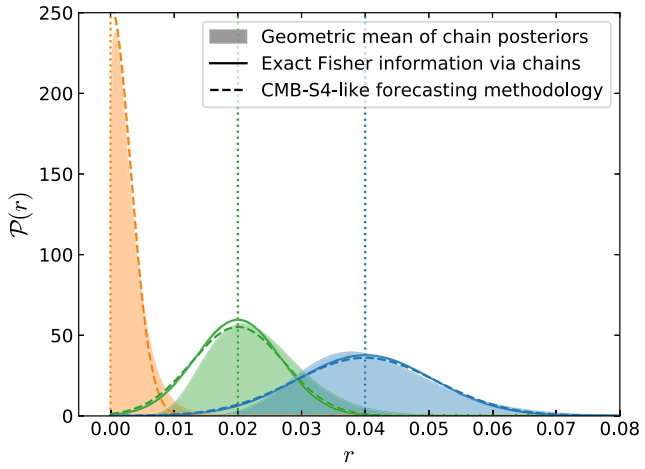

FIG. 10. A comparison of different methods for forecasting constraints on r, assuming configuration `MANY` (see Table II). Three different possible true values of r are explored, indicated by vertical dotted lines. The dashed lines show expected Gaussian constraints forecasted with a method very similar to that used for CMB-S4. In solid lines, we show Gaussian distributions with standard deviation computed from the Fisher information on r. This work is the first time the Fisher information on r from lensed CMB data has been calculated without approximation. In filled contours, we show the geometric mean of the posteriors from several chains. The excellent agreement between all of these is an important validation of the CMB-S4 r forecasting methodology even in the presence of instrumental effects and masking, as is considered here.

We would like to compare against CMB-S4-like forecasts. These types of forecasts are broken up into two steps: 1) first, a postdelensing residual lensed B power is computed, then 2) this is treated as Gaussian noise in a second step to estimate r. For the forecasts in [39], the first step has been based on the method given in Smith *et al.* [27]. This method follows the heuristic idea that to perform optimal delensing, one iterates computing the EB quadratic estimate for ϕ, delenses the data by this ϕ, then recomputes the ϕ estimate, which should now be lower variance because part of the contribution to the error of this estimate, namely the lensed B modes, have been reduced. We note that this computation works only to first order in ϕ, ignores ℓ-to-ℓ correlations and non-Gaussianities in both the ϕ

noise and the residual lensed B modes, and ignores pixel masking. So that information is not double counted, only modes at $\ell \gtrsim 150$ are used in step (1), and only modes at $\ell \lesssim 150$ are used in step (2). Although conceptually the procedure is very reasonable, Smith *et al.* [27] do not explicitly check these simplifications, but rather validate the entire approximation by comparing their residual lensed B amplitude against a more exact computation given for several configurations in Table I of Seljak and Hirata [26] and finding agreement at the ≈10% level. The numbers computed in Seljak and Hirata [26] in turn come from computing an approximate marginal MAP estimate of ϕ and using this for delensing, with error bars on the delensed B power computed via Monte Carlo. Carron and Lewis [34] further sharpen up this result by performing the same test with their exact maximization procedure rather than an approximate one, finding good agreement. Once the residual lensed B mode power spectrum is computed, the residual modes are approximated as isotropic and Gaussian, and a traditional power spectrum Fisher forecast is computed for r [27], or a more sophisticated simulated power spectrum analysis is performed [39].

Our procedure allows us to validate the CMB-S4 forecasting procedure in a much more direct and straightforward way than the long chain of validation steps above, by simply comparing against the Fisher information on r that we derive. This also tests a few remaining assumptions in the CMB-S4-like method, mainly that the residual B modes are Gaussian, that minimal information is lost by the $\ell \lesssim 150$ filter, and that the impact of masking is only a reduction in the number of modes which can be captured by an $f_{\rm sky}$ factor. This latter assumption has never been checked but is particularly worrisome, since masking couples modes across ℓ and will leak E into B, mimicking lensing.

For configuration MANY, we have computed forecasts using the CMB-S4-like procedure described above, accounting for all experimental details listed in Table II, except for the mask, which is instead treated with an $f_{\rm sky}$ factor. Our results are summarized in Fig. 10. One can see the excellent visual agreement between the results from our chains and those from the CMB-S4-like forecast for all values of $r_{\rm fid}$ tested. For $r_{\rm fid} = [0.02, 0.04]$, where we can compute accurate numerical derivatives, our exact Fisher calculation gives $\sigma_r = 1/\sqrt{\mathcal{F}_{rr}} = [0.0067, 0.0106]$ as compared to the CMB-S4-like forecasts which give $\sigma_r = 1/\sqrt{\mathcal{F}_{rr}} = [0.0072, 0.0111]$, or a difference of only 4% and 8%, respectively. This excellent agreement is further proof of the fidelity of existing r forecasts for CMB-S4 [39] and of other current and future forecasts using this same method. We note, though, that this does not necessarily imply that implementing a real analysis pipeline following the heuristic CMB-S4-like treatment would yield an unbiased estimate of r, only that this gives very accurate error bars as a forecasting procedure. Our chains, however, could be used to check this in the future.

VI. CONCLUDING REMARKS

A. The CMBLensing.jl package

Throughout the development of our sampling algorithm, we have used two branches of code in parallel. The first code was initially used to produce the chains presented in the previous sections. The second code, CMBLensing.jl, was developed for wider-spread use and is now faster and is what we recommend for anyone wishing to use, reproduce, or extend our results. The two have been checked for agreement.

The design of CMBLensing.jl was motivated by the desire for: 1) the ability to transparently run the code on CPUs or GPUs, 2) access to automatic differentiation so that gradients of our posterior or of any future modifications do not need to be hand coded, and 3) no sacrifice on performance. To our knowledge, only two truly practical avenues exist to achieve this: either describing the posterior as a neural networklike graph in a machine learning library such as TensorFlow or writing our code in JULIA [70]. We have chosen the latter as it allows writing normal high-level code and avoids the additional complexity involved in translating our algorithm into the language of computational graphs.

As a simple example of the ease of this approach, consider the first order Taylor series expansion for lensing, i.e., $f(x + \nabla\phi) \approx f + \nabla\phi \cdot \nabla f$. This can be written succinctly and true to the underlying mathematical expression in CMBLensing.jl as

```
lense(f,ϕ) = f + Diagonal.(Map(∇*ϕ)) · (∇*f)
```

and the resulting function is no slower than having written out the necessary FFTs and array multiplications by hand. The arguments of this function are CMBLensing.jl field objects, which are just thin wrappers around arrays storing the maps or Fourier coefficients for the fields. Depending on a user setting, these arrays can reside on CPU or NVIDIA GPU, and the above code works transparently in either case. JULIA GPU integration is such that only 30 lines of GPU-specific code are needed in the entire codebase. Figure 11 summarizes the timing for each step in our Gibbs sampler and compares the CPU and GPU performance. We reach improvements in performance of factors of several when running on GPUs,[4] and, encouragingly, the relative improvement grows as we go to larger maps. Additionally, the GPU code is not particularly optimized yet so we expect room for significant improvement, despite it already outperforming the highly optimized CPU code.

Once a function like `lense` is defined, source-to-source reverse-mode automatic differentiation can be used to

[4]There is a large dependence on GPU hardware; for example, our experience is that laptop-grade GPUs offer little to no improvement, in contrast to the more performant GPU used in Fig. 11.

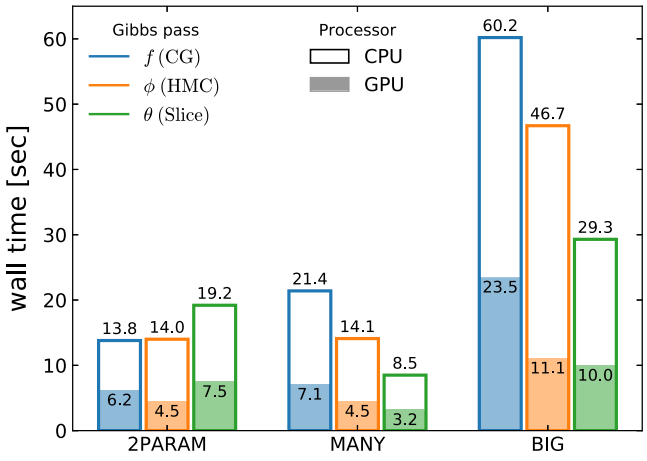

FIG. 11. The wall time in seconds for each Gibbs pass, for each configuration (see Table II), and for running on CPU vs GPU. The CPU benchmarks utilize a full NERSC Cori Haswell node (Intel Xeon Processor E5-2698 v3), and the GPU benchmarks a single NVIDIA GTX 1080Ti GPU. Although our CPU code is highly optimized, our GPU code likely has room for significant improvement, despite already being faster.

compute gradients for most functions on $\mathbb{R}^n \to \mathbb{R}^1$ m, which use `lense` anywhere within their evaluation [71]. Here is a very simple example which takes a gradient with respect to ϕ, evaluated at $\phi = 0$,

```
gradient(ϕ -> norm(lense(f,ϕ)), 0ϕ)
```

Both above code snippets are unmodified from what could be run in a real JULIA session. The flexibility afforded by this system is invaluable to the type of quick exploration which was necessary in arriving at the results in this paper and which will be necessary for applying these methods to increasingly complex data sets moving forward. This package should serve as a useful tool for the CMB lensing community in the future, or as a "black-box" target function (and gradient) for a broader audience wishing to try other inference methods on the CMB lensing problem.

B. Brief summary of main results

In this work, we have developed a method for joint inference of cosmological parameters, unlensed CMB fields, and the gravitational lensing potential, from CMB temperature and polarization data. By working with the Bayesian posterior, we are guaranteed to have extracted all available information from the data; hence, our procedure is "optimal" in some sense and (very) loosely corresponds to what is sometimes referred to as "iterative delensing." Although several methods exist which can produce point estimates of the lensing potential which are lower-variance than the current state-of-the-art quadratic estimate (see Sec. I), our method is unique in making it completely straightforward how to actually extract cosmological information including uncertainty estimates from the lensing potential or from the delensed fields. Specifically, any methods based on a power spectrum point estimate need to quantify cosmology-depend biases and covariances, something which has yet to be demonstrated is feasible in general for lensing except in the quadratic estimate case. Conversely, the Bayesian approach fully extracts all lensing information while implicitly and without approximation handling the impact of such biases and covariances.

We have demonstrated this ability by jointly estimating r and A_ϕ from simulated data. The analysis hinges on three key pieces, and without any one of them our results would not be possible. These are 1) reparametrizing the posterior to a new set of variables whose posterior distribution is more Gaussian and less degenerate 2) tuning our Monte Carlo sampler, in particular making use of HMC to sample the very high dimensional posterior which remains mildly non-Gaussian even after reparametrizing and 3) numerically implementing the lensing operation with LenseFlow, which gives us the needed gradients through the inverse lensing operation and allows to us to avoid an otherwise prohibitive determinant calculation.

We have used this method to arrive at two useful scientific results. First, we have explicitly demonstrated that the correlation between r and A_ϕ is small ($\rho = 0.10$), showing that r inference is not strongly limited by knowledge of the true lensing power spectrum amplitude. Second, we have given the first-ever exact computation of the Fisher information on r in the context of delensing, even including several real instrumental effects, notably pixel masking. Using this, we have validated the r forecasting procedure used for experiments such as CMB-S4, which has never been checked in the presence of pixel masking. Encouragingly, we find that the standard procedure yields results very close (within 8% in terms of the uncertainty on r) to our exact Fisher calculation, giving further evidence that CMB-S4 delensing will work as expected.

C. Future work and new possibilities

The algorithm presented in this work is ready to be applied to current generation CMB data targeting deep observations over patches of sky of several hundreds of square degrees. There is ongoing work to apply these methods to South Pole Telescope data, and, as mentioned previously, they could also be applied to POLARBEAR data, where it would be expected that the delensing effeciency achieved in Adachi *et al.* [35] could be even further improved.

The Bayesian sampling solution still has some challenges that need to be overcome before analyzing a data set of the complexity expected from CMB-S4. One main future challenge is simply scaling up the number of pixels and moving beyond the flat-sky approximation to deal with sky curvature. Conceptually, it is completely straightforward to include sky curvature in our method. In terms of performance, the chains presented here run in 24–48 hours

on one GPU, and scaling up to nearly full-sky observations will likely require improving this by a few factors of 10. Part of this can be trivially gained by running more chains in parallel, which we have not done here but should work well given that we do not find very significant chain burn-in time is necessary. It seems very possible that the remaining improvements could come from some combination of optimizing the GPU code, discovering even better reparametrizations, accelerating Wiener filtering, and going beyond the very basic HMC sampling algorithm we have used.

Another challenge which must be tackled is the inclusion of foregrounds. A simple solution which may work well is simply to run our procedure on component separated maps. A more ambitious approach would be to compute a full forward model for the foregrounds and jointly infer them. This sounds difficult, but at least in the medium to small scale regime in polarization (which will be almost solely responsible for lensing reconstruction in the future), expected foregrounds are surprisingly small and simple. The only component expected to be significantly present is shot noise from radio galaxies [72], which may be quite simple to forward model. We note that forward modeling the foregrounds may put an even bigger requirement on us to work with the joint posterior, because the analytic marginalization in Eq. (12) is likely impossible in the presence of other non-Gaussian components.

One interesting extension to this work is to infer other cosmological parameters besides r or A_ϕ, or even the theoretical spectra themselves. In our work, the shape of all theory spectra has been assumed perfectly known, only the amplitudes r or A_ϕ are uncertain quantities to be inferred. Given this model, we gave the first explicit confirmation that the estimate of r is largely uncorrelated with A_ϕ, but it would be interesting to also confirm that uncertainties in the exact shape of the relevant theory spectra do not impact r inference (something which we would expect to be the case since the leading order effect is simply the total B-mode foreground power generated by lensing but which has never been shown explicitly). For example, it would be straightforward to swap A_ϕ for something like the sum of neutrino masses, Σm_ν, which affects the shape of the lensing potential. We expect the reparametrizations discussed in this work to be sufficiently general to handle this case as well without modification.

Finally, we note that sampling is not the unique way to explore a Bayesian posterior, and many other methods exist which could potentially be accurate enough while being cheaper computationally. Some examples (but by no means an exhaustive list) include "variational inference" methods [73–75], Laplace or higher-order approximations [76], or fall under the category of "likelihood-free inference." [77,78] Many or all of these methods, however, rely on approximations that are extremely difficult to check in the context of the very high dimensional and non-Gaussian CMB lensing problem. By having explored and built intuition about the lensing posterior, and by having developed a sampling method which can be used to compute an approximation-free answer for a realistic-sized and nontrivial data model, these other methods can, for the first time, be explicitly validated for lensing. If they prove to be sufficiently accurate, then perhaps they offer an advantageous way to perform this analysis in the future.

ACKNOWLEDGMENTS

E. A. acknowledges support from NSF Grants No. DMS-1252795, No. DMS-1812199, and a CARMIN research fellowship at Institut des Hautes Études Scientifiques and IHP. B. D. W. acknowledges support from the BIG4 project, Grant No. ANR-16-CE23-0002 of the French Agence Nationale de la Recherche (ANR). The Center for Computational Astrophysics is supported by the Simons Foundation. M. M. and B. D. W. acknowledge the France-Berkeley Fund for support. MM thanks Uros Seljak and Bill Holzapfel for useful discussions.

[1] B. A. Benson *et al.*, Proc. SPIE Int. Soc. Opt. Eng. **9153**, 91531P (2014).
[2] A. J. Anderson *et al.*, J. Low Temp. Phys. **193**, 1057 (2018).
[3] S. W. Henderson *et al.*, J. Low Temp. Phys. **184**, 772 (2016).
[4] A. Suzuki *et al.*, J. Low Temp. Phys. **184**, 805 (2016).
[5] P. Ade *et al.* (T. S. O. Collaboration), J. Cosmol. Astropart. Phys. 02 (2019) 056.
[6] M. H. Abitbol *et al.*, arXiv:1706.02464.
[7] S. Hanany *et al.*, arXiv:1902.10541.
[8] K. Abazajian *et al.*, Bull. Am. Astron. Soc. **51**, 209 (2019), https://arxiv.org/abs/1908.01062.
[9] M. Zaldarriaga and U. Seljak, Phys. Rev. D **59**, 123507 (1999).
[10] W. Hu and T. Okamoto, Astrophys. J. **574**, 566 (2002).
[11] K. M. Smith, O. Zahn, and O. Dore, Phys. Rev. D **76**, 043510 (2007).
[12] S. Das *et al.*, Phys. Rev. Lett. **107**, 021301 (2011).
[13] D. Hanson *et al.*, Phys. Rev. Lett. **111**, 141301 (2013).
[14] Planck Collaboration XVII, Astron. Astrophys. **571**, A17 (2014).
[15] Planck Collaboration XVIII, Astron. Astrophys. **571**, A18 (2014).

[16] Planck Collaboration XV, Astron. Astrophys. **594**, A15 (2016).
[17] Planck Collaboration Int. XLI, Astron. Astrophys. **596**, A102 (2016).
[18] N. Aghanim *et al.* (Planck Collaboration VIII), Astron. Astrophys. **641**, A8 (2020).
[19] A. van Engelen *et al.*, Astrophys. J. **756**, 142 (2012).
[20] P. A. R. Ade *et al.* (Polarbear Collaboration), Phys. Rev. Lett. **113**, 021301 (2014).
[21] K. T. Story *et al.*, Astrophys. J. **810**, 50 (2015).
[22] P. A. R. Ade *et al.* (BICEP2 and Keck Array Collaborations), Astrophys. J. **833**, 228 (2016).
[23] Y. Omori *et al.*, Astrophys. J. **849**, 124 (2017).
[24] W. L. K. Wu *et al.*, Astrophys. J. **884**, 70 (2019).
[25] J. Carron, A. Lewis, and A. Challinor, J. Cosmol. Astropart. Phys. 05 (2017) 035.
[26] U. Seljak and C. M. Hirata, Phys. Rev. D **69**, 043005 (2004).
[27] K. M. Smith, D. Hanson, M. LoVerde, C. M. Hirata, and O. Zahn, J. Cosmol. Astropart. Phys. 06 (2012) 014.
[28] B. Horowitz, S. Ferraro, and B. D. Sherwin, Mon. Not. R. Astron. Soc. **485**, 3919 (2019).
[29] M. Mirmelstein, J. Carron, and A. Lewis, Phys. Rev. D **100**, 123509 (2019).
[30] B. Hadzhiyska, B. D. Sherwin, M. Madhavacheril, and S. Ferraro, Phys. Rev. D **100**, 023547 (2019).
[31] J. Caldeira, W. L. K. Wu, B. Nord, C. Avestruz, S. Trivedi, and K. T. Story, Astron. Comput. **28**, 100307 (2019).
[32] C. M. Hirata and U. Seljak, Phys. Rev. D **68**, 083002 (2003).
[33] C. M. Hirata and U. Seljak, Phys. Rev. D **67**, 043001 (2003).
[34] J. Carron and A. Lewis, Phys. Rev. D **96**, 063510 (2017).
[35] S. Adachi *et al.*, Phys. Rev. Lett. **124**, 131301 (2020).
[36] J. Carron, Phys. Rev. D **99**, 043518 (2019).
[37] M. Betancourt, arXiv:1701.02434.
[38] M. Millea, E. Anderes, and B. D. Wandelt, Phys. Rev. D **100**, 023509 (2019).
[39] K. N. Abazajian *et al.*, arXiv:1610.02743.
[40] South Pole Observatory Collaboration (to be published).
[41] https://github.com/marius311/CMBLensing.jl
[42] D. Beck, G. Fabbian, and J. Errard, Phys. Rev. D **98**, 043512 (2018).
[43] V. Böhm, B. D. Sherwin, J. Liu, J. C. Hill, M. Schmittfull, and T. Namikawa, Phys. Rev. D **98**, 123510 (2018).
[44] A. Lewis, A. Hall, and A. Challinor, J. Cosmol. Astropart. Phys. 08 (2017) 023.
[45] G. Pratten and A. Lewis, J. Cosmol. Astropart. Phys. 08 (2016) 047.
[46] http://camb.info
[47] A. Gelman, Bayesian Anal. **1**, 515 (2006).
[48] B. D. Wandelt, D. L. Larson, and A. Lakshminarayanan, Phys. Rev. D **70**, 083511 (2004).
[49] J. Jasche and B. D. Wandelt, Mon. Not. R. Astron. Soc. **432**, 894 (2013).
[50] G. Lavaux, J. Jasche, and F. Leclercq, arXiv:1909.06396.
[51] D. K. Ramanah, G. Lavaux, J. Jasche, and B. D. Wandelt, Astron. Astrophys. **621**, A69 (2019).
[52] E. Anderes, B. D. Wandelt, and G. Lavaux, Astrophys. J. **808**, 152 (2015).
[53] J. B. Jewell, H. K. Eriksen, B. D. Wandelt, I. J. O'Dwyer, G. Huey, and K. M. Górski, Astrophys. J. **697**, 258 (2009).
[54] B. Racine, J. B. Jewell, H. K. Eriksen, and I. K. Wehus, Astrophys. J. **820**, 31 (2016).
[55] M. Münchmeyer and K. M. Smith, arXiv:1905.05846.
[56] F. Elsner and B. D. Wandelt, Astron. Astrophys. **549**, A111 (2013).
[57] D. Kodi Ramanah, G. Lavaux, and B. D. Wandelt, Mon. Not. R. Astron. Soc. **490**, 947 (2019).
[58] W. Kahan, Commun. ACM **8**, 40 (1965).
[59] R. M. Neal, arXiv:bayes-an/9506004.
[60] R. M. Neal, Ann. Stat. **31**, 705 (2003).
[61] R. M. Neal, arXiv:bayes.
[62] D. Barron, Y. Chinone, A. Kusaka, J. Borril, J. Errard, S. Feeney, S. Ferraro, R. Keskitalo, A. T. Lee, N. A. Roe, B. D. Sherwin, and A. Suzuki, J. Cosmol. Astropart. Phys. 02 (2018) 009.
[63] A. Lewis, arXiv:1910.13970.
[64] W. R. Coulton, P. D. Meerburg, D. G. Baker, S. Hotinli, A. J. Duivenvoorden, and A. van Engelen, Phys. Rev. D **101**, 123504 (2020).
[65] P. D. Meerburg, J. Meyers, A. van Engelen, and Y. Ali-Haïmoud, Phys. Rev. D **93**, 123511 (2016).
[66] G. Box, in *Robustness in Statistics*, edited by R. L. Launer and G. N. Wilkinson (Academic Press, 1979), pp. 201–236.
[67] D. B. Rubin, Ann. Stat. **12**, 1151 (1984).
[68] S. M. Feeney, H. V. Peiris, A. R. Williamson, S. M. Nissanke, D. J. Mortlock, J. Alsing, and D. Scolnic, Phys. Rev. Lett. **122**, 061105 (2019).
[69] J. Goodman and J. Weare, Commun. Appl. Math. Comput. Sci. **5**, 65 (2010).
[70] J. Bezanson, A. Edelman, S. Karpinski, and V. Shah, SIAM Rev. **59**, 65 (2017).
[71] M. Innes, arXiv:1810.07951.
[72] A. T. Crites *et al.*, Astrophys. J. **805**, 36 (2015).
[73] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, J. Am. Stat. Assoc. **112**, 859 (2017).
[74] U. Seljak and B. Yu, arXiv:1901.04454.
[75] J. Knollmüller and T. A. Enßlin, arXiv:1901.11033.
[76] U. Seljak, G. Aslanyan, Y. Feng, and C. Modi, J. Cosmol. Astropart. Phys. 12 (2017) 009.
[77] J.-M. Marin, P. Pudlo, C. P. Robert, and R. Ryder, arXiv:1101.0955.
[78] L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott, J. Comput. Graph. Stat. **27**, 1 (2018).