Bayesian estimates of CMB gravitational lensing

Ethan Anderes[*1] **Benjamin Wandelt**[2] **and Guilhem Lavaux**[2]

[1] *Department of Statistics, University of California, Davis CA 95616, USA.*

[2] *Institut d' Astrophysique de Paris.*

Abstract: The Plank satellite, along with ground based telescopes such as the Atacama Cosmology Telescope (ACT) and the South Pole Telescope (SPT), have mapped the cosmic microwave background (CMB) at such an unprecedented resolution as too allow a detection of the subtle distortions due to the gravitational influence of intervening dark matter. This distortion is called gravitational lensing and has become a powerful probe of cosmology and the nature of dark matter. Estimating gravitational lensing is important for two reasons. First, the weak lensing estimates can be used to construct a map of dark matter which would be invisible otherwise. Second, weak lensing estimates can, in principle, un-lense the observed CMB to construct the original CMB radiation fluctuations. Both of these maps, the unlensed CMB radiation field and the dark matter field, are deep probes into the nature of cosmology and cosmic structure. Bayesian techniques seem a perfect fit for the statistical analysis of lensing and the CMB. One reason is that the priors for the unlensed CMB and the lensing potential are—very nearly—realizations of Gaussian random fields. The Gaussianity coming from physically predicted quantum randomness in the early universe. However, challenges associated with a full Bayesian analysis have prevented previous attempts at developing a working Bayesian prototype. In this paper we solve many of these obstacles with a re-parameterization of the naïve lensing problem. This allows us to obtain approximate draws from the Bayesian lensing posterior of both the lensing potential and the unlensed CMB which converges remarkably fast.

Keywords and phrases: CMB, gravitational lensing, Bayesian, Gibbs sampler, Ancillary Gibbs chain, Sufficient Gibbs chain.

Over the past few years, data from ground based telescopes (ACT, SPT and Bicep2) and the Plank satellite and have resulted in an unprecedented detection of weak gravitational lensing of the cosmic microwave background (CMB) [add citations]. This new data not only probes the nature of dark matter but also constrains cosmological models of gravity waves and dark energy [add citations]. The state-of-the-art estimator of CMB gravitational lensing, the quadratic estimator developed by Hu and Okomoto (2001, 2002), works in part through a delicate cancellation of terms in an infinite Taylor expansion of the lensing effect on the CMB. The effect of this cancellation is particularly sensitive to foreground contaminants and sky masking, which if not fully accounted for, limits the statistical inferential power of this new data.

Possibly the most promising alternative to the quadratic estimator is Bayesian lensing. Indeed, Bayesian techniques applied to the lensed CMB observations have the potential for drastically changing the way lensing is estimated and used for inference. Current frequentest estimators of the unknown lensing potential treat the unlensed CMB as a source of shape noise which is marginalized out. Conversely, a Bayesian lensing posterior treats the lensing potential *and* the unlensed CMB as joint unknowns, whereby obtaining scientific constrains jointly rather than marginally. Moreover, the posterior distribution is easier to interpret and sequentially update with additional data. From the geometry of weak lensing, most of the lensing power comes from matter at a redshift $z \approx 2$ (check the facts here). At these distances the matter distribution is well approximated by linear theory which predicts the matter density fluctuations are nearly Gaussian. In addition, the unlensed CMB is, at present, indistinguishable from an isotropic Gaussian random field. From a statistical perspective, this is a perfect scenario for Bayesian methods in that both the observations and the unknown lensing potential are *physically predicted* to be Gaussian random fields.

Physicists have known, for some time, that Bayesian methods could potentially provide next-generation lensing estimates. In their seminal review [1], authors Lewis and Challinor discuss the

*Research supported by: NSF DMS-1007480, NSF CAREER DMS-1252795

†Code available at url: https://github.com/EthanAnderes/BayesianCmbLensing.git

possibility of obtaining posterior draws from the lensing potential and the unlensed CMB jointly. However, they acknowledge the main obstacle for naive Gibbs implementations:

> "... given a particular lensing potential the delensed sky is given essentially by a delta function. This means that naive Gibbs iterations will not converge within a reasonable time. At the time of writing there are no known practical methods for sampling from the full posterior distribution."

In this paper we show that, indeed, there does exist a practical way to obtain Gibbs iterations which converge quickly. The solution is through a re-parameterization of CMB lensing problem. Instead of treating the lensing potential as unknown we work with inverse-lensing or what we call anti-lensing. Surprisingly, the slowness of naive Gibbs translates to fast convergence of the re-parameterized Gibbs chain.

In Section 2 we motivate our re-parameterization by analyzing a simple two parameter statistical problem. The concepts are then applied to the Bayesian lensing problem in Section 3. The two conditional distributions in our Gibbs implementation are discussed in Section 4 and Section 5. We finish with some simulation examples in Section 6.

1. Weak lensing primer and a Bayesian challenge

In this section we describe the basics of CMB lensing and Bayesian estimation. The effect of weak lensing is to simply remap the CMB, preserving surface brightness. Up to leading order, the remapping displacements are given by $\nabla\phi$, where ϕ denotes a lensing potential and is the planer projection of a three dimensional gravitational potential (see Dodelson, S. (2003), for example). Therefore the lensed CMB can be written $T(\boldsymbol{x} + \nabla\phi(x))$ where $T(x)$ denotes the unlensed CMB temperature fluctuations projected to the observable sky and x represents an observational direction on the unit sphere. For this paper we will be focusing on the small angle limit so that x is assumed to vary in a small patch of $\mathbb{R}^2$. The lensed CMB is observed with additive noise (denoted $n(x)$) due to instrumental noise and a beam effect so that the data can be written in the following form

$$\mathrm{data}(x) = T(x + \nabla\phi(x)) + n(x). \tag{1}$$

The goal of weak lensing surveys is to use the data in (1) to estimate ϕ, T and possibly the spectral densities of T and ϕ.

A very natural approach to develop a Bayesian lensing estimator is to generate posterior samples through a Gibbs algorithm which iteratively samples from the two conditionals: $P(T|\phi, \mathrm{data})$ and $P(\phi|T, \mathrm{data})$. Sampling from $P(T|\phi, \mathrm{data})$ is simply a Gaussian random field prediction problem since conditioning on ϕ models the data as

$$\mathrm{data}(x) = T(\underbrace{x + \nabla\phi(x)}_{\text{known obs locations}}) + n(x).$$

In other words, the data is a noisy version of T observed on an irregular grid. Conversely, when sampling from $P(\phi|T, \mathrm{data})$ the data is of the form

$$\mathrm{data}(x) = \underbrace{T}_{\text{known}}(x + \nabla\phi(x)) + n(x).$$

To see how one might approximate this conditional notice first that the CMB field $T(x)$ is very smooth. Indeed, Silk damping [cite] predicts a exponentially [check the facts here] decaying power spectrum. Therefore a linear Taylor approximation, $\mathrm{data(x)} \approx T(x) + \nabla T(x)\cdot\nabla\phi(x) + n(x)$, may be useful. In fact, the derivation of the quadratic estimator explicitly uses this linear approximation. If one is willing to use this linear approximation then the conditional $P(\phi|T, \mathrm{data})$ is simply a Bayesian regression problem since T (and thus ∇T) are both known with a Gaussian prior on $\nabla\phi$.

Unfortunately, the structure of both of these conditionals make the Gibbs very slow to converge. The case is exacerbated in the situation when noise level is small. For example, in the second conditional, if T is known and fixed, the extent of the possible ϕ's under $P(\phi|T, \mathrm{data})$ is very small compared to the possible ϕ's marginalized out of $P(\phi, T|\mathrm{data})$. This suggests a highly dependent posterior $P(\phi, T|\mathrm{data})$.

2. Two parameter analogy

To motivate our solution to the Bayesian lensing problem we start with a simple two parameter statistical problem. This system has two unknown parameters t, φ with a single data point given by

$$\text{data} = t + \varphi + n$$

where n denotes additive noise. In the Bayesian setting, the posterior distribution is computed as

$$P(t, \varphi|\text{data}) \propto P(\text{data}|t, \varphi)P(t, \varphi) \tag{2}$$

where $P(\text{data}|t, \varphi)$ denotes the likelihood of the data given t, φ and $P(t, \varphi)$ denotes the prior on t, φ. The Gibbs sampler is a widely used algorithm for generating (asymptotic) samples from $P(t, \varphi|\text{data})$ [add citations]. The algorithm generates a Markov chain of parameter values $(t^1, \varphi^1), (t^2, \varphi^2), \ldots$ generated by iteratively sampling from the conditional distributions:

$$\begin{aligned} t^{i+1} &\sim P(t|\varphi^i, \text{data}) \\ \varphi^{i+1} &\sim P(\varphi|t^{i+1}, \text{data}). \end{aligned}$$

A useful heuristic for determining the convergence rate of a Gibbs chain is the extent to which the two parameters t and φ are dependent in $P(t, \varphi|\text{data})$. A highly dependent posterior $P(t, \varphi|\text{data})$ leads to a slow Gibbs chain, near independence leads to a fast Gibbs chain. Indeed, exact independence gives a sample of the posterior after one Gibbs step. A technique for accelerating the convergence of a Gibbs sampler is to find a re-parameterization of t and φ in a way which makes the posterior less dependent. In the remainder of this section we discuss a specific re-parameterization which, by analogy, can be applied to Bayesian lensing.

The relevant situation for Bayesian lensing is the case that t and φ are highly negatively correlated in $P(t, \varphi|\text{data})$. This motivates re-parameterizing (t, φ) to $(\widetilde{t}, \varphi)$ where $\widetilde{t} \equiv t + \varphi$ so that

$$\text{data} = \widetilde{t} + n.$$

In the statistics literature, (t, φ) is commonly referred to as an **ancillary parameterization** whereas $(\widetilde{t}, \varphi)$ is referred to as a **sufficient parameterization** [add citations]. Figure 2 illustrates the difference between an ancillary versus sufficient posterior distribution for our simple two parameter model. The left plot shows the posterior density contours for the ancillary parameterization (t, φ), along with 40 steps of a Gibbs sampler. Conversely, the right plot shows the posterior density contours for the sufficient chain $(\widetilde{t}, \varphi)$ with 40 Gibbs steps. Notice that negative correlation in the ancillary parameterization manifests in near independence for the sufficient chain. Indeed, the slower the ancillary chain the faster the sufficient chain and vice-versa.

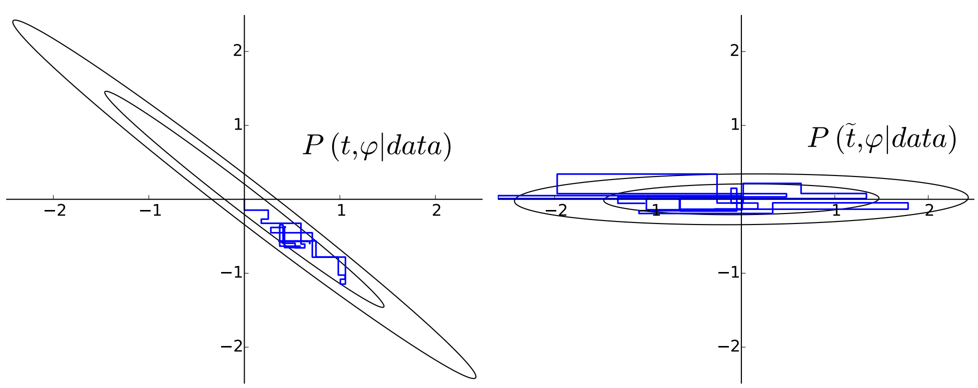

FIG 1. Left: *density contours of the* **ancillary** *chain $P(t, \varphi|data)$ with 40 steps of a Gibbs sampler.* Right: *density contours of the* **sufficient** *chain $P(\widetilde{t}, \varphi|data)$ with 40 steps of a Gibbs sampler.*

3. Ancillary versus sufficient parameters for the lensed CMB

The ancillary parameterization presented in the previous section is analogous to the lensed CMB problem as follows

$$\text{data}(x) = T(x + \nabla\phi(x)) + n(x) \quad \textit{\textbf{analogous to}} \quad \text{data} = t + \varphi + n$$

where the unlensed CMB temperature field T and the lensing potential ϕ are the two unknown parameters. As was discussed in Section 1 the Gibbs chain based on the ancillary parameters $T(x)$ and $\phi(x)$ is exceedingly slow. This clearly motivates the following re-parameterization to sufficient parameters for the lensed CMB problem

$$\text{data}(x) = \widetilde{T}(x) + n(x) \quad \textit{\textbf{analogous to}} \quad \text{data} = \widetilde{t} + n$$

where now $\widetilde{T}$ denotes the lensed CMB temperature field with no noise or beam. The sufficient chain then proceeds as

$$\widetilde{T}^{i+1} \sim P(\widetilde{T}|\phi^i, \text{data}) \tag{3}$$

$$\phi^{i+1} \sim P(\phi|\widetilde{T}^{i+1}, \text{data}). \tag{4}$$

In Section 4 we derive a Hamiltonian Markov Chain algorithm to sample from (4). In Section 5 we adapt an iterative message passing algorithm, originally developed in [add citation], for sampling from (3). Both these algorithms rely heavily on an approximation—motivated again by the two parameter system—we call *anti-lensing*.

3.1. Anti-lensing approximation

In the two parameter analogy from Section 2, the relation between the sufficient parameter $\widetilde{t}$ and the ancillary parameter t is given by $\widetilde{t} - \varphi = t$. The corresponding relation for CMB lensing we refer to as *anti-lensing*

$$\widetilde{T}(\underbrace{x - \nabla\phi(x)}_{\text{anti-lensing}}) \approx T(x). \tag{5}$$

Indeed, anti-lensing is a key concept in this paper and allows sampling from both distributions (3) and (4). We distinguish between *inverse lensing* and *anti-lensing*. Inverse lensing denotes the true coordinate displacement which, when applied to $\widetilde{T}$, recovers the unlensed T. Conversely, anti-lensing is given by $-\nabla\phi$ and approximates inverse lensing. To examine the difference between these two operations, start with a Helmholtz decomposition of the inverse lensing displacement: $-\nabla\phi^{\text{inv}}(x) - \nabla^{\perp}\psi^{\text{inv}}(x)$, where $\nabla^{\perp} \equiv \big(-\frac{\partial}{\partial y}, \frac{\partial}{\partial x}\big)$ and ψ^{inv} denotes a stream function potential which models a field rotation. Now ϕ^{inv} and ψ^{inv} are used to convert a lensed CMB $\widetilde{T}$ to T as follows

$$\widetilde{T}\big(x - \nabla\phi^{\text{inv}}(x) - \nabla^{\perp}\psi^{\text{inv}}(x)\big) = T(x).$$

Due to the fact that the expected size of the lensing displacement $\nabla\phi$ is much smaller than the correlation length scale of ϕ we have

$$-\nabla\phi \approx -\nabla\phi^{\text{inv}} \approx -\nabla\phi^{\text{inv}} - \nabla^{\perp}\psi^{\text{inv}}. \tag{6}$$

Indeed, the typical displacement size $\nabla\phi(x)$ is less than 3 arcmin whereas the correlation length scale of ϕ is on the order of degrees. In Figure 3.1 we show a simulation of $-\phi$ (upper left) with the corresponding inverse lensing potential $-\phi^{\text{inv}}$ (upper right). The difference $\phi - \phi^{\text{inv}}$ is also shown (bottom left) along with the stream function $-\psi^{\text{inv}}$. Clearly, the magnitude of the difference $\phi^{\text{inv}} - \phi$ and $-\psi^{\text{inv}}$ is sub-dominant to estimation error expected in current lensing experimental conditions.

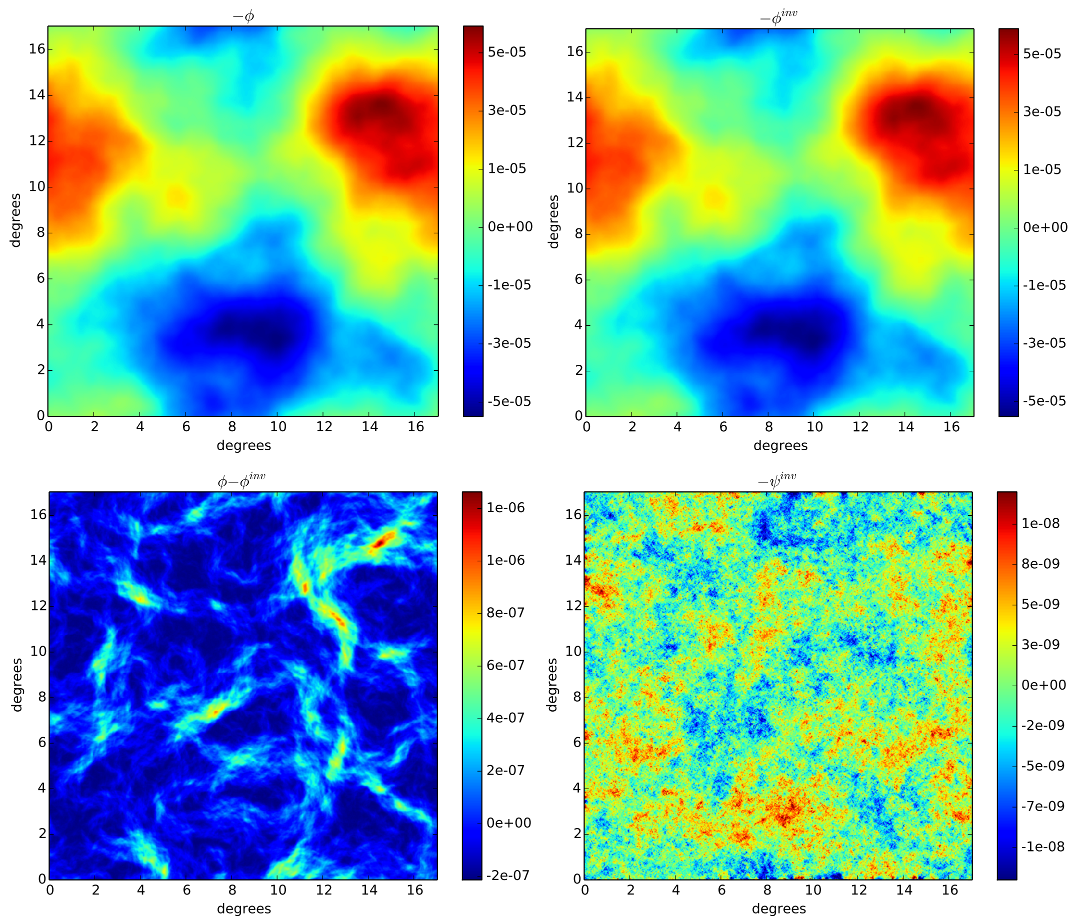

Fig 2. *The difference between anti-lensing and inverse lensing.* Upper left: *anti-lensing potential* $-\phi$. Upper right: *The inverse lensing potential* $-\phi^{inv}$. Bottom left: *The difference* $\phi^{inv} - \phi$. Bottom right: *The inverse lensing stream function* $-\psi^{inv}$.

Regardless of the fact that $-\nabla\phi$ is a good approxiamtion to inverse lensing, the Hamiltonian sampler for $P(\phi|\widetilde{T}, \text{data})$, described below in Section 4, can be easily adjusted to sample from the inverse lensing stream and potential functions jointly $P(\phi^{\text{inv}}, \psi^{\text{inv}}|\widetilde{T}, \text{data})$. However, we have excluded it from our analysis since the magnitude of the lensing potential is at least an order of magnitude smaller than estimation error.

4. Hamiltonian Monte Carlo sampler for $P(\phi|\widetilde{T}, \text{data})$

The Hamiltonian Monte Carlo (HMC) algorithm is an iterative sampling algorithm designed to mitigate the low-acceptance rate of the Metropolis-Hastings algorithm when working in high dimension. A nice review of HMC can be found in [cite Neal's paper]. For applications of HMC in cosmology see [add citations]. In the present case we utilize the HMC algorithm to produce samples of ϕ from $P(\phi|\widetilde{T}, \text{data})$. The key to making HMC work for lensing is to parameterize ϕ in terms of it's Fourier transform. One can then utilize Claim 1, presented below, to efficiently compute the gradient of the log conditional density of $P(\phi|\widetilde{T}, \text{data})$, which is a the necessary component of the HMC algorithm.

Notation: Throughout the remainder of this paper, the Fourier transform of any function $f(x)$ will be denoted by f_l or f_k so that $f_l = \int_{\mathbb{R}^2} e^{-ix\cdot l} f(x) \frac{dx}{2\pi}$ and $f(x) = \int_{\mathbb{R}^2} e^{ix\cdot l} f_l \frac{dl}{2\pi}$ where $l \in \mathbb{R}^2$ is a two dimensional frequency vector and $x \in \mathbb{R}^2$ is a two dimensional spatial coordinate.

To describe the HMC algorithm let $\boldsymbol{\phi}$ denote the concatenation of the real and imaginary parts of ϕ_l as l ranges through discrete frequencies l ranging up to a pre-specified $|l|_{max}$ (but excluding half of the Fourier frequencies due to the Hermitian symmetries associated with the Fourier transform of a real field). Note that $\boldsymbol{\phi}$ is a vector of real numbers. Let $P(\boldsymbol{\phi}|\widetilde{T}, \text{data})$ denote the density of $\boldsymbol{\phi}$ given $\widetilde{T}$ and the data. Let $\boldsymbol{p}$ denote a 'momentum' vector and $\boldsymbol{m}$ denote a the 'mass' vector, which are both the same length as $\boldsymbol{\phi}$. The Hamiltonian is a function of $\boldsymbol{\phi}$ and $\boldsymbol{p}$ and is defined as follows

$$H(\boldsymbol{\phi}, \boldsymbol{p}) := -\log P(\boldsymbol{\phi}|\widetilde{T}, \text{data}) + \sum_k \frac{\boldsymbol{p}_k^2}{2\boldsymbol{m}_k^2}.$$

This Hamiltonian generates a time-dependent evolution of $\boldsymbol{\phi}$ and $\boldsymbol{p}$ given by

$$\begin{aligned} \frac{d\boldsymbol{\phi}^t}{dt} &= \quad \nabla_{\boldsymbol{p}} H(\boldsymbol{\phi}^t, \boldsymbol{p}^t) \\ \frac{d\boldsymbol{p}^t}{dt} &= -\nabla_{\boldsymbol{\phi}} H(\boldsymbol{\phi}^t, \boldsymbol{p}^t). \end{aligned}$$

The HMC is a discrete version of this time-dynamic equation, using a leapfrog method, which produces a Markov chain $(\boldsymbol{\phi}_1, \boldsymbol{p}_1), (\boldsymbol{\phi}_2, \boldsymbol{p}_2), \ldots$ where the $i+1$ iteration is given by the following algorithm

Algorithm 1 i^{th} step of the Hamiltonian Markov Chain

1: Set $\boldsymbol{\phi}^0 := \boldsymbol{\phi}_{i-1}$ and simulate $\boldsymbol{p}^0 \sim \mathcal{N}(0, \Lambda_m^2)$ where Λ_m is diagonal with $diag(\Lambda_m) = \boldsymbol{m}$.

2: Recursively compute $\boldsymbol{\phi}^{k\epsilon}$ and $\boldsymbol{p}^{k\epsilon}$ for $k = 1, \ldots, n$ using the following equations:

$$\begin{aligned} \boldsymbol{\phi}^{t+\epsilon} &:= \boldsymbol{\phi}^t + \epsilon\Lambda_m^{-1}\left[\boldsymbol{p}^t - \frac{\epsilon}{2}\nabla_{\boldsymbol{\phi}} H(\boldsymbol{\phi}^t, \boldsymbol{p}^t)\right]. \\ \boldsymbol{p}^{t+\epsilon} &:= \boldsymbol{p}^t - \frac{\epsilon}{2}\Big[\nabla_{\boldsymbol{\phi}} H(\boldsymbol{\phi}^t, \boldsymbol{p}^t) + \nabla_{\boldsymbol{\phi}} H(\boldsymbol{\phi}^{t+\epsilon}, \boldsymbol{p}^t)\Big] \end{aligned}$$

3: Simulate $u \sim \mathcal{U}(0,1)$, and define $p := \min\left(1, e^{-H(\boldsymbol{\phi}^{n\epsilon}, \boldsymbol{p}^{n\epsilon})}/e^{-H(\boldsymbol{\phi}^0, \boldsymbol{p}^0)}\right)$.

4: If $u < p$, set $\boldsymbol{\phi}_i := \boldsymbol{\phi}^{n\epsilon}$, otherwise set $\boldsymbol{\phi}_i := \boldsymbol{\phi}_{i-1}$.

The HMC algorithm is notoriously sensitive to tuning parameters. The prevailing wisdom is that one should set $\boldsymbol{m}$ to match the posterior variance. For the problems given in the simulation simply set $\boldsymbol{m}$ to be nearly proportional to $C_l^{\phi\phi}$ (with slightly more attenuation at low wavenumber). As is advocated in [cite it] we use a random ϵ to avoid resonant frequencies.

The key computational difficulty in using algorithm 1 is the computation of the $\nabla_{\boldsymbol{\phi}} H(\boldsymbol{\phi}^t, \boldsymbol{p}^t)$, or equivalently the computation of $\nabla_{\boldsymbol{\phi}} \log P(\boldsymbol{\phi}|\widetilde{T}, \text{data})$. The number of frequencies is extrememly large and therefore, any slow computation of the gradients will present a serious bottleneck. The follow claim shows that the gradient of the log density of $P(\phi|\widetilde{T}, \text{data})$, with respect to the Fourier basis of ϕ, can be computed quickly with Fourier and inverse Fourier transforms. This claim makes the Hamiltonian Monte Carlo (HMC) algorithm possible.

Claim 1. *Under the anti-lensing approximation (5) for any nonzero frequecy vector $l \equiv (l_1, l_2) \in \mathbb{R}^2$*

$$\frac{\partial}{\partial \phi_l} \log P(\phi|\widetilde{T}, \text{data}) \propto -\frac{\phi_l}{C_l^{\phi\phi}} - \sum_{q=1,2} i l_q \int_{\mathbb{R}^2} e^{-ix\cdot l} A^q(x) B(x) \frac{dx}{2\pi} \tag{7}$$

where $\phi_l = \text{re}\phi_l + i\,\text{im}\phi_l$, $\frac{\partial}{\partial \phi_l} \equiv \frac{\partial}{\partial\, \text{re}\phi_l} + i\frac{\partial}{\partial\, \text{im}\phi_l}$ *and*

$$B_l \equiv \frac{1}{C_l^{TT}} \int e^{-ix\cdot l}\widetilde{T}(x - \nabla\phi(x))\frac{dx}{2\pi} \tag{8}$$

$$A^q(x) \equiv \frac{\partial \widetilde{T}}{\partial x_q}\big(x - \nabla\phi(x)\big). \tag{9}$$

Remark: It is interesting to note that the above gradient evaluated at $\phi = 0$, equals the un-normalized quadratic estimate in the case of a noise free observation. Indeed, an approximate Newton step, using (7), is an accurate approximation to the regular quadratic estimate. This gives an indication that the parameterization $(\widetilde{T}, \phi)$ results in fast convergence of the corresponding Gibbs chain.

5. Iterative message passing algorithm for $P(\widetilde{T}|\phi, \text{data})$

There are two typical ways to model the lensed CMB $\tilde{T}$. If one marginalizes out ϕ, then $\widetilde{T}$ is modeled as a *non-Gaussian* but isotropic random field. Conversely, if one conditions on ϕ the field $\widetilde{T}$ is modeled as a *non-isotropic* but Gaussian random field. The latter case is relevant for sampling from $P(\widetilde{T}|\phi, \text{data})$ which is, therefore, simply a Gaussian conditional simulation problem. Unfortunately, the non-isotropic (indeed, non-stationary) nature of the conditional distribution of $\widetilde{T}$ presents serious computational challenges. In what follows we utilize a new iterative algorithm developed by Wandelt, Elsner and Lavaux for Gaussian conditional sampling when the signal is diagonalized in harmonic space that the noise is diagonalized in pixel space [add citations].

Start by transforming each pixel location x to it's pre-lensed location $x+\nabla\phi(x)$, while simultaneously preserving the data associated with that pixel. This effectively de-lenses $\text{data}(x) = T(x+\nabla\phi(x)) + n(x)$ but produces observations on an irregular grid. In particular, one may switch to the lensed coordinates $y = x + \nabla\phi(x)$ so that

$$\underbrace{(x + \nabla\phi(x),\ \ \text{data}(x))}_{\text{(pixel, data) tuple}} = (y,\, T(y) + \tilde{n}(y))$$

where $\tilde{n}(x + \nabla\phi(x)) = n(x)$. Now the data $(y,\, T(y) + \tilde{n}(y))$ is arranged on an irregular grid in y. This irregular grid is then embedded into a high resolution regular grid by nearest neighbor interpolation. The points y which do not get assigned an observation $T(y) + \tilde{n}(y)$ under the interpolation we consider to be masked. Figure 3 illustrates this situation. The left hand plot shows the irregularly sampled data $(x + \nabla\phi(x),\ \text{data}(x))$ and the right hand plot shows the grid embedding. The filled dots represent observations of $T(y) + \tilde{n}(y)$ whereas the empty dots correspond to a masked observation of $T(y)$. Finally we extend the definition of $\tilde{n}(y)$ to have infinite variance over the masked region, whereby producing data $T(y) + \tilde{n}(y)$ over a dense regular grid in y.

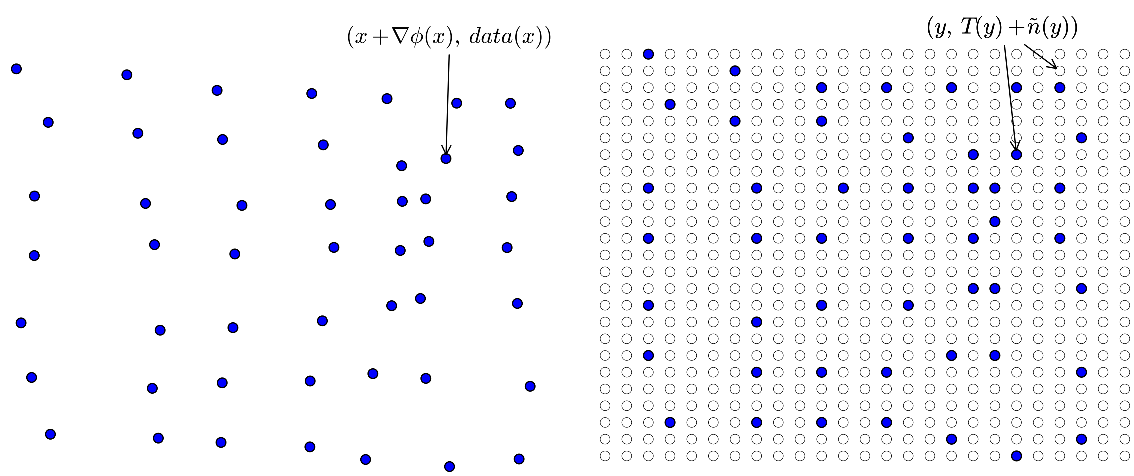

FIG 3. *The embedding of the lensed grid into a high resolution regular grid. The open circles represent masking.*

As a intermediate step in producing a sample from $P(\widetilde{T}|\phi, \text{data})$ we produce a conditional sample of $T(y)$ given the observations $T(y)+\tilde{n}(y)$. The difficulty of this step is that $\tilde{n}(y)$ is non-homogeneous noise—from the masking and any inhomogeneity in $n(x)$—and therefore it is not decorrelated by the Fourier transform. To handle this situation we adapted a new method developed by Wandelt, Elsner and Lavaux for Gaussian conditional sampling. This method works particularly well for observations with large amounts of irregular masking, as in our case. The algorithm utilizes a messenger field which effectively behaves as a latent—signal plus white noise—model.

Algorithm 2 Iterative message passing for sampling $P(T|T+\tilde{n})$

1: Set cooling schedule $\lambda^1 \geq \ldots \geq \lambda^n = 1$.
2: Simulate $Z^1(y),\ldots,Z^n(y)$ as independent realizations of standard Gaussian noise, in the pixel domain, so that $E\big[Z^k(y)Z^k(y')\big]=\delta_{yy'}$.
3: Simulate $W_l^1,\ldots,W_l^n$ as independent realizations of complex Gaussian white noise, in the Fourier domain, so that $E\big[W_l^n W_{l'}^{n\,*}\big]=\delta_{l-l'}$.
4: Initialize the fields $M^0(y)$ and $T^0(y)$ to be zero.
5: Decompose the noise variance $\mathrm{var}(\tilde{n}(y))$ into a homogeneous part $\bar{\sigma}^2$ and a non-homogeneous part $\tilde{\sigma}_y^2$ so that $\mathrm{var}(\tilde{n}(y))=\bar{\sigma}^2+\tilde{\sigma}_y^2$ (setting $\tilde{\sigma}_y^2=\infty$ on all masked pixels y).
6: Recursively compute the fields T^k, M^k for $k=1,\ldots,n$ as follows:

$$M^k(y) := \big[T(y)+\tilde{n}(y)\big]\frac{\lambda^k\bar{\sigma}^2}{\lambda^k\bar{\sigma}^2+\tilde{\sigma}_y^2} + T^{k-1}(y)\frac{\tilde{\sigma}_y^2}{\lambda^k\bar{\sigma}^2+\tilde{\sigma}_y^2} + Z^k(y)\frac{1}{\sqrt{1/(\lambda^k\bar{\sigma}^2)+1/\tilde{\sigma}_y^2}} \tag{10}$$

$$T_l^k := M_l^k\frac{C_l^{TT}}{C_l^{TT}+dy\,\lambda^k\bar{\sigma}^2} + W_l^k\frac{1}{\sqrt{1/C_l^{TT}+1/(dy\lambda^k\bar{\sigma}^2)}} \tag{11}$$

where dy denotes the pixel grid area.

The following algorithm describes how to use Algorithm 2 to produce a sample from $P(\widetilde{T}|\phi,\mathrm{data})$.

Algorithm 3 Sampling from $P(\widetilde{T}|\phi,\mathrm{data})$

1: Embedded the de-lensed pixel/data pairs $(x+\nabla\phi(x),\mathrm{data}(x))$, as illustrated in Figure 3, into observations of the form $(y,T(y)+\tilde{n}(y))$ where y ranges over a high resolution regular grid.
2: Use Algorithm 2 to produce a sample $T\sim P(T\,|\,T+\tilde{n})$
3: Return $\widetilde{T}(x)=T(x+\nabla\phi(x))$.

Remark: At present, Algorithm 3 only works when the noise $n(x)$ has a beam which is sufficiently small compared to the magnitude of $\nabla\phi(x)$. The reason is that one needs to be able to transform the data $(x+\nabla\phi(x),\mathrm{data}(x))$ to $(y,T(y)+\tilde{n}(y))$ in such a way that T is diagonalized in harmonic space and $\tilde{n}$ is diagonalized in pixel space. The presence of a large beam implies that $\tilde{n}(y)$ has complicated noise structure which is not diagonalized in either pixel space or harmonic space.

6. Simulation examples

Work this section and be sure to add details about the cooling schedule.

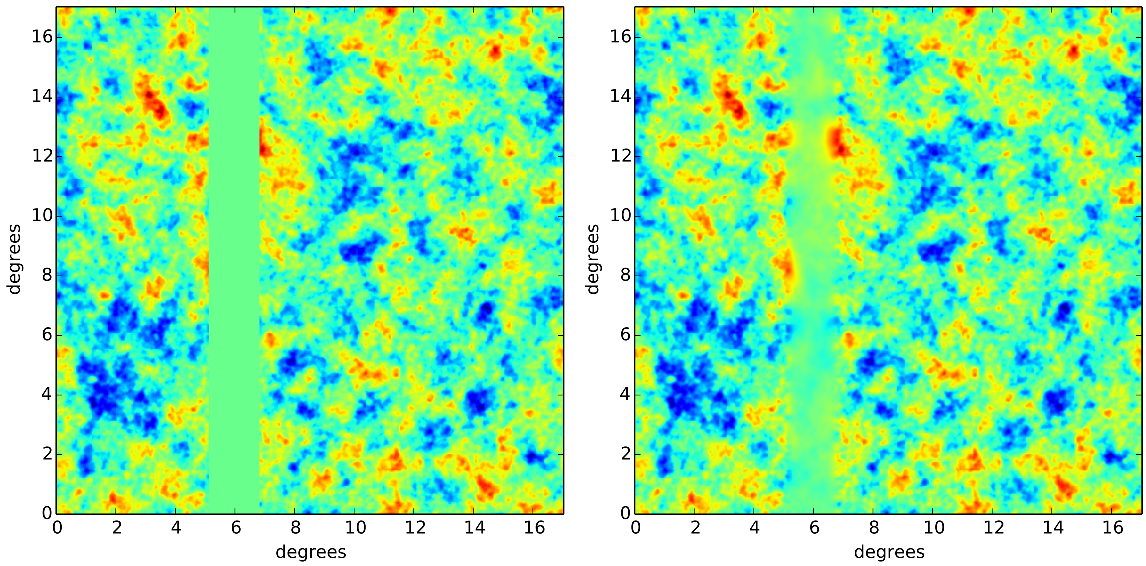

FIG 4. Left: *Data*. Middle: *posterior mean* $E(\widetilde{T}(x)|data)$.

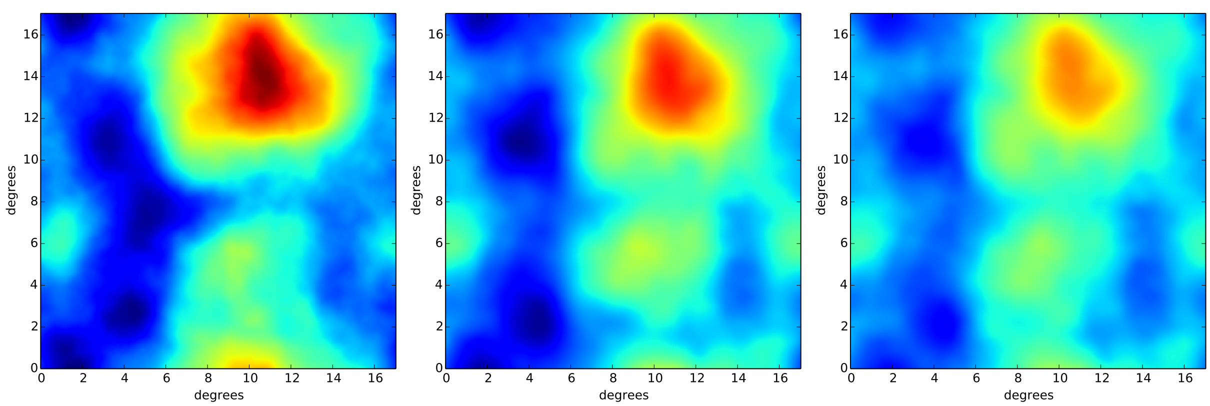

FIG 5. Left: *simulation truth* $\phi(x)$. Middle: *posterior mean* $E(\phi(x)|data)$. Right: Quadratic estimate with additional masked data.

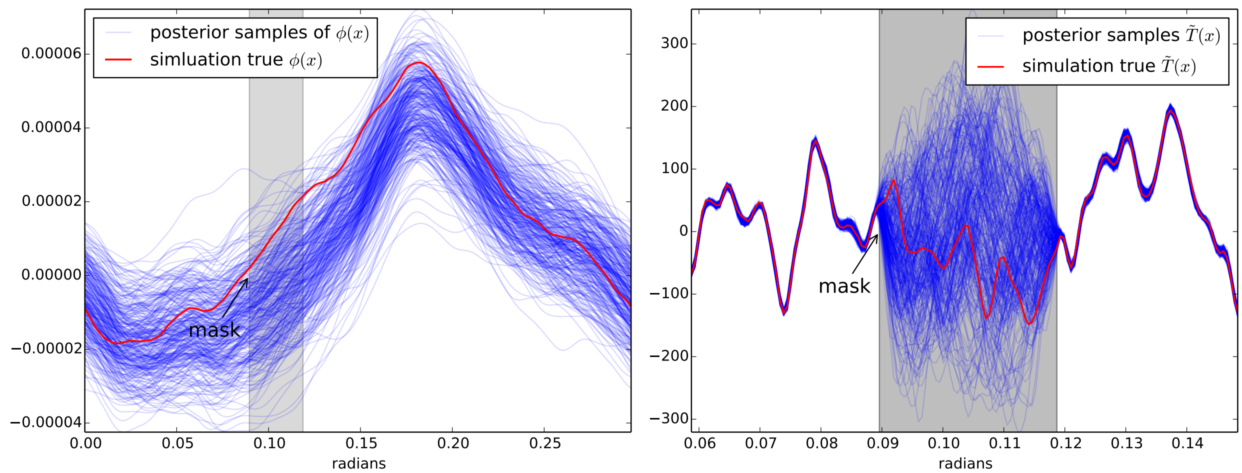

FIG 6. *sliced posterior draws.*

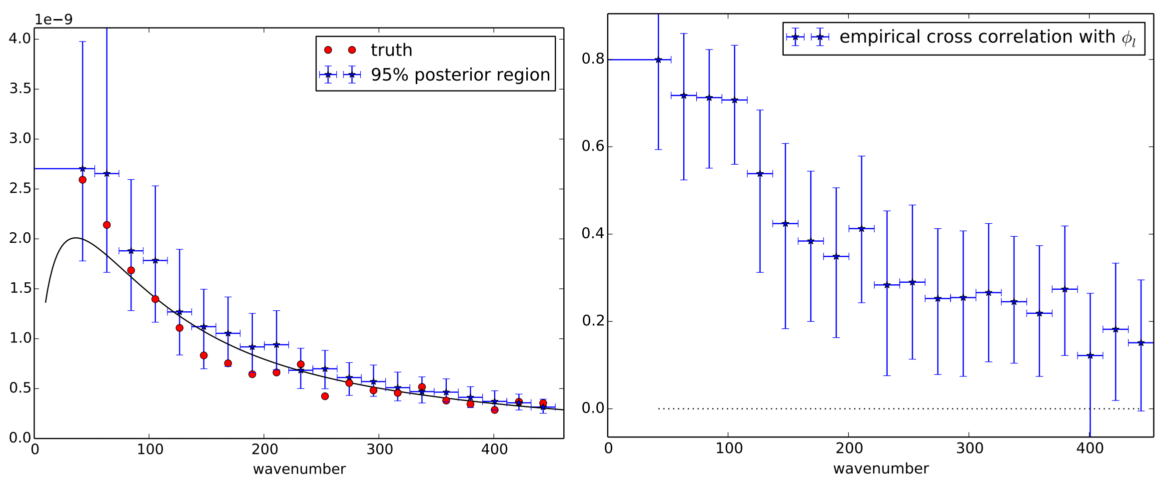

FIG 7. *Spectral estimates.* Right:*Mode by mode correlation.*

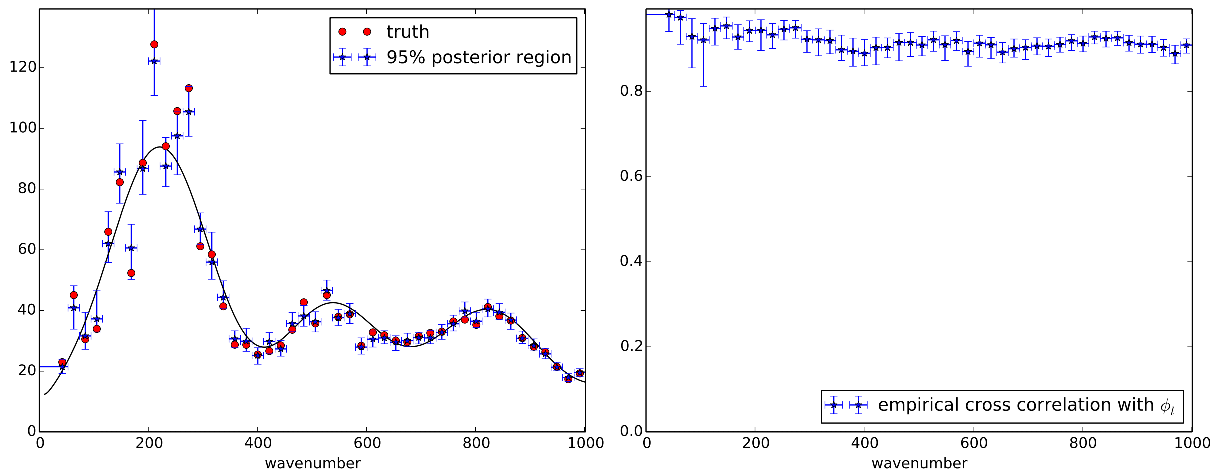

FIG 8. *Spectral estimates.* *put a mode- by mode correlation*

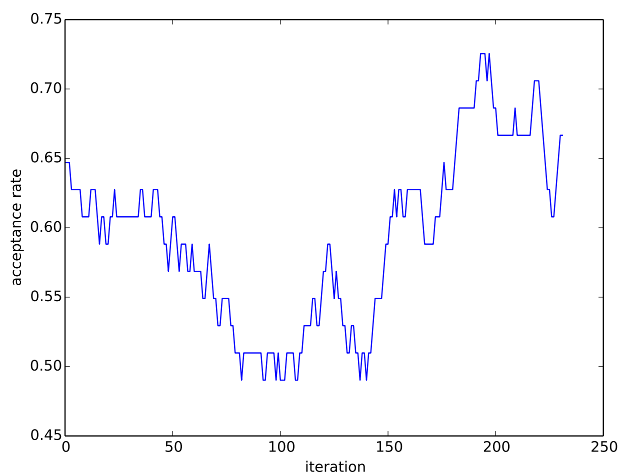

FIG 9. *Acceptance rate on a sliding window of size* 50*.*

7. Concluding remarks

Work this section

What we have accomplished

- Solved one of the fundamental obstacles in a Gibbs implementation of the Bayesian lensing (that the naive parameterization is extremely slow)
- Found a fast FFT based way to compute the gradient of ... which allows HMC to run
- Utilized a messenger algorithm to get high resolutions conditional Gaussian samples under irregular sampling.
- We have generated a prototype Bayesian posterior sampler, implemented in Julia and publically available, which demonstrates success

What needs to be done.

- High resolution embedding does not scale to Planck (the native resolution is ~10 million pixels)

- Approximate Gaussian conditional sampling on Planck data resolution (using previous ϕ^i to seed a conj gradient?)
- Need fast anti-lensing operations in frequency space to compute $A^q(x)$ and $B(x)$ that does not requires high res $\widetilde{T}$
- Incorporate spectral density uncertainty
- Incorporate polarization
- A full solution to this problem would handle non-stationary noise, non-stationary beam, cut sky or masking, In this paper we ... one of the main obstacles for the Bayesian lensing problem is ...
- Possible application to lensing of the 21 cm where the taylor approximation of lensing doesn't lend itself to a quadratic estimate

References

[1] Antony Lewis and Anthony Challinor. Weak gravitational lensing of the cmb. *Physics Reports*, 429(1):1–65, 2006.

Appendix A

Before we proceed to the proofs we say a few words regarding notation. Firstly, we do not differentiate, notationally, from the case of smooth random field with periodic boundary conditions defined on $(-L/2, L/2]^2$ and the case where $L\to\infty$ so that the Fourier series $\sum_{l\in\frac{2\pi}{L}\mathbb{Z}} e^{ix\cdot l} f_l \frac{2\pi/L}{2\pi}$ converges to the continuous Fourier transform $\int_{\mathbb{R}^2} e^{ix\cdot l} f_l \frac{dl}{2\pi}$. For example, at times we will refer to an infinitesimal area element dl or dk in Fourier space, which simply equals $\frac{2\pi}{L}$ for large L. In this case δ_l denotes a discrete dirac delta function which we equate with $1/dk$ when $l=0$ and zero otherwise. Secondly, for any function $f(x)$ let $f^\phi(x) = f(x-\nabla\phi(x))$ denote anti-lensing of f and f_l^ϕ denote the Fourier transform of $f^\phi(x)$.

Proof of Claim 1. Since $\tilde{T}$ is sufficient for the unknown ϕ we have that

$$P(\phi|\widetilde{T}, \text{data}) = P(\phi|\widetilde{T}) \propto P(\widetilde{T}|\phi)P(\phi).$$

Since $\phi(x)$ is an isotropic random field with spectral density $C_l^{\phi\phi}$ we have that $E(\phi_l\,\phi_{l'}^*) = \delta_{l-l'}C_l^{\phi\phi}$. Therefore $E(\phi_l\,\phi_l^*) = \delta_0 C_l^{\phi\phi}$ and $E(\phi_l\phi_l) = 0$ one gets that the random variables $\text{re}\phi_l$, $\text{im}\phi_l$ are independent $\mathcal{N}(0, \frac{1}{2}\delta_0 C_l^{\phi\phi})$ for each fixed l. Moreovoer $\phi(x)$ takes values in $\mathbb{R}$ so that $\phi_l = \phi_{-l}^*$. This implies that ϕ_l and are independent random variables over all l which are restricted to the a Hermitian half of the Fourier grid, denoted $\mathbb{H}$. In particular, if we exclude the zero frequency $l=0$ we get

$$\log P(\phi) - c_1 = -\frac{1}{2}\sum_{k\in\mathbb{H}\setminus\{0\}}\left[\frac{(\text{re}\phi_k)^2}{\frac{1}{2}\delta_0 C_k^{\phi\phi}} + \frac{(\text{im}\phi_k)^2}{\frac{1}{2}\delta_0 C_k^{\phi\phi}}\right] = -\frac{1}{2}\int_{\mathbb{R}^2}\frac{|\phi_k|^2}{C_k^{\phi\phi}}dk \tag{12}$$

$$\log P(\widetilde{T}|\phi) - c_2 = -\frac{1}{2}\sum_{k\in\mathbb{H}\setminus\{0\}}\left[\frac{(\text{re}\widetilde{T}_k^\phi)^2}{\frac{1}{2}\delta_0 C_k^{TT}} + \frac{(\text{im}\widetilde{T}_k^\phi)^2}{\frac{1}{2}\delta_0 C_k^{TT}}\right] = -\frac{1}{2}\int_{\mathbb{R}^2}\frac{\left|\widetilde{T}_k^\phi\right|^2}{C_k^{TT}}dk \tag{13}$$

where c_1 and c_2 are constants and $\widetilde{T}^\phi(x) \equiv \widetilde{T}(x-\nabla\phi(x))$. Taking derivatives in (12) gives

$$\frac{\partial}{\partial\phi_l}\log P(\phi) = -2(dl)\frac{\phi_l}{C_l^{\phi\phi}}. \tag{14}$$

Taking derivatives in (13) gives

$$\frac{\partial}{\partial\,\mathrm{re}\phi_l}\log P(\widetilde T|\phi) = -\,\mathrm{re}\int_{\mathbb{R}^2}\frac{\partial \widetilde T^\phi_k}{\partial\,\mathrm{re}\phi_l}\,\frac{\widetilde T^{\phi^*}_k}{C^{TT}_k}\,dk \tag{15}$$

$$\frac{\partial}{\partial\,\mathrm{im}\phi_l}\log P(\widetilde T|\phi) = -\,\mathrm{re}\int_{\mathbb{R}^2}\frac{\partial \widetilde T^\phi_k}{\partial\,\mathrm{im}\phi_l}\,\frac{\widetilde T^{\phi^*}_k}{C^{TT}_k}\,dk. \tag{16}$$

Taking linear combinations of the two equalities in Lemma 1 we get

$$\frac{\partial \widetilde T^\phi_k}{\partial\,\mathrm{re}\phi_l} = \frac{1}{2}\frac{\partial \widetilde T^\phi_k}{\partial \phi_l} + \frac{1}{2}\frac{\partial \widetilde T^\phi_k}{\partial \phi^*_l} = \frac{dk}{2\pi}\sum_{q=1,2} il_q\left\{[(\nabla^q\widetilde T)^\phi]_{k-l} - [(\nabla^q\widetilde T)^\phi]_{k+l}\right\} \tag{17}$$

$$\frac{\partial \widetilde T^\phi_k}{\partial\,\mathrm{im}\phi_l} = \frac{-i}{2}\frac{\partial \widetilde T^\phi_k}{\partial \phi_l} + \frac{i}{2}\frac{\partial \widetilde T^\phi_k}{\partial \phi^*_l} = \frac{dk}{2\pi}\sum_{q=1,2} l_q\left\{-[(\nabla^q\widetilde T)^\phi]_{k-l} - [(\nabla^q\widetilde T)^\phi]_{k+l}\right\}. \tag{18}$$

Now the above two equations establish, by Lemma 2, that both integrals $\int_{\mathbb{R}^2}\frac{\partial \widetilde T^\phi_k}{\partial\,\mathrm{re}\phi_l}\,\frac{\widetilde T^{\phi^*}_k}{C^{TT}_k}\,dk$ and $\int_{\mathbb{R}^2}\frac{\partial \widetilde T^\phi_k}{\partial\,\mathrm{im}\phi_l}\,\frac{\widetilde T^{\phi^*}_k}{C^{TT}_k}\,dk$ are real which implies

$$\begin{aligned}
\frac{\partial}{\partial\phi_l}\log P(\widetilde T|\phi) &= -\int_{\mathbb{R}^2}\frac{\partial \widetilde T^\phi_k}{\partial \phi_l}\,\frac{\widetilde T^{\phi^*}_k}{C^{TT}_k}\,dk \\
&= -\frac{dk}{\pi}\sum_{q=1,2} il_q\int_{\mathbb{R}^2}[(\nabla^q\widetilde T)^\phi]_{k+l}\,\frac{\widetilde T^{\phi^*}_k}{C^{TT}_k}\,dk \\
&= -i2(dk)\sum_{q=1,2} l_q\int_{\mathbb{R}^2}[(\nabla^q\widetilde T)^\phi]_{k+l}\,\frac{\widetilde T^{\phi^*}_k}{C^{TT}_k}\,\frac{dk}{2\pi} \\
&= -i2(dk)\sum_{q=1,2} l_q\int_{\mathbb{R}^2} e^{-ix\cdot l}A^q(x)\,B(x)\,\frac{dx}{2\pi}, \quad \text{by Lemma 3}
\end{aligned}$$

where $A^q(x) \equiv (\nabla^q\widetilde T)^\phi(x)$ and $B_k \equiv (\widetilde T^\phi_k)^*/C^{TT}_k$. □

Lemma 1.

$$\frac{\partial \widetilde T^\phi_k}{\partial \phi_l} = \frac{dk}{\pi}\sum_{q=1,2} -il_q[(\nabla^q\widetilde T)^\phi]_{k+l} \tag{19}$$

$$\frac{\partial \widetilde T^\phi_k}{\partial \phi^*_l} = \frac{dk}{\pi}\sum_{q=1,2} il_q[(\nabla^q\widetilde T)^\phi]_{k-l} \tag{20}$$

where $\nabla^q\widetilde T \equiv \frac{\partial \widetilde T}{\partial x_q}$.

Proof. First notice

$$\frac{\partial}{\partial\,\mathrm{re}\phi_l}\frac{\partial\phi(x)}{\partial x_q} = \int_{\mathbb{R}^2} ik_q e^{ix\cdot k}\frac{\partial\phi_k}{\partial\,\mathrm{re}\phi_l}\frac{dk}{2\pi} = \left[il_q e^{ix\cdot l} - il_q e^{-ix\cdot l}\right]\frac{dk}{2\pi} \tag{21}$$

$$\frac{\partial}{\partial\,\mathrm{im}\phi_l}\frac{\partial\phi(x)}{\partial x_q} = \int_{\mathbb{R}^2} ik_q e^{ix\cdot k}\frac{\partial\phi_k}{\partial\,\mathrm{im}\phi_l}\frac{dk}{2\pi} = \left[-l_q e^{ix\cdot l} - l_q e^{-ix\cdot l}\right]\frac{dk}{2\pi}. \tag{22}$$

This implies

$$\begin{aligned}
\frac{\partial \widetilde{T}^\phi_k}{\partial \phi_l} &= \frac{\partial}{\partial \phi_l}\int_{\mathbb{R}^2} e^{-ix\cdot k}\widetilde{T}(x+\nabla\phi(x))\frac{dx}{2\pi} \\
&= \sum_{q=1,2}\int_{\mathbb{R}^2} e^{-ix\cdot k}\nabla^q\widetilde{T}(x+\nabla\phi(x))\left[\frac{\partial}{\partial\,\mathrm{re}\phi_l}\frac{\partial\phi(x)}{\partial x_q} + i\frac{\partial}{\partial\,\mathrm{im}\phi_l}\frac{\partial\phi(x)}{\partial x_q}\right]\frac{dx}{2\pi} \\
&= \sum_{q=1,2}\frac{-il_q dk}{\pi}\int_{\mathbb{R}^2} e^{-ix\cdot(k+l)}\nabla^q\widetilde{T}(x+\nabla\phi(x))\frac{dx}{2\pi}, \quad \text{by (21) and (22)} \\
&= \sum_{q=1,2}\frac{-il_q dk}{\pi}[(\nabla^q\widetilde{T})^\phi]_{k+l} \qquad (23)
\end{aligned}$$

Similarly

$$\begin{aligned}
\frac{\partial \widetilde{T}^\phi_k}{\partial \phi^*_l} &= \sum_{q=1,2}\frac{il_q dk}{\pi}\int_{\mathbb{R}^2} e^{-ix\cdot(k-l)}\nabla^q\widetilde{T}(x+\nabla\phi(x))\frac{dx}{\pi} \qquad (24)\\
&= \sum_{q=1,2}\frac{il_q dk}{\pi}[(\nabla^q\widetilde{T})^\phi]_{k-l}. \qquad (25)
\end{aligned}$$

□

Lemma 2. *If $A(x)$ and $B(x)$ are real scalar fields then the two integrals, $\int_{\mathbb{R}^2} i\big\{A_{k-l} - A_{k+l}\big\}B^*_k dk$ and $\int_{\mathbb{R}^2}\big\{A_{k-l} + A_{k+l}\big\}B^*_k dk$, are both real numbers.*

Proof. By a simple change of variables it is clear that $\int_{\mathbb{R}^2}\big(i\big\{A_{k-l} - A_{k+l}\big\}B^*_k\big)^* dk = \int_{\mathbb{R}^2} i\big\{A_{k'-l} - A_{k'+l}\big\}B^*_{k'}dk'$ and $\int_{\mathbb{R}^2}\big(\big\{A_{k-l} + A_{k+l}\big\}B^*_k\big)^* dk = \int_{\mathbb{R}^2}\big\{A_{k'-l} + A_{k'+l}\big\}B^*_{k'}dk'$.

□

Lemma 3. *If $A(x)$ and $B(x)$ are real scalar fields then $\int_{\mathbb{R}^2} A_{k+l}B^*_k\frac{dk}{2\pi} = \int_{\mathbb{R}^2} e^{-ix\cdot l}A(x)B(x)\frac{dx}{2\pi}$.*

Proof.

$$\begin{aligned}
\int_{\mathbb{R}^2} A_{k+l}B^*_k\frac{dk}{2\pi} &= \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} e^{-x\cdot(k+l)}A(x)B^*_k\frac{dx}{2\pi}\frac{dk}{2\pi} \\
&= \int_{\mathbb{R}^2} e^{-ix\cdot l}A(x)\left[\int_{\mathbb{R}^2} e^{-x\cdot k}B^*_k\frac{dk}{2\pi}\right]\frac{dx}{2\pi} \\
&= \int_{\mathbb{R}^2} e^{-ix\cdot l}A(x)\left[\int_{\mathbb{R}^2} e^{x\cdot k}B_k\frac{dk}{2\pi}\right]^*\frac{dx}{2\pi} \\
&= \int_{\mathbb{R}^2} e^{-ix\cdot l}A(x)B^*(x)\frac{dx}{2\pi} \\
&= \int_{\mathbb{R}^2} e^{-ix\cdot l}A(x)B(x)\frac{dx}{2\pi}, \quad \text{since } B(x) \text{ is real.}
\end{aligned}$$

□