Title
Marginal unbiased score expansion and application to CMB lensing

Journal
Physical Review D, 105(10)

ISSN
2470-0010

Authors
Millea, Marius
Seljak, Uroš

Publication Date
2022-05-15

DOI
10.1103/physrevd.105.103531

Peer reviewed

MUSE: Marginal Unbiased Score Expansion and Application to CMB Lensing

Marius Millea*
Department of Physics, University of California, Berkeley, CA 94720, USA and
Department of Physics, University of California, Davis, CA 95616, USA

Uroš Seljak
Department of Physics, University of California, Berkeley,
and Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

We present the marginal unbiased score expansion (MUSE) method, an algorithm for generic high-dimensional hierarchical Bayesian inference. MUSE performs approximate marginalization over arbitrary non-Gaussian latent parameter spaces, yielding Gaussianized asymptotically unbiased and near-optimal constraints on global parameters of interest. It is computationally much cheaper than exact alternatives like Hamiltonian Monte Carlo (HMC), excelling on funnel problems which challenge HMC, and does not require any problem-specific user supervision like other approximate methods such as Variational Inference or many Simulation-Based Inference methods. MUSE makes possible the first joint Bayesian estimation of the delensed Cosmic Microwave Background (CMB) power spectrum and gravitational lensing potential power spectrum, demonstrated here on a simulated data set as large as the upcoming South Pole Telescope 3G 1500 $\deg^2$ survey, corresponding to a latent dimensionality of ~ 6 million and of order 100 global bandpower parameters. On a subset of the problem where an exact but more expensive HMC solution is feasible, we verify that MUSE yields nearly optimal results. We also demonstrate that existing spectrum-based forecasting tools which ignore pixel-masking underestimate predicted error bars by only $\sim 10\%$. This method is a promising path forward for fast lensing and delensing analyses which will be necessary for future CMB experiments such as SPT-3G, Simons Observatory, or CMB-S4, and can complement or supersede existing HMC approaches. The success of MUSE on this challenging problem strengthens its case as a generic procedure for a broad class of high-dimensional inference problems.

I. INTRODUCTION

Bayesian inference is a highly successful paradigm for quantifying uncertainty in the face of observed data. The procedure centers on updating a prior probability distribution, $\mathcal{P}(\theta)$, with the likelihood of some observed data, $\mathcal{P}(x\,|\,\theta)$, where θ represent some parameters of interest and x represents the data. Bayes' Theorem describes the form of this update:

$$\mathcal{P}(\theta\,|\,x) = \frac{\mathcal{P}(x\,|\,\theta)\,\mathcal{P}(\theta)}{\mathcal{P}(x)}. \tag{1}$$

The posterior, $\mathcal{P}(\theta\,|\,x)$, then summarizes the entirety of the information on θ provided by the observations.

In many problems of interest, the data do not depend just on θ, but also on a set of unobserved latent variables, z, which themselves depend stochastically on θ. In this case, the likelihood in Eqn. (1) involves a marginalization over the latent space z,

$$\mathcal{P}(x\,|\,\theta) = \int \mathrm{d}^N z\; \mathcal{P}(x, z\,|\,\theta) = \int \mathrm{d}^N z\; \mathcal{P}(x\,|\,z, \theta)\mathcal{P}(z\,|\,\theta), \tag{2}$$

where $\mathcal{P}(x, z\,|\,\theta)$ is the joint likelihood of both data and of z, given θ. These problems are considered "hierarchical" Bayesian problems, as there can be a hierarchy of latent variables, each depending probabilistically on the previous set. For the purposes of our work, we will consider z to encompass the entire latent space, defined as all parameters other than those of interest, θ.

The fundamental challenge of hierarchical Bayesian inference is performing the integral in Eqn. (2). Closed-form solutions rarely exist for all but the simplest problems, and the latent space can often be very high-dimensional and non-Gaussian, making numerical integration costly or intractable.

Some intermediate quantities which are easy to calculate can be useful for diagnosis or as pieces of other algorithms, but by themselves are not helpful for inferring parameters. For example, maximizing the joint likelihood or joint posterior of (θ, z) to produce the joint maximum likelihood estimate (JMLE) or joint maximum a posteriori estimate (JMAP), respectively, do not give useful estimates of θ since they only represent the peak of the integrand in Eqn. (2) but have not performed the necessary integral. Attempting to use the JMAP or JMLE yields estimates of θ which are biased or not meaningful [e.g. 1, 2]. Only the marginal maximum likelihood estimate (MMLE) of θ, which maximizes $\mathcal{P}(x\,|\,\theta)$, is asymptotically unbiased, but of course still involves the difficult integral in Eqn. (2).

Several popular solutions to the marginalization problem exist, each with various advantages and tradeoffs. Methods such as Hamiltonian Monte Carlo (HMC) are asymptotically exact, but can become very slow for high-dimensional problems or even for moderate dimensions

* mariusmillea@gmail.com

with sufficiently non-Gaussian latent spaces (see e.g. [3] for a review). Variational Inference (VI) forms another popular class of methods, which are deemed to be faster than HMC. However, VI is approximate and generically requires the user to chose a family of variational distributions, rendering the method less automatic (see e.g. [4] for a review). For cosmological applications, simplifications such as mean-field VI are inaccurate due to the correlations between the modes induced by survey masks, while full rank VI is equally infeasible due to the high dimensionality of the problem, which would require estimation of a very high dimensional covariance matrix. The difficulty of evaluating the marginal likelihood of Eqn. (2) has popularized the use of methods where the likelihood is not used at all, which go under the name of Likelihood Free Inference (LFI) or Simulation-Based Inference (SBI). These have attracted recent interest, but struggle for data and/or latent spaces which are very high-dimensional (see e.g. [5] for a review).

Here, we present the Marginal Unbiased Score Expansion (MUSE) method. It provides an often highly accurate approximation to Eqn. (2) (or rather, to its gradient, as we will discuss), and is much faster to compute than exact methods. The approximation was developed by Seljak *et al.* [6], with some applications to Gaussian problems in Horowitz *et al.* [7]. In this work, we extend previous results by quantifying the properties of this approximation for general non-Gaussian latent spaces, rendering it applicable to a much wider range of problems[1].

We show that regardless the structure of the latent space, MUSE provides estimates of parameters which are asymptotically unbiased, meaning they are unbiased as long as many different data modes contribute to constraints on each θ. This condition is quite often met automatically for the high-dimensional problems where MUSE is useful over alternatives anyway. In the context of cosmology, this is typical, as one often seeks to infer constraints on a small number of parameters or bandpowers from the aggregate of a huge number of data modes.

Additionally, as both a demonstration and as a novel solution in its own right, we apply MUSE to the challenging problem of Cosmic Microwave Background (CMB) gravitational lensing (see e.g. [8] for a review). Specifically, we use it to infer parameters and/or bandpowers of the gravitational lensing potential joint with the unlensed CMB. This problem is hierarchical because the parameters of interest control the statistics of the unobserved (latent) unlensed CMB and lensing potential. It is non-Gaussian due to the nature of lensing distortions. And it is high-dimensional because the size of the latent space is on the order of the number of map pixels, which can be millions for modern observations. As such, it is a perfect test-bed for MUSE.

Traditionally, analysis of CMB lensing has relied on the so-called "quadratic estimator," (QE), which is an estimator formed from quadratic combinations of the data [9, 10]. The QE is near-optimal for present day instrumental noise levels, and is based on an explicit semi-analytic formula which does not involve marginalization over any latent space. It has been widely used in almost all CMB lensing analyses to-date. However, as first shown by Hirata and Seljak [11, 12] and Seljak and Hirata [13], at noise thresholds which are currently being crossed by the most sensitive experiments, the QE ceases to be near-optimal and Bayesian methods which fully extract all-orders information from the data can yield significantly better results. At the noise levels of the planned CMB-S4 experiment [14], this includes reconstructing the gravitational lensing field to ~ 10 times lower noise levels [15] and yielding delensed maps of B modes which allow ~ 3 times better constraints on the amplitude of primordial gravitational waves, r [16].

The original work by Hirata and Seljak [11, 12] gave a method applicable to idealized mask-free data and set the stage for a number of ensuing improvements. The issue of masking, a necessity for any real analysis which must excise contaminated or unobserved parts of the sky, is central to the challenge of optimal lensing. While it is easy by comparison to devise estimates which assume full-sky or periodic flat-sky boundary conditions without masking, the impact of masking is drastic as it transforms the correlation structure of the high-dimensional latent space from sparse to dense, causing the breakdown of many simple approximations which could otherwise be used. Carron and Lewis [17] extended the original work in [11, 12] and computed a MAP estimate of ϕ for realistic data conditions which included masking, but did not attempt the integral in Eqn. (2), which would be needed to infer parameter constraints. A power spectrum estimate based on this MAP was recently given in Legrand and Carron [18], but the approximations are not validated in the presence of masking. Carron [16] also used the estimate to perform a brute-force integration of Eqn. (2) in the case that θ is one-dimensional, but the method does not scale to higher dimensions. Machine-learning estimates of ϕ have been demonstrated by Caldeira *et al.* [19] and Guzman and Meyers [20], but which likewise do not attempt the integral in Eqn. (2). All of these extensions also at present lack a demonstrated way to infer power spectra of the unlensed CMB, needed for full parameter extraction. Full-sky tools have been developed by Green *et al.* [21] and Hotinli *et al.* [22] which consider optimal joint lensing reconstruction and delensing, and while they serve as very useful power spectrum forecasting tools, they do not correspond to a map-level procedure which could be applied to real data

To-date, the only tractable method for performing the latent-space integration while considering lensing, delensing, and realistic data conditions, is based on the HMC sampling procedure of Millea *et al.* [2], Anderes *et al.* [23], Millea *et al.* [24]. This method was applied to South Pole Telescope data to demonstrate the first joint parameter estimation from an optimally reconstructed ϕ

[1] We also dub this generalized version "MUSE".

field and delensed CMB by Millea *et al.* [15]. While the HMC lensing approach has several appealing features, its downside is that it is slow and the Monte Carlo sampler is necessarily sequential and cannot be trivially sped up. For example, the $100\,\mathrm{deg}^2$ of polarization data analyzed by [15] took roughly 4 wall-hours on GPU, with a naive scaling to the entire SPT-3G $1500\,\mathrm{deg}^2$ survey suggesting HMC chains would require a week to converge. Here we will demonstrate an analysis of this $1500\,\mathrm{deg}^2$ dataset which completes in hours, and which is very amenable to trivial parallelization. The output of the estimate is a set of ϕ and delensed E-mode bandpowers as well as their joint covariance, making this a familiar data product for cosmologists to then use in a subsequent parameter estimation step.

Alongside this paper, we provide a software package, MuseInference.jl[2], which is a generic implementation of MUSE that can be used on any hierarchical Bayesian problem. This package has an interface into the probabilistic programming language (PPL) Turing.jl [25], and can immediately be applied to existing models and compared against HMC or VI. Interfaces to other PPLs like Stan or PyMC are planned. The interface also includes the existing CMBLensing.jl[3] code for application to the CMB lensing problem.

We begin in Sec. II with a description of the MUSE procedure aimed at a general audience, and demonstrate it on a generic toy problem in Sec. III. In Sec. IV we introduce the CMB lensing problem and present our main lensing results, before summarizing and giving concluding comments in Sec. V.

II. THE ALGORITHM

We begin with a generic description of the MUSE algorithm, applicable to any hierarchical Bayesian problem. In the later sections, we will turn to our specific application of CMB gravitational lensing.

A. The MUSE approximation

The MUSE algorithm is based on an approximation to the gradient of the marginal log-likelihood, a quantity usually called the marginal score:

$$s_i(\theta, x) \equiv \frac{d}{d\theta_i} \log \mathcal{P}(x \,|\, \theta) \tag{3}$$

The marginal score represents a lossless compression of the data, containing all information on parameters which can theoretically be extracted [26]. While many approximate Bayesian methods seek to approximate the value of the marginal or joint posterior, approximating the marginal score is just as good, and potentially more direct if in the end we are only interested in inferences of θ anyway. For example, if we had access to the marginal score, we could use exact Bayesian methods such as HMC to infer θ directly, since the relevant Hamiltonian trajectories would depend on just the marginal score. Similarly, optimal estimators like the MMLE can be defined entirely in terms of the marginal score, since the estimate is the parameter vector, $\hat{\theta}^{\mathrm{MMLE}}$, which solves

$$s_i(\hat{\theta}^{\mathrm{MMLE}}, x) = 0 \tag{4}$$

It is thus well-motivated to search for efficient ways to compute or approximate the marginal score.

Of course, the exact marginal score still requires performing the difficult integral in Eqn. (2). The MUSE solution involves an approximation which is extremely fast to compute in comparison to exact integration. As motivation, consider a Taylor series expansion of the joint likelihood,

$$\begin{aligned} &\log \mathcal{P}(x, z \,|\, \theta) = \\ &\quad = \log \mathcal{P}(x, \hat{z}_{\theta,x} \,|\, \theta) + \frac{1}{2}(z - \hat{z}_{\theta,x})^\dagger H_{\theta,x} (z - \hat{z}_{\theta,x}) + ..., \end{aligned} \tag{5}$$

where $\hat{z}_{\theta,x}$ is the maximum a posteriori (MAP) estimate of the latent space variables given x and fixed θ,

$$\hat{z}_{\theta,x} \equiv \underset{z}{\mathrm{argmax}} \; \log \mathcal{P}(x, z \,|\, \theta), \tag{6}$$

and $H_{\theta,x}$ is the Hessian matrix at this point. In terms of this expansion, the marginal score is then,

$$\begin{aligned} &s_i(\theta, x) = s_i^{\mathrm{MAP}}(\theta, x) + \\ &\frac{d}{d\theta_i} \log \int \mathrm{d}^N z \, \exp\left[\frac{1}{2}\,(z - \hat{z}_{\theta,x})^\dagger H_{\theta,x} (z - \hat{z}_{\theta,x}) + ...\right] \end{aligned} \tag{7}$$

where we have defined the gradient evaluated at the MAP as,

$$s_i^{\mathrm{MAP}}(\theta, x) \equiv \frac{d}{d\theta_i} \log \mathcal{P}(x, \hat{z}_{\theta,x} \,|\, \theta). \tag{8}$$

Note that the chain rule term which would appear above involving $d\hat{z}_{\theta,x}/d\theta$ never needs to be computed because it is multiplied by the gradient of the distribution, which is by definition zero at the MAP.

One common approach for approximating Eqn. (7) is to keep only the quadratic term in the exponential, yielding a Gaussian integral with an analytic solution. This

[2] https://github.com/marius311/MuseInference.jl
[3] https://github.com/marius311/CMBLensing.jl

[4] Although typical HMC implementations would also need the value of the posterior for the error-correcting Metropolis-Hastings step, this is only a practical issue of symplectic integration error, and is not theoretically needed in the limit of infinitely small step-size.

is the well-known Laplace approximation. However, this still requires obtaining the trace and inverse of the Hessian matrix, which in practice may also be extremely difficult due to the high dimensionality of the latent space and hence of this matrix. The key insight of MUSE is not to attempt to perform the remaining integral at all, but rather approximate it with its data-averaged value. This can in turn easily be obtained from the "unbiased score equation," which is the fact that any arbitrary score function (under regularity conditions) obeys

$$\Big\langle s_i(\theta, x) \Big\rangle_{x \sim \mathcal{P}(x \,|\, \theta)} = 0. \tag{9}$$

Enforcing this condition and solving for the expected value of the integral yields the MUSE marginal score approximation [6]:

$$\boxed{s_i^{\text{MUSE}}(\theta, x) \equiv s_i^{\text{MAP}}(\theta, x) - \Big\langle s_i^{\text{MAP}}(\theta, x') \Big\rangle_{x' \sim \mathcal{P}(x' \,|\, \theta)}.} \tag{10}$$

In practice, the second term in Eqn. (10) is computed via a Monte Carlo average over a suite of forward data simulations generated at the given value of θ.

Note that the MUSE approximation has the desirable property that in the limiting case of a Gaussian joint likelihood, where the latent space is Gaussian with a data-independent Hessian, it becomes exact (an explicit example of this is given in Appendix A). Even for mildly non-Gaussian latent spaces, one expects the data-dependence of the integral to be small, with most of the data-dependence instead captured by the MAP term. Additionally, regardless of whether the latent space is Gaussian or not, s^{MUSE} always obeys the unbiased score equation by construction. This feature will turn out to be key allowing Bayesian or frequentist estimates built from s^{MUSE} to remain unbiased.

We emphasize that the MUSE approximation does not correspond to the Laplace approximation for the joint likelihood. In fact, it may not correspond to *any* approximation for the joint likelihood, because s^{MUSE} is not, in general, a conservative vector field. This means it cannot be written as the gradient of some scalar function, which would then be interpretable as the approximate distribution. Exceptions to this include the Gaussian problem, where the non-conservative terms in s^{MUSE} cancel, and the case of a one dimensional θ, where this distinction does not exist. This technical detail will have a few important consequences, discussed below.

B. The frequentist view

We now describe both a frequentist and Bayesian approach for parameter inference which make use of s^{MUSE}. Ultimately, both correspond to performing the identical computation, and differ only in interpretation. This is not surprising since the regime where MUSE is best applicable is where the θ are well-constrained relative to the prior, and hence where Bayesian and likelihood-based frequentist methods agree. It is instructive, however, to follow the assumptions inherent in each description, which might point to different types of future extensions of the method. We begin with the frequentist version.

In the frequentist approach, we are interested in building an estimator for θ. In analogy to the MMLE defined in Eqn. (4), it is natural to define the MUSE estimate as the parameter vector which solves,

$$\boxed{s_i^{\text{MUSE}}(\hat{\theta}^{\text{MUSE}}, x) = 0.} \tag{11}$$

In this way, if the latent space is Gaussian where MUSE is exact, we recover the MMLE, and MUSE is therefore asymptotically unbiased and minimum variance.

Note that while Eqn. (4) can be rephrased as maximizing a scalar function (i.e., maximizing the marginal likelihood), the fact that s^{MUSE} may be non-conservative means MUSE must generically be regarded as a vector-valued root-finding problem. Such problems are not guaranteed to have a solution, and if no solution is found for some particular case, then MUSE cannot be used. Having noted this, we have not found it to be a typical concern except in some pathological instances.

We next need to determine the bias and covariance of the MUSE estimator. We will consider the asymptotic limit of a large number of observations, N, since we are targeting problems where a large N has driven the estimator distribution towards Gaussian by the central limit theorem. The log-likelihood of N data, $x_1, ..., x_N$, drawn independently from $\mathcal{P}(x \,|\, \theta)$, is the sum of the log-likelihoods of each. Accordingly, the MUSE estimate for N data is implicitly defined by the solution to

$$s_i^{\text{MUSE}}(\hat{\theta}, \{x_n\}) \equiv$$
$$\frac{1}{N} \sum_{n=1}^{N} s_i^{\text{MAP}}(\hat{\theta}, x_n) - \Big\langle s_i^{\text{MAP}}(\hat{\theta}, x') \Big\rangle_{x' \sim \mathcal{P}(x' \,|\, \hat{\theta})} = 0. \tag{12}$$

where we have defined the MUSE gradient for N data as $s_i^{\text{MUSE}}(\hat{\theta}, \{x_n\})$ for later use, and have used $\hat{\theta} \equiv \hat{\theta}^{\text{MUSE}}$ for brevity. In the limit of $N \to \infty$, the summation in Eqn. (12) becomes an expectation value over $x \sim \mathcal{P}(x \,|\, \theta)$, and the equation is trivially solved when $\hat{\theta} = \theta$, demonstrating that the MUSE estimate is asymptotically unbiased (for any Gaussian or non-Gaussian latent space).

Note that for finite N, the difference between the sum in the term in brackets in Eqn. (12) and its asymptotic limit will scale as $1/\sqrt{N}$ by the central limit theorem, thus the estimator bias scales like $1/\sqrt{N}$ times its standard deviation, similarly as for the MMLE. We also note that the MUSE estimate is trivially unbiased for *any* N when the θ which generates the simulations in Eqn. (10) happens to be the truth. Thus, if some prior knowledge suggests some regularization which brings the MUSE es-

timate closer to truth[5], the bias can actually be smaller than for the MMLE.

Finally, we consider the covariance of the MUSE estimate. We first note that it is straightforward to compute the estimator covariance at some θ via Monte Carlo, by running the MUSE estimate on a suite of simulated data and taking the empirical covariance. If the computational cost of running a sufficient number of simulations is not prohibitive, this is likely the easiest approach in practice, and is guaranteed to give an exact answer up to Monte Carlo errors. We can, however, significantly reduce the computational cost. If the latent space is known to be Gaussian, Horowitz *et al.* [7] demonstrated a fast and exact approach. For the non-Gaussian latent spaces of interest in this work, we supersede the suggestions in Seljak *et al.* [6] with one which works for more general non-Gaussian distributions.

First, expand Eqn. (12) to first order around the true value, denoted θ^*,

$$s_i^{\mathrm{MUSE}}(\theta^*, \{x_n\}) \,+\, (\hat{\theta}_j - \theta_j^*)\, h_{ij}^{\mathrm{MUSE}}(\theta^*, \{x_n\}) = 0 \quad (13)$$

where we have defined the Jacobian

$$h_{ij}^{\mathrm{MUSE}} = \frac{d s_i^{\mathrm{MUSE}}}{d\theta_j}. \quad (14)$$

Introducing a factor of N and rearranging terms yields

$$\sqrt{N}(\hat{\theta}_j - \theta_j^*) = - \left[\frac{1}{N} h_{ij}^{\mathrm{MUSE}}(\theta^*, \{x_n\})\right]^{-1} \left[\sqrt{N}\frac{1}{N} s_i^{\mathrm{MUSE}}(\theta^*, \{x_n\})\right]. \quad (15)$$

Assuming suitable regularity conditions, the second term in brackets can be shown by the central limit theorem to converge in probability as $N \to \infty$ to a normal distribution with zero mean and with covariance given by J_{ij}, and the quantitiy in the first term in brackets to converge to H_{ij}, where these matrices are defined as

$$J_{ij} = \left\langle s_i^{\mathrm{MUSE}}(\theta^*, \{x_n\})\, s_j^{\mathrm{MUSE}}(\theta^*, \{x_n\}) \right\rangle_{x_n \overset{\mathrm{iid}}{\sim} \mathcal{P}(x\,|\,\theta^*)} \quad (16)$$

$$H_{ij} = \left\langle h_{ij}^{\mathrm{MUSE}}(\theta^*, \{x_n\}) \right\rangle_{x_n \overset{\mathrm{iid}}{\sim} \mathcal{P}(x\,|\,\theta^*)}. \quad (17)$$

These expressions can be further simplified to be written only in terms of averages of gradients at the MAP (see Appendix B):

$$\boxed{\begin{aligned} J_{ij} &= \left\langle s_i^{\mathrm{MAP}}(\theta^*, x)\, s_j^{\mathrm{MAP}}(\theta^*, x) \right\rangle_{x \sim \mathcal{P}(x\,|\,\theta^*)} \\ &\quad - \left\langle s_i^{\mathrm{MAP}}(\theta^*, x) \right\rangle \left\langle s_j^{\mathrm{MAP}}(\theta^*, x) \right\rangle_{x \sim \mathcal{P}(x\,|\,\theta^*)} \end{aligned}} \quad (18)$$

$$\boxed{H_{ij} = \frac{d}{d\theta_j}\left[\left\langle s_i^{\mathrm{MAP}}(\theta^*, x)\right\rangle_{x \sim \mathcal{P}(x\,|\,\theta)}\right]\Big|_{\theta=\theta^*}} \quad (19)$$

The interpretation of J is straightforward: it is the covariance of the MAP gradient simulations. As these are already computed for the purposes of computing s^{MUSE} itself, they do not add any extra computational cost. H involves one extra derivative of these gradient simulations, and we note that careful attention should be given to which variables are held fixed; the derivative does not act on the argument of s^{MAP}, which is held fixed at θ^*, instead only acting on the parameters which generate the simulated data, x. Supposing that finite differences were used to compute H highlights another interpretation of this term: H is computed by injecting infinitesimal parameter shifts into simulated data, and observing how the MAP gradient changes in response.

As long as H is invertible, the continuous mapping theorem then gives the final result for the asymptotic covariance of the MUSE estimate:

$$\Sigma_{ij}^{\mathrm{MUSE}} \equiv \langle \Delta\hat{\theta}_i^{\mathrm{MUSE}} \Delta\hat{\theta}_j^{\mathrm{MUSE}} \rangle = (H^{-1} J\, H^{-\dagger})_{ij} \quad (20)$$

Note that one can show that for Gaussian problems, $J = H = \mathcal{F}^{-1}$ where $\mathcal{F}$ is the Fisher information matrix, such that the estimator saturates the Cramér-Rao bound and is considered optimal. This follows from the fact that for Gaussian problems, MUSE becomes the MMLE, which itself is known to asymptotically saturate this bound.

C. The Bayesian view

In the Bayesian approach, instead of a point estimate, we are interested in exploring the posterior distribution, $\mathcal{P}(\theta\,|\,x)$. Although we can use the MUSE score in conjunction with the prior to approximate the gradient of the log-posterior,

$$g_i^{\mathrm{MUSE}}(\theta, x) \equiv s_i^{\mathrm{MUSE}}(\theta, x) + \frac{d}{d\theta_i} \log \mathcal{P}(\theta), \quad (21)$$

the fact that s^{MUSE} and hence g^{MUSE} are not guaranteed to be conservative vector fields means there is no trivial way to use this to back out an approximation to $\mathcal{P}(\theta\,|\,x)$. Accordingly, HMC is not in theory applicable to g^{MUSE} because the Hamiltonian equations underpinning HMC assume two trajectories starting and ending at the same points in parameter space yield the same change in conjugate momenta, which would not be the case for a non-conservative g^{MUSE}. It may be the case that this is not

[5] In the context of cosmology, one such regularization arises if the θ represent a spectral density which is expected from physical arguments to vary slowly with scale, such as the CMB or matter power spectrum. In this case, θ can be regularized by applying some chosen smoothing kernel.

a problem in practice, or that the Hamiltonian dynamics can be modified to account for it, but we have not explored this avenue.

One option to proceed is a Gibbs sampling approach, considering one parameter at a time. In one dimension, issues of conservativeness are irrelevant, and we can integrate $s^{\rm MUSE}$ on a grid to obtain a full posterior shape, draw a sample, then continuously repeat for subsequent dimensions. We give an example of the grid evaluation in the next section, but have not pursued this further.

Another natural Bayesian approach is to consider $\hat{\theta}^{\rm MUSE}$ as a summary statistic, and explore the posterior $\mathcal{P}(\theta \,|\, \hat{\theta}^{\rm MUSE})$. In summarizing the data with $\hat{\theta}^{\rm MUSE}$, we have potentially lost some information, and this will be reflected in a potentially less constraining $\mathcal{P}(\theta \,|\, \hat{\theta}^{\rm MUSE})$ as compared to $\mathcal{P}(\theta \,|\, x)$. However, this will only be significant given large latent non-Gaussianity, and regardless, the posterior inference will be valid.

If we suspect that $\mathcal{P}(\theta \,|\, \hat{\theta}^{\rm MUSE})$ is considerably non-Gaussian, we could use any of a number of SBI methods to map out this distribution [e.g. 5]. In this sense, we can view $\hat{\theta}^{\rm MUSE}$ as a near-optimal data-compression step, of the kind required by many SBI methods. Because MUSE is fast to compute, the total computational cost can still be well below the cost of performing full HMC or SBI on the joint posterior.

If we instead assume a near-Gaussian $\mathcal{P}(\theta \,|\, \hat{\theta}^{\rm MUSE})$, we can forego SBI and compute the distribution under some simple assumptions. Writing $\hat{\theta} \equiv \hat{\theta}^{\rm MUSE}$ for clarity, the posterior conditioned on the summary statistic is

$$\mathcal{P}(\theta \,|\, \hat{\theta}) = \frac{\mathcal{P}(\hat{\theta} \,|\, \theta)\, \mathcal{P}(\theta)}{\mathcal{P}(\hat{\theta})}. \tag{22}$$

The first term in the numerator can be written as

$$\mathcal{P}(\hat{\theta} \,|\, \theta) = \int \mathrm{d}^N x\; \mathcal{P}(\hat{\theta}^{\rm MUSE} \,|x) \mathcal{P}(x \,|\, \theta) \tag{23}$$

$$= \int \mathrm{d}^N x\; \delta^P\big(s^{\rm MUSE}(\hat{\theta}, x)\big) \mathcal{P}(x \,|\, \theta) \tag{24}$$

where δ is the Dirac delta function, P is the dimensionality of θ and N the dimensionality of x. We can use the first and second moments to approximate the mean and covariance of this Gaussian. The mean is,

$$\mu_i = \int \mathrm{d}^P \hat{\theta}\; \hat{\theta}_i \int \mathrm{d}^N x\; \delta^P\big(s^{\rm MUSE}(\hat{\theta}, x)\big) \mathcal{P}(x \,|\, \theta). \tag{25}$$

Assuming the likelihood is adequately peaked, it suffices to Taylor expand the MUSE score around θ,

$$s_i^{\rm MUSE}(\hat{\theta}, x) \approx s_i^{\rm MUSE}(\theta, x) + (\hat{\theta}_j - \theta_j)\, h_{ij}^{\rm MUSE}(\theta, x) \tag{26}$$

then perform the integral over $\hat{\theta}$ to obtain

$$\int \mathrm{d}^N x\; \mathcal{P}(x \,|\, \theta) \Big[h_{ji}^{\rm MUSE}(\theta, x)^{-1} s_i^{\rm MUSE}(\theta, x) + \theta \Big]. \tag{27}$$

Assuming $h^{\rm MUSE}$ is independent of data and using the fact that $s^{\rm MUSE}$ obeys the unbiased score equation yields simply $\mu = \theta$. We note that if desired, this assumption can be explicitly checked in practice since Eqn. (27) is just an average over forward data simulations and can be computed via Monte Carlo.

The covariance will be

$$\tilde{\Sigma}_{ij}^{\rm MUSE} = \int \mathrm{d}^P \hat{\theta}\; (\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j) \int \mathrm{d}^N x\; \delta^P\big(s_k^{\rm MUSE}(\hat{\theta}, x)\big) \mathcal{P}(x \,|\, \theta), \tag{28}$$

where the tilde differentiates this covariance from the covariance of the MUSE estimator defined in the previous section. Simplifying similarly as above yields

$$\int \mathrm{d}^N x\; \mathcal{P}(x \,|\, \theta) \Big[h_{ij}^{\rm MUSE}(\theta, x)^{-1} s_j^{\rm MUSE}(\theta, x) \times s_k^{\rm MUSE}(\theta, x) h_{kj}^{\rm MUSE}(\theta, x)^{-1} \Big], \tag{29}$$

which can also be explicitly computed via Monte Carlo. Note that if we assume $h^{\rm MUSE}$ is realization-independent, then $\tilde{\Sigma}_{ij}^{\rm MUSE} = \Sigma_{ij}^{\rm MUSE}$, meaning the Bayesian and frequentist estimates are identical.

With the Gaussian approximation to $\mathcal{P}(\hat{\theta} \,|\, \theta)$ ascertained in this way, one can combine it with any prior desired to obtain the full posterior $\mathcal{P}(\theta \,|\, \hat{\theta})$. In the special case that the prior is also a Gaussian, $\mathcal{N}(\theta_{\rm p}, \Sigma_{\rm p})$, we have that

$$\mathcal{P}(\theta \,|\, \hat{\theta}) = \mathcal{N}\Big(\Sigma_{\rm tot}(\Sigma^{-1}\hat{\theta} + \Sigma_{\rm p}^{-1}\theta_{\rm p}),\; \Sigma_{\rm tot}\Big). \tag{30}$$

where $\Sigma_{\rm tot} = (\Sigma^{-1} + \Sigma_{\rm p}^{-1})^{-1}$. Note also that the mean of this distribution can be equivalently calculated by simply solving $g^{\rm MUSE}(\hat{\theta}, x) = 0$ rather than $s^{\rm MUSE}(\hat{\theta}, x) = 0$.

D. Practical considerations

In practice, either Bayesian or frequentist views of MUSE involve first iteratively solving the vector equation $g_i^{\rm MUSE}(\hat{\theta}, x) = 0$ or $s_i^{\rm MUSE}(\hat{\theta}, x) = 0$ for $\hat{\theta}$, respectively. We use Broyden's method [27], a standard choice for this type of problem. This requires an initial guess for the Jacobian, whose exact value would be $h^{\rm MUSE}$. We find it sufficient to approximate this as J^{-1} or even Diagonal(J^{-1}), which can be computed for free from simulations already performed for the first step. This choice impacts only the speed of convergence, not the final solution.

At each step of Broyden's method, we compute $s^{\rm MUSE}$ at the current value of θ. This involves generating M data simulations given θ, and finding the MAP solution of each. We note that this step is completely amenable to trivial parallelization, and on GPUs, is ideally performed efficiently with batching. These MAP solutions dominate the runtime of the algorithm, and the fact that they are parallelizable in this way is a particular strength of MUSE. The MAP solutions can be found with standard

methods like L-BFGS, although individual problems may have domain-specific solutions as well (e.g. the CMB lensing problem discussed later features an optimized solver based on coordinate descent). For performance, it is very advantageous to use the same random seeds for these simulations throughout each iterative Broyden step, and to start the MAP solver for each step from the solution for the previous step. Since θ is not changing much by the final Broyden iterations, the MAP solver will require very few steps.

The choice of how many simulations, M, to use, is informed by how much Monte Carlo error one is willing to incur. The error between $\hat{\theta}$ computed using M simulations and its true value in the limit $M \to \infty$, relative to its standard deviation, scales as $1/\sqrt{M}$. Thus, using 100 simulations incurs a possible error of $0.1\,\sigma$, which for many purposes is sufficient. Even 10 simulations yields an error of $0.3\,\sigma$, which is often acceptable. If one is using MUSE to compare the impact of two choices of modeling assumptions, one can use the same simulation random seeds for both runs, in which case the error cancels out to first order and the impact can be determined to much better than $0.1\,\sigma$ even with 100 simulations (and potentially even much fewer). Alternatively, if one is interested in the error between $\hat{\theta}$ and the true value of θ, the scaling is instead $\sqrt{1 + 1/M}$. Thus, with only 10 simulations, the error relative to the true value is only increased by about $0.05\,\sigma$. Different choices of which metric to use to chose M can be valid in different circumstances.

Computing the J and H matrices can vary in difficulty depending on the problem and depending on if second order Automatic Differentiation (AD) is available. The J matrix can be computed from the same MAP solutions which went into the computation of s^{MUSE} on the final Broyden iteration. If needed, additional simulations can be performed just at the final θ for the purposes of recovering J and its inverse to better accuracy. Additionally, since J is a covariance, shrinkage estimators can be used if some particular structure is expected. Computing H involves propagating a second-order derivative through the MAP solver, which, with second-order AD, can be done in the same pass as the MAP solutions needed for J. If the AD library does not provide higher-order derivatives, finite differences can be used, which may still be quite fast as the dimensionality of θ is generally small. We also note that H is typically quite realization-independent, thus does not require averaging over a large number of simulations (and this can of course be checked in practice). Finally, for many problems, particularly in cosmology where a viable fiducial model is already known, J and H can be computed once at the fiducial model and only need to be recomputed under significant changes to modeling assumptions.

III. MUSE ON A TOY PROBLEM

We now demonstrate the MUSE algorithm on a toy problem. A scenario which arises in many different contexts in Bayesian hierarchical modeling is the so-called "funnel problem" [28], which exhibits both non-Gaussianity and is particularly difficult to sample with HMC without additional tricks [29]. The funnel problem is a standard benchmark for inference algorithms, and is a useful problem to build intuition about the MUSE estimate.

We also slightly extend the funnel problem for the purposes of demonstration. The extended version embeds the funnel within one additional hierarchical layer, and contains some tunable parameters which can further increase the non-Gaussianity. It is defined by

$$\theta \sim \text{Normal}(0, 3) \tag{31}$$

$$z_i \sim \text{Normal}(0, e^{\theta/2}) \tag{32}$$

$$x_i \sim \text{Normal}\big(\beta \tanh(z/\beta), \sigma\big) \tag{33}$$

The nomenclature is the same as in the previous section: θ is the parameter of interest, z are the latent space variables, and x are the observed variables. The tunable parameter β can increase the non-Gaussianity of the latent space; in the limit $\beta \to \infty$, the latent space is Gaussian when conditioned on x and θ, whereas for $\beta \sim 1$, non-Gaussianity is induced even just within the latent space conditional. The latent space and data are indexed by $i = 1...N$ where N controls the number of data points and hence how closely we approach the asymptotic limit. The parameter σ controls "noise" in the observations. In the limit $\sigma \to \infty$ and $\beta \to \infty$, we recover exactly the standard funnel problem.

As a short aside from the general discussion, we note that funnel-like problems are extremely common in cosmological inference problems since theories generally predict the statistical properties of some field, rather than the field itself. For example, the CMB power spectrum predicts the distribution of fluctuations in the CMB temperature and polarization fields, and the galaxy power spectrum predicts the distribution of fluctuations in the galaxy density field. Inferring either power spectrum amounts to inferring a variance parameter analogous to θ in the funnel problem, with the field serving as z.

We consider three cases of our toy problem, pictured in the three columns in Fig. 1, which demonstrate important regimes for the MUSE estimate. In each case, we compare the MUSE estimate with the exact posterior computed with the more expensive but exact HMC procedure. In all cases, we set the true value of θ to zero and generate a sample of x from the forward model, then use MUSE or HMC to estimate $\mathcal{P}(\theta \,|\, x)$. Because this toy problem features a prior on θ, we are using the Bayesian version of MUSE described in the previous section.

We benchmark the two algorithms by comparing the number of joint likelihood gradient evaluations needed to estimate the posterior mean of θ to similar accuracy.

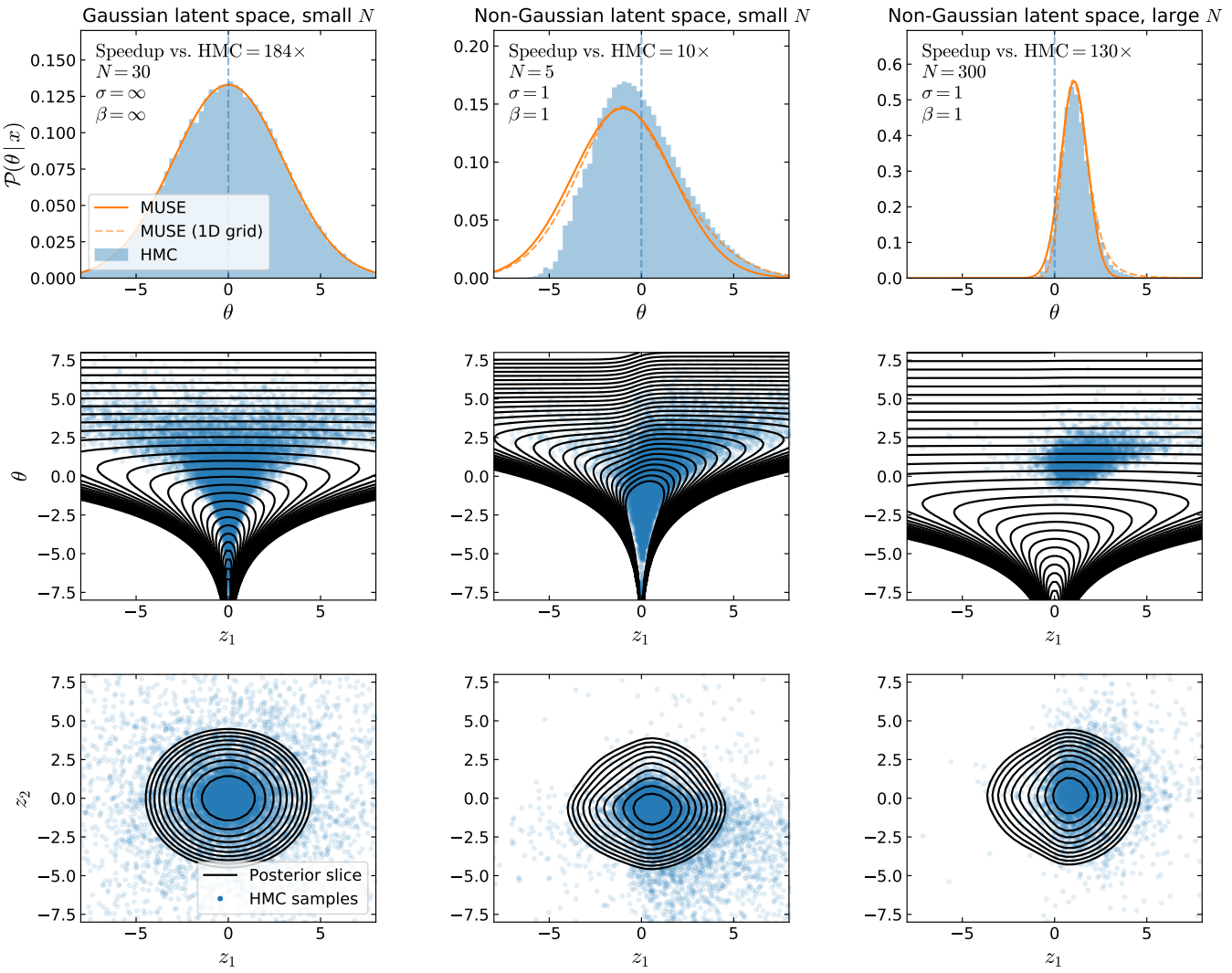

FIG. 1. MUSE applied to the toy "funnel"-like problem described in Sec. III. Each column corresponds to a run using the N, σ, and α parameters indicated in the top panel. The top row compares the MUSE estimate (solid orange) with the exact posterior determined via HMC (histogram). Also shown (dashed orange) is the result of integrating over the MUSE score evaluated on a grid, which is possible in 1D. In the middle and bottom rows, the black contours show a slice through the joint likelihood with parameters which are not pictured fixed to zero, and the blue dots show HMC samples and map out the marginal distribution. MUSE is most useful in cases like the final column where the joint likelihood can be non-Gaussian in any parameter, but the posterior on a parameter of interest has been driven to near-Gaussianity by the central limit theorem. Conversely, the middle column is a failure mode of MUSE due to low dimensionality and latent non-Gaussianity. In the cases here spanning $N = 30 - 300$ latent dimensions, MUSE outperforms HMC in terms of the number of posterior gradient evaluations by one to two orders of magnitude.

Of course, HMC can always produce the exact posterior mean whereas MUSE will sometimes be approximate, so this comparison should only be regarded as fair for the problems with near-Gaussian θ posteriors that are targeted by MUSE. In the case of HMC, the error on the posterior mean relative to the standard deviation will scale as $1/\sqrt{M}$ where M is the number of statistically independent samples in the chain (which can be computed from the effective sample size). For MUSE, the error will scale as $1/\sqrt{M}$ where M is the number of simulations used to compute the expectation value in s^{MUSE}. We thus compare the number of gradient evaluations needed for one independent sample vs. the number needed for one of the MUSE simulations throughout the course of the entire iterative procedure. For MUSE, the computation of J does not add additional cost since the relevant sims are already be computed for the purposes of the estimate itself, and H is usually a subdominant cost since it is only weakly data-dependent, needs very few sims, and can often be speed up with AD.

For HMC, we use the NUTS algorithm as implemented in Turing.jl [25]. The target acceptance rate is set to

0.65 (the optimal choice for a Gaussian [30]), except for the first case where it is set to 0.999 to adequately sample the tails of the distribution [31]. For MUSE, we use our package MuseInference.jl, with a step size $\alpha = 0.7$, relative error tolerance on θ of 1% of the standard deviation, and an absolute tolerance on the gradient at the MAP solution of 10^{-4}. We now describe the three cases:

($N = 30$, $\sigma = \infty$, $\beta = \infty$) This case is the standard funnel problem. It features a Gaussian latent space and an N which is too small to achieve the asymptotic guarantees of MUSE. Nevertheless, MUSE still recovers the true posterior perfectly because s^{MUSE} is always exact for Gaussian latent spaces. The black contours in the bottom row show a slice through the (z_1, z_2) posterior with all other z_i and θ set to zero, demonstrating the latent Gaussianity. Note that the (θ, z_i) posterior can still be (highly) non-Gaussian, as shown by the black contours in the middle row, which instead exhibit the "funnel" shape which is namesake of the problem. The top panel shows the near-perfect overlap between a histogram of the HMC samples and the MUSE posterior estimate. Blue dots in the lower two panels also show HMC samples for reference (note the difference between the blue points which sample the posterior *marginalized* over the other variables, vs. the black contours which fix these other variables to zero). In terms of benchmarks, the funnel problem is nearly trivial for MUSE to solve. Because here the data is uninformative, the MAP solution is $\hat{z}_{\theta,x} = 0$ for any x or θ, which is achieved with only a few posterior gradient evaluations (we do not put in the solution by hand, instead letting the optimizer proceed as usual). The MUSE posterior is then simply given by the Gaussian expansion of the prior, which gives the exact answer here. Conversely, HMC struggles on this problem due to the wide range of scales which must be traversed by the symplectic integrator in the mouth vs. neck of the funnel. Overall, we find MUSE requires 184 times fewer gradient evaluations than HMC on this case. We stress that while this toy problem is certainly expected to highlight the benefits of MUSE, it can be regarded as fair since no special information was input to MUSE beyond the standard generic procedure described in the previous section. While there are known ways to greatly improve HMC performance on the funnel problem [24, 29], these require manual intervention, whereas MUSE will automatically perform well on funnel-like problems, or subspaces of more complex problems which exhibit funnel-like behavior, even if these subspaces are not known a priori.

($N = 5$, $\sigma = 1$, $\beta = 1$) This case sets $\sigma = \beta = 1$ and reduces the dimensionality to $N = 5$. We are now even further from the asymptotic limit, and the latent space is no longer Gaussian. This is the main failure of mode of MUSE. Here, there are no guarantees that the inference is unbiased or that it has the correct variance. Indeed, in the top center panel we see neither are correct when compared to the HMC posterior. If for some real-world problem one suspects that this failure mode is being hit, one diagnostic is to run MUSE on a suite of simulations with a known input truth and to check if, on average, the truth is recovered and if the empirical covariance matches Eqn. (30).

($N = 300$, $\sigma = 1$, $\beta = 1$) This is like the previous case except we have increased the dimensionality to $N = 300$. This is large enough that the asymptotic guarantees of MUSE kick in, and the top-right panel of Fig. 1 shows that we again recover the HMC posterior near-perfectly. In doing so, we find that MUSE uses 130 times fewer posterior gradient evaluations than HMC. This case is the closest to where we envisage MUSE is most useful in the real-world: high-dimensional problems where the joint posterior need not be Gaussian in any variable, but where the marginal posterior on parameters of interest is asymptotically driven to near-Gaussianity. The CMB lensing problem which we will discuss in the next section is most similar to this third example, featuring a latent space with dimensionality $N \sim 10^5 - 10^7$.

For this last case, we also show in Fig. 1 the result of computing s^{MUSE} on a grid of θ values and integrating the result to produce an exact MUSE posterior approximation (possible only in 1D). This allows us to visualize the quality of the MUSE approximation separately from its Gaussianization around the peak, and note that it tracks the true posterior extremely well. Although computationally costlier, this can be a valid way to deal with cases where there is only a single parameter of interest.

IV. MUSE CMB LENSING

A. The CMB lensing problem

We now describe the CMB lensing problem, which we use to demonstrate MUSE in a challenging real-world scenario, and for which MUSE provides a novel solution. The goal of this analysis is to estimate the power spectrum of the CMB lensing potential and the power spectrum of the unlensed E-mode polarization given noisy lensed CMB data. We ignore CMB temperature because this leads to smaller differences between Bayesian and QE methods, and we do not estimate B-mode bandpowers since including these is qualitatively the same as the E-modes in terms of MUSE (B-mode polarization maps do enter the algorithm, just with a theory spectrum which is assumed perfectly known). It is straightforward to include either of these components if desired. The lensing problem can be summarized as

$$(A_E^b, A_\phi^b) \sim \text{Uniform}\big(0, \infty\big) \tag{34}$$

$$f \sim \mathbb{C}_f(A_E^b) \tag{35}$$

$$\phi \sim \mathbb{C}_\phi(A_\phi^b) \tag{36}$$

$$x \sim \text{Normal}\big(\mathbb{A}\,\mathbb{L}(\phi) f,\, \mathbb{C}_n\big), \tag{37}$$

with the corresponding joint likelihood distribution

$$\mathcal{P}(x, \underbrace{f, \phi}_{z} \mid \underbrace{A^b_E, A^b_\phi}_{\theta}) \propto \frac{\exp\left\{-\frac{\left(x - \mathbb{A}\,\mathbb{L}(\phi)\,f\right)^2}{2\,\mathbb{C}_n}\right\}}{\det \mathbb{C}_n^{1/2}} \frac{\exp\left\{-\frac{f^2}{2\,\mathbb{C}_f(A^b_E)}\right\}}{\det \mathbb{C}_f(A^b_E)^{1/2}} \frac{\exp\left\{-\frac{\phi^2}{2\,\mathbb{C}_\phi(A^b_\phi)}\right\}}{\det \mathbb{C}_\phi(A^b_\phi)^{1/2}} \mathcal{P}_S(\phi) \tag{38}$$

where

- $\mathbb{A} \equiv \mathbb{M} \cdot \mathbb{T} \cdot \mathbb{B}$ is a linear operator containing the instrumental beam, $\mathbb{B}$, transfer function, $\mathbb{T}$, and any masking, $\mathbb{M}$
- f is the map of the unlensed CMB polarization
- ϕ is the map of gravitational lensing potential
- x is the data
- $\mathbb{L}(\phi)$ is a linear operator which lenses a map (and whose dependence on ϕ is non-linear)
- $\mathbb{C}_n$, $\mathbb{C}_f$, and $\mathbb{C}_\phi$ denote the covariances for the noise, unlensed CMB, and ϕ, respectively
- A^b_E and A^b_ϕ are bandpower amplitudes which control the covariances $\mathbb{C}_f$ and $\mathbb{C}_\phi$, respectively. Specifically, they scale the isotropic CMB polarization and lensing power spectra as

$$C^{\phi\phi}_\ell = \left[1 + \sum_b (A^b_\phi - 1)\, W^\phi_{b\ell}\right] C^{\phi\phi,\mathrm{fid}}_\ell \tag{39}$$

$$C^{EE}_\ell = \left[1 + \sum_b (A^b_E - 1)\, W^E_{b\ell}\right] C^{EE,\mathrm{fid}}_\ell, \tag{40}$$

  where bandpower window functions $W_{b\ell} = 1$ if ℓ falls within bin b, and is zero otherwise. We chose to estimate amplitudes relative to a fiducial model rather than the power spectra themselves for simplicity and without loss of generality. The form of W is arbitrary and chosen for simplicity, and should not matter as long as the spectra do not vary significantly across the bin. Here we use a typical binning used in previous SPT analyses which features $\Delta\ell = 50$ for E and 10 logarithmically spaced bins for ϕ.
- $\mathcal{P}_S(\phi)$ is a "super-sample" prior on ϕ which we discuss below.

As indicated in Eqn. (38), A^b_ϕ and A^b_E form the θ parameters which wish to infer, and the maps of the unlensed CMB and of the lensing potential, f and ϕ, form the latent space, z. Depending on the pixelization and area of sky observed, these maps can easily contain more than a million pixels, and although the likelihood is Gaussian in f, it is a non-Gaussian function of ϕ. This high-dimensionality and non-Gaussianity makes the CMB lensing problem an excellent test-bed for MUSE.

The joint likelihood shown in Eqn. (38) is exactly as described in previous works which have attempted to maximize or sample this distribution [2, 23, 24]. Here, we make one additional and simple change which we find is very helpful in making the problem more amenable to MUSE estimation. As discussed in [2], in cases where a pixel mask is present (which is the case for any real analysis), the MAP estimate when maximizing jointly over both f and ϕ incurs a contribution from a "mean-field." The mean-field approximately manifests as an additive offset to the magnification, $\kappa \equiv -\nabla^2\phi/2$, across unmasked pixels. It arises because the mean magnification is otherwise very unconstrained due to aliasing from the mask. The presence of the mean-field can imprint a bias in the MUSE estimate because these unconstrained modes are then aliased into the bandpowers of interest, but do not benefit from guarantees of asymptotic unbiasedness present for other modes. A simple resolution which we have found is to impose a prior

$$\mathcal{P}_S(\phi) = \exp\left(-\frac{\left(\langle -\nabla^2\phi/2\rangle_{\text{unmasked pixels}}\right)^2}{2\sigma^2_\kappa}\right), \tag{41}$$

which recenters the mean κ in the joint MAP estimate to near zero. The prior almost perfectly removes the mean-field from the joint MAP estimate, and otherwise does not require any modification to the generic MUSE procedure (it is simply an additional contribution to the $\mathcal{P}(z \mid \theta)$ term in Eqn. 2). The prior can be motivated physically by noting that its need arises from the data failing to constrain modes in the κ map that are larger than the observed field. However, theoretically we know these "super-sample" modes should be near zero. Alternatively, for a given region of sky, full-sky *Planck* observations may give a good estimate of what the magnification actually is. It is thus valid to impose this as a prior, with the added benefit that at the same time it remedies possible biases in the MUSE estimate from this effect.

Finally, we note that instead of the joint distribution in Eqn. (38), it would in theory be possible to analytically marginalize over f, then perform MUSE on just the remaining part of the distribution, $\mathcal{P}(\phi, \theta \mid x)$. This form of the posterior was explored by [12, 17], with some further discussion in [2]. While it may seem advantageous to perform as much of the marginalization as possible analytically, in terms of speed it would actually be a detriment because the subsequent gradients and MAP estimates of

Simulation	CMBS4-150D	SPT-3G
Map size	256×256	1024×2048
Pixel width	3 arcmin	2.1 arcmin
Total area	$\sim$150 deg^2	$\sim$1500 deg^2
Noise level in T	1 μK-arcmin	2.3 μK-arcmin
$(\ell_{\rm knee}, \alpha_{\rm knee})$	(100,3)	(100,3)
Beam FWHM	2 arcmin	1 arcmin
Fourier masking ($\mathbb{K}$)	$2 < \ell < 3300$	$2 < \ell < 5000$
Pixel masking ($\mathbb{M}$)	varying	SPT-3G
Fiducial r	$r = 10^{-1}$	$r = 10^{-1}$
MUSE estimation		
Estimated θ	A^b_ϕ and A^b_E	A^b_ϕ and A^b_E
# of simulations	100	100
Step size	0.7	0.7
# of steps	10	10
MUSE runtime	5 minutes	60 minutes
HMC runtime	5 hours	$\sim$ week (est.)

TABLE I. Simulation parameters and MUSE solution parameters for the different configurations of simulated data used in this work. Listed runtimes are wall-time using four Tesla A100 GPUs. HMC runtime refers to the time to run an MCMC chain sampling just a single bandpower parameter whereas the MUSE runtime estimates the full set of bandpowers.

the marginal distribution are far more computationally costly than of the joint. It is instead much faster to work with the original joint distribution and simply allow the integral over f to be performed implicitly as part of the MUSE procedure. Because MUSE is exact for Gaussian conditional slices (e.g. Appendix A), this will not introduce any extra approximations. One can in fact view the marginal MAP ϕ estimate of [17] as exactly equivalent to running MUSE with $\theta = \phi$, $z = f$, and fixed cosmological parameters.

B. Lensing bandpowers

As a first check, we run MUSE on a suite of simulated 256×256 pixel 150 deg^2 patches of sky with noise levels similar to those planned for CMB-S4. Although a smaller sky area than the CMB-S4 footprint, we do not expect the accuracy of the MUSE estimate to depend significantly on the size of the observed sky area, since lensing is a fairly local operation (lensing deflections are coherent across only a few degrees). Limiting the size of the dataset allows us to run a large number of simulations and very accurately quantify the properties of the MUSE estimate. To start, we will fix $A^b_E = 1$ and estimate only A^b_ϕ. In the next subsection we will demonstrate estimating both simultaneously. We also consider two versions of this case, one without a pixel mask, and another with a 1° mask around the edges of the field. We refer to this dataset as the CMBS4-150D data, and exact simulation parameters are given in Table I.

We run MUSE as described in Sec. II. Here, the $\hat{z}_{\theta,x}$ which appears in the procedure is the best-fit (f, ϕ) at fixed A^b_ϕ:

$$(\hat{f}_J, \hat{\phi}_J) = \underset{f,\phi}{\rm argmax}\ \log \mathcal{P}(x, f, \phi \,|\, A^b_\phi), \tag{42}$$

We follow previous CMB lensing literature in denoting the joint MAP as $(\hat{f}_J, \hat{\phi}_J)$. The maximization is performed iteratively using the coordinate descent algorithm presented in [2]. We then solve the MUSE equation to obtain the estimate of the bandpower mean, $\hat{A}^b_\phi$, which solves

$$s^{\rm MUSE}(\hat{A}^b_\phi) = 0. \tag{43}$$

For convenience we use AD to compute the gradients of Eqn. (38) with respect to A^b_ϕ, although the analytic gradient is simple to derive as well. We use 100 simulations to compute the expectation value in $s^{\rm MUSE}$, corresponding to a Monte Carlo error of $\lesssim 0.1\sigma$. We find $\sim$15 Broyden iterations are sufficient for the bandpowers to converge.

We note that previous works based on HMC sampling found it necessary to work with a reparameterized form of Eqn. (38) which decorrelated f and ϕ, and removed funnel-like correlation between them and A^b_ϕ (this was denoted the "mixed posterior"). Although we still use mixing for speeding up the computation of the joint MAP estimate, we highlight that no such reparameterization is needed for the MUSE estimate itself, which performs identically whether or not we use mixing. This is due to the excellent performance of MUSE on funnel-like problems which was highlighted in the previous section.

The majority of the runtime is spent computing the joint MAP for the data and for the simulations in each MUSE iteration. As mentioned earlier, it is crucial for performance to use the joint MAP estimate from the previous iteration as a starting point for the joint MAP estimate at the next iteration, both for the data and for the suite of same-seeded simulations. This reduces the overall runtime by almost an order of magnitude, since in later iterations, θ is not changing much and the starting points are extremely close to the final estimate. The entire procedure runs in about 5 minutes on four Tesla A100 GPU for this problem size.

In Fig. 2, we plot the resulting $\hat{A}^b_\phi$ estimates after each Broyden iteration for a simulated masked CMBS4-150D dataset. This shows a typical evolution of the bandpower parameters as the MUSE solution is iteratively obtained. Error bars have been plotted at each step for demonstration, although in practice they would only be calculated for the last step. The final result scatters around the input fiducial model, which is a simple sanity check that the estimate is unbiased. More quantitatively, we check the bias is sufficiently small using simulations. Although the bias is asymptotically zero, it may be non-zero for any finite data vector, and scales as $1/\sqrt{N}$ as demonstrated

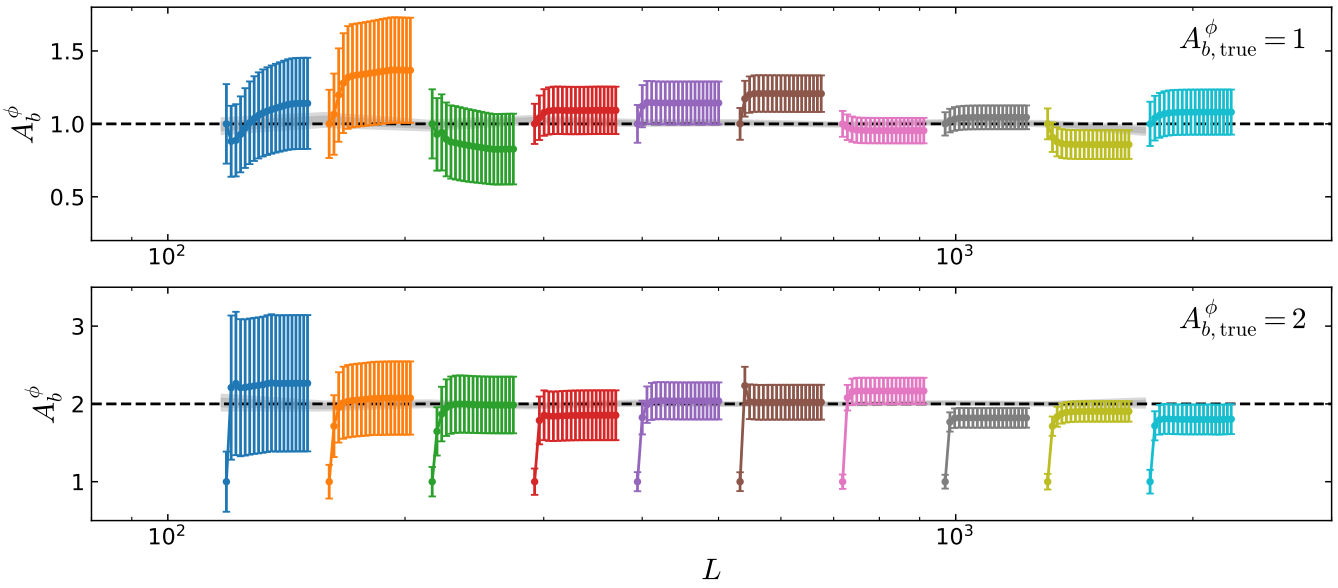

FIG. 2. Examples of the MUSE estimate of the binned bandpower parameters, A^b_ϕ, for a simulated CMBS4-150D dataset, including masking. Each color represents a different bandpower, and within each bandpower, each iterative MUSE step is arbitrarily offset in ℓ for clarity. In the top panel, the true value is $A^b_\phi = 1$ (for all b), and the bottom panel the true value is $A^b_\phi = 2$. Both cases start from an initial guess of $A^b_\phi = 1$. Each panel also includes a gray band showing the $1\,\sigma$ bound on the bias in each bandpower, computed from an average over many simulated realizations (the bias is asymptotically zero, and demonstrated above to be sufficiently near-zero as well for the finite number of modes constraining each bandpower here).

in Sec. II B. Since N is non-trivial to estimate, here we empirically determine the bias directly. The gray band in Fig. 2 shows the mean over 512 simulated MUSE analyses and its $1\,\sigma$ standard error. We find no evidence for any bias at the level of $0.05\,\sigma$ which is afforded by this number of simulations. The bottom panel of Fig 2 shows a case where the fiducial $A^b_\phi = 2$, but the initial starting guess for the MUSE estimate is $A^b_\phi = 1$. The colored error bars show how the estimate iteratively moves towards the higher value of A^b_ϕ. The gray band similarly shows an average over 512 realizations, and demonstrates the bias is consistent with zero for this alternate fiducial model as well. We conclude that down to the noise level of CMB-S4, for similarly wide bandpowers, and for sky area of $150\,\mathrm{deg}^2$ or larger (the number of modes would grow with larger sky area, reducing the estimator bias), the MUSE lensing estimate is effectively unbiased.

Next, we consider the optimality of the estimate. Because the bias is effectively zero, the covariance must satisfy the Cramér-Rao bound. Writing $\Sigma \equiv \Sigma^{\mathrm{MUSE}}$ for brevity, the bound states that for all θ,

$$\Sigma_{bb'} - \mathcal{F}^{-1}_{bb'} \geq 0, \tag{44}$$

where the inequality represents that the left-hand side must be a positive semi-definite matrix, and for an optimal estimate this becomes an equality. At present, there is no exact way to calculate $\mathcal{F}_{bb'}$ for the full set of bandpowers.[6] However, we can obtain $\mathcal{F}$ for an overall amplitude parameter which is derived as a minimum variance combination of $\hat{A}^b_\phi$

$$\hat{A}_\phi \equiv w_b \hat{A}^b_\phi \tag{45}$$

where

$$w_b = \frac{\sum_{b'} C^{-1}_{bb'}}{\sum_{b,b'} C^{-1}_{bb'}} \tag{46}$$

The posterior distribution for this same parameter, $\mathcal{P}(A_\phi \,|\, x)$, can be computed via MCMC using the method described in [24]. If a flat prior on A_ϕ is assumed, this also equals the likelihood, $\mathcal{P}(x \,|\, A_\phi)$. With the likelihood obtained in this manner, the Fisher information can be computed as

$$\mathcal{F} = \left\langle \frac{d^2}{dA^2_\phi} \log \mathcal{P}(x \,|\, A_\phi) \right\rangle_{x \sim \mathcal{P}(x \,|\, A_\phi)} \tag{47}$$

by explicitly averaging over the curvature of the log-likelihood from several chains on different simulated data.

[6] Sometimes the iterative power spectrum forecasts of [32] are taken as a benchmark of optimality. While these are extremely useful, we note that they are only approximations rather than formal calculations of Fisher information, and it is unknown how they perform in the presence of effects like masking which we wish to check here.

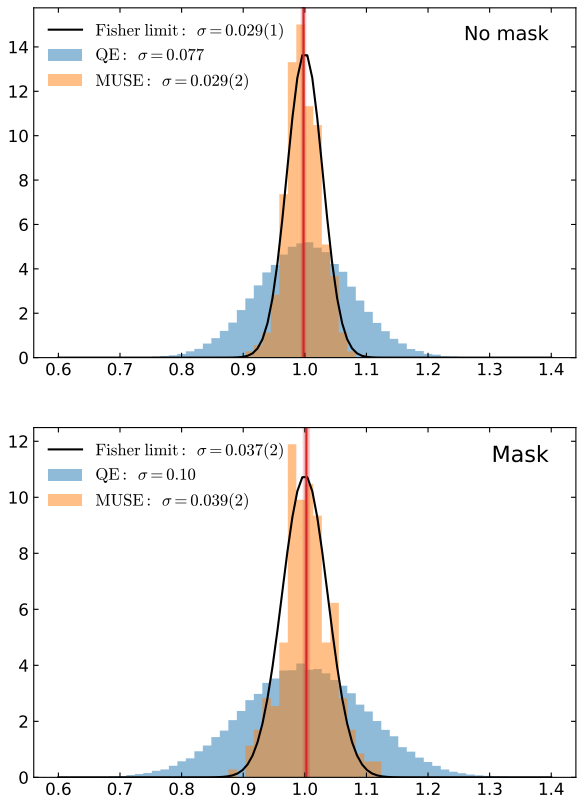

FIG. 3. Empirical scatter from simulations in A_ϕ estimates computed from MUSE lensing bandpowers (orange bars) or quadratic estimate lensing bandpowers (blue bars), as compared to the exact Fisher information computed from MCMC chains (black line). The numbers in parenthesis in the legend give the sampling error on the final digit of these standard deviations due to the finite number of simulations which entered these calculations. The red vertical band gives the mean of the MUSE estimates and the 1 and 2 σ standard error on the mean. This plot demonstrates the MUSE result is consistent with being unbiased and saturates the Cramér-Rao bound (i.e. is effectively optimal) to better than $\sim 10\%$ even in the presence of masking.

This can then be compared with variance of the $\hat{A}_\phi$ estimator computed from simulations.[7]

Fig. 3 shows MUSE $\hat{A}_\phi$ estimates from a suite of 256 simulations. For comparison, results from the quadratic estimator (QE) are also shown for a suite of 10^5 simulations (these are computationally inexpensive to compute). The black curve is a Gaussian with standard deviation given by $\mathcal{F}^{-1/2}$, computed via MCMC chains as just described. Because the scatter in the curvature of the log-likelihood is small (put another way, the error bars on A_ϕ from each chain are only mildly realization-dependent), only a small number of chains are needed to compute the Fisher information to within acceptable Monte Carlo error, here only ten chains.

Independent of whether we apply pixel masking or not (top and bottom panels in Fig. 3), we find the MUSE results are consistent with saturating the Cramér-Rao bound. This is encouraging, because the pixel mask induces additional non-Gaussianity of the latent space posterior (beyond that present due simply to lensing), which one could have been suspected to cause sub-optimality in the MUSE estimator. Evidently, however, these are not large enough to significantly degrade its optimality.

Because $\hat{A}_\phi$ saturates the Cramér-Rao bound, an even stronger statement can be derived, mainly that $\hat{A}^b_\phi$ also saturates the bound. Intuitively, this is because Fisher information is always positive, so if the Fisher information for any function derived from the bandpowers is maximized, the Fisher information in each individual bandpower must also be maximized. More rigorously, consider the covariance and Fisher information for A_ϕ and A^b_ϕ,

$$\Sigma \equiv \sum_{b,b'} w_b \Sigma_{bb'} w_{b'} \tag{48}$$

$$\mathcal{F}^{-1} \equiv \sum_{b,b'} w_b \mathcal{F}^{-1}_{bb'} w_{b'} \tag{49}$$

Since we have empirically verified that $\Sigma = \mathcal{F}^{-1}$ (up to Monte Carlo error), we have that

$$\sum_{b,b'} w_b \big(\Sigma_{bb'} - \mathcal{F}^{-1}_{bb'}\big) w_{b'} = 0 \tag{50}$$

$$\sum_{b} v_b^2 \big(\tilde{\Sigma}_b - \tilde{\mathcal{F}}_b^{-1}\big) = 0 \tag{51}$$

where in the second equation we have simultaneously diagonalized the covariance and Fisher matrix (possible because both are positive definite), yielding new weights, v and diagonal entries, $\tilde{\Sigma}_b$ and $\tilde{\mathcal{F}}_b$. Since the Cramér-Rao bound guarantees that $\tilde{\Sigma}_b - \tilde{\mathcal{F}}_b^{-1}$ is a positive number, it follows that every term in the sum in Eqn. (51) must be zero individually, and hence that $\Sigma_{bb'} = \mathcal{F}^{-1}_{bb'}$.

C. Joint lensing and delensed bandpower estimates

Next, we consider simultaneously estimating the lensing potential power spectrum and the unlensed E mode bandpowers. Doing so is nearly identical to the previous section except that now A^b_E is an estimated parameter in addition to A^b_ϕ. This means the MAP estimates that are part of MUSE now take place at varying values of A^b_E as well,

$$\hat{f}_J, \hat{\phi}_J \equiv \underset{f,\phi}{\text{argmax}}\ \log \mathcal{P}(x, f, \phi \,|\, A^b_\phi, A^b_E), \tag{52}$$

[7] In theory, one could use HMC sampling to infer the full bandpower posterior, $\mathcal{P}(A^b_\phi \,|\, d)$, and estimate the Fisher information matrix in this way. This may be possible in practice, but has yet to be demonstrated, and was not attempted here due to its additional computational cost.

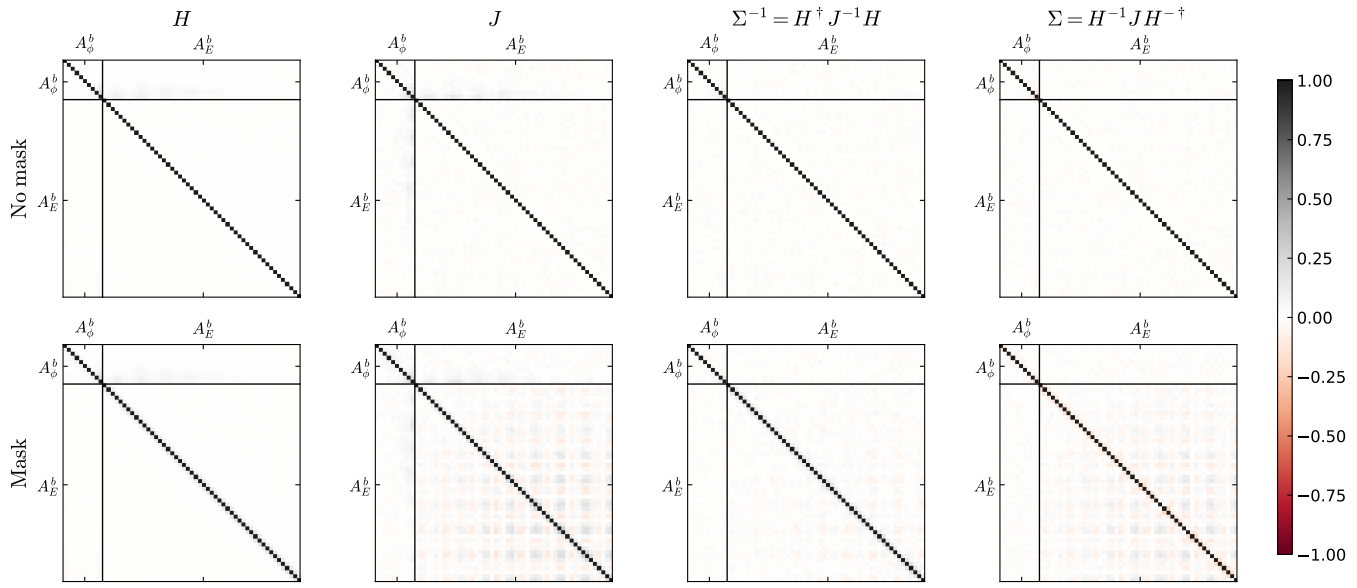

FIG. 4. The J and H matrices (defined in Eqs. 16 and 17), and the bandpower covariance and its inverse, Σ and Σ^{-1}. The CMBS4-150D configuration is assumed, and bottom and top rows correspond to with and without masking, respectively. Note the asymmetry visible in H, demonstrating that $s^{\rm MUSE}$ is non-conservative for the lensing problem. Additionally, we show in the text that the checkerboard pattern in the far off-diagonals does not impact cosmological parameter inferences at more than a few percent σ.

and the estimate is the solution jointly over both sets of bandpowers:

$$s^{\rm MUSE}(\hat{A}_\phi^b, \hat{A}_E^b) = 0 \tag{53}$$

We begin by showing in Fig. 4 the H, J, Σ and Σ^{-1} matrices computed via Eqns. (19-20) for the joint (A_ϕ^b, A_E^b) parameter space, with masked and unmasked cases given in the bottom and top rows. These have been computed using 2048 simulations for the J matrix and 64 simulations for the H matrix (the latter which we find is only very weakly realization-dependent). Determining the joint covariance, Σ, has been a challenging problem for the field even when only the quadratic estimate is used to estimate the lensing potential [33–35]. Solutions when using more optimal estimates of the lensing potential have thus far been developed only under simple forecasting assumptions [21, 22]. The result here shows this is now also feasible for a map-level procedure which accounts for masking.

In the top row (unmasked case) we see there are very few significantly non-zero entries in any of the off-diagonals for any of the matrices. The H matrix does have some small non-zero entries in the upper $A_\phi^b \times A_E^b$ block. As per Eqn. (17), this corresponds to having injected power into the unlensed signal in some particular E mode bandpower, and the gradient at the MAP having responded instead with a change to a ϕ bandpower. This is empirical proof that $s^{\rm MUSE}$ is non-conservative (due to the lensing-induced posterior non-Gaussianity), as otherwise the H matrix would be symmetric. In the bottom row (masked case) we additionally see two features: 1) a negative cross-covariance between neighboring E-mode bandpowers, and 2) a "checkerboard" pattern even for distant E-mode bandpower bins. The former effect is typical of induced mode coupling due to the mask. The latter effect is also expected and has been noted by [21, 22]. It arises because the biggest effect of lensing on the E-mode power spectrum is a smoothing of the peaks, and this effect is sourced mainly by lensing modes near the peak of the lensing potential power spectrum ($L \sim 100$). Thus, depending on if these lensing modes fluctuate high or low, the entire E-mode spectrum will be over or under smoothed with respect to the mean theory expectation. This then induces these correlations across very distant neighbors.

Although with MUSE it is entirely possible to quantify these distant off-diagonal correlations, it is of interest to what extent they actually impact parameter inference and whether they can simply be ignored. To do so, we propagate the MUSE bandpower covariance, $\Sigma_{bb'}$, to the Fisher information matrix on cosmological parameters, $\mathcal{F}_{\alpha\beta}$, via

$$\mathcal{F}_{\alpha\beta} = \frac{d\log C_b}{d\alpha} \Sigma^{-1}_{bb'} \frac{d\log C_{b'}}{d\beta} \tag{54}$$

where α, β represent cosmological parameters from a standard set $\{\omega_{\rm b}, \omega_{\rm m}, \Sigma m_\nu, \theta_s, A_{\rm s}, n_{\rm s}\}$, and the log appears because in our definition, Σ is the covariance matrix for bandpower amplitudes, rather than bandpowers themselves. We then invert $\mathcal{F}_{\alpha\beta}$ for the case where the full $\Sigma_{bb'}$ is used vs. where only the first off-diagonals are kept, and examine the square root of the diagonal

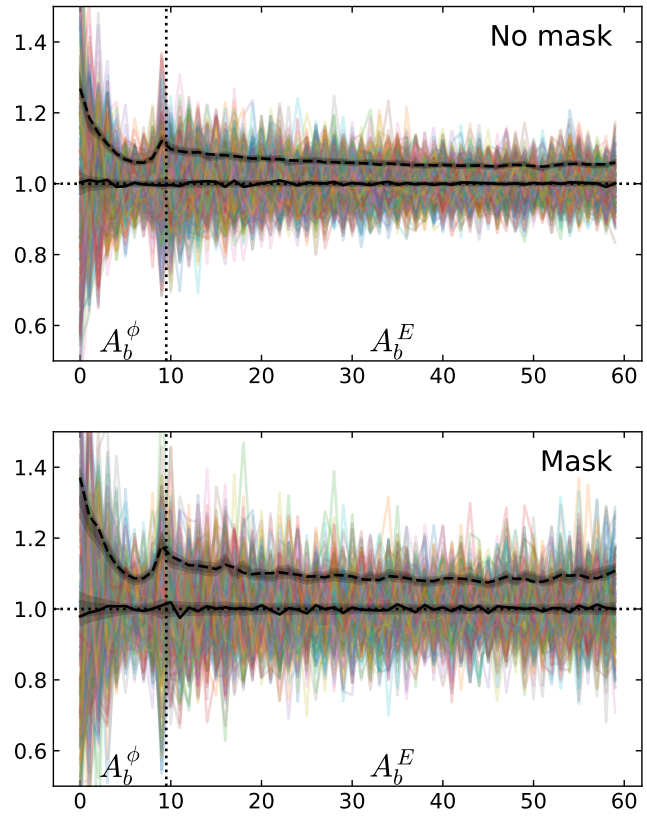

FIG. 5. Colored lines show MUSE estimates of (A_ϕ^b, A_E^b) for a large suite of simulations. The CMBS4-150D configuration is assumed, and bottom and top panels correspond to with and without masking, respectively. The black line around 1 is the mean over all simulations, and the black band around this is standard error on the mean, confirming that MUSE provides an unbiased estimate of lensing bandpowers joint with delensed E modes. The black dashed lines are the bandpower errors computed from $\sqrt{\text{diag}(\Sigma^{\text{MUSE}})}$, and the black bands around this are the empirical scatter and its standard error, demonstrating that the MUSE covariance prescription accurately reproduces the observed scatter.

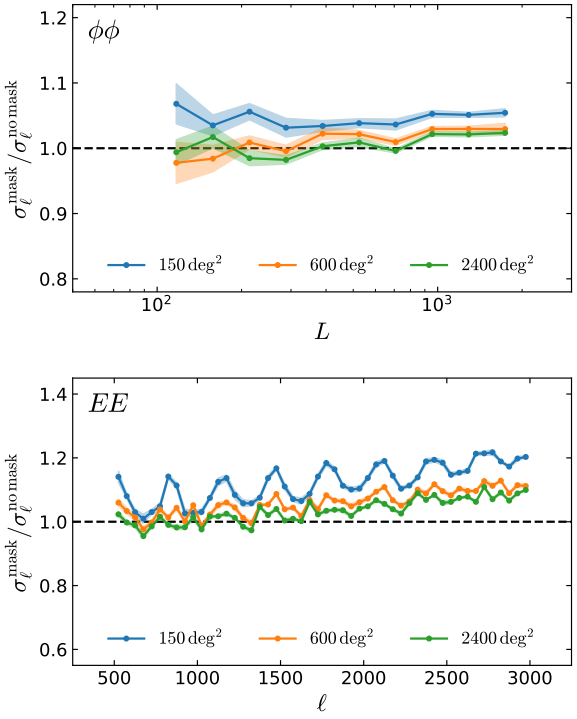

FIG. 6. Demonstration of how masking leads to excess variance in inferences of lensing bandpowers and delensed E mode bandpowers as compared to expectation from a simplistic f_{sky}-scaled periodic sky forecast. Each panel shows the ratio of error bars computed from $\sqrt{\text{diag}(\Sigma^{\text{MUSE}})}$ between a masked case and a periodic sky case with identical effective f_{sky}. Error bands show the $1\,\sigma$ standard error on these quantities estimated via bootstrap resampling, as the covariances are themselves computed via Monte Carlo. As the size of the field grows and the mask boundary is a smaller percentage of the total field, any differences reduce. By a 2400 deg^2 (close to the smallest planned CMB-S4 field), the impact is $\lesssim 10\%$, providing important confirmation of such forecasts even if masking was ignored.

entries (i.e. the forecasted standard deviation error on parameters). We find that no parameter error changes by more than 1%. This can also be interpreted as that any potential biases due to ignoring these entries would be less than 1% of the parameter error. We thus conclude that for the CMBS4-150D configuration, it would be safe to ignore these entries. Note that this includes ignoring the $A_\phi^b \times A_E^b$ cross-covariance entirely. Additionally, because lensing is a relatively local operation, this statement is not expected to depend significantly on the size of the field, thus likely holds for the larger sky fractions targeted by CMB-S4. We also check whether keeping only the diagonal entries of the covariance could be a sufficiently good approximation. Here, we find a 10% change in parameter error and corresponding 10% possible bias. This is likely too large to be acceptable, therefore (unsurprisingly) one cannot ignore the nearby-bin mode coupling induced due to the mask.

We also wish to verify empirically that the joint estimate of (A_ϕ^b, A_E^b) is unbiased and that its covariance, Σ^{MUSE}, calculated based on Eqn. (19), accurately represents the empirical scatter of the estimate. This is demonstrated in Fig. 5. Colored lines show a suite of simulated joint estimates, and the solid and dashed lines show that their mean recovers the input theory model on average, and that their scatter matches the prediction based on Σ^{MUSE}, respectively. This is the case whether masking is used or not (top and bottom panels).

Thus far, all forecasting results, including all performed for CMB-S4, have only approximately accounted for the effects of pixel masking by assuming that constraints scale with $\sqrt{f_{\text{sky}}}$, where f_{sky} is the fraction of the total sky which is observed. While this is expected to be reasonably accurate, particularly for small-scale power spectrum estimation, the story is more compli-

cated for lensing reconstruction and delensing, where masking mixes some of the same lensing modes which in turn also impact delensing. Using MUSE, we perform an important check of the accuracy of the $f_{\rm sky}$ scaling assumed in these forecasts.

Specifically, we compare the MUSE covariance for two cases, here in the CMBS4-150D configuration. In one case we apply a pixel mask and in the other we shrink the field so that it is the same effective $f_{\rm sky}$ as the masked case, but otherwise do not apply any masking. In Fig. 6, we show the ratio of $\sqrt{\rm diag}(\Sigma^{\rm MUSE})$ for the two cases. Additionally, we also scale up the size of the field to larger more realistic sizes (while keeping the size of the 1° border mask unchanged). We expect that as we increase the field size, any masking effects beyond $f_{\rm sky}$ should reduce, as the masking impacts a smaller and smaller fraction of the total modes which enter the estimate. This is indeed what we find, showing that while for the smallest $\sim 150\,{\rm deg}^2$ field, an $f_{\rm sky}$-scaled forecast could be as large as 10% too optimistic in the lensing reconstruction error bars and 20% optimistic in the delensed E mode error bars at high-ℓ, by the time we reach realistic field sizes of a few thousand square degrees, the impact is almost nothing for ϕ and $\lesssim 10\%$ for delensed E. This is an encouraging result which suggests masking will not pose any unexpected problems for lensing analyses of CMB-S4 data. In the future, it will be interesting to compare full-sky methods like those of [21, 22] directly to MUSE to quantify the impact of realistic data effects even further.

D. Realistic example on SPT-3G data

The previous section examines the properties of the MUSE lensing estimate on a relatively small patch of sky where large suites of simulated analyses are computationally convenient. We now demonstrate the feasibility of MUSE on a much larger region of sky, representative of the deep fields of ongoing and upcoming CMB surveys. Specifically, we will consider a $1500\,{\rm deg}^2$ region with the noise levels and masking expected for the upcoming SPT-3G survey [36, specification given in Table. I]. This is also similar in size to the region expected to be probed (to deeper noise levels) with the CMB-S4 survey for the main purpose of primordial gravitational wave detection [14]. We continue to work in the flat-sky approximation, which begins to break down near patches of sky of this size. We will comment later on a simple way to account for this within the MUSE procedure.

For visualization, Fig. 7 shows a typical SPT-3G pixel mask, as well as typical $\hat{f}_J$, $\hat{\phi}_J$, and $\hat{\kappa}_J = -\nabla^2\hat{\phi}_J/2$ maps (only the E component of $\hat{f}_J$ is shown since we do not estimate bandpowers of B, but the MAP reconstruction of B is also non-zero and used in the algorithm). In the bottom two panels, the left half of the image shows the result when imposing the super sample prior defined in Eqn. (41), whereas the right half show the result with-

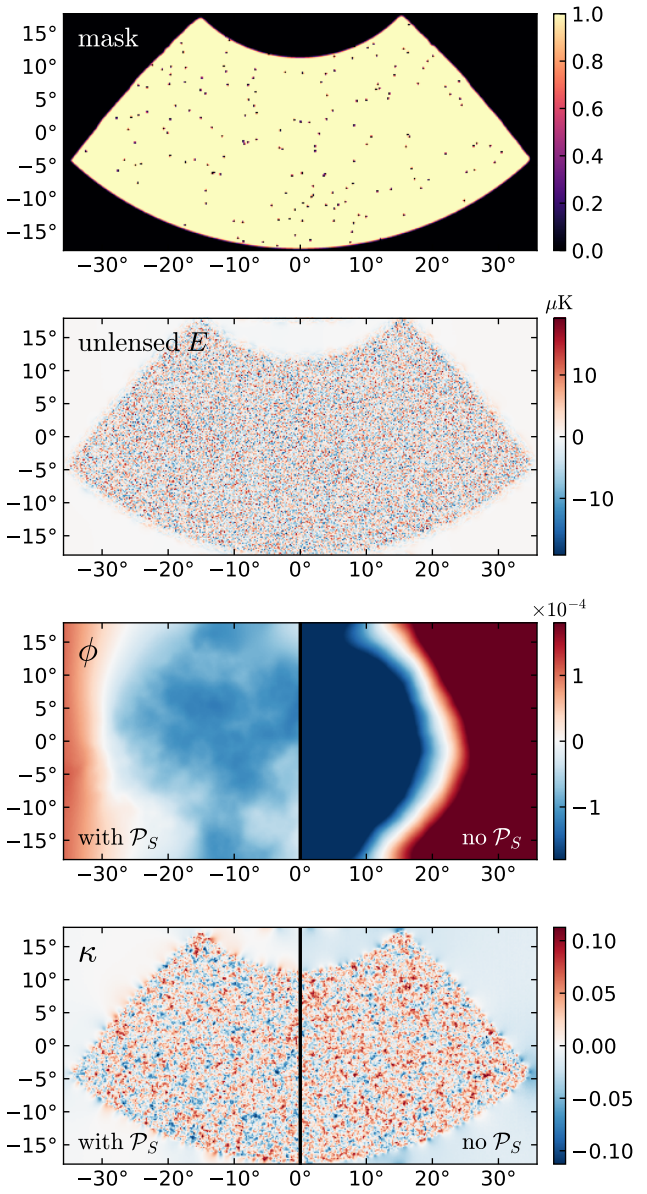

FIG. 7. (Top panel) The pixel mask used in the analysis presented in Sec. IV D, which features $1500\,{\rm deg}^2$ of simulated SPT-3G data. (Next two panels) Typical joint MAP estimates of unlensed E and ϕ, which enter the MUSE estimate as the point in the latent parameter space at which gradients with respect to CMB and lensing bandpowers are taken. (Bottom panel) The $\kappa = -\nabla^2\phi/2$ corresponding to the ϕ panel. In the bottom two panels, the left half of the image shows the result when imposing the super-sample prior (Eqn. 41), whereas the right half shows the result without any additional prior. The impact of this prior is to reduce the mean-field feature to levels small enough (though not necessary zero) as to render the MUSE estimate unbiased and near-optimal.

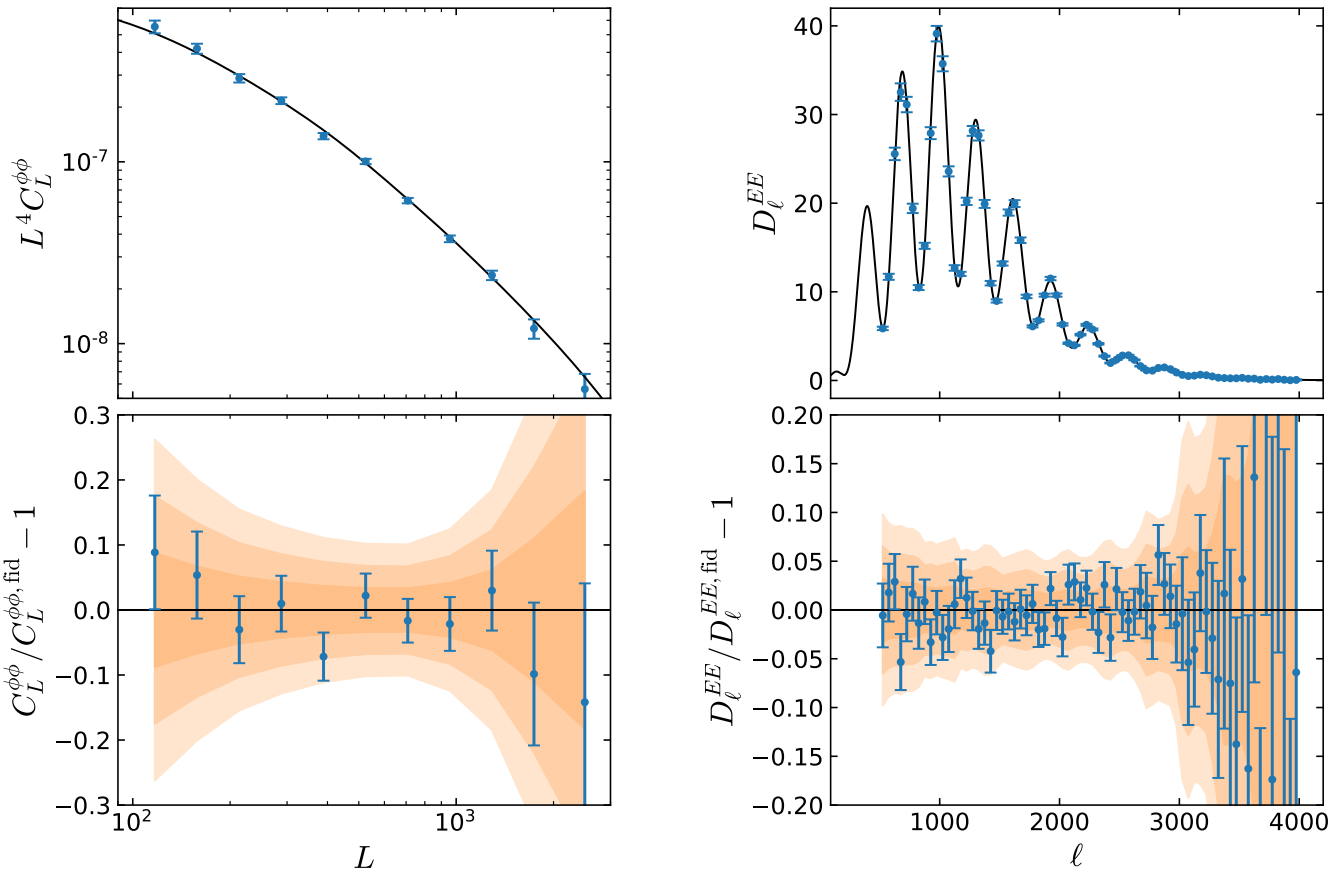

FIG. 8. MUSE estimates of the lensing potential and unlensed E-mode bandpowers for a 1500 deg^2 simulated SPT-3G dataset. The black line shows the fiducial theory. The top panels show the absolute estimate, while the bottom panels show the fractional difference against the truth in units of percent. The orange shaded regions in the bottom panel are 1, 2, and 3 σ error bands (they are the same as the size of the error bars, but easier to compare across different multipoles). This full analysis required around one hour of computing time, a drastic improvement over HMC sampling.

out this prior. The impact of the prior is to reduce the mean-field feature, which is visible as a large residual in ϕ, and an additive offset in κ (note the negative average κ in the masked regions, and the positive average κ in the unmasked region, barely visible as a slight preponderance for hot spots as opposed to cold spots there). The mean-field is a real feature of the joint MAP and arises due to any masking or filtering that breaks statistical isotropy or introduces non-Gaussianity (it is also present in the masked CMBS4-150D analyses in the previous subsections, and not unique to the SPT-3G mask). It poses problems for the MUSE estimate as it leaks a large number of unconstrained modes into bandpowers of interest. Put another way, a κ offset corresponds to an overall magnification which brings modes from outside of the mask into view, but which are unconstrained by data and hence lack the asymptotic guarantees needed by MUSE. The super sample prior stops enough of these modes from entering and potentially leading to biases. We also note that the removal is not perfect—nor does it need to be—to render MUSE effectively unbiased and near-minimum variance, as demonstrated in the previous subsection.

Computing estimates of $\hat{A}_\phi^b$ and $\hat{A}_E^b$ for the SPT-3G simulation proceeds identically as in the previous subsection. A typical MAP solution without a starting guess takes around two minutes to complete on a Tesla A100, and the entire MUSE estimate is under an hour on a handful of GPUs. By contrast, a naive scaling of existing HMC methodology would predict over a week for a sufficiently converged chain. An estimate and error-bands for one particular simulation are shown in Fig. 8.

V. CONCLUSION

In this paper we have described the MUSE algorithm, a generic method for hierarchical Bayesian inference. MUSE is based on an approximation to the marginal score which is extremely fast to compute relative to exact methods. It is most applicable to problems where one needs to marginalize over a very high-dimensional

latent space, but the final constraints of interest involve a small number of fairly well-constrained parameters. In the limit of a perfectly Gaussian latent space, MUSE becomes exact and is equivalent to the MMLE. It also performs extremely well on funnel problems, which are challenging for other methods such as HMC. The only technical requirements of the MUSE estimate are that one has access to gradients of the joint likelihood function and one can generate simulations from the forward model. This requirement is usually satisfied for a broad range of problems, including the common case of problems defined via probabilistic programming languages. In its current form, MUSE is interpretable as either a frequentist estimator for parameters of interest which is asymptotically normal and unbiased, or as an approximate Bayesian procedure yielding Gaussianized marginal posterior inferences.

Computing the MUSE estimate is straight forward. To summarize, it involves:

1. Finding the MAP estimate of the latent space variables, z, given a starting guess for θ, and computing the gradient with respect to θ at this z.

2. Computing this same quantity on suite of data simulations from the forward model generated given the current θ, and differencing the average of these from the data value.

3. Iteratively solving for where this difference is zero, yielding the MUSE estimate.

4. Computing the H and J matrices according to Eqn. (19), then combining them to form the covariance, which involves only more MAP gradients of the same kind used in the estimate itself (and which can be sped up if second order AD is available).

One can view the effectiveness of MUSE as stemming from having replaced the difficult problem of high-dimensional integration with the much more tractable problem of high-dimensional optimization. In this sense, it is similar in spirit to both VI and EM, although differs from VI in that no surrogate distributions need to be chosen by the user. Of course, the ability to choose a well-tailored surrogate distribution is, in another sense, a strength of VI which makes it more generic. However, for VI to be tractable in high-dimensions, the surrogate distribution must often necessarily be a Gaussian, or in the case of mean-field VI, an uncorrelated Gaussian. Conversely, MUSE implicitly deals with all correlations in the latent space, and does not necessarily correspond to a Gaussian latent space approximation. One can also view MUSE as an approximate EM procedure, where the averaging over simulations in Eqn. (10) serves as the expectation step, and the root-finding over θ is analogous to the maximization step. MUSE is also similar, but not equivalent, to the Laplace approximation, differing in that the Hessian of the latent space never needs to be computed. In general, MUSE excels on high-dimensional problems because no dense high-dimensional operators are ever needed. In a more general sense, one could consider MUSE a form of SBI, as it requires only the ability to generate forward simulations, along with the ability to compute gradients of the joint likelihood, which is often automatically satisfied if forward simulations are available.

We then applied MUSE to the problem of estimating the unlensed CMB power spectra and gravitational lensing potential power spectra from realistic CMB data. The original works demonstrating that improvements over the standard QE procedure are possible [11, 12] have motivated many followups attempting to render the method ready to apply to data. To consider the optimal lensing data analysis problem solved, we view it as necessary to demonstrate 1) joint estimation of bandpowers or parameters controlling both the lensing potential and the unlensed CMB spectra 2) proof that the estimate is optimal, for example by comparing against exact posterior distributions or against exact Fisher matrix calculations 3) demonstration that the procedure works and remains optimal even in the presence of masking and 4) is reasonably fast. Some methods have satisfied a subset of these requirements, but thus far MUSE is the first to satisfy them all. Coupled with its conceptual simplicity, we therefore believe it is a very promising path forward for CMB lensing and delensing analysis.

Many CMB cosmologists intuitively view optimal lensing analysis as performed by an "iterative quadratic estimate," based on popular discussion in [32]. Although there is no unique definition of the iterative QE, and the original discussion was only a heuristic forecasting procedure, it is true that many optimal lensing results take the form of iterating an estimate that is quadratic in the data. MUSE can be afforded this interpretation too, since each gradient step used in obtaining the joint MAP that is part of the calculation is quadratic in the data, and this is iterated until convergence. The additional pieces of MUSE lensing, mainly debiasing the score with simulations, performing the root-finding iterations, and the covariance prescription, can be considered as implicitly performing a cosmology-dependent power spectrum debiasing and noise quantification which is sometimes imagined as part of what a map-level iterative QE would entail.

Our results here demonstrated that MUSE is effectively unbiased and optimal for the noise levels and sky areas of all upcoming surveys, and in the presence of masking. A key but simple development which allowed MUSE to work on CMB lensing was the addition to the super-sample prior discussed in Sec. IV, which reduces the mean-field feature in the joint MAP, but is otherwise very non-informative in terms of the marginal posterior. We showed that at CMB-S4 noise levels, correlations between bandpowers of ϕ and bandpowers of unlensed E can be ignored with minimal impact on resulting parameters even when masking is present. However, we caution

that we did not consider higher noise levels nor correlations with lensed E estimates which are known to be larger [22]. We also compared simple $f_{\rm sky}$-scaled forecasts given CMB-S4 noise levels with ones more exactly accounting for pixel masking. We found that $f_{\rm sky}$-scaled forecast are overly optimistic by as much as 20% given very small patches of sky, but this reduces to $\lesssim 10\%$ for larger realistic fields like the SPT-3G survey or even the smallest CMB-S4 fields. This is an assuring confirmation of forecasted next-generation constraints.

Our code for this work assumed the flat-sky approximation, but the MUSE algorithm is generic and directly generalizes to the curved sky as well. Although HMC sampling on the curved sky is at present slightly out of reach, MUSE is much faster and runs more easily on the curved sky, and we expect to update the software accordingly in the near future. It is interesting to note, however, that there is an alternate and easier way to deal with sky curvature. The proof that MUSE is asymptotically unbiased does not rely on $s^{\rm MAP}$ being exact. Instead, the only requirement is that the simulations which are averaged over in $s^{\rm MUSE}$ accurately describe the data distribution. Thus, as long as the forward simulations are generated on the curved sky, the much-costlier MAP solutions can still be computed on the flat-sky. This will not add any additional bias to the MUSE estimate, only potentially excess variance, which will be captured in the H and J matrices. This can serve as a fast substitute for intermediate sky areas where the flat-sky approximation fails but any excess variance is still small. Additionally, the same argument allows us incorporate any of a number of other effects, such as foregrounds or other instrumental effects, in the simulations, but not in the MAP calculation, and still recover unbiased results. Future extensions can proceed by first incorporating these effects (possibly with free parameters which we will then infer) in the simulation model, and only accounting for them in the MAP when the excess variance is deemed too large.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Numbers 1814370 and NSF 1839217, by NASA under Grant Number 80NSSC18K1274, and by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research under Contract No. DE-AC02-05CH11231 at Lawrence Berkeley National Laboratory. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank Emmanuel Schaan, Vanessa Böhm, and Selim Hotinli for useful discussions, and Ethan Anderes for the particular interpretation of the mean-field given in Sec. IV D.

Appendix A: MUSE is exact for a Gaussian latent space

The MUSE gradient in Eqn. (10) is approximate in general, but is exact if the Hessian is constant, which requires a Gaussian likelihood with linear dependence of the data model on the latent space, as well as a Gaussian prior. A typical case where this arises is when estimating parameters which control a signal covariance in the "Wiener filter" problem. To help build intuition, here we demonstrate explicitly that MUSE gives an exact answer to this problem, which also has an analytic solution.

Consider the case where some data, x, is the sum of signal and noise, s and n, both of which are Gaussian with covariances $\mathbb{S}(\theta)$ and $\mathbb{N}$, respectively, the former depending on parameters which we wish to infer, θ. The signal s plays the role of latent variables z, i.e. these are the variables that need to be marginalized over. In this case, the marginal is analytic and the posterior distribution for θ is

$$2\log\mathcal{P}(\theta\,|\,x) \propto -x^\dagger\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}x - \log\det\big(\mathbb{S}(\theta)+\mathbb{N}\big) \tag{A1}$$

The gradient of the RHS, which can be used to iteratively step to the MMLE, is

$$2g = \frac{d}{d\theta}\left[-x^\dagger\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}x - \log\det\big(\mathbb{S}(\theta)+\mathbb{N}\big)\right] = \tag{A2}$$

$$x^\dagger\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}\frac{d\mathbb{S}}{d\theta}\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}x - \mathrm{tr}\left[\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}\frac{d\mathbb{S}}{d\theta}\right]. \tag{A3}$$

Next, consider the MUSE gradient from Eqn. (10) for the same problem. This uses the joint distribution of x and s,

$$2\log\mathcal{P}(s,x\,|\,\theta) = -\frac{(d-s)^2}{\mathbb{N}} - s^\dagger\mathbb{S}(\theta)^{-1}s - \left[\log\det\mathbb{S}(\theta) + \log\det\mathbb{N} + (n_s+n_d)\log 2\pi\right]. \tag{A4}$$

The MAP estimate at fixed θ, recognizable as the Wiener filter of the data, is

$$\hat{s}_{\mathrm{MAP}|\theta} = \underset{s}{\mathrm{argmax}}\ \log\mathcal{P}(s,x\,|\,\theta) = \mathbb{S}(\theta)\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}x \tag{A5}$$

The first term in Eqn. (10) is the gradient of Eqn. (A4) evaluated at the MAP estimate from Eqn. (A5). This is

$$\frac{d}{d\theta}\left[-\frac{(x-s)^2}{\mathbb{N}}-\frac{s^2}{\mathbb{S}(\theta)}-\log\det\mathbb{S}(\theta)\right]\Bigg|_{\hat{s}_{\mathrm{MAP}|\theta}} = \left[s^\dagger\mathbb{S}(\theta)^{-1}\frac{d\mathbb{S}}{d\theta}\mathbb{S}(\theta)^{-1}s-\mathrm{tr}\,\mathbb{S}(\theta)^{-1}\frac{d\mathbb{S}}{d\theta}\right]\Bigg|_{\hat{s}_{\mathrm{MAP}|\theta}} \tag{A6}$$

$$= x^\dagger\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}\frac{d\mathbb{S}}{d\theta}\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}x-\mathrm{tr}\,\mathbb{S}(\theta)^{-1}\frac{d\mathbb{S}}{d\theta} \tag{A7}$$

The second term in Eqn. (10) subtracts the average of Eqn. (A7) over data, x. Since the trace in Eqn. (A7) does not depend on data, this piece will cancel. The remaining piece is,

$$\left\langle x^\dagger\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}\frac{d\mathbb{S}}{d\theta}\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}x\right\rangle_{x\sim\mathcal{N}(0,\mathbb{S}(\theta)+\mathbb{N})} = \mathrm{tr}\left[\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}\frac{d\mathbb{S}}{d\theta}\right] \tag{A8}$$

where we have made use of the identity that $\mathrm{tr}\,\mathbb{A} = \langle z^\dagger\mathbb{A}z\rangle_{z\sim\mathcal{N}(0,\mathbb{I})}$. This gives a full MUSE gradient of

$$2s^{\mathrm{MUSE}} = x^\dagger\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}\frac{d\mathbb{S}}{d\theta}\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}x-\mathrm{tr}\left[\big(\mathbb{S}(\theta)+\mathbb{N}\big)^{-1}\frac{d\mathbb{S}}{d\theta}\right] \tag{A9}$$

We see this is identical to Eqn. (A2), confirming that in this Gaussian case, MUSE gives the exact gradient and hence gives the exact MMLE upon iteratively stepping in the gradient direction. The key feature of MUSE, made explicit in this example, is the use of Monte-Carlo to compute the gradient of the log-determinant which appears in the marginal posterior of θ, phrased in a conceptually straightforward way which only requires computing MAP estimates and gradients of the joint posterior of θ and s.

It is also useful to show that marginal over the latent space is required and that MLE/MAP fails in these examples where the dimensionality of latent space equals that of the data. One may for example think that evaluating MAP+MLE of s and θ simultaneously would be sufficient. One can see from equation A4 that maximizing all at the same time leads to solution $s^2 = \mathbb{S} = 0$, such that $s^2 \propto \mathbb{S}$, which is clearly the wrong solution. This is because MLE/MAP estimators are only asymptotically unbiased, which is not satisfied if $n_d = n_s$. Once we marginalize out s we are left with n_θ parameters where $n_\theta \ll n_d$, and we can use MLE in the asymptotic limit.

Appendix B: Simplification of J and H computation

In this appendix we show how to simplify the J and H matrices defined in Eqs. (16) and (17) to their final form in Eqn. (19). The J matrix is defined in Eqn. (16) as

$$J_{ij} = \left\langle s_i^{\mathrm{MUSE}}(\theta^*,\{x_n\})\, s_j^{\mathrm{MUSE}}(\theta^*,\{x_n\})\right\rangle_{x_n\overset{\mathrm{iid}}{\sim}\mathcal{P}(x\,|\,\theta^*)}, \tag{B1}$$

where

$$s_i^{\mathrm{MUSE}}(\theta^*,\{x_n\}) = \frac{1}{N}\sum_{n=1}^{N} s_i^{\mathrm{MAP}}(\theta^*,x_n) - \left\langle s_i^{\mathrm{MAP}}(\theta^*,x')\right\rangle_{x'\sim\mathcal{P}(x'\,|\,\theta^*)}. \tag{B2}$$

Substituting this in and dropping the arguments to the summation, to s^{MUSE}, and to the expectation value for clarity yields,

$$\left\langle\left(\tfrac{1}{N}\textstyle\sum s_i^{\mathrm{MAP}} - \left\langle s_i^{\mathrm{MAP}}\right\rangle\right)\left(\tfrac{1}{N}\textstyle\sum g_j^{\mathrm{MAP}} - \left\langle g_j^{\mathrm{MAP}}\right\rangle\right)\right\rangle \;\rightarrow\; \left\langle s_i^{\mathrm{MAP}}\, s_j^{\mathrm{MAP}}\right\rangle - \left\langle s_i^{\mathrm{MAP}}\right\rangle\left\langle s_j^{\mathrm{MAP}}\right\rangle \tag{B3}$$

where in the second step we have taken the limit $N\to\infty$. This yields the result in Eqn. (19) (which is written there with the arguments included).

The H matrix is defined in Eqn. (17) as:

$$H_{ij} = \left\langle\frac{ds_i^{\mathrm{MUSE}}}{d\theta_j}(\theta^*,\{x_n\})\right\rangle_{x_n\overset{\mathrm{iid}}{\sim}\mathcal{P}(x\,|\,\theta^*)}. \tag{B4}$$

Substituting and simplifying, we find

$$\left\langle \frac{1}{N}\sum_{n=1}^{N}\frac{ds_i^{\rm MAP}}{d\theta_j}(\theta^*,x_n)\right\rangle_{x_n\overset{\rm iid}{\sim}\mathcal{P}(x\,|\,\theta^*)} - \frac{d}{d\theta_j}\left\langle s_i^{\rm MAP}(\theta,x')\right\rangle_{x'\sim\mathcal{P}(x'\,|\,\theta)}\Bigg|_{\theta=\theta^*} \tag{B5}$$

$$= \left\langle \frac{ds_i^{\rm MAP}}{d\theta_j}(\theta^*,x)\right\rangle_{x\sim\mathcal{P}(x\,|\,\theta^*)} - \frac{d}{d\theta_j}\left[\left\langle s_i^{\rm MAP}(\theta,x')\right\rangle_{x'\sim\mathcal{P}(x'\,|\,\theta)}\right]\Bigg|_{\theta=\theta^*} \tag{B6}$$

$$= -\frac{d}{d\theta_j}\left[\left\langle s_i^{\rm MAP}(\theta^*,x)\right\rangle_{x\sim\mathcal{P}(x\,|\,\theta)}\right]\Bigg|_{\theta=\theta^*} \tag{B7}$$

where the last line follows from noting that there are two chain rule terms arising from the term in brackets in Eqn. (B6): one where the derivative acts on the θ inside the expectation value, and another when it acts on the θ controlling the distribution over which the expectation value is taken. The first of these chain rules terms is canceled by the first the term in Eqn. (B6), leaving only the second, which yields the result in Eqn. (19).

[1] J. Neyman and E. L. Scott, Econometrica **16**, 1 (1948).
[2] M. Millea, E. Anderes, and B. D. Wandelt, Physical Review D **100**, 023509 (2019).
[3] M. Betancourt, arXiv:1701.02434 [stat] (2017), arXiv:1701.02434 [stat].
[4] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt, arXiv:1711.05597 [cs, stat] (2018), arXiv:1711.05597 [cs, stat].
[5] K. Cranmer, J. Brehmer, and G. Louppe, Proceedings of the National Academy of Sciences **117**, 30055 (2020).
[6] U. Seljak, G. Aslanyan, Y. Feng, and C. Modi, Journal of Cosmology and Astroparticle Physics **2017**, 009 (2017), arXiv:1706.06645.
[7] B. Horowitz, U. Seljak, and G. Aslanyan, Journal of Cosmology and Astroparticle Physics **2019**, 035 (2019), arXiv:1810.00503.
[8] A. Lewis and A. Challinor, Physics Reports **429**, 1 (2006), arXiv:astro-ph/0601594.
[9] M. Zaldarriaga and U. Seljak, Physical Review D **59**, 123507 (1999), arXiv:astro-ph/9810257.
[10] W. Hu and T. Okamoto, The Astrophysical Journal **574**, 566 (2002).
[11] C. M. Hirata and U. Seljak, Physical Review D **67** (2003), 10.1103/PhysRevD.67.043001, arXiv:astro-ph/0209489.
[12] C. M. Hirata and U. Seljak, Physical Review D **68** (2003), 10.1103/PhysRevD.68.083002, arXiv:astro-ph/0306354.
[13] U. Seljak and C. M. Hirata, Physical Review D **69**, 043005 (2004).
[14] K. N. Abazajian, P. Adshead, Z. Ahmed, S. W. Allen, D. Alonso, K. S. Arnold, C. Baccigalupi, J. G. Bartlett, N. Battaglia, B. A. Benson, C. A. Bischoff, J. Borrill, V. Buza, E. Calabrese, R. Caldwell, J. E. Carlstrom, C. L. Chang, T. M. Crawford, F.-Y. Cyr-Racine, F. De Bernardis, T. de Haan, S. di Serego Alighieri, J. Dunkley, C. Dvorkin, J. Errard, G. Fabbian, S. Feeney, S. Ferraro, J. P. Filippini, R. Flauger, G. M. Fuller, V. Gluscevic, D. Green, D. Grin, E. Grohs, J. W. Henning, J. C. Hill, R. Hlozek, G. Holder, W. Holzapfel, W. Hu, K. M. Huffenberger, R. Keskitalo, L. Knox, A. Kosowsky, J. Kovac, E. D. Kovetz, C.-L. Kuo, A. Kusaka, M. Le Jeune, A. T. Lee, M. Lilley, M. Loverde, M. S. Madhavacheril, A. Mantz, D. J. E. Marsh, J. McMahon, P. D. Meerburg, J. Meyers, A. D. Miller, J. B. Munoz, H. N. Nguyen, M. D. Niemack, M. Peloso, J. Peloton, L. Pogosian, C. Pryke, M. Raveri, C. L. Reichardt, G. Rocha, A. Rotti, E. Schaan, M. M. Schmittfull, D. Scott, N. Sehgal, S. Shandera, B. D. Sherwin, T. L. Smith, L. Sorbo, G. D. Starkman, K. T. Story, A. van Engelen, J. D. Vieira, S. Watson, N. Whitehorn, and W. L. Kimmy Wu, ArXiv e-prints **1610**, arXiv:1610.02743 (2016).
[15] M. Millea, C. M. Daley, T.-L. Chou, E. Anderes, P. A. R. Ade, A. J. Anderson, J. E. Austermann, J. S. Avva, J. A. Beall, A. N. Bender, B. A. Benson, F. Bianchini, L. E. Bleem, J. E. Carlstrom, C. L. Chang, P. Chaubal, H. C. Chiang, R. Citron, C. Corbett Moran, T. M. Crawford, A. T. Crites, T. de Haan, M. A. Dobbs, W. Everett, J. Gallicchio, E. M. George, N. Goeckner-Wald, S. Guns, N. Gupta, N. W. Halverson, J. W. Henning, G. C. Hilton, G. P. Holder, W. L. Holzapfel, J. D. Hrubes, N. Huang, J. Hubmayr, K. D. Irwin, L. Knox, A. T. Lee, D. Li, A. Lowitz, J. J. McMahon, S. S. Meyer, L. M. Mocanu, J. Montgomery, T. Natoli, J. P. Nibarger, G. Noble, V. Novosad, Y. Omori, S. Padin, S. Patil, C. Pryke, C. L. Reichardt, J. E. Ruhl, B. R. Saliwanchik, K. K. Schaffer, C. Sievers, G. Smecher, A. A. Stark, B. Thorne, C. Tucker, T. Veach, J. D. Vieira, G. Wang, N. Whitehorn, W. L. K. Wu, and V. Yefremenko, arXiv e-prints **2012**, arXiv:2012.01709 (2020).
[16] J. Carron, arXiv:1808.10349 [astro-ph] (2018), arXiv:1808.10349 [astro-ph].
[17] J. Carron and A. Lewis, Physical Review D **96**, 063510 (2017).
[18] L. Legrand and J. Carron, arXiv:2112.05764 [astro-ph] (2021), arXiv:2112.05764 [astro-ph].
[19] J. Caldeira, W. L. K. Wu, B. Nord, C. Avestruz, S. Trivedi, and K. T. Story, arXiv:1810.01483 [astro-ph] (2018), arXiv:1810.01483 [astro-ph].
[20] E. Guzman and J. Meyers, arXiv:2101.01214 [astro-ph, stat] (2021), arXiv:2101.01214 [astro-ph, stat].
[21] D. Green, J. Meyers, and A. van Engelen, arXiv:1609.08143 [astro-ph, physics:hep-ph, physics:hep-th] (2016), arXiv:1609.08143 [astro-ph, physics:hep-ph, physics:hep-th].
[22] S. C. Hotinli, J. Meyers, C. Trendafilova, D. Green, and A. van Engelen, arXiv:2111.15036 [astro-ph, physics:gr-qc, physics:hep-ph] (2021), arXiv:2111.15036 [astro-ph, physics:gr-qc, physics:hep-ph].

[23] E. Anderes, B. D. Wandelt, and G. Lavaux, The Astrophysical Journal **808**, 152 (2015).
[24] M. Millea, E. Anderes, and B. D. Wandelt, arXiv e-prints **2020**, arXiv:2002.00965 (2020).
[25] H. Ge, K. Xu, and Z. Ghahramani, in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2018) pp. 1682–1690.
[26] J. Alsing and B. Wandelt, Monthly Notices of the Royal Astronomical Society **476**, L60 (2018).
[27] C. G. Broyden, Mathematics of Computation **19**, 577 (1965).
[28] R. M. Neal, The Annals of Statistics **31**, 705 (2003).
[29] O. Papaspiliopoulos, G. O. Roberts, and M. Sköld, Statistical Science **22**, 59 (2007).
[30] A. Beskos, N. S. Pillai, G. O. Roberts, J. M. Sanz-Serna, and A. M. Stuart, (2010).
[31] C. Modi, A. Barnett, and B. Carpenter, (2021).
[32] K. M. Smith, D. Hanson, M. LoVerde, C. M. Hirata, and O. Zahn, Journal of Cosmology and Astroparticle Physics **2012**, 014 (2012), arXiv:1010.0048.
[33] M. M. Schmittfull, A. Challinor, D. Hanson, and A. Lewis, Physical Review D **88**, 063012 (2013), arXiv:1308.0286.
[34] J. Peloton, M. Schmittfull, A. Lewis, J. Carron, and O. Zahn, Physical Review D **95**, 043508 (2017).
[35] D. Han, N. Sehgal, A. MacInnis, A. van Engelen, B. D. Sherwin, M. S. Madhavacheril, S. Aiola, N. Battaglia, J. A. Beall, D. T. Becker, E. Calabrese, S. K. Choi, O. Darwish, E. V. Denison, J. Dunkley, S. Ferraro, A. E. Fox, M. Hasselfield, J. C. Hill, G. C. Hilton, M. Hilton, R. Hložek, J. Hubmayr, J. P. Hughes, A. Kosowsky, J. Van Lanen, T. Louis, K. Moodley, S. Naess, T. Namikawa, F. Nati, J. P. Nibarger, M. D. Niemack, L. A. Page, B. Partridge, F. J. Qu, A. Schillaci, D. N. Spergel, S. Staggs, E. Storer, and E. J. Wollack, arXiv:2007.14405 [astro-ph, physics:hep-ph] (2020), arXiv:2007.14405 [astro-ph, physics:hep-ph].
[36] B. A. Benson, P. A. R. Ade, Z. Ahmed, S. W. Allen, K. Arnold, J. E. Austermann, A. N. Bender, L. E. Bleem, J. E. Carlstrom, C. L. Chang, H. M. Cho, S. T. Ciocys, J. F. Cliche, T. M. Crawford, A. Cukierman, T. de Haan, M. A. Dobbs, D. Dutcher, W. Everett, A. Gilbert, N. W. Halverson, D. Hanson, N. L. Harrington, K. Hattori, J. W. Henning, G. C. Hilton, G. P. Holder, W. L. Holzapfel, K. D. Irwin, R. Keisler, L. Knox, D. Kubik, C. L. Kuo, A. T. Lee, E. M. Leitch, D. Li, M. McDonald, S. S. Meyer, J. Montgomery, M. Myers, T. Natoli, H. Nguyen, V. Novosad, S. Padin, Z. Pan, J. Pearson, C. L. Reichardt, J. E. Ruhl, B. R. Saliwanchik, G. Simard, G. Smecher, J. T. Sayre, E. Shirokoff, A. A. Stark, K. Story, A. Suzuki, K. L. Thompson, C. Tucker, K. Vanderlinde, J. D. Vieira, A. Vikhlinin, G. Wang, V. Yefremenko, and K. W. Yoon, arXiv:1407.2973 [astro-ph] , 91531P (2014), arXiv:1407.2973 [astro-ph].