DeepCMB: Lensing Reconstruction of the Cosmic Microwave Background with Deep Neural Networks

J. Caldeira[a,d,*], W. L. K. Wu[b], B. Nord[b,c,d], C. Avestruz[a,b], S. Trivedi[e], K. T. Story[f]

[a] *Enrico Fermi Institute & Kadanoff Center for Theoretical Physics, University of Chicago, Chicago, IL 60637, USA*
[b] *Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[c] *Department of Astronomy and Astropysics, University of Chicago, 5640 S. Ellis Ave., Chicago, IL 60134*
[d] *Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL 60510, USA*
[e] *Institute for Computational and Experimental Research in Mathematics, Brown University, Providence, RI 02903, USA*
[f] *Descartes Labs, Santa Fe, NM 87501, USA*

Abstract

Next-generation cosmic microwave background (CMB) experiments will have lower noise and therefore increased sensitivity, enabling improved constraints on fundamental physics parameters such as the sum of neutrino masses and the tensor-to-scalar ratio r. Achieving competitive constraints on these parameters requires high signal-to-noise extraction of the projected gravitational potential from the CMB maps. Standard methods for reconstructing the lensing potential employ the quadratic estimator (QE). However, the QE is known to perform suboptimally at the low noise levels expected in upcoming experiments. Other methods, like maximum likelihood estimators (MLE), are under active development. In this work, we demonstrate reconstruction of the CMB lensing potential with deep convolutional neural networks (CNN) — i.e., a ResUNet. The network is trained and tested on simulated data, and otherwise has no physical parametrization related to the physical processes of the CMB and gravitational lensing. We show that, over a wide range of angular scales, ResUNets recover the input gravitational potential with a higher signal-to-noise ratio than the QE method, reaching levels comparable to analytic approximations of MLE methods. We demonstrate that the network outputs quantifiably different lensing maps when given input CMB maps generated with different cosmologies. We also show we can use the reconstructed lensing map for cosmological parameter estimation. This application of CNNs provides a few innovations at the intersection of cosmology and machine learning. First, while training and regressing on images, this application predicts a continuous-variable field rather than discrete classes. Second, we are able to establish uncertainty measures for the network output that are analogous to standard methods. Beyond this first demonstration, we expect this approach to excel in capturing hard-to-model non-Gaussian astrophysical foreground and noise contributions.

Keywords: cosmic microwave background, cosmology, deep learning, convolutional neural networks

1. Introduction

The earliest light we can observe in the Universe is the cosmic microwave background (CMB), which was emitted $\sim$ 400,000 years after the Big Bang during a period called recombination and encodes a wealth of information about the state of the Universe at and before that time. The CMB is a strong probe of both the geometry and the content of the Universe, as shown through a number of experiments over the past two decades — e.g., COBE, Boomerang, WMAP, Planck, SPT, ACT (Mather et al., 1994; Lange et al., 2001; Bennett et al., 2013; Planck Collaboration et al., 2018; Louis et al., 2017; Henning et al., 2018). In particular, measurements within the last five years have provided strong evidence for the standard cosmological ΛCDM paradigm (e.g., Hinshaw et al., 2013; Planck Collaboration et al., 2016a; Louis et al., 2017; Henning et al., 2018). Upcoming and proposed CMB experiments are designed to reach unprecedentedly low levels of map noise ($<$ few μK-arcmin) (Benson et al., 2014; Matsumura et al., 2014; Abazajian et al., 2016; The Simons Observatory Collaboration et al., 2018). At these noise levels, CMB Stage-4, for example, is projected to be able to constrain the tensor-to-scalar ratio r to a precision of $\sigma(r) \sim 5 \times 10^{-4}$, the number of relativistic species $N_{\rm eff}$ to $\sigma(N_{\rm eff}) \sim 0.03$, and the sum of neutrino masses M_ν to $\sigma(M_\nu) \sim 20$

*Corresponding author
Email address: caldeira@fnal.gov (J. Caldeira)

meV (Abazajian et al., 2016). Tight constraints on these parameters are key to the potential discovery of primordial gravitational waves from inflation (r), extra degrees of freedom in the early universe ($N_{\rm eff}$), and differentiating the neutrino mass hierarchy (M_ν).

Constraining these parameters at these levels of precision relies on high signal-to-noise ratio reconstruction of the lensing potential — the projected weighted gravitational potential along the line-of-sight between us and the CMB. As CMB photons travel to us, their paths get deflected by the intervening mass distributions. The lensing potential is therefore a source of information about the universe, as it is sensitive to the matter power spectrum (and therefore the sum of neutrino masses). On the other hand, lensing of the CMB distorts the CMB at recombination and degrades our ability to constrain early universe physics that made imprints on the CMB at that time. As a result, reconstruction of the lensing potential and removal of the effects of lensing (delensing) from observed CMB maps are key for decoding early-universe physics. The quadratic estimator (QE; Hu and Okamoto, 2002) is commonly used for lensing reconstruction for the current generation of CMB experiments (Story et al., 2015; Planck Collaboration et al., 2016b; Sherwin et al., 2017; Ade et al., 2014; BICEP2 Collaboration et al., 2016a), and is close to optimal at current noise levels. However, when the CMB map noise is reduced to a few μK-arcmin, QE will no longer be optimal (Millea et al., 2017), meaning that solutions exist with lower noise. Therefore, maximum likelihood methods will be required in order to improve the signal-to-noise of the lensing potential reconstruction from QE (Hirata and Seljak, 2003; Millea et al., 2017) — though they have yet to be demonstrated on data.

In this work, we investigate and demonstrate the usage of neural networks as an alternative for lensing reconstruction of the CMB. The model is learned through supervision of a training set that contains the relevant physics. This training set consists of a set of simulated maps, including observed lensed maps, corresponding unlensed maps, and maps of the gravitational convergence (related to the lensing potential). The observed maps are the inputs to the neural network, and the unlensed and convergence maps are the output. To learn a function from one set of images to another suggests an architecture with two distinct steps — one for encoding the input map information into an efficient parametrization, and one for decoding those parameters back into the output.

In neural networks, this encoder-decoder design pattern is ubiquitous, and can be used in various tasks such as learning efficient representations of the inputs (Rumelhart et al., 1986; Elman and Zipser, 1988; Hinton and Salakhutdinov, 2006), machine language translation (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2016), and semantic image segmentation (Noh et al., 2015; Shelhamer et al., 2016). We employ an instance from a family of network architectures, ResUNets (Kayalıbay et al., 2017; Zhang et al., 2017), which learns a transformation between images. While these architectures were designed with image segmentation in mind, where the desired outcome is the assignment of a discrete set of labels to the pixels, they can be adapted to image-to-image regression, where the outputs are a continuous function of the inputs. This is an adaptation of ResUNets that is more suited for physics applications.

Standard approaches treat lensing reconstruction and delensing as separate steps. Recent work (Millea et al., 2017) employed maximum likelihood methods that jointly output the lensing potential and unlensed CMB maps and demonstrated the technique on simulations. In this work, we apply ResUNets to both the lensing reconstruction and delensing problems simultaneously, similarly to Millea et al. (2017). We will focus on characterizing the efficacy of the lensing recovery.

The paper is organized as follows. In §2, we present a basic background for the CMB and lensing, concluding with a statement of the problem. We then outline traditional CMB analysis tools for lensing reconstruction in §3. In §4, we describe convolutional autoencoders and ResUNets, and present the simulated data sets with which we train and test our algorithms in §5. We then describe the results of the new algorithm and its comparison with standard algorithms in §6, with a discussion of the results and their potential in §7. We conclude and present an outlook for future work in §8.

We use the following notation conventions.

- X: unlensed field; "true" CMB field.
- $\tilde{X}$: lensed field
- $\hat{\hat{X}}$: unbiased estimate/prediction of X
- $\hat{X}$: biased estimate/prediction
- $\langle X \rangle$: mean over population sample
- X^*: complex conjugate

2. The CMB and Gravitational Lensing

In this section, we discuss the physical underpinnings of the CMB maps that are used for the development and testing of lensing reconstruction algorithms.

Modern CMB experiments observe the temperature anisotropies and polarization of CMB photons (and any other foregrounds) in millimeter wavelengths. Temperature anisotropies are $\sim 0.01\%$ ($\sim 300\mu$K) deviations from the mean CMB temperature of ~ 2.7K. They arise from the acoustic oscillation of the photon-baryon fluid before the cosmological epoch of recombination. This oscillation can be sourced by both density fluctuations and primordial gravitational waves. Given the quadrupole anisotropies in the temperature, Thomson scattering of photons with free electrons during recombination causes the CMB photons to acquire a net polarization. For reviews, see Dodelson (2003); Lewis and Challinor (2006).

CMB polarization maps are commonly represented in two distinct bases. The (Q, U) basis corresponds to Stokes parameters and is convenient for mapping onto from the CMB instruments' polarization detector coordinates. Alternatively, there is the (E, B) basis, which is helpful for connecting the measurements to the physics of the source of polarization (Seljak and Zaldarriaga, 1997; Kamionkowski et al., 1997). In particular, scalar perturbations from inflation (density fluctuations) can only source the even-parity E-mode polarization, while tensor perturbations (gravitational waves) can source both E-mode and the odd-parity B-mode polarizations at recombination. Polarization signals, however, are more than an order of magnitude fainter than the temperature anisotropies, and therefore have only been mapped to high signal-to-noise on small patches of sky by ground-based CMB experiments (e.g. Louis et al., 2017; POLARBEAR Collaboration et al., 2017; Henning et al., 2018; BICEP2 Collaboration et al., 2018).

As the CMB photons travel from the last scattering surface to us, their paths are deflected by the gradient of the gravitational potential ϕ, an effect called gravitational lensing:

$$\tilde{X}_{\pm}(\hat{n}) = X_{\pm}(\hat{n} + \nabla\phi(\hat{n})), \tag{1}$$

where X is the unlensed field and $\tilde{X}$ denotes the lensed field (e.g., Hu, 2000) and $X_{\pm} = Q \pm iU$. for the polarization fields. These deflections generate distorted versions of the Q and U maps from the surface of last scattering. As a result, when transformed to the (E, B) basis, some E modes get converted to B modes. We call these lensing B modes. This means that we are therefore guaranteed to observe some B modes when we observe the CMB even if there were no primordial B modes at the surface of last scattering. These B modes have been detected only in recent years (BICEP2 Collaboration et al., 2016b; Keisler et al., 2015; Louis et al., 2017; POLARBEAR Collaboration et al., 2017), and are about $10\times$ fainter than E modes.

We can reconstruct the lensing potential from lensed CMB maps by leveraging the cross-multipole correlations that lensing introduces into the CMB maps. With a measurement of the lensing potential in hand, we can also remove the effect of lensing from CMB maps to recover primordial signals. Observed CMB maps from various experiments have been used to reconstruct the projected gravitational potential, whose power spectrum yields constraints on cosmological parameters (e.g., Ω_m, see Planck Collaboration et al., 2016b; Sherwin et al., 2017; Omori et al., 2017; Simard et al., 2017). To reach new levels of precision, high signal-to-noise reconstructions of the lensing potential play a crucial role (Manzotti et al., 2017): the shape of the lensing power spectrum is sensitive to M_ν, whereas both r and $N_{\rm eff}$ require delensing for their parameter uncertainties to reach the projected levels.

For Stage 4 CMB experiments, polarization information is expected to dominate the signal-to-noise of the lensing reconstruction (Abazajian et al., 2016). Unlike the T anisotropy maps and the E-mode maps, primordial signal in the B-mode map in the standard ΛCDM cosmology would require non-zero r for fitting observations. With $r = 0$, any observed B modes (in the absence of foregrounds and noise) would come from lensing itself. Therefore, using the E and B maps for lensing reconstruction is extremely clean. In the following, we anchor our comparisons to lensing reconstruction using the EB estimator.

Gravitational lensing can also be quantified using the *gravitational convergence* κ, a scalar field that physically corresponds to weighted overdensities integrated along the line-of-sight. Throughout this text, we will represent the lensing field using either κ or ϕ, as one can move between the two fields using Poisson equation. In Fourier space, using the flat-sky approximation, this is

$$\kappa(\boldsymbol{\ell}) = -\frac{1}{2}\ell^2\phi(\boldsymbol{\ell}), \tag{2}$$

where $\boldsymbol{\ell}$ is the two-dimensional vector of multipole moments.

We defer investigations of impacts from galactic and extragalactic foregrounds to lensing reconstruction and delensing to later work. Therefore, all the simulations involved only contain information from the CMB maps (both lensed and unlensed) and κ. We set unlensed $B = 0$, as our focus is on lensing reconstruction and primordial B-modes have not yet been discovered. We do add basic realism by adding noise, beam, and apodization mask, which we describe in more detail in §5. The task we set ourselves is to recover the unlensed E map as well

as the lensing convergence map κ from the lensed $(\tilde{Q}, \tilde{U})$ maps. This can be treated as an image-to-image regression problem and is summarized in Fig. 1.

3. Standard CMB Lensing Reconstruction Methods

We will compare the neural network approach to current standards in analysis — the quadratic estimator, and more futuristic iterative methods. We quantify the efficiency of the neural network approach by comparing the algorithms in terms of a noise proxy, defined in §3.1. In the following we briefly describe the QE lensing reconstruction scheme, define the noise that we use to compare different methods, and outline the maximum likelihood noise estimate.

3.1. Quadratic Estimator (QE)

The primordial CMB is well-approximated as Gaussian random fields and therefore can be completely described by 2-point statistics (e.g. the power spectrum, denoted by the multipole moments C_ℓ). Lensing of the CMB by the intervening gravitational potential introduces correlations between angular scales corresponding to the size of the lenses. Therefore, the covariance of the CMB map is no longer diagonal in ℓ, as it would have been if it were not lensed.

The QE method for lensing reconstruction uses the off-diagonal covariance of the lensed CMB maps to estimate the lensing potential ϕ. Specifically, the covariance is proportional to ϕ:

$$\langle X(\boldsymbol{\ell}) X'(\mathbf{L}-\boldsymbol{\ell})\rangle \propto \phi(\mathbf{L}), \tag{3}$$

where X, X' denote lensed CMB fields and $\langle\rangle$ denotes the average over CMB realizations. This relationship can be written down explicitly for pairs of CMB fields (T, E, B) (e.g. Hu and Okamoto, 2002) and can therefore be used to construct estimators that extract ϕ from the lensed fields. Indeed, the covariance would be zero if $\phi = 0$. In our case, the maps are the observed (lensed) $\tilde{E}$- and $\tilde{B}$-mode maps, which are converted from the Stokes $(\tilde{Q}, \tilde{U})$ space. For further reference, the equations in this section are based on (Hu and Okamoto, 2002) for the EB estimator.

An unnormalized ϕ map in Fourier space, $\hat{\phi}$, can be estimated as

$$\hat{\phi}_{\mathbf{L}} = \int \frac{d^2\boldsymbol{\ell}}{(2\pi)^2} w^{\phi}_{\mathbf{L},\boldsymbol{\ell}} \bar{E}_{\boldsymbol{\ell}} \bar{B}_{\mathbf{L}-\boldsymbol{\ell}} \tag{4}$$

where $\mathbf{L}$ and $\boldsymbol{\ell}$ are two-dimensional multipole vectors, and $(\bar{E}, \bar{B})$ are Wiener-filtered $(\tilde{E}, \tilde{B})$ maps; the filter is defined as

$$\bar{X} \equiv \frac{\tilde{X}}{\left(\tilde{C}^{XX}_\ell + N^{XX}_\ell\right)}, \tag{5}$$

where N^{XX}_ℓ is the noise power spectrum for the field X. The weight w is given by

$$w^{\phi}_{\mathbf{L},\boldsymbol{\ell}} = -\boldsymbol{\ell}\cdot(\mathbf{L}-\boldsymbol{\ell})\, \tilde{C}^{EE}_\ell \sin(2\psi), \tag{6}$$

where ψ is the angle between $\boldsymbol{\ell}$ and $\mathbf{L}-\boldsymbol{\ell}$ (Hu and Okamoto, 2002).

To obtain an unbiased estimate of ϕ, $\hat{\hat{\phi}}$, we subtract the mean-field that arises from masking and normalize it by $1/R$, where R is the response:

$$\hat{\hat{\phi}}_{\mathbf{L}} = \frac{1}{R}\left(\hat{\phi}_{\mathbf{L}} - \langle\hat{\phi}_{\mathbf{L}}\rangle\right) \tag{7}$$

$$\frac{1}{R} = \int \frac{d^2\boldsymbol{\ell}}{(2\pi)^2} \frac{|w^{\phi}_{\mathbf{L},\boldsymbol{\ell}}|^2}{(\tilde{C}^{EE}_\ell + N^{EE}_\ell)(\tilde{C}^{BB}_\ell + N^{BB}_\ell)} \tag{8}$$

Generically assuming $\hat{\phi} = R\phi + n_\phi$, where one can think of R as a multiplicative bias and n_ϕ as an additive bias, we can write the power spectrum of $\hat{\hat{\phi}}$ as follows:

$$\langle\hat{\hat{\phi}}^*(\mathbf{L})\hat{\hat{\phi}}(\mathbf{L}')\rangle = \delta(\mathbf{L}-\mathbf{L}')(C^{\phi\phi}_L + N^{\phi\phi}_L), \tag{9}$$

where $C^{\phi\phi}_L$ is the input ϕ field's power spectrum and $N^{\phi\phi}_L$ is the noise spectrum. $N^{\phi\phi}_L$ is the noise term that we compare between the various algorithms.

The noise spectrum provides a numerical point of comparison, because it is informative for parameter estimation. For example, when constraining the sum of neutrino masses, the noise spectrum directly enters the Fisher forecast of 1-σ uncertainty (e.g. Wu et al., 2014). Additionally, in the limit of noise reachable by CMB experiments in the next decade, the delensing efficiency is mostly a function of cross-correlation of the $\hat{\hat{\phi}}$ to the true underlying ϕ, $\rho_L = C^{\hat{\hat{\phi}}\phi}_L / \sqrt{(C^{\phi\phi}_L + N^{\phi\phi}_L) C^{\phi\phi}_L}$, which is also directly related to the noise spectrum.

In this work, we extract $N^{\phi\phi}_L$ from the QE by

$$N^{\phi\phi}_L = \langle C^{\hat{\hat{\phi}}\hat{\hat{\phi}}}_L\rangle - \langle C^{\phi\phi}_L\rangle, \tag{10}$$

the difference between the ensemble average of the auto-spectra of the estimated ϕ and that of the input ϕ. $N^{\phi\phi}_L$ is typically estimated from simulations in standard analysis, calculated as N_0 and N_1 noise biases. N_0 denotes the disconnected 4-point term that is 0th order in $C^{\phi\phi}_L$, while N_1 is 1st order in $C^{\phi\phi}_L$ (Kesden et al., 2003). N_0 is the largest contribution to $N^{\phi\phi}_L$. Here, since we are working with simulations, instead of calculating N_0 and N_1 directly, the difference as defined in Eqn. (10) suffices for comparison. It would capture N_0, N_1 and any other noise source that does not correlate with the input.

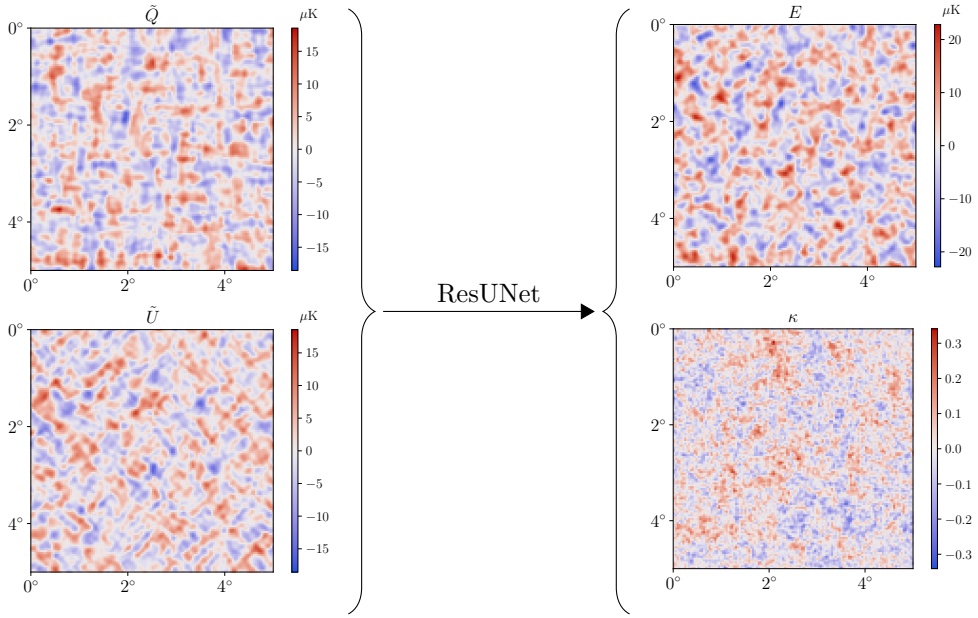


Figure 1: We train neural networks to learn a mapping from the lensed $(\tilde{Q}, \tilde{U})$ maps into the unlensed E map and the gravitational convergence map κ, extracting the underlying fields from the observed quantities. Here we illustrate this mapping using one of the realizations in the training set. The maps correspond to a patch of the sky five degrees across.

3.2. Iterative Estimator

We use the algorithm presented in Smith et al. (2012) to estimate the noise achievable by maximum likelihood estimators. The idea is based on iterating the quadratic estimator for ϕ with CMB maps that are delensed with the estimated ϕ. Supposing that the lensing B modes are the only B-mode contribution, one can reconstruct ϕ from E- and B-mode maps, and the ϕ map can be used to remove the lensing B modes in the input map. One can then use the delensed B-mode map in combination with the E-mode map to estimate the remaining ϕ field, and then use this estimated ϕ to delens the B-mode map. These two steps can be iterated until the residual B modes no longer get reduced.

It was found that the N_0 noise computed in this approach asymptotes towards maximum likelihood estimators, as presented in Hirata and Seljak (2003). We therefore compare the noise proxy from our neural network approach to this N_0 to get a sense of how closely the neural network estimator gets to maximum likelihood ϕ estimators. This is a reasonable comparison because the analytic estimate provides a theoretical lower limit on the noise of the reconstructed lensing potential power spectrum. If the neural network recovered ϕ's noise spectrum approaches this, it would provide an argument for the utility of this estimator for next-generation CMB experiments.

Note that comparing the analytic N_0 from the iterative estimator to the noise spectrum from the output of the network is not strictly an apples-to-apples comparison. In particular, we expect the analytic N_0 to perform better than N_0 extracted from realistic simulations, as the analytic N_0 does not capture effects like masking. On the other hand, the comparison between the QE noise spectrum and the network output noise spectrum is an apples-to-apples comparison.

4. Deep Learning Reconstruction Method: Residual U-Nets (ResUNet)

A growing number of physics tasks utilize machine learning techniques (see Mehta et al. (2018) for a recent review). Our task is to create a network that can learn the mapping from the observed lensed $(\tilde{Q}, \tilde{U})$ images to unlensed CMB images and the gravitational convergence map κ. In this application, we use a type of *feed-forward deep neural network* called a Residual U-Net (ResUNet) (Kayalıbay et al., 2017; Zhang et al., 2017).

The fundamental building block of a feed-forward neural network is a neuron, which receives (typically scalar) inputs, and outputs a real number. Much of the power of neural networks stems from how neurons are connected to each other. A neural network is organized in a number of layers, each layer comprising a set of neurons. The first layer's inputs are the inputs of the network, while the final layer gives the network output. For instance, in our problem the output layers have a total of $2 \times 128 \times 128$ neurons, each corresponding to a pixel of one output map. *Deep* neural networks typically

refer to neural networks having more than (usually much more than) 3 to 4 layers.

Information is propagated forward layer-by-layer through the network from the inputs to the outputs, hence the terminology *feed-forward*. Every neuron takes a linear combination of the outputs of a subset of neurons in the previous layer, and then applies a non-linear function, known as the *activation function*, to that combination. The composition of these simple operations from the first to the final layer can result in a highly non-linear mapping between the input and the output images. The multiple weights in each of the linear combinations are optimized using gradient descent applied to the error in the output. As the gradient uses the chain rule to move back from the outputs to each layer, this process is called *backpropagation*.

When working with images, the most ubiquitous type of neural network is a *Convolutional Neural Network* (CNN), which is defined as a feed-forward neural network with at least one convolutional layer, named so because it implements a discrete convolution. Each neuron in a convolutional layer takes input from neurons in the previous layer located inside a $n \times n$ window centered at its position. Typical values for n range within $n = 3, 5, 7$, and the transformation performed by the convolutional layer on the window is a *filter*. Crucially, the network keeps weights of the linear combinations independent of the position in the image. This parameter sharing reduces the complexity of the network and explicitly encodes translational equivariance.

The output size of a convolutional layer is controlled by three parameters: Number of convolutional filters applied, *stride*, and amount of zero padding. The stride may be defined as the distance in pixels between the centers of adjacent filters. With appropriate zero padding around the image, and sliding the convolution filter with a stride of 1, the output map will be of the same size as the input. Likewise, we can reduce the size of the output map into half by choosing a stride of 2. The size of the output map can also be doubled by up-sampling the input, i.e., introducing zeros between pixels of the input. A review of convolutional layers and their arithmetic can be found in Dumoulin and Visin (2016).

Convolutional layers are a natural way to take spatial context into account, as each pixel is only a function of the pixels in the previous layer that are contained inside the window defined by the convolutional filter. As we stack convolutional layers on top of each other, the region of the input that any given pixel is a function of increases. The size of this region at any specific layer is called *receptive field* of the layer. The fact that the receptive field increases as we move from layer to layer makes it so that each layer is sensitive to features at increasingly larger scales (corresponding to lower ℓ modes), allowing for both local and global information to propagate through the network.

We choose the architecture of our neural network such that the network first encodes relevant information from the input maps into smaller maps, and then decodes that information to form the output maps. The canonical example of this design pattern is an autoencoder (Rumelhart et al., 1986; Hinton and Salakhutdinov, 2006; Elman and Zipser, 1988), for which the desired outputs are equal to the inputs, and which have applications in dimensionality reduction, compression, and unsupervised feature learning. More generally, the encoder-decoder strategy can be employed to learn compact representations of mappings that are not necessarily the identity function: The encoder part of the network learns the important features of the input at different scales, and the decoder combines these features into more and more complicated representations. This is our goal in this work.

To achieve this, we implement a UNet (Ronneberger et al., 2015), which takes the simple encoder-decoder with convolutional layers, and adds extra shortcuts (*skip connections*) between the encoding and decoding layers to allow for propagation of small-scale information that might be lost when the size of the images decreases. UNets were first introduced as a method for image segmentation in a biomedical context (Ronneberger et al., 2015), and have a recent application in physics, where they were used to process sea surface temperature measurements and aid in the prediction of future sea surface temperature (de Bezenac et al., 2017). Notably, physics applications of these architectures, such as in de Bezenac et al. (2017) and this work, use UNets for a regression task with a continuous output variable. This is distinct from the typical image semantic segmentation, where the outputs are discrete labels applied to each pixel.

All neurons have Scaled Exponential Linear Unit (SELU) activation functions (Klambauer et al., 2017), except for the last layer which uses the identity function, as usual in regression problems. This activation function was chosen because it led to better results on the validation set than other possibilities, such as ReLU, leaky ReLU and ELU. The size of the convolutional filters is chosen to be 5 $\times$ 5. We also add dropout layers to avoid overfitting to the training set, and batch normalization to ensure that the input of each layer is appropriately normalized, facilitating smoother training. Most layers have stride 1, leading to output images with the same dimensions as the inputs, but we set stride to 2 in some layers in the encoding phase,

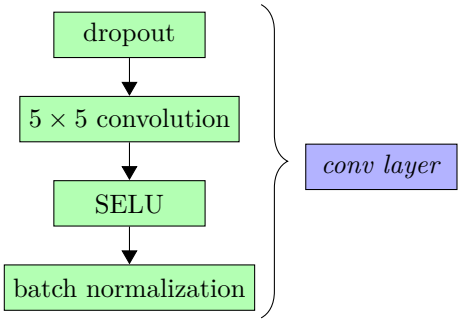


Figure 2: Basic building block in our neural network ("*conv layer*"; blue). It consists of (in green) a dropout layer to prevent overfitting, a convolutional layer to convolve the input images, the application of an activation function (i.e., SELU), and a batch normalization layer.

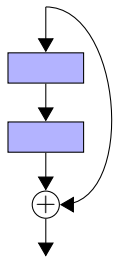


Figure 3: Illustration of a residual connection amongst "conv" layers. If the two inputs to the sum have different dimensions, an extra convolutional layer with no activation function is added to the shortcut path before the sum.

down-sampling the images at those points. In the decoding phase, we up-sample the input back to its original size. The basic building block for our network considering the above design choices is illustrated in Fig. 2.

Finally, we also use *residual connections* in our network (He et al., 2015). To construct a residual connection, we take the inputs of a given layer and sum it to the outputs of the layer after that one, as seen in Fig. 3. In our network, we connect the inputs to the outputs of the second layer, those to the outputs of the fourth layer, and so on. Residual connections are known to improve the training performance of deep neural networks and were instrumental for recent artificial intelligence breakthroughs, such as AlphaGo Zero (Silver et al., 2017). Residual connections have been used in UNets to form ResUNets (Kayalıbay et al., 2017; Zhang et al., 2017). We found that the introduction of residual connections dramatically decreased the final output error in the task at hand.

The assemblage of the building blocks into our full network is depicted in Fig. 4, where we chose to omit the residual connections. The ResUNet provides a non-linear mapping between $\mathbb{R}^{2\times128\times128}$ (representing $(\tilde{Q},\tilde{U})$), to $\mathbb{R}^{2\times128\times128}$ (representing (E,κ)). The representation in the middle of the bottleneck of the network is an element of $\mathbb{R}^{256\times32\times32}$, and it will form a processed version of the information in the input maps, optimized for generation of the output maps.

5. Data

We use simulated data to develop and compare the gravitational lensing reconstruction algorithms. We prepare 11200 independent realizations of simulated CMB maps ($\tilde{Q}$, $\tilde{U}$, E, κ), each 5 deg × 5 deg in size on sky, by extracting 160 patches from each of 70 full-sky maps. Note that the training, validation, and test sets are selected from separate sets of full-sky maps, so no contamination from large-scale information in the training set is possible. The images are pixelized into smaller images that are 128 × 128 pixels using the Lambert azimuthal equal-area projection. These simulations are created given the E and κ power spectra generated based on Planck 2013 best-fit ΛCDM cosmology: $\Omega_b h^2 = 0.0222$, $\Omega_{\rm CDM} h^2 = 0.1185$, $A_s = 2.21 \times 10^{-9}$, $n_s = 0.9624$, $\tau = 0.0943$, $H_0 = 67.94$, using CAMB (Lewis et al., 2000) for the theory spectra, and HEALPix[1] for synthesizing the $a_{\ell m}$'s. We project the $a_{\ell m}$'s to an equirectangular projection and lens them with the QUICKLENS[2] package.

We add complexity and realism to the simulations by including various white noise levels of 1, 2, 5 μK-arcmin to the $(\tilde{Q},\tilde{U})$ maps, a 1 arcmin beam smoothing, and an apodization mask. Examples of these images with varying noise levels can be seen in Fig. 5.

5.1. Data preparation and network optimization

The 11200 simulations were separated in 80 : 10 : 10 proportion into training, validation and test sets. A different network was trained for each noise level, beam smoothing and mask configuration. All results presented here come from running a trained network on the test set, which was only used once the architecture had been optimized with respect to its performance on the validation set. Some deeper architectures than the final one used here were tried with no performance improvement, but we do not claim to have explored the full parameter set. Training is done using the Adam optimizer on mini-batches of 32 samples, with initial learning rate 0.25 which is halved every time the validation error has not improved for three consecutive

[1] http://healpix.sourceforge.net
[2] https://github.com/dhanson/quicklens

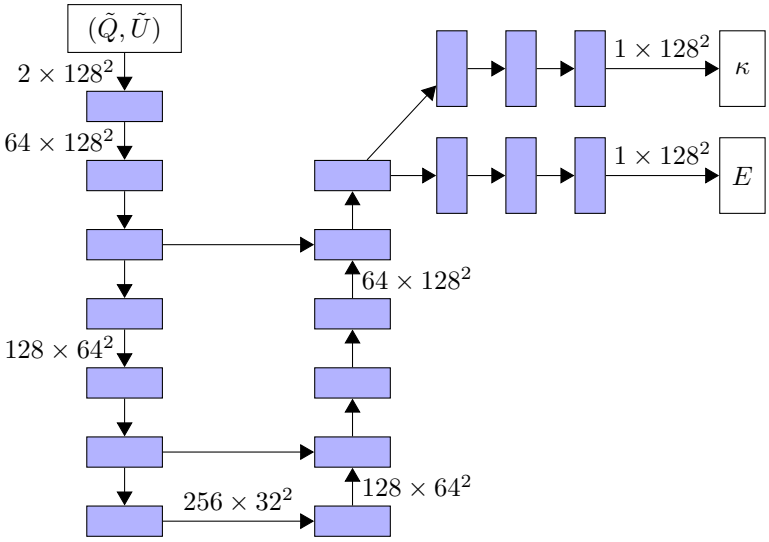


Figure 4: Complete network architecture, with residual connections omitted. Each blue layer contains the components shown in Fig. 2, except for the last conv layers before the outputs to which no activation function is applied. Arrows coming out of a network layer always carry the outputs of that layer, and arrows coming into a layer are always inputs to it. When more than one arrow comes out of a layer, the outputs are duplicated on each arrow. If more than one arrow comes into a layer, the several inputs are concatenated. The shape of the outputs of a layer is omitted if it is equal to the shape of its inputs. The network has just over 5.4 million parameters in total, and the receptive field for each pixel of the output has a size of 101×101 pixels.

epochs. The dropout rate is set to 0.3. The network is considered to have converged and training is stopped if the validation error does not improve for ten consecutive epochs. Networks were trained on a single NVIDIA P100 GPU, using Keras with a TensorFlow backend. Training took roughly 200 seconds per epoch, for a total of three to five hours. Running the trained network on each sample of the test set takes 11 ms, or a total of about one minute to run over all 1120 realizations in the test set if we include the time to load the network and simulations from memory.

Both inputs and outputs will be images with 128×128 pixels. Before training, we calculate the standard deviation of pixel values across all $\tilde{Q}$ maps in the training set, and normalize all $\tilde{Q}$ inputs to the network in the training, validation, and test sets by this value. The corresponding normalizing factors are also applied to each of $(\tilde{U}, E, \kappa)$.

We employ mean squared error in image space as the loss function for training. Since both outputs are normalized to have unit standard deviation, errors on the outputs are equally weighted. We choose to use the noise spectrum of the output convergence map κ defined in Eqn. (10) as a metric of the network performance, allowing for a direct comparison to the performance of standard methods. We have tried to introduce loss functions closer to this metric in the training of the network, such as mean squared error in Fourier space, but found no improvement on the performance.

6. Results

In this section, we first compare the trained network's output $\hat{E}$ and $\hat{\kappa}$ with the true E and κ. We then compare the κ noise spectra between traditional methods and neural network outputs. We perform null tests of the networks by passing unlensed CMB maps through them. Finally, we perform checks on the robustness of the neural network approach against different input cosmologies, and use a simple toy scenario to demonstrate how to carry out parameter estimation from the network's results.

6.1. Recovering unlensed E and κ

We apply the network to the problem of recovering the unlensed CMB maps and the convergence map from observed polarization maps, as described in §2. We will always take as inputs the lensed Q and U maps, while the desired outputs are the unlensed E map as well as κ.

Figs. 5, 6 and 7 show an example of the input $(\tilde{Q}, \tilde{U})$ maps, the target (E, κ) maps, and the predicted $(\hat{E}, \hat{\kappa})$ maps from the network for one realization in the test set. From visual inspection we can tell that for noiseless inputs, the network recovers structures in both the E-mode map and the κ map fairly well – the red and blue clumps trace each other in the true vs. the predicted maps. The bottom panels show the differences of the predicted maps from the input maps. When the inputs are

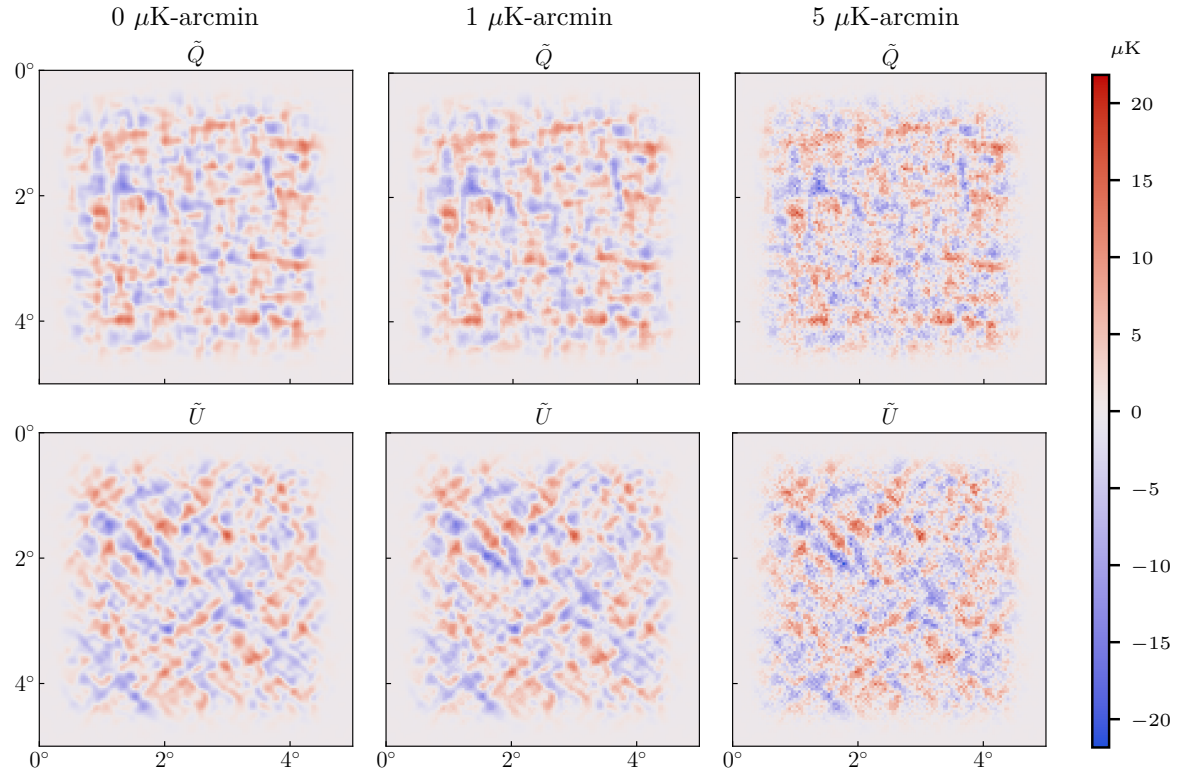


Figure 5: Example of the input maps $\tilde{Q}$ (top) and $\tilde{U}$ (bottom), with apodization applied and some of the different amounts of noise used in this work (increasing left to right), for one realization of the test set. The difference between noise levels of 0 and 1 μK-arcmin is difficult to see by eye, but 5 μK-arcmin noise is clearly visible.

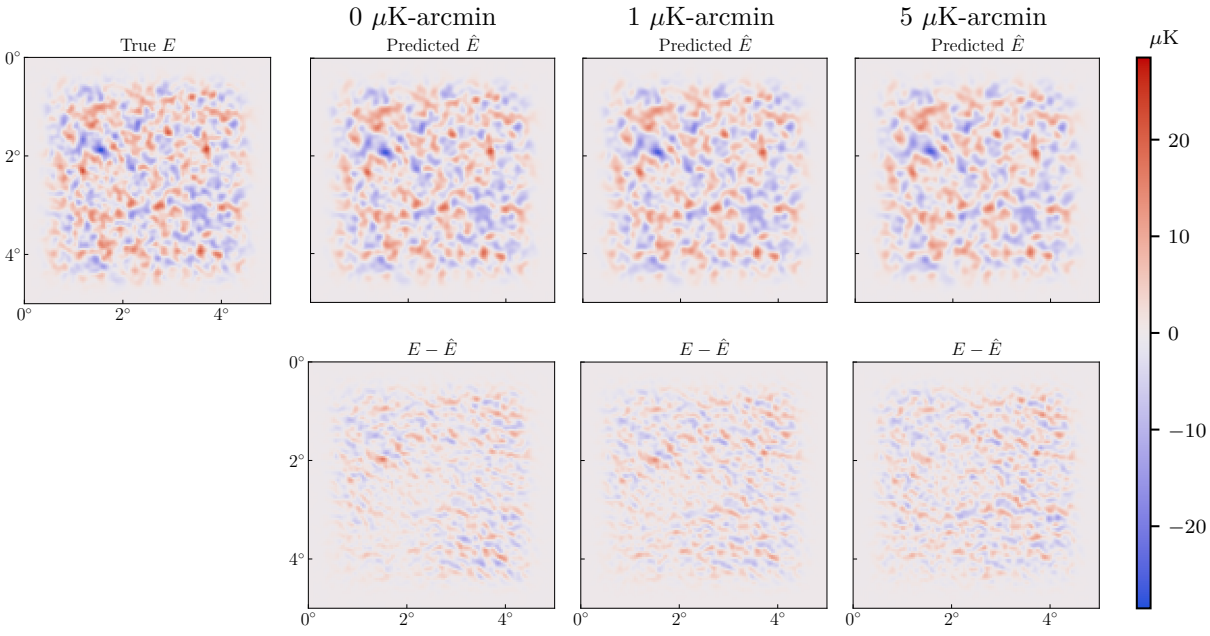


Figure 6: Example of E-mode maps for the realization corresponding to the $(\tilde{Q}, \tilde{U})$ maps shown in Fig. 5. The true map (E) is shown on the left. The ResUNet predictions $\hat{E}$ (top) and the related residuals $E - \hat{E}$ (bottom) are shown for increasing levels of input noise (0, 1, 5μK-arcmin; left to right). Comparing the true and predicted maps, most of the larger-scale structure is recovered, but some visible structure remains in the residual maps. While the amplitudes of the residual maps increase with noise, the difference between the different levels of noise is not immediately visible from the predicted maps.

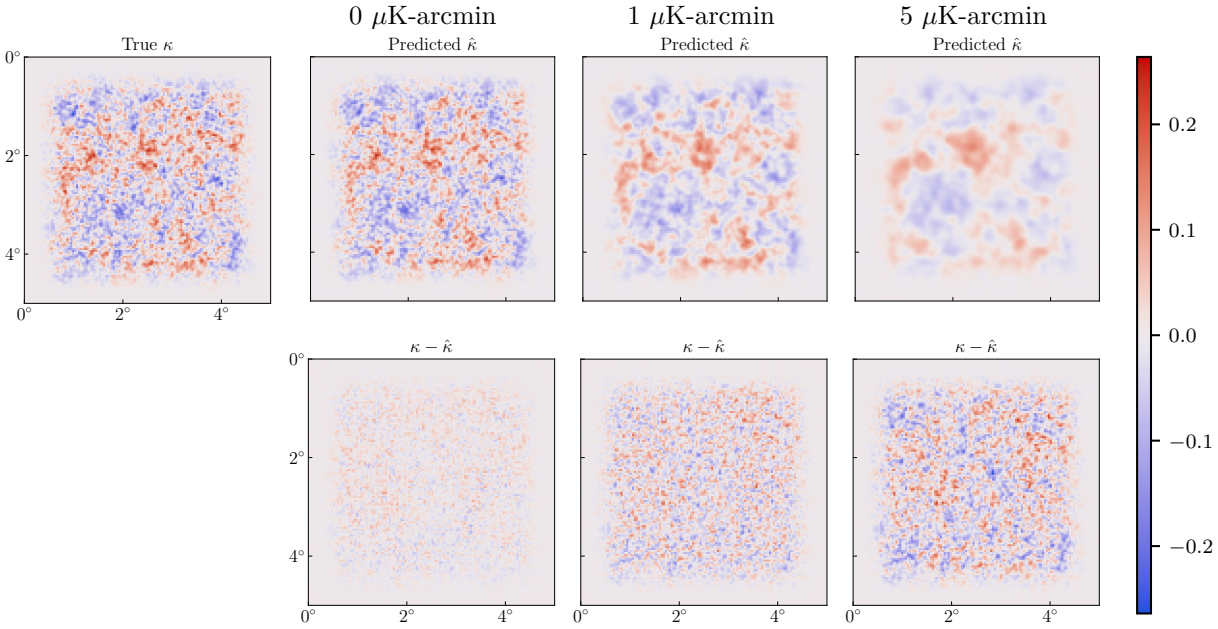


Figure 7: Example of gravitational convergence κ maps for the realization corresponding to the $(\tilde{Q}, \tilde{U})$ maps shown in Fig. 5. The true map (κ) is shown on the left. The ResUnet predictions of $\hat{\kappa}$ (top) and the related residuals $\kappa - \hat{\kappa}$ (bottom) are shown with increasing levels of noise (0, 1, 5μK-arcmin; left to right). Without noise, κ recovery is better than E recovery, and this is reflected here by the lack of large-scale structure in the left-most residual map. However, κ recovery suffers much more from the addition of noise to the inputs than E recovery, and once we reach 5 μK-arcmin only large-scale structure is visible in the predicted map.

noiseless, both the residuals from the recovered E and κ maps are within a few tens of percent of the input maps' maximum pixel value. From the residual maps, it is apparent that most large scale structure present in the κ map was captured by the network predicted map, while the residual E-mode map has visible structure that is not being captured. However, noise has a smaller effect on E than on κ, whose recovery very visibly degrades to the extent that at 5 μK-arcmin the network can recover structure at only the larger scales.

To make these observations more precise and to help quantify the efficacy of the network, we compute the power spectrum of the recovered images and compare them to the true E and κ, as shown in Fig. 8.

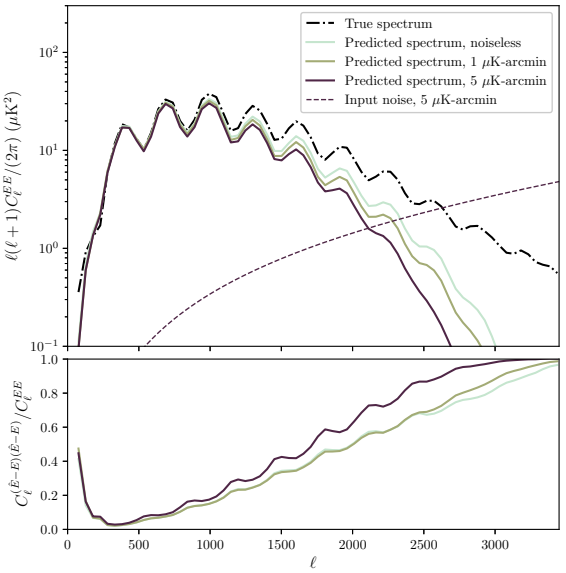
Fig. 8 shows the power spectra of the recovered E and κ maps for three input map noise levels: noiseless, 1 μK-arcmin, and 5 μK-arcmin. For the E-mode map, the mode recovery gets systematically worse as ℓ increases, but adding noise to the input maps does not degrade E-mode recovery significantly from the noiseless case. For the κ map, on the other hand, from noiseless inputs the network is able to recover more than 90% of the κ map fluctuations (80% in power spectrum) for the entire L range we consider. When noise is added to the input maps, the recovery visibly degrades. We note that even in the noiseless case, the E-mode recovery is worse than the κ recovery both on large angular scales and small angular scales. This is slightly surprising because the mathematical conversion between (Q, U) and E is very simple, compared to that between $(\tilde{Q}, \tilde{U})$ and κ. This may mean that recovery of maps that have oscillatory amount of correlation across different angular scales is a more challenging problem than that of maps whose correlation across different angular scales is smooth. One might speculate that since the conversion of (Q, U) to (E, B) is non-local, it will not be straightforward for the network to recover. To investigate this, we built a network that converts $(\tilde{Q}, \tilde{U})$ to (Q, U) and κ and found no improvement in the E recovery. While this is an intriguing problem, in this article we focus on κ recovery, so we leave optimizing for E recovery for future work.

One way to compare this performance against that of traditional reconstruction methods is to compare the noise-per-mode in the recovered lensing convergence. To do that, we estimate the equivalent of the noise spectrum in Eqn. (10) for the network. To compare $\hat{\kappa}$ recovered with the network directly to QE-reconstructed $\hat{\kappa}$, we normalize it by $1/R$ to get the equivalent of the unbiased κ,

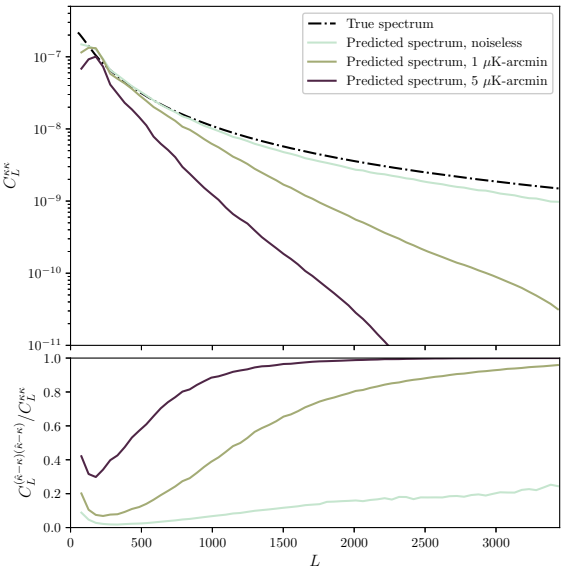
$$\hat{\bar{\kappa}} = \frac{1}{R}\hat{\kappa}, \tag{11}$$

where

$$R = \frac{\langle \kappa \hat{\kappa}^* \rangle}{\langle \kappa \kappa^* \rangle} \tag{12}$$



(a) E spectra. For comparison, we also show the spectrum of the 5 μK-arcmin white noise applied to the inputs.



(b) κ spectra.

Figure 8: In the top panels, we see the power spectra of true and recovered E and κ maps, averaged over all realizations in the test set. Note that the recovered spectra degrade as noise is increased in the input maps. In the noiseless case, κ recovery is more successful than that of E across a larger L range, as we had anticipated in Fig. 7. However noise has a much larger effect on κ recovery, strongly degrading its quality while E recovery stays qualitatively similar. Bottom panels show the ratio of the difference-map auto-spectrum to the input map auto-spectrum. The difference maps constitute the difference between the network-predicted outputs and the input maps for each noise level.

averaged over the entire validation set.[3] This is analogous to the response in QE ϕ reconstruction (Omori et al., 2017), describing how much of the true map is correctly estimated. Note that by construction we will then have

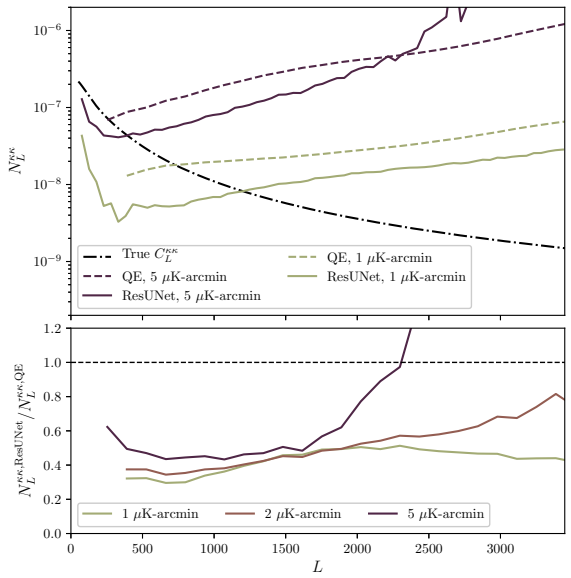
$$\langle \kappa \hat{\hat{\kappa}}^* \rangle = \langle \kappa \kappa^* \rangle \tag{13}$$

on the validation set.

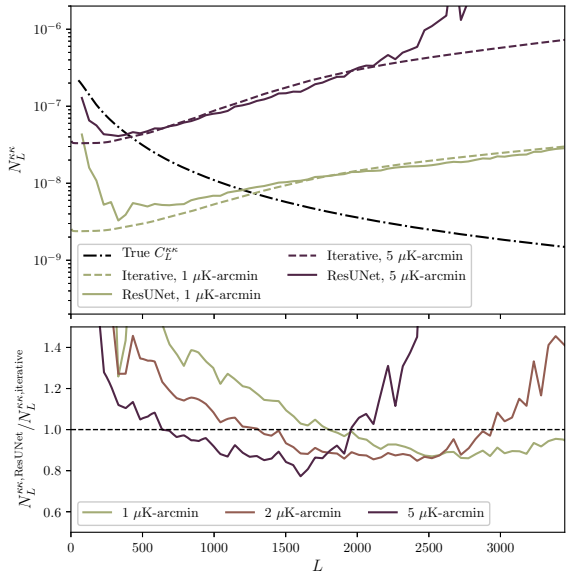
After normalizing with R, we compute the power spectrum of $\hat{\hat{\kappa}}$ and extract the noise spectrum by differencing the auto-spectrum of $\hat{\hat{\kappa}}$ and κ, as in Eqn. (10). The noise spectra from κ reconstructed through the QE and from the ResUNet are shown in Fig. 9a. We see that at angular scales below L of 2500, the noise levels from the ResUNet are lower than the QE's noise levels, regardless of the input maps' noise levels. While 5 μK-arcmin is a high enough noise level that the QE is in principle close to optimal, the actual reconstruction by the QE on these maps does not provide as low as noise level as the QE analytic N_0 estimates, mainly because of E-B mode mixing due to boundary effects. We choose to highlight 1 and 5 μK-arcmin as these are realistic next-generation noise levels, but note that a similar performance difference is obtained for noiseless maps. From these results, we conclude that ResUNets outperform the QE for input map noise levels below 5 μK-arcmin.

In Fig. 9b, we also show the N_0 noise curves from the iterative estimator using the formalism of Smith et al. (2012) outlined in §3.2. Once again we highlight 1 and 5 μK-arcmin noise levels, as with noiseless inputs the iterative estimator can reach zero error. We see that the neural networks provide comparable performance to the iterative estimator across a wide range of angular scales. This means that the neural network approach is able to extract information at efficiency close to the iterated EB estimator. Since the inputs to the network are CMB $(\tilde{Q}, \tilde{U})$ polarization maps, the neural network should also include information from the EE estimator of standard methods. In other words, the noise levels in the κ maps extracted by the neural network are higher than the combined N_0 from iterated EB and EE N_0. With that said, the iterated EB estimator provides the lowest reconstruction noise among individual QE's for future CMB experiment noise levels, so the neural network κ noise being not far off from

[3] There are other ways one can define noise in the $\hat{\kappa}$ maps. For example, it can be extracted through the correlation coefficient of κ and $\hat{\kappa}$. We chose to define noise this way to symmetrize R in Eqn. (8) and R in Eqn. (11). In ideal conditions (i.e. white noise, azimuthally-symmetric filtering), R as defined in (12) would give identical results as R as defined in (8).



(a) To further evaluate the quality of κ reconstruction by ResUNets at different input noise levels, we compare the noise spectra from ResUNets to those from quadratic estimators. We see that the results have $50-70\%$ less noise than quadratic estimator reconstructions across a wide range of angular scales L. For input noise of 5 μK-arcmin, performance quickly degrades for $L \gtrsim 2000$.



(b) We also compare noise spectra of κ reconstruction using ResUNets to expected noise levels from iterative estimators (which approach maximum-likelihood results), as in Smith et al. (2012). The iterative method noise levels are taken from an EB estimator only. For all noise levels used here, ResUNets and iterative methods have comparable noise levels across a wide range of L. Significant performance differences mainly occur for the smallest and largest scales pictured.

Figure 9: We compare κ reconstruction using ResUNets to current standard methods. The noise spectra are calculated by taking the average spectrum over all realizations in the test set.

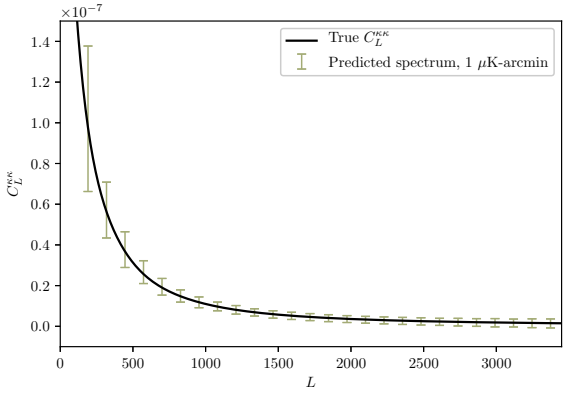


Figure 10: To quantify the variance of the network output on different simulation realizations, we rescale and then remove the noise spectrum from each recovered $\hat{\kappa}$ with 1 μK-arcmin input $(\tilde{Q}, \tilde{U})$ maps. We then find the 1-σ deviation around the mean spectrum. This variation comes both from the difference in spectrum between each simulation (cosmic variance) and from the noise in the measurements. It is the uncertainty in a measurement of the spectrum from a single simulation realization or on real data.

that demonstrates its viability for beyond-QE ϕ reconstruction. We should also note that once the network has been trained on simulations, it can be applied to real data very quickly when compared to other maximum likelihood methods under development.

In Fig. 10, we seek to quantify the variance of the network output when given simulations from the test set. To obtain these error bars, we calculate the spectrum from each $\hat{\kappa}$, rescale it by R, and subtract the average noise spectrum $N_L^{\kappa\kappa}$ shown in Fig. 9. We use bins of width 127 for Fig. 10. The error bars represent one-standard-deviation variations from the mean power spectrum of the test set. This calculation provides a measure of the effects of cosmic variance and noise variance between different patches in the test set on the neural network predictions. For noiseless inputs, the relative uncertainty in the first bin is 33%. The uncertainty decreases to 13% at $L \sim 1500$ and 10% at $L \sim 3000$. With 1 μK-arcmin noise, as shown in Fig. 10, the relative uncertainty in the first bin remains at 33%, but the increased $N_L^{\kappa\kappa}$ for higher L brings it up to 33% at $L \sim 1500$ and 117% at $L \sim 3000$. With 5 μK-arcmin noise, the first bin suffers from the increase in noise, so the relative error bar goes up to 48%. At $L \sim 1500$, we reach 298%. Beyond that L the noise shoots up so the reconstruction of κ is no longer meaningful. This variation would form a part of the uncertainty budget when the neural networks are applied to real data.

6.2. Null test

A basic test to check that the network encodes a sensible mapping of the input lensed maps to the underlying lensing convergence is to pass unlensed maps through the network and compare the output to the noise spectrum. For this network to be useful for cosmology, we need to know that when fed with a map with no lensing, the network recovers a field that is uncorrelated with the convergence. Therefore, we feed unlensed versions of the (Q, U) maps (that is, maps with $\kappa = 0$) through the network trained on lensed $(\tilde{Q}, \tilde{U})$ maps.

As a first test, to check that the network has not overfit the training set, we run unlensed versions of the noiseless (Q, U) maps in the training set through, and calculate the cross-spectrum of the true κ (present in the training set, but not applied to the maps we run here) and the output field $\hat{\kappa}$. We obtain

$$\langle \kappa \hat{\kappa}^* \rangle < 10^{-4} \langle \kappa \kappa^* \rangle, \tag{14}$$

showing no significant amount of overfitting.

Next, in defining the noise proxy as the difference between the auto-spectrum of $\hat{\hat{\kappa}}$ and κ, we have assumed that we can model the $\hat{\hat{\kappa}}$ as $\hat{\hat{\kappa}} = \kappa + n_\kappa$, where n_κ is an uncorrelated noise term. If this model is accurate, we expect the auto-spectrum of $\hat{\hat{\kappa}}$ maps output when given unlensed (Q, U) maps to be consistent with the noise spectrum $N_L^{\kappa\kappa}$ we used in the previous section. In other words,

$$C_L^{\hat{\hat{\kappa}}\hat{\hat{\kappa}}} = \langle \hat{\hat{\kappa}} \hat{\hat{\kappa}}^* \rangle = \langle \kappa \kappa^* \rangle + \langle n_\kappa n_\kappa^* \rangle = C_L^{\kappa\kappa} + N_L^{\kappa\kappa}. \tag{15}$$

To test this, we run unlensed versions of the input maps (Q, U) in the test set, with noise levels of 1 and 5 μK-arcmin, through the networks trained on lensed inputs at the same level of noise. We then rescale the outputs of the network using the same factor $1/R$, and compare $C_L^{\hat{\hat{\kappa}}\hat{\hat{\kappa}}}$ from unlensed inputs to the noise spectra that we extracted from the tests on lensed inputs. For both white noise levels, the $C_L^{\hat{\hat{\kappa}}\hat{\hat{\kappa}}}$ from unlensed inputs align well with the noise spectrum with percent-level differences, as we can see in Fig. 11. The difference can be attributed to two potential causes: (1) a subdominant part of the noise n_κ that correlates with the lensing convergence field, similar to N_1 in standard QE methods; (2) the R computed from the training set fluctuates high/low and biases $\hat{\hat{\kappa}}$. For future work, it will be important to characterize how the network interacts with higher-order noise terms. In conclusion, this check confirms the assumption of $\hat{\hat{\kappa}} = \kappa + n_\kappa$ to be good to within a few percent.

6.3. Tests on cosmology

To test that we can apply the network on actual data, we should check whether it will be sensitive to changes in the input (Q, U) maps' cosmology, and has not simply learned to reproduce the cosmology in the training set. To this end, we give noiseless $(\tilde{Q}, \tilde{U})$ maps that are generated with different parameters as inputs to the network trained with the fiducial set of $(\tilde{Q}, \tilde{U}, E, \kappa)$ maps. The two different cosmologies have $\Omega_{\rm CDM} h^2 = 0.1085$ and $\Omega_{\rm CDM} h^2 = 0.1285$ respectively (while $\Omega_{\rm CDM} h^2 = 0.1185$ for the fiducial cosmology), with all the other parameters fixed to the fiducial. We found that the recovered κ spectrum is significantly different from the spectrum recovered from the fiducial set.

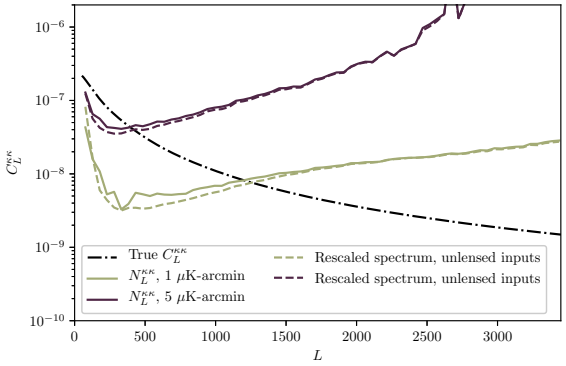


Figure 11: To test how robust the network is to inputs with different levels of lensing, we look at the lensing maps predicted by the network when we pass unlensed versions of the test set inputs through it. We see that the rescaled outputs have a spectrum that is within a few percent of the noise spectra we found above. This confirms that our model of uncorrelated noise is accurate.

To quantify the difference, we pose the null hypothesis: "if we apply the network trained on maps generated from the fiducial cosmology on real data with cosmologies different from the fiducial cosmology, we will get the same output as the fiducial cosmology." We calculate the χ^2 of each sample of the recovered κ spectrum from the two different input cosmologies compared against the average recovered κ spectrum from the fiducial cosmology, as

$$\chi^2 = (\boldsymbol{d} - \boldsymbol{\mu})^\dagger \mathbf{C}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}), \tag{16}$$

where $\boldsymbol{d}$ is the binned noise-debiased $\hat{\hat{\kappa}}$ power spectrum, $\boldsymbol{\mu}$ is the κ spectrum in the fiducial cosmology, and $\mathbf{C}$ is the covariance matrix that describes the L to L' bin covariance. We construct $\boldsymbol{\mu}$ and $\mathbf{C}$ from the outputs of the network for all realizations in the test set (generated with the fiducial cosmology). We note the error bars shown in Fig. 10 are the square roots of the diagonal elements in $\mathbf{C}$. Using χ^2, we rule out the null hypothesis at 2.9 ± 0.9 and 4.5 ± 1.4 σ respectively. This demonstrates that the network is sensitive to differences in the input maps' cosmology.

To demonstrate that we can use the recovered $\hat{\tilde{\kappa}}$ from neural networks for extracting cosmological parameters, we use the $\hat{\kappa}$ spectrum for parameter estimation in a simple case, where we only fit for the parameter $\Omega_{\rm CDM}h^2$. We construct a Gaussian likelihood for the parameter $\theta = \Omega_{\rm CDM}h^2$ given the $\hat{\tilde{\kappa}}$ spectrum:

$$\mathcal{L}(\theta|\boldsymbol{d}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu}(\theta))^\dagger \mathbf{C}^{-1}(\boldsymbol{d}-\boldsymbol{\mu}(\theta))\right), \tag{17}$$

where $\boldsymbol{\mu}(\theta)$ is the κ spectrum given different $\Omega_{\rm CDM}h^2$ values while $\mathbf{d}$ and $\mathbf{C}$ are as in Eqn. (16), with the covariance constructed from simulations. When noise debiasing the $\hat{\tilde{\kappa}}$ spectrum, we always subtract the noise spectrum estimated from the fiducial set from the $\hat{\tilde{\kappa}}$ spectrum recovered from input $(\tilde{Q}, \tilde{U})$ maps of different cosmologies.

From the set of simulations with the fiducial cosmology, the mean and standard deviation of the distribution of maximum-likelihood $\Omega_{\rm CDM}h^2$ values from 320 realizations are 0.1183 ± 0.0016, recovering the input $\Omega_{\rm CDM}h^2 = 0.1185$ at within about 0.1σ. From the set of simulations that has different input $\Omega_{\rm CDM}h^2$ values, the mean of the distribution of maximum-likelihood $\Omega_{\rm CDM}h^2$ values are 0.1096 for input $\Omega_{\rm CDM}h^2 = 0.1085$ and 0.1260 for $\Omega_{\rm CDM}h^2 = 0.1285$, respectively. The low $\Omega_{\rm CDM}h^2$ set is biased high, whereas the high $\Omega_{\rm CDM}h^2$ is biased low. This is due to the noise debias term being too small for the low $\Omega_{\rm CDM}h^2$ case and too large for the high $\Omega_{\rm CDM}h^2$ case. In standard cosmology parameter estimation from the lensing power spectrum, since the noise debias terms are cosmology-dependent, one can correct for the difference in the noise terms between the fiducial cosmology and the sampled cosmology to avoid this bias (see e.g. Appendix C of Planck Collaboration et al. (2016b)). In our case, when we apply the network to data and use $\hat{\tilde{\kappa}}_{\rm data}$ for parameter estimation, we can apply similar corrections by computing the noise terms for many different input cosmologies and obtaining the numerical derivatives of the noise terms with respect to the CMB power spectra. But that is outside the scope of current work. From this simple test, we conclude that the neural network recovered κ map and the input $\Omega_{\rm CDM}h^2$ have a one-to-one mapping. While more detailed calibration has to be done, there is no conceptual roadblock for parameter estimation from these maps.

7. Discussion

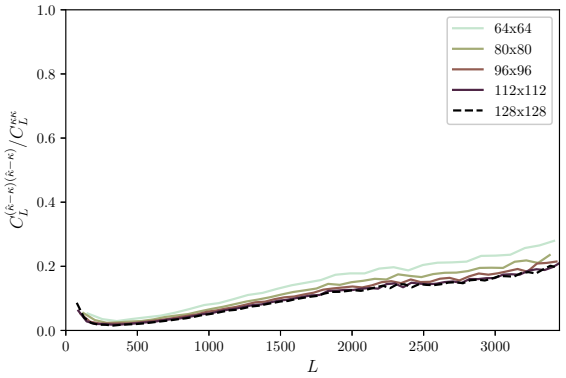
The results outlined in the previous section provide a strong argument for the viability of using neural networks to reconstruct the κ map for next-generation low-noise CMB experiments. In this section, we comment on some of the possible issues and what can be done to mitigate them.

One clear feature of Fig. 9 is the κ reconstruction noise shooting up for $L \gtrsim 2000$ when the inputs have 5 μK-arcmin noise. A similar phenomenon also happens for 2 μK-arcmin input noise after $L \gtrsim 3000$. These angular scales are approximately where the RMS of the noise added to the input (signal) maps dominates the signal's RMS, thus submerging information contained in these modes of the inputs. This issue affects the performance of the neural networks much more sharply than it affects that of traditional methods. The correlation between the angular scales at which the input's signal-to-noise ratio falls below 1 and the angular scales at which the output κ power spectrum degrades leads us to conjecture that the network is using information that is more local in angular scales than standard methods.

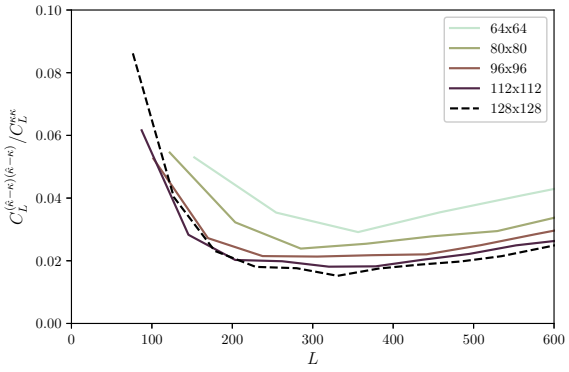
One way to resolve this would be to impose a physical model on the κ power spectrum: for example, one can deduce the smaller angular scale mode information from the structure contained in the larger angular scales (lower-L) modes, given our knowledge of the shape of the lensing convergence power spectrum. However, this is out of the scope of this work where we are interested in a model-free inference of κ and primordial E.

One other feature of the results, already apparent in Fig. 8 even on the noiseless data, is a decrease in recovered power for the low-L/ℓ modes in both κ and E. This manifests itself in Fig. 9 as a sharp increase in the noise levels. This feature is an effect of the finite size of the maps we have used (5 degrees across). We have tested this hypothesis by training networks to learn the same task on smaller maps, obtained by cutting out sections of the simulations used in this paper. We found that the recovered power starts to drop similarly at smaller angular scales (larger L/ℓ values) when we decrease the map size. This can be seen in Fig. 12. Extrapolating this tendency, we could improve the results in the low-L/ℓ region by performing lensing reconstruction on a larger patch of sky. We could also increase the receptive field of the networks used, although deeper networks in the same family resulted in no improvement in the validation phase of this work.

One final possible issue is the inherent randomness associated with the training of a neural network. Since the weights are randomly initialized, and the batches used to calculate the gradient at each step are randomly selected from the training set, the final mappings learned by two networks with different initializations will not be exactly the



(a) Ratio between the spectrum of the difference-maps and spectrum of the true maps for different cutout sizes.



(b) Same ratio, zoomed into the low L region and capped at 10%.

Figure 12: We vary the cutout size and look at the ratio plotted in the bottom panel of Fig. 8b, each line corresponding to a new network trained on images of that size. We can see that performance degrades overall as the cutouts become smaller, as might be expected since there is less information contained in the inputs. On the bottom we zoom in on the low-L region, showing that the uptick in relative difference occurs at larger L for smaller images. This suggests that low-L performance could be improved by working on a larger patch of sky.

same. To test how important this effect is, we trained 20 networks with different initializations on the noiseless data, and calculated the power spectra for the κ maps predicted by each network for each realization in the test set.

In order to evaluate the spread of predictions from different networks, we calculate the power spectrum $C^{\kappa\kappa}_{L,i\alpha}$ of realization i obtained by network α, and bin it in the same way as we did for plotting. Denoting the binned spectra as $C^{\kappa\kappa}_{b,i\alpha}$, we then evaluate

$$R_{b,i\alpha} = \frac{C^{\kappa\kappa}_{b,i\alpha}}{\left\langle C^{\kappa\kappa}_{b,i} \right\rangle_\alpha}, \tag{18}$$

where $\left\langle C^{\kappa\kappa}_{b,i} \right\rangle_\alpha$ is the binned power spectrum for each realization averaged over the results of all 20 networks. $R_{b,i\alpha}$ are dimensionless quantities whose spread around 1 gives us a measure of the uncertainty due to the randomness of the neural network training algorithm. We found that the standard deviation of $R_{b,i\alpha}$ over all networks, realizations, and bins was 3.0%. If a lower variability is desirable, we could use an ensemble of networks to find a final result, or use weight averaging as introduced in Izmailov et al. (2018). We leave these refinements to future work.

8. Conclusion and Outlook

In this work, we demonstrated that deep learning algorithms (in this work, Residual UNets) can be used to recover the lensing convergence and unlensed CMB maps in simulated data. The networks were trained on 5×5 deg^2 sized simulated CMB maps. We first compare our predicted maps (Figs. 6, 7) and power spectra (Fig. 8) to the true signals for various noise levels, showing that modes between ℓ =100 and ℓ =600 are predicted with errors lower than 10% for E and 20% for κ for input noise of up to 1 μK-arcmin. We then compare our results for the lensing convergence maps to the standard quadratic estimator method at noise levels between 1 and 5 μK-arcmin. We show that ResUNets outperform the QE by $50-70$% across a wide range of L values in this comparison. This is reflected in the power spectra, and in the ratios of noise spectra in Fig. 9a. In fact, the results approximate maximum likelihood EB estimator results, as we can see in Fig. 9b.

There are some challenges still present in the use of these methods. We found that the discrepancies between the true and the recovered maps, even in noiseless cases, increased with the multipole number: small-scale features are more challenging to recover. While this is also the case in standard

methods, we note that neural networks tend to perform even worse. We speculate that standard methods perform better because of their ability to provide a physical model of the signal and the noise.

In future work, we plan to apply a similar network to recover primordial B modes as well as E and κ, with more attention paid to parameter estimation from the recovered delensed E- and B-mode maps. These are equivalent to delensed CMB polarization maps from standard methods, and will be important for constraining r and $N_{\rm eff}$. One interesting route would be to directly estimate cosmological parameters from the input CMB maps themselves, as introduced in e.g. Ravanbakhsh et al. (2017). In addition, we plan on including simulations of galactic and extragalactic foregrounds in the input maps to both extract the foreground components and study their effects on κ and unlensed CMB recovery. To extend this network usage for actual data that are often taken from larger regions of the sky, we also need to use simulations from larger sky patches. This might necessitate the use of different network architectures such as group-equivariant convolutional networks (in particular, the spherical convolutional networks in Cohen et al., 2018; Kondor et al., 2018), as the flat-sky approximation will no longer be valid. During the revision stages of this manuscript, other architectures have been proposed to address this issue (Perraudin et al., 2019; Krachmalnicoff and Tomasi, 2019). It would also be interesting to apply techniques similar to those in this work to the removal of other foregrounds which are hard to model explicitly. We expect the inherent non-linearities of deep neural networks to be helpful in such tasks.

The CMB is a potentially powerful data set with which to explore and develop deep learning techniques. Because standard techniques to analyze the CMB are quite mature and rich with physical insights, we can develop a better picture of what can be understood with deep learning approaches by comparing the information recovered using neural networks to the standard methods. This helps us uncover opportunities to improve on standard analyses. An area where this is especially true is extraction of information contained in the input maps that is not as well-modeled as the CMB itself, such as that coming from galactic foregrounds, instrumental noise, and systematics.

Machine learning can be extremely effective for cosmological data analyses. However, in order for us to fully leverage its power we first need to elucidate how standard statistical quantities like signal/noise (co-)variance are extracted in each specific application. This is particularly relevant as machine learning tools are increasingly utilized for scientific analysis and to aid in gaining physical insight. This work represents a small step in working out one example of connecting standard physical analyses of gravitational lensing to a neural network approach.

Acknowledgments

Author Contributions

Caldeira: Performed all neural net computational work; innovated the choice of NN architecture; performed all NN diagnostic analysis; kept the work alive and pushed it through the challenging stages.
Wu: Directed scientific analysis related to CMB; established the noise measure as a method to compare between QE, max likelihood and NN; performed QE analysis of maps; performed CMB Simulations; performed parameter estimation; guided NN diagnostics; guided analysis.
Nord: Initiated problem concept; initiated NN architecture; performed initial network tests on toy data; guided narrative of paper; guided analysis.
Avestruz: Consultation on analysis methods; writing manuscript.
Trivedi: Consultation on NN methods.
Story: Initiated problem concept; performed initial network tests on toy data; generated initial set of CMB simulations.

The authors thank Daniel Gruen for discussions early on. We thank Wayne Hu for useful discussions. We thank earlier paper reviewers, Renée Hložek, Alessandro Manzotti, and Eyjólfur Guðmundsson for their comments.

This work is supported by the Deep Skies Community (`deepskieslab.com`), which helped to bring together the authors and reviewers. The authors of this paper have committed themselves to performing this work in an equitable, inclusive, and just environment, and we hold ourselves accountable, believing that the best science is contingent on a good research environment.

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

JC was supported in part by NSF Grant No. PHY-1720480. WLKW, BN, CA were supported in part by the Kavli Institute for Cosmological Physics at the University of Chicago through an endowment from the Kavli Foundation and its founder Fred Kavli. CA was in addition supported by the Kavli Institute for Cosmological Physics at the University of Chicago through NSF Grant No. PHY-1125897 and the Enrico Fermi Institute. ST was supported by the National Science Foundation under Grant

No. DMS-1439786 while the author was in residence at the Institute for Computational and Experimental Research in Mathematics in Providence, RI, during the non-linear algebra program.

References

Abazajian, K.N., Adshead, P., Ahmed, Z., et al., 2016. CMB-S4 science book, first edition. arXiv:1610.02743 [astro-ph.CO] .

Ade, P.A.R., Akiba, Y., Anthony, A.E., et al., 2014. Measurement of the Cosmic Microwave Background Polarization Lensing Power Spectrum with the POLARBEAR Experiment. Physical Review Letters 113, 021301. doi:10.1103/PhysRevLett.113.021301, arXiv:1312.6646.

Bahdanau, D., Cho, K., Bengio, Y., 2016. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations .

Bennett, C.L., Larson, D., Weiland, J.L., et al., 2013. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results. Astrophys. J. Suppl. 208, 20. doi:10.1088/0067-0049/208/2/20, arXiv:1212.5225.

Benson, B.A., Ade, P.A.R., Ahmed, Z., et al., 2014. SPT-3G: a next-generation cosmic microwave background polarization experiment on the South Pole telescope, in: Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy VII, p. 91531P. doi:10.1117/12.2057305, arXiv:1407.2973.

de Bezenac, E., Pajot, A., Gallinari, P., 2017. Deep learning for physical processes: Incorporating prior scientific knowledge. CoRR abs/1711.07970. URL: http://arxiv.org/abs/1711.07970, arXiv:1711.07970.

BICEP2 Collaboration, Keck Array Collaboration, Ade, P.A.R., et al., 2018. Constraints on Primordial Gravitational Waves Using Planck, WMAP, and New BICEP2/Keck Observations through the 2015 Season. Phys. Rev. Lett. 121, 221301. doi:10.1103/PhysRevLett.121.221301, arXiv:1810.05216.

BICEP2 Collaboration, Keck Array Collaboration, Ade, P.A.R., et al., 2016a. BICEP2/Keck Array VIII: Measurement of Gravitational Lensing from Large-scale B-mode Polarization. Astrophys. J. 833, 228. doi:10.3847/1538-4357/833/2/228, arXiv:1606.01968.

BICEP2 Collaboration, Keck Array Collaboration, Ade, P.A.R., et al., 2016b. Improved Constraints on Cosmology and Foregrounds from BICEP2 and Keck Array Cosmic Microwave Background Data with Inclusion of 95 GHz Band. Physical Review Letters 116, 031302. doi:10.1103/PhysRevLett.116.031302, arXiv:1510.09217.

Cho, K., Merrienboer, B.v., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8) .

Cohen, T.S., Geiger, M., Koehler, J., Welling, M., 2018. Spherical CNNs. ArXiv e-prints arXiv:1801.10130.

Dodelson, S., 2003. Modern Cosmology. Academic Press.

Dumoulin, V., Visin, F., 2016. A guide to convolution arithmetic for deep learning. ArXiv e-prints arXiv:1603.07285.

Elman, J.L., Zipser, D., 1988. Learning the hidden structure of speech. J. Acoust Soc Am 83, 1615–1626.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. CoRR abs/1512.03385. URL: http://arxiv.org/abs/1512.03385, arXiv:1512.03385.

Henning, J.W., Sayre, J.T., Reichardt, C.L., et al., 2018. Measurements of the Temperature and E-mode Polarization of the CMB from 500 Square Degrees of SPTpol Data. Astrophys. J. 852, 97. doi:10.3847/1538-4357/aa9ff4, arXiv:1707.09353.

Hinshaw, G., Larson, D., Komatsu, E., et al., 2013. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results. Astrophys. J. Suppl. 208, 19. doi:10.1088/0067-0049/208/2/19, arXiv:1212.5226.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.

Hirata, C.M., Seljak, U., 2003. Reconstruction of lensing from the cosmic microwave background polarization. Phys. Rev. D 68, 083002. doi:10.1103/PhysRevD.68.083002, arXiv:astro-ph/0306354.

Hu, W., 2000. Weak lensing of the CMB: A harmonic approach. Phys. Rev. D 62. doi:10.1103/PhysRevD.62.043007.

Hu, W., Okamoto, T., 2002. Mass Reconstruction with Cosmic Microwave Background Polarization. Astrophys. J. 574, 566–574. doi:10.1086/341110, arXiv:astro-ph/0111606.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D.P., Wilson, A.G., 2018. Averaging weights leads to wider optima and better generalization. CoRR abs/1803.05407. URL: http://arxiv.org/abs/1803.05407, arXiv:1803.05407.

Kamionkowski, M., Kosowsky, A., Stebbins, A., 1997. A Probe of Primordial Gravity Waves and Vorticity. Physical Review Letters 78, 2058–2061. doi:10.1103/PhysRevLett.78.2058, arXiv:astro-ph/9609132.

Kayalıbay, B., Jensen, G., van der Smagt, P., 2017. CNN-based segmentation of medical imaging data. CoRR abs/1701.03056. URL: http://arxiv.org/abs/1701.03056, arXiv:1701.03056.

Keisler, R., Hoover, S., Harrington, N., et al., 2015. Measurements of Sub-degree B-mode Polarization in the Cosmic Microwave Background from 100 Square Degrees of SPTpol Data. Astrophys. J. 807, 151. doi:10.1088/0004-637X/807/2/151, arXiv:1503.02315.

Kesden, M., Cooray, A., Kamionkowski, M., 2003. Lensing reconstruction with CMB temperature and polarization. Phys. Rev. D 67, 123507. doi:10.1103/PhysRevD.67.123507, arXiv:astro-ph/0302536.

Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S., 2017. Self-normalizing neural networks. Advances in Neural Information Processing Systems 30 (NIPS 2017) .

Kondor, R., Lin, Z., Trivedi, S., 2018. Clebsch-Gordan Nets: a Fully Fourier Space Spherical Convolutional Neural Network. ArXiv e-prints arXiv:1806.09231.

Krachmalnicoff, N., Tomasi, M., 2019. Convolutional Neural Networks on the HEALPix sphere: a pixel-based algorithm and its application to CMB data analysis. arXiv e-prints , arXiv:1902.04083arXiv:1902.04083.

Lange, A.E., Ade, P.A., Bock, J.J., et al., 2001. Cosmological parameters from the first results of Boomerang. Phys. Rev. D 63, 042001–+.

Lewis, A., Challinor, A., 2006. Weak gravitational lensing of the CMB. Phys. Rep. 429, 1–65. doi:10.1016/j.physrep.2006.03.002, arXiv:astro-ph/0601594.

Lewis, A., Challinor, A., Lasenby, A., 2000. Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models. Astrophys. J. 538, 473–476. doi:10.1086/309179, arXiv:astro-ph/9911177.

Louis, T., Grace, E., Hasselfield, M., et al., 2017. The Atacama Cosmology Telescope: two-season ACTPol spectra and parameters. JCAP 6, 031. doi:10.1088/1475-7516/2017/06/031, arXiv:1610.02360.

Manzotti, A., Story, K.T., Wu, W.L.K., et al., 2017. CMB Polarization B-mode Delensing with SPTpol and Herschel. Astrophys. J. 846, 45. doi:10.3847/1538-4357/aa82bb, arXiv:1701.04396.

Mather, J.C., Cheng, E.S., Cottingham, D.A., et al., 1994. Measurement of the cosmic microwave background spectrum by the COBE FIRAS instrument. Astrophys. J. 420, 439–444. doi:10.1086/173574.

Matsumura, T., Akiba, Y., Borrill, J., et al., 2014. Mission Design of LiteBIRD. Journal of Low Temperature Physics 176, 733–740. doi:10.1007/s10909-013-0996-1, arXiv:1311.2847.

Mehta, P., Bukov, M., Wang, C.H., et al., 2018. A high-bias, low-variance introduction to Machine Learning for physicists. ArXiv e-prints arXiv:1803.08823.

Millea, M., Anderes, E., Wandelt, B.D., 2017. Bayesian delensing of CMB temperature and polarization. ArXiv e-prints arXiv:1708.06753.

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. IEEE International Conference on Computer Vision (ICCV) , 1520–1528.

Omori, Y., Chown, R., Simard, G., et al., 2017. A 2500 deg^2 CMB Lensing Map from Combined South Pole Telescope and Planck Data. Astrophys. J. 849, 124. doi:10.3847/1538-4357/aa8d1d, arXiv:1705.00743.

Perraudin, N., Defferrard, M., Kacprzak, T., Sgier, R., 2019. DeepSphere: Efficient spherical convolutional neural network with HEALPix sampling for cosmological applications. Astronomy and Computing 27, 130. doi:10.1016/j.ascom.2019.03.004, arXiv:1810.12186.

Planck Collaboration, Ade, P.A.R., Aghanim, N., et al., 2016a. Planck 2015 results. XIII. Cosmological parameters. Astr. & Astroph. 594, A13. doi:10.1051/0004-6361/201525830, arXiv:1502.01589.

Planck Collaboration, Ade, P.A.R., Aghanim, N., et al., 2016b. Planck 2015 results. XV. Gravitational lensing. Astr. & Astroph. 594, A15. doi:10.1051/0004-6361/201525941, arXiv:1502.01591.

Planck Collaboration, Aghanim, N., Akrami, Y., et al., 2018. Planck 2018 results. VI. Cosmological parameters. ArXiv e-prints arXiv:1807.06209.

POLARBEAR Collaboration, Ade, P.A.R., Aguilar, M., et al., 2017. A Measurement of the Cosmic Microwave Background B-mode Polarization Power Spectrum at Subdegree Scales from Two Years of polarbear Data. Astrophys. J. 848, 121. doi:10.3847/1538-4357/aa8e9f, arXiv:1705.02907.

Ravanbakhsh, S., Oliva, J., Fromenteau, S., et al., 2017. Estimating Cosmological Parameters from the Dark Matter Distribution. ArXiv e-prints arXiv:1711.02033.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. CoRR abs/1505.04597. URL: http://arxiv.org/abs/1505.04597, arXiv:1505.04597.

Rumelhart, D.E., McClelland, J.L., Group, P.R., 1986. Parallel distributed processing: explorations in the microstructure of cognition, volumes 1 and 2. MIT Press.

Seljak, U., Zaldarriaga, M., 1997. Signature of Gravity Waves in the Polarization of the Microwave Background. Physical Review Letters 78, 2054–2057. doi:10.1103/PhysRevLett.78.2054, arXiv:astro-ph/9609169.

Shelhamer, E., Long, J., Darrell, T., 2016. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Sherwin, B.D., van Engelen, A., Sehgal, N., et al., 2017. Two-season Atacama Cosmology Telescope polarimeter lensing power spectrum. Phys. Rev. D 95, 123529. doi:10.1103/PhysRevD.95.123529, arXiv:1611.09753.

Silver, D., Schrittwieser, J., Simonyan, K., et al., 2017. Mastering the game of Go without human knowledge. Nature 550, 354–359. doi:10.1038/nature24270.

Simard, G., Omori, Y., Aylor, K., et al., 2017. Constraints on Cosmological Parameters from the Angular Power Spectrum of a Combined 2500 deg^2 SPT-SZ and Planck Gravitational Lensing Map. ArXiv e-prints arXiv:1712.07541.

Smith, K.M., Hanson, D., LoVerde, M., Hirata, C.M., Zahn, O., 2012. Delensing CMB polarization with external datasets. JCAP 6, 014. doi:10.1088/1475-7516/2012/06/014, arXiv:1010.0048.

Story, K.T., Hanson, D., Ade, P.A.R., et al., 2015. A Mea-

surement of the Cosmic Microwave Background Gravitational Lensing Potential from 100 Square Degrees of SPTpol Data. Astrophys. J. 810, 50. doi:10.1088/0004-637X/810/1/50, arXiv:1412.4760.

Sutskever, I., Vinyals, O., Le, Q., 2014. Sequence to sequence learning with neural networks. Neural Information Processing Systems .

The Simons Observatory Collaboration, Ade, P., Aguirre, J., et al., 2018. The Simons Observatory: Science goals and forecasts. ArXiv e-prints arXiv:1808.07445.

Wu, W.L.K., Errard, J., Dvorkin, C., et al., 2014. A Guide to Designing Future Ground-based Cosmic Microwave Background Experiments. Astrophys. J. 788, 138. doi:10.1088/0004-637X/788/2/138, arXiv:1402.4108.

Zhang, Z., Liu, Q., Wang, Y., 2017. Road extraction by deep Residual U-Net. CoRR abs/1711.10684. URL: http://arxiv.org/abs/1711.10684, arXiv:1711.10684.