

# Line of Sight Lensing Mass Reconstructions in the Millennium Simulation

Spencer Everett

*Department of Physics, DePaul University\**

(Dated: September 6, 2016)

As dark matter does not absorb or emit light, its distribution in the universe must be inferred through indirect effects such as the gravitational lensing of distant galaxies. While most sources are only weakly lensed, the systematic alignment of background galaxies around a foreground lens can constrain the mass of the lens which is largely dark matter. In this thesis, I have implemented a framework to reconstruct lensing mass along lines of sight in a 144 arcmin<sup>2</sup> field of the Millennium Simulation by predicting the weak lensing of 1440 generated source galaxies using a best-case dark matter halo model in which the halo mass is known. The lensed source ellipticities are characterized by the ellipticity-ellipticity and galaxy-mass correlation functions and compared to the same statistic using the ‘true’ ray-traced ellipticities. In the ellipticity-ellipticity correlation function, I find that the halo model systematically underpredicts the correlation on scales above 0.2 arcminutes by a mean NRMSE of 0.22 using an average of 4836 halos per lensing prediction, although this can be decreased by including more halo contributions. The model predicted galaxy-mass correlation function is in agreement with the ray-traced statistic on scales from 0.2 to 2 arcminutes but over-predicts the correlation on scales below 0.2 arcminutes by a mean NRMSE of 0.36 due to strong lensing effects. Both best-case statistics were found to be well approximated by using only the most ‘relevant’ halos, which on average only needed the 300 and 70 most relevant halos to predict the correlation functions with a NRMSE below 5% and 10% respectively. The feasibility of using the halo model to infer hyperparameters of the Millennium Simulation with either maximum likelihood estimation or approximate Bayesian computation is explored, with preliminary results favoring likelihood estimation. Optimization of the framework code has reduced the mean CPU time per lensing prediction by 59% to  $57 \pm 8$  ms, although this is approximately two orders of magnitude slower than required for a planned scale-up to a 100 deg<sup>2</sup> field. Physical and computational limitations of the framework are discussed as well as potential improvements for future work.

## 1 INTRODUCTION

In a universe teeming with galaxies and light, it came as a shock when 20<sup>th</sup> century astronomers discovered that most of the mass in the universe is in fact dark; the ‘normal’ matter made of atoms that we interact with in everyday life, called baryonic matter, accounts for less than 20% of the mass in the observable universe [1]. The remaining mass takes the form of an exotic dark matter that does not absorb or emit light rendering it invisible to our telescopes. While this claim sounds bizarre, there has been an abundance of indirect evidence in recent decades for the existence of dark matter including the flattening of galaxy rotation curves [2], velocity dispersions of galaxies [3], the power spectrum of the cosmic microwave background radiation [1], acoustic peaks in the correlation function of luminous red galaxies [4], and galaxy cluster collisions [5].

One of the most successful probes of dark matter has been gravitational lensing. The path of light from distant ‘background’ galaxies is bent when traveling through regions of space containing large amounts of ‘foreground’ mass. Light from different origins in a source galaxy is subject to different bending which results in a distortion of the galaxy image. As the foreground mass is known to be largely dark matter, gravitational lensing supplies

a direct constraint on the mass of dark matter in that region [6].

While the effects of gravitational lensing can be dramatic, it is extremely rare for background galaxies to be perfectly aligned with large foreground mass to produce strongly lensed images. However, as Tyson et al. postulated in [7] and confirmed in [8], background sources are also ‘weakly’ lensed by lesser foreground mass nearby the line of sight (LOS) by what is now called galaxy-galaxy lensing. As the shape of most galaxies is only distorted by a few percent, and the intrinsic shape is not known, weak lensing must be detected statistically using a large number of sources.

Predicting accurate dark matter mass maps from weak lensing requires understanding the correlation between galaxies and dark matter structure. Numerous N-body simulations have predicted dark matter structure to have evolved from overdense regions in the early universe into ‘halos’ of virialized matter that eventually collapsed from self-gravity [9], where galaxies could then form in the large gravitational potentials of the halos. However, there is still extensive research being done on the correct density profile of the halos (isothermal spheres [10], NFW [11], BMO [12], Einasto [13]), identifying the stellar mass - halo mass relation [14], abundance matching [15], halo substructure [16], and the lensing by LOS halos [17] and by dark matter structures not associated with galaxies [18]. If a model for the distribution of dark matter in a region of foreground mass can accurately predict the statistical signal of the weak lensing of background

---

\* spencerweverett@gmail.com

sources, then the model can be used on galaxies in existing sky survey data to extrapolate the amount of dark matter in the region and construct large-scale maps of the dark matter in the universe.

Guided by these results, in this thesis I build upon the work of Collett et al. in [17] to apply a simple dark matter halo model to reconstruct lensing mass along lines of sight in the Millennium Simulation in order to predict the weakly lensed ellipticities of generated background sources. I then compare my predicted lensed ellipticities to the ‘true’ lensing calculated by Hilbert et al. in [19] by ray-tracing light through the Millennium Simulation. By doing so, I attempt to answer the following questions:

- Under the best possible circumstances of having noiseless redshifts and halo masses, how accurately can I reconstruct LOS dark matter mass structures using a halo model?
- How many halos are needed to make these predictions?
- How feasible is an inference of the Millennium Simulation’s cosmological hyperparameters, such as the stellar mass - halo mass relation, using all available halo data?
- How long does such an inference take?

A brief introduction to the halo model, the effects of strong and weak gravitational lensing, representing galaxy ellipticities, and cosmological correlation functions is discussed in Section 2. The implementation of the **Pangloss** framework to make LOS lensing predictions in the Millennium Simulation is described in Section 3, and the results of the framework on 1440 galaxies on a field size of 144 arcmin<sup>2</sup>, along with a comparison of the predicted lensed ellipticities to ray-traced ellipticities, is given in Section 4. Section 5 explores the feasibility of using the presented **Pangloss** framework to make parameter inferences using all available halo data either by likelihood estimation or approximate Bayesian computation. Section 6 discusses limitations of the used framework as well as potential physical and computational improvements that can be made for upcoming work before concluding remarks in Section 7.

## 2 COSMOLOGICAL BACKGROUND

Before describing the developed model for LOS dark matter mass map reconstructions from weak lensing measurements, I briefly review a few topics in astronomy and cosmology that are relevant for this thesis such as the dark matter halo model, gravitational lensing, and correlation functions.

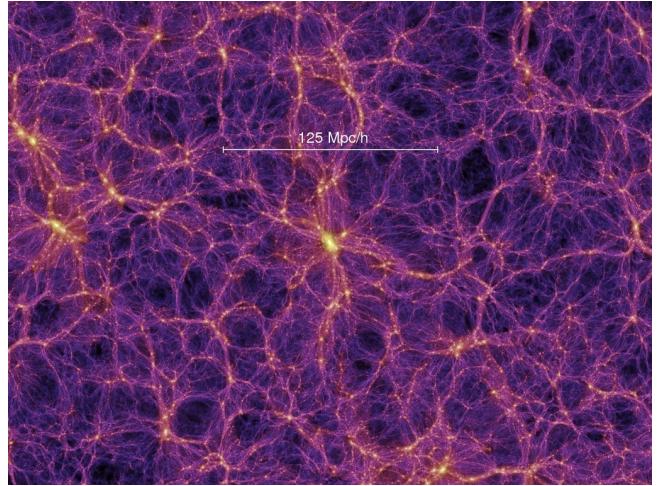


FIG. 1: A visualization of the large-scale dark matter distribution in the Millennium Simulation. The largest bright spots correspond to superclusters of galaxies which are connected by thread-like structures of matter called filaments. Taken from [20].

### 2.1 Dark Matter Halos

Cosmological simulations like the Millennium Simulation demonstrate that the underlying matter distribution of our universe is likely breathtakingly complex; rather than a largely disordered collection of galaxies and interstellar gas, matter appears to condense into clusters and superclusters of galaxies threaded together by large filaments pocketed with under-dense regions called voids, as shown in Figure 1 [20]. While the exact relation between the distribution of galaxies and dark matter is not known, simulations predict that early galaxies formed in the large gravitational potentials of over-dense regions of dark matter [9]. This suggests that, while we cannot map dark matter directly, galaxies should trace out at least some of the underlying dark matter structures [15]. A comparison of the predicted galaxy and dark matter distribution from a small region of the simulation is shown in Figure 2.

The simplest way to model the relationship between galaxies and dark matter is by enveloping each galaxy in a spherically symmetric dark matter ‘halo’ of a certain mass  $M_h$ . These halos extend far beyond the edge of the visible galaxy that they enclose, with the Milky Way’s own halo estimated to have a radius of between 300,000 and 400,000 light-years [21]. While the density profile of the halos may be complex, numerous simulations have shown that it can be well approximated locally by the Navarro-Frenk-White (NFW) profile which has the form

$$\rho_{NFW}(r) = \frac{\rho_0}{\frac{r}{R_s} \left(1 + \frac{r}{R_s}\right)^2} \quad (1)$$

where the constant  $\rho_0$  and the scale radius  $R_s$  are parameters that vary from halo to halo [11]. However, the

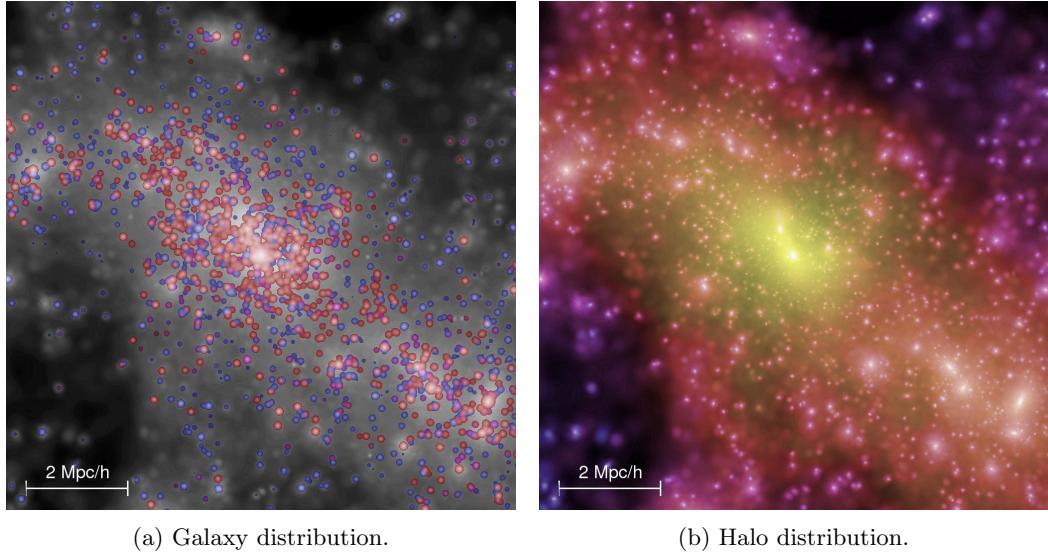


FIG. 2: A comparison of the galaxy and dark matter distribution for a single cluster in the Millennium Simulation. Galaxies tend to be located in the centers of clumps of dark matter called ‘halos’, which suggests that galaxies at least partially trace out the underlying dark matter distribution. Taken from [20].

NFW profile is not entirely physical as the total mass diverges as the density is integrated out to an infinite radius. Instead, this work uses a truncated NFW profile called the Baltz-Marshall-Oguri (BMO) profile given by

$$\rho_{BMO}(r) = \rho_{NFW} \cdot \left( \frac{r_t^2}{r^2 + r_t^2} \right)^2 \quad (2)$$

where  $r_t$  is the truncation radius as it has been shown to be a better fit to simulated data [12].

## 2.2 Filaments, Voids, and the Smooth Component Correction

The halo model only accounts for dark matter structures associated with galaxies. However, this neglects other mass structures such as the filaments and voids that can be seen in Figure 1. Modeling these features accurately would be difficult; we could try, for example, to model filaments by attaching cylinders of dark matter between massive clusters of galaxies. While an intriguing idea, rigorously testing for the appropriate filament density distributions, selection criteria, and frequency to statistically match simulations could constitute its own thesis.

Instead, as the mean mass density of large enough ( $\sim 100$  Mpc) regions should approximate the mean mass density of the universe, we can make a ‘first-order’ approximation by assuming that all additional mass (or absence of mass) is contained in a uniform ‘smooth-component’ density that ensures the total mass density of any sufficiently large region equals the mean density of the universe at that redshift. This assumption can be

expressed as

$$\rho_{\text{matter}}(z) = \rho_{\text{halos}}(z) + \rho_{\text{smooth}}(z) \quad (3)$$

where  $z$  is the redshift and  $\rho_{\text{matter}}(z)$ ,  $\rho_{\text{halos}}(z)$ , and  $\rho_{\text{smooth}}(z)$  are the mean matter densities of the universe, halos, and smooth-component correction respectively. For simplicity, we rename these quantities  $\rho_m(z)$ ,  $\rho_h(z)$ , and  $\rho_s(z)$  so that Equation (3) now reads as

$$\rho_m(z) = \rho_h(z) + \rho_s(z). \quad (4)$$

To solve for the desired smooth-component correction, we first need to calculate  $\rho_m(z)$  and  $\rho_h(z)$  and then subtract:

$$\rho_s(z) = \rho_m(z) - \rho_h(z). \quad (5)$$

Note that, unlike the halo density, each smooth-component mass sheet at a given  $z$  can have positive *or* negative mass density as the halos at that redshift could happen to be in a local overdensity or underdensity.

### 2.2.1 Calculating $\rho_m(z)$ :

Using the fluid equation and equation of state [22], we can show that the energy density of non-relativistic matter  $\varepsilon_m(z)$  evolves as

$$\varepsilon_m(z) = \frac{\varepsilon_{m,0}}{a(z)^3} = \varepsilon_{m,0}(1+z)^3, \quad (6)$$

where  $a(z) = (1+z)^{-1}$  is the cosmic scale factor ( $a(0) = 1$  by convention) and a subscript of 0 means the value at the present time, or equivalently a redshift of 0. This

is physically intuitive as the (mostly constant) mass is being diluted by an expanding volume. As the mass is non-relativistic, its energy density can be approximated by

$$\varepsilon_m(z) \approx \rho_m(z)c^2$$

and thus

$$\rho_m(z) \approx \rho_{m,0}(1+z)^3. \quad (7)$$

The current matter density  $\rho_{m,0}$  is usually reported in literature in terms of the present matter density parameter  $\Omega_{m,0}$  defined as

$$\Omega_{m,0} = \frac{\rho_{m,0}}{\rho_{c,0}} = \frac{8\pi G}{3H_0^2} \rho_{m,0}, \quad (8)$$

where  $\rho_{c,0} = \frac{3H_0^2}{8\pi G}$  is the present critical density. Solving for  $\rho_{m,0}$  gives

$$\rho_{m,0} = \frac{3H_0^2}{8\pi G} \Omega_{m,0}$$

and so we may write the non-relativistic mass density of the universe as a function of redshift as

$$\rho_m(z) \approx \left( \frac{3H_0^2}{8\pi G} \Omega_{m,0} \right) (1+z)^3. \quad (9)$$

### Calculating $\rho_h(z)$ :

The halo mass density  $\rho_h(z)$  at a given redshift is simply the sum of the individual halo masses  $M_{h,i}$  over a physical volume  $dV_p$ . However, we want a proper volume element that is expressed in terms of the solid angle  $d\Omega$  and the redshift  $dz$ . To do this, we can start with the standard spherical volume element  $dV = r^2 dr d\Omega$ . However, one must be careful to (1) use proper distances, (2) account for the possibility of a curved universe, and (3) use the correct distance measure for  $r$  (in this case, the angular diameter distance). From this we see that the proper volume element must have the form of

$$dV_p = d_A(z)^2 dr_p d\Omega \quad (10)$$

where  $d_A(z)$  is the angular diameter distance and the subscript  $p$  denotes proper distance. When the angular diameter distance  $d_A(z)$  is multiplied by an angle (i.e.  $d_A(z)d\theta$  and  $d_A(z)d\phi$ ), it gives an approximation of the proper distance between two objects at the same redshift.  $d_A(z)$  can be expressed in terms of the transverse comoving distance  $d_M(z)$ ,

$$d_A(z) = a(z)d_M(z) = \frac{d_M(z)}{1+z} \quad (11)$$

which is the comoving analog of  $d_A(z)$ . The transverse comoving distance depends on the assumed cosmology of

the universe and takes the form of

$$d_M(z) = \begin{cases} d_H/\sqrt{\Omega_k} \sinh [\sqrt{\Omega_k} d_C(z)/d_H] & : \Omega_k > 0 \\ d_C(z) & : \Omega_k = 0 \\ d_H/\sqrt{|\Omega_k|} \sin [\sqrt{|\Omega_k|} d_C(z)/d_H] & : \Omega_k < 0 \end{cases} \quad (12)$$

where  $d_H = c/H_0$  is the Hubble distance,  $\Omega_k$  is the curvature parameter ( $\Omega_k = 0$  for a flat universe), and  $d_C(z)$  is the LOS comoving distance given by

$$d_C(z) = d_H \int_0^z \frac{dz'}{E(z')} \quad (13)$$

where  $E(z)$  is the dimensionless Hubble parameter

$$E(z) = \sqrt{\Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}. \quad (14)$$

Substituting Equation (11) into (10) leads to

$$\begin{aligned} dV_p &= d_a(z)^2 dr_p d\Omega \\ &= \frac{d_m(z)^2}{(1+z)^2} dr_p d\Omega \end{aligned} \quad (15)$$

and so all that remains is to express  $dr_p$  in terms of  $dz$ .

As the proper distance between two objects  $r_p(z)$  is related to the comoving distance  $d_C(z)$  by

$$r_p(z) = a(z)d_C(z) = \frac{d_H}{1+z} \int_0^z \frac{dz'}{E(z')} \quad (16)$$

it follows that an infinitesimal proper displacement  $dr_p$  is given by

$$dr_p = \frac{d_H}{E(z)(1+z)} dz. \quad (17)$$

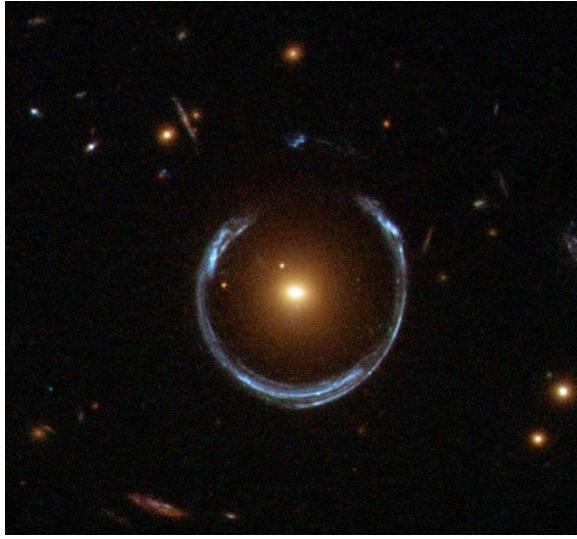
Combining this result with Equation (15) leads to the desired proper volume element

$$dV_p = \frac{d_H d_M(z)^2}{E(z)(1+z)^3} dz d\Omega. \quad (18)$$

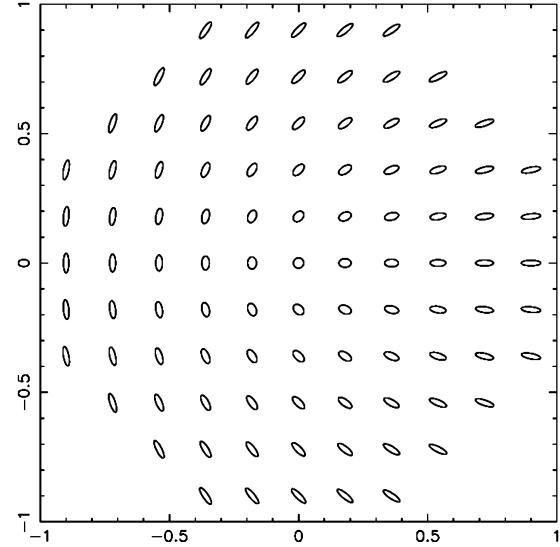
The proper volume is found by integrating the above volume element over the solid angle  $d\Omega$  and the desired  $dz$  range, so the halo density  $\rho_h(z)$  is then given by

$$\begin{aligned} \rho_h(z) &= \frac{\sum_i M_{h,i}}{V(z)} \\ &= \sum_i M_{h,i} \cdot \left( d_H \int \frac{d_M(z')^2}{E(z')(1+z')^3} dz' d\Omega \right)^{-1} \end{aligned} \quad (19)$$

and the smooth-component correction is simply the difference between Equations (9) and (19).



(a) An image of the luminous red galaxy LRG 3-757 along with a strongly lensed background galaxy, called the ‘Cosmic Horseshoe’ [23].



(b) The shape of a galaxy image for various ellipticity components  $\varepsilon_1$  and  $\varepsilon_2$  on the  $x$  and  $y$ -axes respectively. Taken from [24].

FIG. 3

### 2.3 Gravitational Lensing

While dark matter does not absorb or emit light, it does interact with light indirectly through gravitational lensing. The presence of matter curves spacetime locally which causes light to follow curved paths called geodesics. The deviation in path is usually negligible, but can be significant in regions containing clusters of galaxies and their massive dark matter halos. By observing how the magnitudes and shapes of distant galaxies are distorted when traveling through regions of matter, we can infer the masses of the dark matter halos and reconstruct the hidden dark matter mass distribution.

A full mathematical treatment of the gravitational lensing of galaxies due to the gravitational fields of massive objects requires general relativity (see [22] for details). However, the important results can be summarized as follows: Foreground mass distorts the image of a background galaxy in two distinct ways; the image is magnified and sheared tangentially about the foreground mass making it more elliptical. The magnification of the image is determined by the convergence  $\kappa$  defined as

$$\kappa(\vec{\theta}) = \frac{\sum(\vec{r}, z_l)}{\sum_{cr}(z_l, z_s)} \quad (20)$$

where  $\sum(\vec{r}, z_l)$  is the projected surface mass density of the lensing mass and  $\sum_{cr}(z_l, z_s)$  is the critical surface density for a given source redshift  $z_s$  and lens redshift  $z_l$  [24]. The shearing of the source is most often described by the complex shear  $\gamma$  defined to be

$$\gamma = \gamma_1 + i\gamma_2 = |\gamma| e^{2i\varphi} \quad (21)$$

where  $|\gamma|$  is the magnitude of the shear and  $\varphi$  is the orientation of the shear. While the intrinsic ellipticities of source galaxies are randomly oriented near the foreground mass before lensing, they will be systematically more aligned with the shear field after lensing.

The quantity usually of interest in lensing calculations is the reduced shear, defined as

$$g = \frac{\gamma}{1 - \kappa}. \quad (22)$$

Then using the thin lens approximation for the lensing of a background source of intrinsic ellipticity  $\varepsilon_i$  by foreground mass at a point with reduced shear  $g$ , the lensed ellipticity  $\varepsilon$  is given by

$$\varepsilon = \begin{cases} \frac{\varepsilon_i + g}{1 + g^* \varepsilon_i} & : |g| \leq 1 \\ \frac{1 + g \varepsilon_i^*}{\varepsilon^* + g^*} & : |g| > 1 \end{cases} \quad (23)$$

where an asterisk denotes the complex conjugate [24]. The behavior of the distortion relies strongly on the magnitude of  $g$ ; the effect is called *strong* lensing if  $|g| \gtrsim 1$  and *weak* lensing if  $|g| \ll 1$ . The effects of strong lensing can be quite dramatic, distorting sources into large arcs, multiple images, or even an Einstein ring as shown in Figure 3a.

While strong lenses are rare as the alignment of the source and foreground mass must be nearly perfect, *all* sources are weakly lensed when traveling through foreground halos. The effect is small, usually an ellipticity distortion of only a few percent, but can be detected locally by averaging the ellipticities of all sources in a small

region. As the orientations of the sources should be random, it would be expected that

$$\langle \varepsilon \rangle = 0.$$

However, as sources in the same small region are sheared in (approximately) the same way, this implies that

$$\langle \varepsilon \rangle \approx g. \quad (24)$$

Finally, as in the weak lensing regime  $\kappa \ll 1$  and  $|\gamma| \ll 1$ , it follows that

$$\gamma \approx g \approx \langle \varepsilon \rangle \quad (25)$$

which provides a method of detecting the shear observationally.

The analytical form of the convergence and shear contribution from a BMO-profile halo is quite messy, but is derived and explained in detail in [12]. Calculating the lensing contribution of the smooth-component correction is much simpler; the projected surface mass density at a given redshift is found by integrating  $\rho_s(z)$  over a sufficiently small  $dz$  interval, and then the mass sheet's convergence is found using Equation (20). There is no shear contribution as a uniform sheet of mass causes only magnification due to the mass sheet degeneracy [25]. However, Equation (23) shows that the lensed ellipticities depend on the *reduced* shear, which does depend on the additional convergence terms resulting from the smooth-component correction.

## 2.4 Galaxies as Ellipses

To calculate the lensed ellipticities in the previous section, galaxies must first be translated from pixel intensities into elliptical representations of the galaxies. This is not always possible as galaxies are diverse in type and shape and not all are well approximated by ellipses. However for this thesis, all galaxy ellipticities are simulated so this issue is avoided. For a more thorough analysis of measuring image ellipticities, see [24].

Consider a galaxy image that can be well approximated as an ellipse at an angle  $\phi$  above an arbitrary reference line. The galaxy's complex ellipticity is defined to be

$$\varepsilon = \varepsilon_1 + i\varepsilon_2 = |\varepsilon| e^{2i\phi} \quad (26)$$

where the magnitude of the galaxy's ellipticity  $|\varepsilon|$  is defined as

$$|\varepsilon| = \frac{1-r}{1+r} \quad (27)$$

and  $r \leq 1$  is the ratio of the semi-minor and semi-major axes of the ellipse. This compact notation combines the eccentricity and orientation of the ellipse into a single object. A plot from [24] showing the shape of elliptical galaxies for various values of  $\varepsilon_1$  and  $\varepsilon_2$  is shown in Figure 3b.

There are many complications to using this scheme in practice, most notably the multiplicative bias resulting from the smearing of galaxy images by the observational point spread function (PSF) [26]. While the effects of a PSF can be complex, in general it causes galaxy images to appear less elliptical than they truly are. To account for this in generated galaxy images, a multiplicative bias parameter  $M$  is often used to lessen the intrinsic ellipticity using

$$\varepsilon_{\text{obs}} = M \cdot \varepsilon_{\text{int}} \quad (28)$$

where  $\varepsilon_{\text{int}}$  is the generated intrinsic ellipticity of the image and  $\varepsilon_{\text{obs}}$  is the ellipticity that would be recorded by a detector.

## 2.5 Correlation Functions

If all galaxies were spherical then the measurement of their ellipticities would give an exact description of the lensing done by foreground mass structures. Unfortunately, the intrinsic ellipticities of background galaxies contribute significant noise as we do not know how any individual galaxy ellipticity has been altered by weak lensing. However as two sources lensed by the same foreground mass will have their ellipticities distorted in a similar way, their lensed ellipticities should be correlated regardless of intrinsic orientation. The average correlation of a sufficient number of galaxy pairs should overcome the noise and lead to a detectable signal. Importantly, this implies that *the correlation of galaxy ellipticities is a probe of the underlying dark matter mass distribution* [6].

The mathematical tool that measures the average correlation as a function of separation distance between two galaxies is called the ellipticity-ellipticity ( $\varepsilon$ - $\varepsilon$ ) correlation function. A quick overview of this function is given below, but readers that are unfamiliar with correlation functions in the context of cosmology or want a visual aid should see Appendix 9.1.

Given two galaxies with ellipticities  $\varepsilon_i$  and  $\varepsilon_j$  whose orientation is offset by polar angle  $\alpha$ , the tangential and cross-component components of the ellipticities,  $\varepsilon_t$  and  $\varepsilon_x$  respectively, for each are defined as

$$\varepsilon_{k_t} = -\text{Re}(\varepsilon_k e^{-2i\alpha}), \quad (29)$$

$$\varepsilon_{k_x} = -\text{Im}(\varepsilon_k e^{-2i\alpha}). \quad (30)$$

Instead of using the autocorrelation and cross-correlation of these quantities, it turns out to be more convenient to define the combinations

$$\begin{aligned} \xi_{\pm}(\Delta\theta) &= \left\langle \varepsilon_{i_t}(\vec{\theta}) \varepsilon_{j_t}(\vec{\theta} + \Delta\theta) \right\rangle \\ &\pm \left\langle \varepsilon_{i_x}(\vec{\theta}) \varepsilon_{j_x}(\vec{\theta} + \Delta\theta) \right\rangle, \end{aligned} \quad (31)$$

$$\xi_x(\Delta\theta) = \left\langle \varepsilon_{i_t}(\vec{\theta}) \varepsilon_{j_x}(\vec{\theta} + \Delta\theta) \right\rangle \quad (32)$$

where  $\vec{\theta}$  is the position vector of the first galaxy and  $\Delta\vec{\theta}$  is the angular separation vector between the pair of galaxies [24]. Gravitational lensing shears ellipticities tangentially, so  $\xi_x(\Delta\theta)$  should vanish with enough galaxies and therefore is a useful estimate of the bias. The amplitude of  $\xi_-$  is significantly lower than  $\xi_+$  as it depends on the initial orientation of *both* ellipticities and not just their offset, which leaves  $\xi_+(\Delta\theta)$  as the desired probe of lensing mass. As galaxies will only be lensed by the same foreground mass at relatively small separation distances, the  $\varepsilon$ - $\varepsilon$  correlation function decreases sharply as the separation increases.

A similar statistical measure often used in weak lensing measurements is the galaxy-mass correlation function which, in this setting, measures the cross-correlation of galaxy ellipticities and halo locations. Simply put, the function measures the correlation of lensed ellipticities around foreground halos as a function of separation distance and only has one component relevant for this work (see [27] for details).

### 3 PANGLOSS: LINE OF SIGHT MASS RECONSTRUCTIONS

To infer foreground dark matter mass structures using weak gravitational lensing, first a model of the relationship between foreground galaxies and dark matter must be established and robustly tested to see if, statistically, it makes the same lensing predictions of background sources as the true underlying dark matter structure. To do this, I built upon the publicly available **Pangloss**<sup>1</sup> framework used by Collett et al. in [17] to reconstruct all the mass along the lines of sight of generated background galaxies in the Millennium Simulation using a halo model with smooth-component correction. The lensing contribution of each halo is calculated, and the total lensing of the background galaxy is the sum of each halo contribution. The implementation of this process is detailed in the following sections.

#### 3.1 Model Assumptions

While **Pangloss** may be used more generally, the present analysis makes some additional strong assumptions to simplify the problem for a first attempt at making weak lensing predictions:

1. The dark matter mass distribution can be approximated by spherically symmetric BMO halo profiles attached to each galaxy, along with a smooth-component correction.

2. The stellar mass of the foreground galaxy is negligible for lensing calculations.
3. The mass of the dark matter halo of each foreground galaxy is known.
4. A spectroscopic redshift of each foreground galaxy is known.

Testing the first assumption is the main goal of this thesis. The second assumption is reasonable as it is estimated that dark matter halos are on average one to two orders of magnitude more massive than their host galaxies [14]. Clearly the third assumption will never be true for any observational data. However, this allows for a best-case estimate of how well the **Pangloss** framework *could* predict weak lensing effects given all possible information. This assumption can be relaxed by sampling a dark matter halo mass from an assumed stellar mass - halo mass relation. The fourth assumption is also unrealistic as most galaxies in sky surveys only have a less-reliable photometric redshift due to time constraints, but this again allows for a best-case estimate. This assumption could be relaxed by instead using photometric redshifts, adding random noise, and repeating the upcoming analysis on many realizations of the photometric redshifts.

#### 3.2 The Millennium Simulation

**Pangloss** cannot be used to make dark matter mass maps using existing galaxy catalogs until it is tested on a simulated universe with known dark matter structure to determine how accurately and precisely it predicts the lensing of background sources. For this purpose, **Pangloss** was tested on galaxy catalogs from the Millennium Simulation, an N-body simulation consisting of over 10 billion dark matter ‘particles’ each representing a billion solar masses and populated with about 20 million galaxies [29]. The simulation contains baryonic and dark matter structure believed to be similar to our own universe, and uses cosmological parameters from WMAP 1<sup>st</sup>-year data analysis which assumed a matter density of  $\Omega_m = 0.25$ , a dark energy density of  $\Omega_\Lambda = 0.75$ , and a Hubble constant of  $H_0 = 73 \text{ km s}^{-1} \text{ Mpc}^{-1}$  [30]. From the work of Hilbert et al. in [19], high resolution maps of the ray-traced shear and convergence of galaxy sources at  $z = 1.3857$  are publicly available. From these maps, an estimate of the ‘true’ lensing of background galaxy ellipticities when traveling through the foreground mass of the Millennium Simulation can be calculated using Equation (23).

#### 3.3 Generating Background Galaxies

With the (0,0,0,0) catalog of Millennium Simulation foreground galaxies chosen, a set of 1440 background

---

<sup>1</sup> <https://github.com/drphilmmarshall/Pangloss>

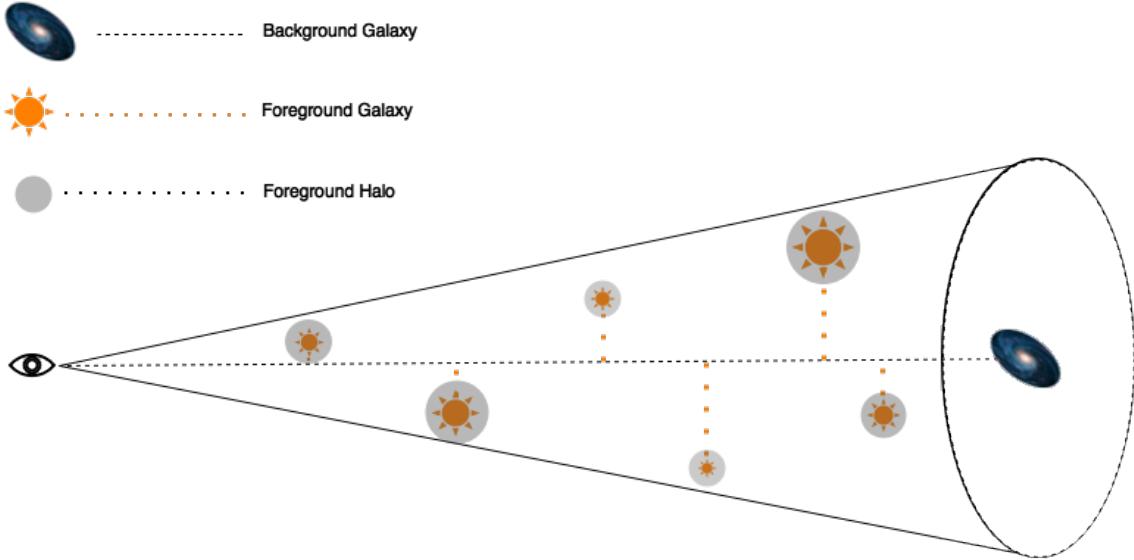


FIG. 4: A cartoon model of how the Pangloss framework uses a dark matter halo model to reconstruct the foreground mass along the LOS of a background source galaxy. A ‘lightcone’ centered at a background source is constructed with radius  $R_L$  and populated with all foreground galaxies contained in this volume. Each foreground object has an attached dark matter halo and physical distance to the LOS (dashed line). The shear and convergence along the LOS contributed by each halo is calculated using [28], and the predicted ellipticity is determined using the sum of these contributions and Equation (23).

galaxies over a field of 144 arcmin<sup>2</sup> was generated, with density 10 galaxies per arcmin<sup>2</sup>. The intrinsic orientation of each galaxy was sampled from a uniform distribution as, without lensing, there should be no preferred orientation due to the assumption of an isotropic universe. The magnitude of the galaxy ellipticities was sampled from a normal distribution with a standard deviation of 0.2, but any ellipticities with magnitude greater than one were re-sampled. Random ellipticity noise was sampled from a normal distribution with a standard deviation of 0.1 and added to the intrinsic ellipticities. Finally, each ellipticity was multiplied by  $M = 0.9$  to account for a multiplicative bias of 10%.

### 3.4 Creating Lightcones

While ideally all foreground mass in a field would be considered when predicting the weak lensing of a background galaxy, it is computationally prohibitive to do so. Instead, all foreground halos contained within a ‘lightcone’ centered along the LOS to the source and extending out to a chosen lightcone radius  $R_L$  were considered when calculating the lensing contributions for the background source. A cartoon model of this process is shown in Figure 4. Unless otherwise specified, experiments in this paper used a radius of 6 arcminutes and halo truncation scale of five times the virial radius.

To calculate the convergence and shear contribution of each halo, the physical distance from the halo to the LOS was needed. To increase the speed of distance calcu-

lations, the foreground halo redshifts were first binned to a grid of 100 equally-sized redshift bins  $\Delta z = 0.013857$  from  $z = 0$  to the sourced redshift of  $z = 1.3857$ , and then converted to physical distances using the cosmology defined by the Millennium Simulation.

### 3.5 Assigning Relevance to Halos

Not all foreground halos are equally relevant to the lensing calculation. To add a significant contribution to the combined shear and convergence, a halo must be massive, close to the LOS, or preferably both. McCully et al. in [18] derived that the correct metric for determining the relevance of a given foreground halo to the overall lensing calculation is given by

$$Rel(M_h, R) \propto \frac{M_h}{R^3} \quad (33)$$

where  $M_h$  is the halo mass and  $R$  is the distance from the halo to the LOS. Numerical values are assigned by comparing a halo’s individual  $M_h$  and  $R$  to a threshold  $M_T$  and  $R_T$ :

$$Rel(M_h, R) = \frac{(M_h/M_T)}{(R/R_T)^3}. \quad (34)$$

In this work, the threshold values used are  $M_T = 10^{12} M_\odot$  and  $R_T = 10$  kpc. The relevance distribution for one background catalog realization of 1440 sources with  $4836 \pm 104$  foreground halos per lightcone

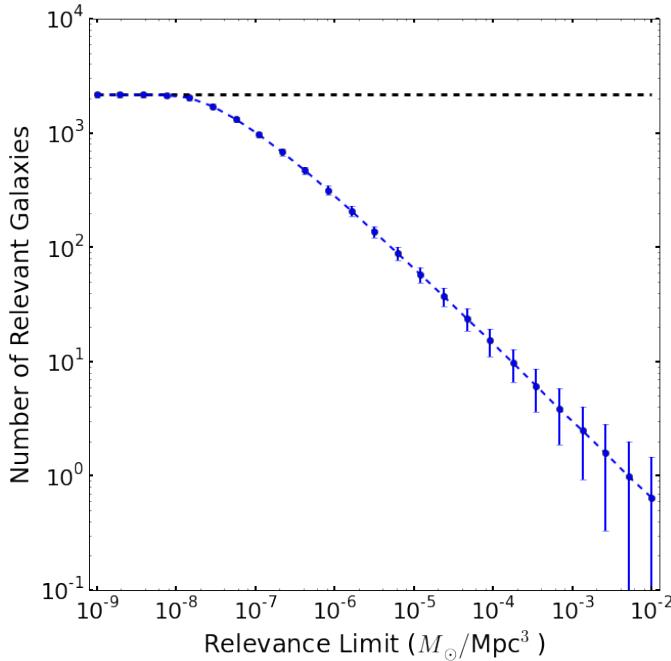


FIG. 5: The mean number of relevant halos in a single realization of 1440 source galaxies, each with  $4836 \pm 104$  halos per lightcone, determined by Equation (34) with a threshold mass and distance of  $M_T = 10^{12} M_\odot$  and  $R_T = 10$  kpc respectively. The error bars correspond to the standard deviation of relevant halos across all lightcones. As expected, most halos are either too far away from the LOS or not massive enough to make a significant contribution, or both. The number of halos falls logarithmically with the relevance limit and reaches a mean of one halo at a relevance of  $10^{-2} M_\odot/\text{Mpc}^3$ .

with radii of 6 arcminutes is shown in Figure 5. As expected, there are very few massive halos along the LOS and therefore only a small percentage of foreground halos should have a high relative contribution to the predicted lensing quantities. This encouraged the intended analysis to be done on lightcones with all halos and those with only a small subset of the most relevant halos to compare the results and see if the computational time saved by using only the most relevant halos was worth the loss in prediction accuracy.

### 3.6 Calculating the Smooth-Component Correction

As discussed in Section 2.2, the halo model does not account for additional mass structures such as filaments. To apply the smooth-component correction to `Pangloss`, the mean mass density  $\rho(z)$  and halo density  $\rho_h(z)$  had to be calculated at each of the 100 redshift bins. For  $\rho_h(z)$ , this was done by summing all halo masses  $M_{h,i}$  in a given

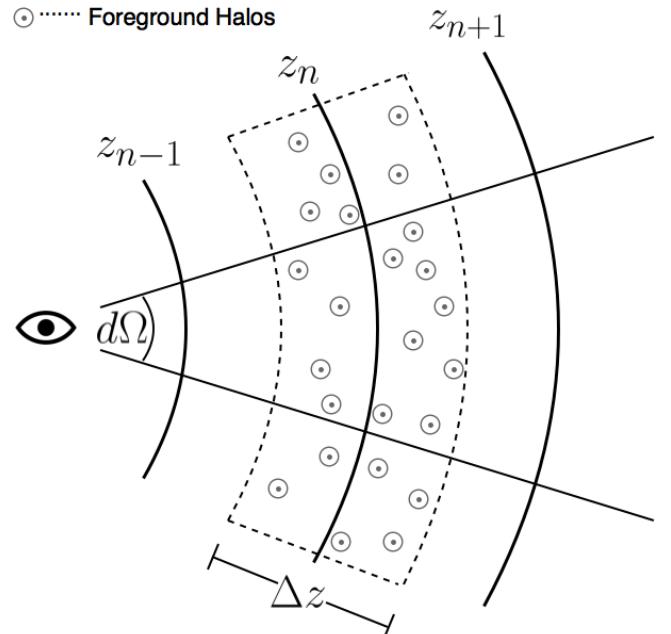


FIG. 6: A diagram of the smooth-component correction implementation in `Pangloss`. The halo density at each redshift slice  $z$  was calculated by summing the masses of all foreground halos within  $\Delta z$  and dividing by the proper volume of the cone slice using Equation (18) and solid angle  $d\Omega$ .

redshift bin and integrating Equation (18) from  $z - \Delta z/2$  to  $z + \Delta z/2$ , except for the boundary bins, and multiplying by the solid angle of a cone  $\Omega = 2\pi(1 - \cos R_L)$ . Once  $\rho_h(z)$  had been calculated for each redshift slice, the projected surface mass density was found by integrating over the redshift interval,

$$\begin{aligned} \sum(z) &= \int_{z - \frac{\Delta z}{2}}^{z + \frac{\Delta z}{2}} \rho_s(z) dz \\ &= \int_{z - \frac{\Delta z}{2}}^{z + \frac{\Delta z}{2}} \rho(z) - \rho_h(z) dz, \end{aligned} \quad (35)$$

and the convergence contribution of the mass sheet was calculated using Equation (20) with  $z_l = z$  and  $z_s = 1.3857$ . The corrected total convergence for a given background source is simply the sum of the halo contributions and the smooth-component contributions. There is no additional shear produced by a uniform sheet of mass [25], but the corrected reduced shear is now calculated using the corrected total  $\kappa$  in Equation (22). A diagram of this process is shown in Figure 6.

The volume and projected surface mass densities of each redshift slice for the same source realization as Figure 5 are shown in Figures 7a and 7b respectively. In each plot, the halo density at each redshift slice is plotted in green, the smooth-component density in blue, and universe mean density in black, as well as a histogram of the galaxies per bin in the background. In most redshift

bins, the halo density is one to two orders of magnitude lower than than the universe mean density leading to a smooth-component density very close to that of the mean density. However, all mass densities are far below the critical density, plotted in red, and so well within the weak-lensing regime. The convergence distribution of each halo and the 100 smooth-components is plotted in Figure 8.

### 3.7 Line of Sight Lensing Predictions

Using the halo mass and physical separation distance to the LOS, the shear and convergence contribution of a single foreground halo with BMO profile is calculated using methods described in Section 2.3 (see [12] and [28] for more details). The total convergence and shear along the LOS of the lightcone is simply the sum of the convergence and shear contributions of each halo contained in the lightcone, as well as the additional convergence terms from the smooth-component correction. The **Pangloss** predicted lensed source ellipticity is then calculated using Equation (23).

## 4 CHECKING THE PANGLOSS MODEL

Using the methodology presented in Section 3, 1440 background galaxies in a 144 arcmin<sup>2</sup> subset of the (0,0,0) Millennium Simulation foreground catalog were generated and lensed by both the ray-traced shear and convergence maps as well as the **Pangloss** framework. With a lightcone radius of 6 arcminutes, the lightcones contained an average of  $4836 \pm 104$  foreground halos. Both sets of lensed ellipticities were characterized with correlation functions, as well as the intrinsic ellipticities of the sources before lensing.

Figure 9 displays different views an example lightcone where the foreground halos are plotted as circles whose size are proportional to their galaxy brightness (top left), angular size (top center), **Pangloss** predicted convergence contributions (top right), and halo mass (bottom). Comparing the center and top right figures demonstrates the importance of proximity to the LOS for significant  $\kappa$  contributions. Figure 10a shows a foreground cluster in a 1.44 arcmin<sup>2</sup> section of the Millennium Simulation along with the ray-traced convergence (plotted in inverse grayscale) and shear maps (plotted as red shear sticks) from [19]. Figure 10b displays the generated background ellipticities in blue, the ‘true’ lensed ellipticities in green, and the halo model lensed ellipticities in purple. While there is a visible difference between the lensed ellipticities, both align tangentially around the foreground mass as expected.

Instead of comparing the ray-traced and **Pangloss** predicted ellipticities for individual galaxies, the lensing is characterized globally with correlation functions. As described in section 2.5, the  $\varepsilon\text{-}\varepsilon$  correlation function mea-

sures how correlated the ellipticities of pairs of galaxies are as a function of separation distance, while the galaxy-mass correlation function measures the correlation of lensed ellipticities around foreground halos as a function of separation distance. Both correlation functions are used in this work to estimate how well the **Pangloss** framework models weak lensing by dark matter structures and are computed using the publicly available **TreeCorr** module written by Mike Jarvis<sup>2</sup>. Note that the  $\varepsilon\text{-}\varepsilon$  correlation function definition used in **TreeCorr** is slightly different than that used in most of the literature; for a derivation of the connection between Jarvis’s definition [31] and the more common Schneider definition [24], see Appendix 9.2.

### 4.1 Ellipticity-Ellipticity Correlation Function

The first test of the halo model was with the  $\varepsilon\text{-}\varepsilon$  correlation function with a lightcone radius of 6 arcminutes and all  $4836 \pm 104$  halos per lightcone which is given in Figure 11a. The statistic measured the average correlation between pairs of ellipticities at separation distances between 0.1 and 2 arcminutes. The cross-component of the correlation function  $\xi_{\times}$  shows no significant deviation from zero as expected, as gravitational lensing only shears galaxies tangentially. For the  $\xi_{+}$  component, the ray-traced (plotted in green) and halo model (plotted in purple) are in relative agreement from separation scales of 0.1 to 0.2 arcminutes, but there is a systematic underprediction of correlation by the halo model on scales larger than 0.2 arcminutes by an average 31% error. This result is consistent with Pace et al. in [32] which found that a purely halo model underpredicted the lensing correlation on average by 40%. This may be indicative that not enough halo mass is being used in the lensing prediction, the halo model does not adequately address large-scale mass structures, or that there is significant dark matter mass not correlated with galaxies. These issues are discussed further in Section 6.1.

Using the same catalog of background sources, this statistic was recalculated for the halo model using various lightcone radii ranging from  $R_L = 1$  to  $R_L = 8$  arcminutes. The result is shown in Figure 11b, where the series of model predicted correlation functions is compared to the ray-traced correlation function. As  $R_L$  increases, the mean number of foreground galaxies contained in each lightcone grows quadratically from  $134 \pm 17$  galaxies when  $R_L = 1$  arcminute to  $8570 \pm 87$  galaxies when  $R_L = 8$  arcminutes. Increasing the number of foreground objects, and thus increasing the mass and structure considered for shear and convergence predictions, systematically increased the predicted correlation on all scales. This result encourages the use of lightcones with radii of

---

<sup>2</sup> <https://github.com/rmjarvis/TreeCorr>

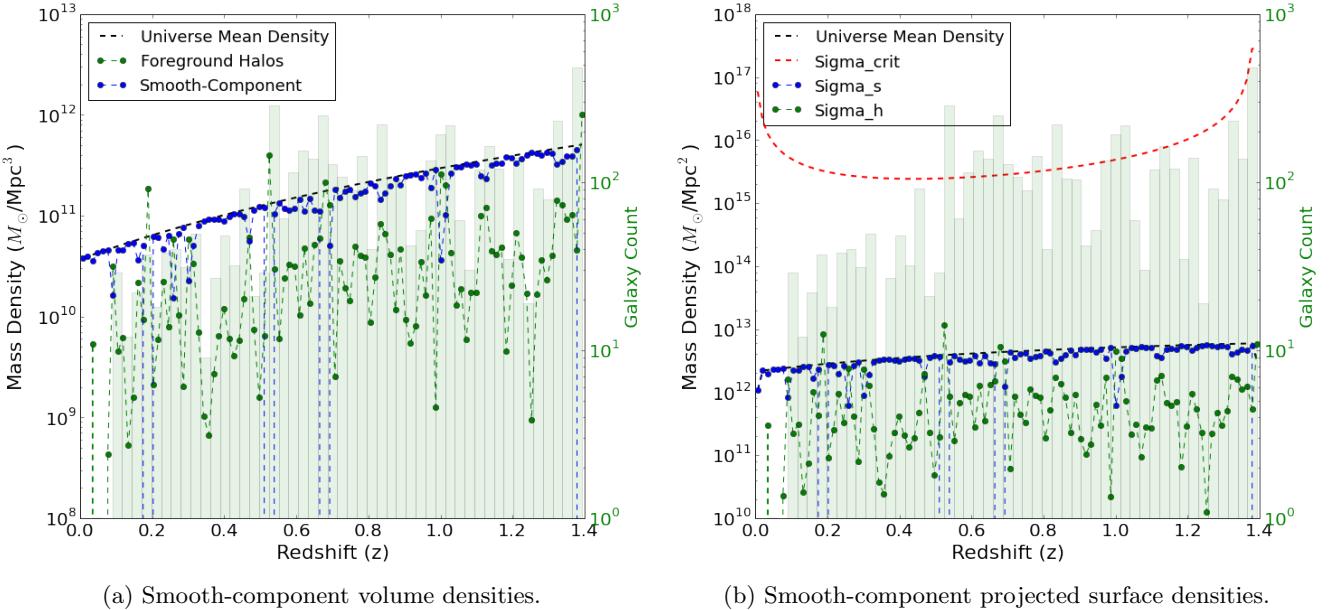


FIG. 7: The surface and volume densities for each of the 100 smooth-component correction and halo densities in a lightcone with radius  $R_L = 6$  arcminutes. The green line is the halo density of each slice and the blue line is the calculated smooth-component density such that the two densities sum to the universe mean density plotted in black. The number of halos in each redshift bin is plotted as a histogram in the background. The red line in plot (b) is the critical surface density.

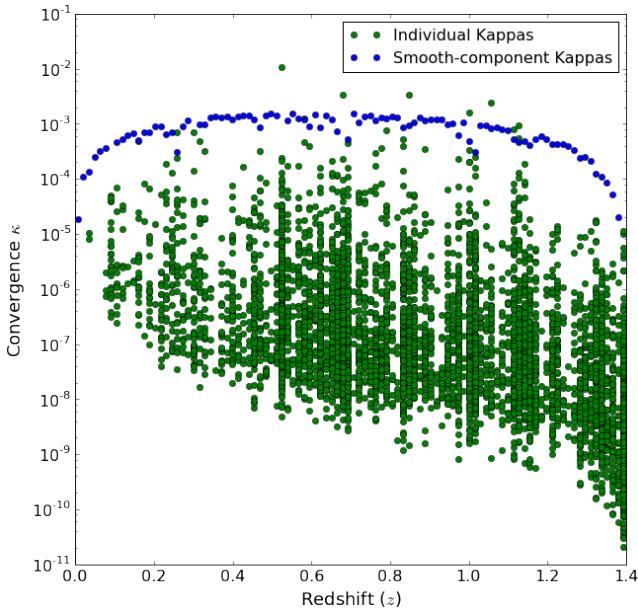


FIG. 8: The convergence distribution of the 4836 individual halos, plotted in green, and the 100 smooth-component slices, plotted in blue, contained in the same lightcone as Figure 7. The mass sheets have a higher convergence contribution than all but the most relevant halos in the lightcone. The curve of smooth component  $\kappa$ 's indicates how foreground mass at the center of the light's path has the highest lensing efficiency.

at least 8 arcminutes, and further emphasizes the need for computational efficiency to allow the timely use of large lightcones.

#### 4.2 Galaxy-Mass Correlation Function

The catalog of lensed ellipticities was also analyzed using the galaxy-mass correlation function with separation distances from 0.1 to 2 arcminutes using the same lightcones and halo number. The result is shown in Figure 12a. The lensed ellipticities from the ray-tracing and halo model both show positive correlation on all calculated scales and largely agree on scales larger than 0.3 arcminutes except for a slight overprediction by the halo model near 0.45 arcminutes. However, a new feature is the large overprediction of correlation on separation scales smaller than 0.2 arcminutes. This is likely because background sources this close to massive foreground halos are often in the strong lensing regime which is not currently accounted for by the Pangloss framework.

The galaxy-mass correlation function was similarly recalculated at increasing lightcone radii from  $R = 1$  to  $R = 8$  arcminutes with the result in Figure 12b. As with the  $\varepsilon$ - $\varepsilon$  correlation, structure larger than  $\sim 1$  arcminute is captured much better by increasingly large lightcone radii as expected. However, there is little to no improvement of the galaxy-mass correlation on scales smaller than 0.3 arcminutes which further points to the need to model strong lensing.

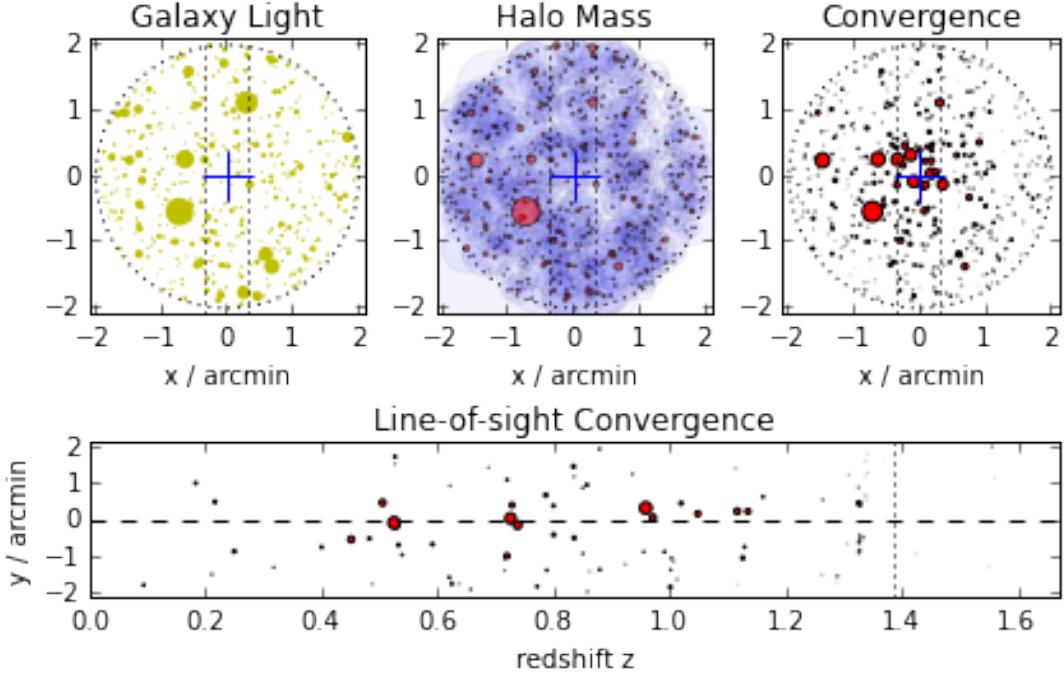


FIG. 9: Four different views of a lightcone with radius  $R_L = 6$  arcminutes centered on the LOS of a particular background source containing 4829 halos. The top left figure plots the positions of the halos with the size proportional to their galaxy brightness in the  $i$ -band, the center figure plots the angular size of foreground halos with red and blue regions representing the NFW scale radius and virial radius of each halo respectively, the top right figure plots the  $\kappa$  contributions of each halo, and the bottom plots the positions of each halo along the redshift axis with size proportional to  $\kappa$  contribution.

### 4.3 Correlation Fraction of Relevant Halos

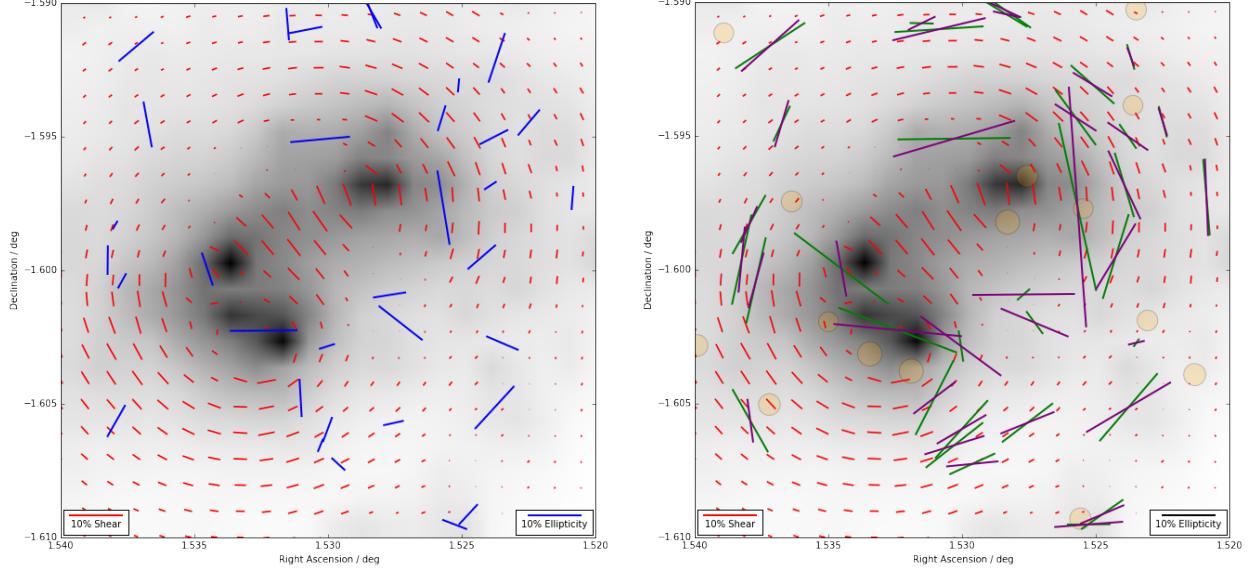
Of intrinsic interest is the relative number of relevant halos needed in each lightcone to accurately predict the halo correlation functions in Figures 11a and 12a, as a low percentage would allow much faster lensing predictions. Figures 13a and 13b display an example  $\varepsilon$ - $\varepsilon$  and galaxy-mass correlation function comparison of the halo model using all halos (plotted in purple) and using only halos above the relevance limit of  $10^{-5} M_\odot/\text{Mpc}^3$  (plotted in blue) with  $R_L = 6$  arcminutes, as well as the percent error of each predicted value between the two implementations. This relevance limit corresponded to  $66 \pm 10$  relevant halos per lightcone and resulted in a normalized root-mean-square error (NRMSE) of 0.164 and 0.082 for the two statistics. Figure 14 plots the  $\varepsilon$ - $\varepsilon$  and galaxy-mass correlation function NRMSE as a function of relevance limits and mean relevant halos, still using a lightcone radius of 6 arcminutes. These results suggest that either correlation function prediction can be approximated on average with an NRMSE below 10% using around the 70 most relevant halos, or an NRMSE of below 5% using around the 300 most relevant halos. Calibration curves such as this can allow the best-case correlation function to be predicted within a desired accuracy at only a fraction of the full lensing calculation

which will be especially helpful in future analyses with much larger fields of view.

### 4.4 Convergence Mass Maps

While correlation functions can describe how well the lensing predictions are being made globally, it is still instructive to analyze the difference between the ray-traced and halo model projected mass maps. This was done by selecting a region of foreground mass, calculating the ray-traced and Pangloss-predicted convergence of a uniform grid of background sources in the field, and binning the convergence maps to a desired resolution. Convergence is proportional to projected surface mass density, so comparing the two convergence maps is equivalent to comparing their respective projected mass distributions.

The result of a particular 9 arcmin<sup>2</sup> field is shown in Figure 15, where the ray-traced convergence is shown on the left, the halo model convergence is in the center, and the ray-traced convergence subtracted from the halo model convergence is shown on the right. While particular foreground objects are almost always predicted, the currently implemented halo model over-predicts the convergence throughout most of the field; the halo model does not appear to accurately account for the voids be-



(a) The ray-traced convergence and shear in a 1.44 arcmin<sup>2</sup> field of the Millennium Simulation, as calculated from [19]. The convergence, plotted in inverse grayscale, is proportional to the projected mass density and represents foreground mass. The shear field is plotted as sparsely-sampled red sticks that have size proportional to their magnitude. A few of the generated background sources are plotted as blue ellipticity sticks with length proportional to their eccentricities.

(b) The same 1.44 arcmin<sup>2</sup> field as 10a, but now with the ‘true’ lensed ellipticities in green, the halo model predicted lensed ellipticities in purple, and a few of the most massive foreground halos plotted as orange circles. Lensed ellipticity predictions tend to be more accurate further from the lensing mass, with significant differences near the centers of massive objects. All are sheared tangentially around the massive structures as expected.

FIG. 10

tween foreground masses and over-predicts the mass of most foreground objects. This could be a result of incorrect halo profiles, a too-simplistic implementation of the smooth-component correction, missing features in the halo model, a significant amount of dark matter mass not being correlated with galaxies, or some combination of all four possibilities.

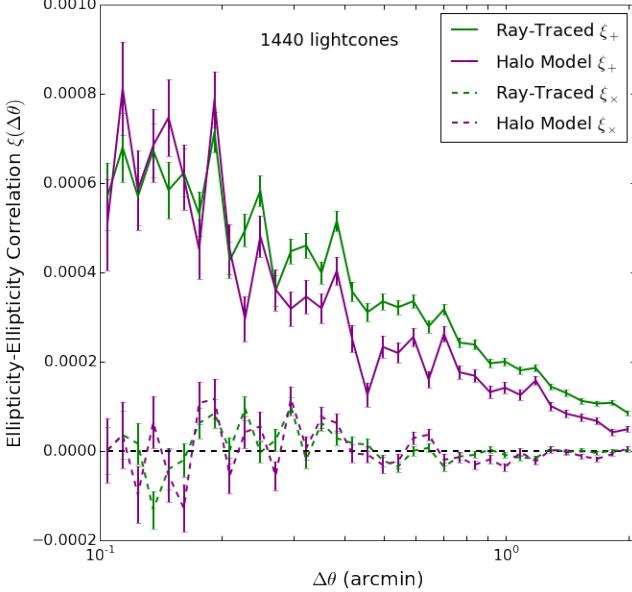
#### 4.5 Computational Performance

A secondary goal of this research was to investigate how quickly the shear and convergence prediction for a single background source could be made. The mean CPU time required for the lensing prediction of a single lightcone with radius  $R_L = 6$  arcminutes using all halos was  $140 \pm 20$  ms for the first run. However, many small improvements to the coding infrastructure of **Pangloss** were made to speed up the prediction. The mean CPU time for a single lightcone lensing prediction for 6 separate runs after various (and cumulative) changes is plotted in Figure 16, with a final run time of  $57 \pm 8$  ms. Run 1 corresponds to the original **Pangloss** framework code, run 2 wrote only the needed quantities for this analysis

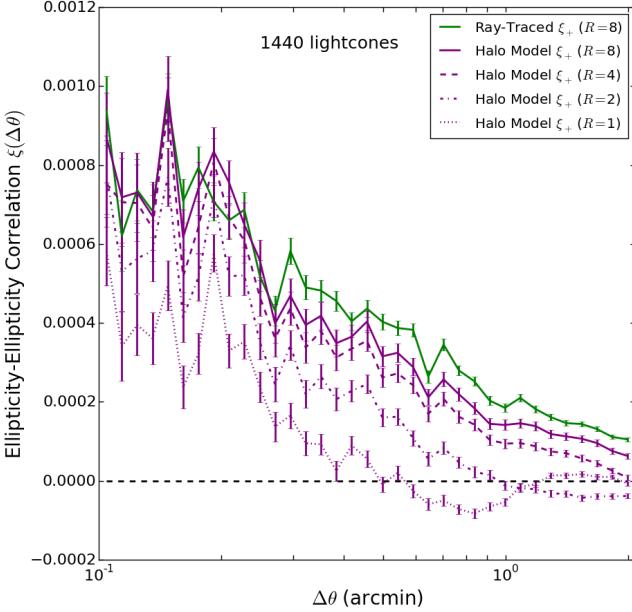
to the catalog **Astropy**<sup>3</sup> tables, run 3 changed the redshift grid binning to be done once for all halos in the drilling rather than for every individual halo in lensing, run 4 optimized various lensing calculations with matrices whenever possible, run 5 replaced the computationally expensive lensing calculation functions with a lookup table, and run 6 used only the  $65 \pm 9$  most relevant halos from a relevance limit of  $10^{-5} M_\odot/\text{Mpc}^3$ . Cumulatively, these changes accounted for a 59% decrease in mean CPU time per lightcone and halved the initial variation. While significant progress has been made, the code needs to be further optimized to allow lensing predictions of fields of many square degrees to be done in a few seconds. Section 6.3 discusses the details of future computational scale-up plans.

Figure 17 displays how the mean CPU runtime per lightcone scales as a function of lightcone radius  $R_L$  with all halos (left) and relevance limit at a constant radius of  $R_L = 4$  and  $R_L = 8$  arcminutes (right). Unsurprisingly, the runtime depends strongly on the number of galaxies per lightcone. Both figures show that the lensing calculation requires an overhead of about  $55 \pm 10$  ms,

<sup>3</sup> <http://www.astropy.org/>

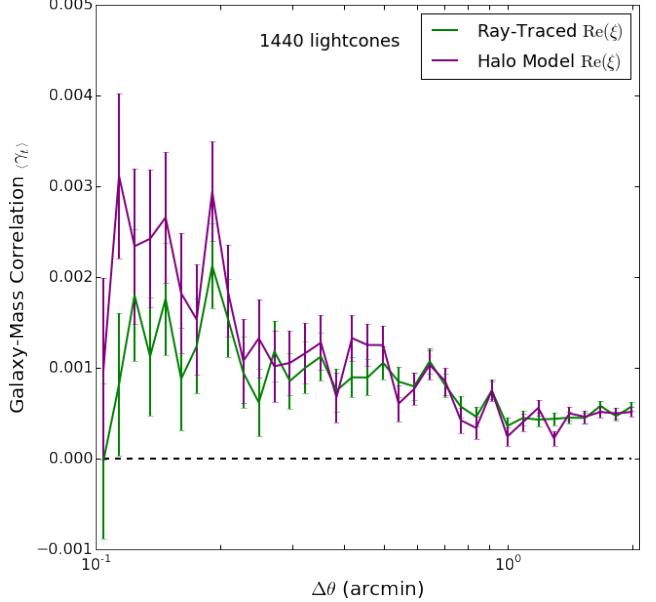


(a) The  $\varepsilon\text{-}\varepsilon$  correlation function for 1440 sources at separation distances from 0.1 to 2 arcminutes using lightcones with radii of  $R_L = 6$  arcminutes, where green is the ‘true’ ray-traced correlation and purple is the halo model predicted correlation. While the cross-component  $\xi_x(\Delta\theta)$  for both is consistent with zero across all scales as expected, the halo model systematically underpredicts the correlation on scales larger than 0.2 arcminutes.

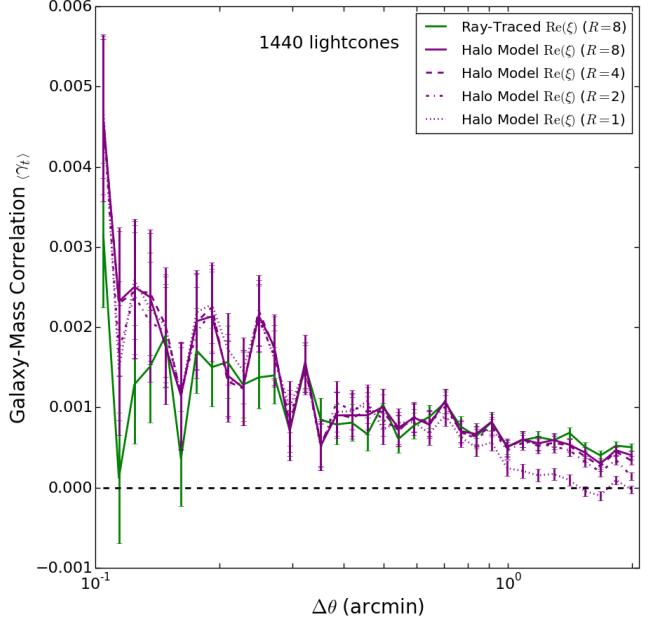


(b) The  $\varepsilon\text{-}\varepsilon$  correlation functions of 1440 sources for a series of halo model frameworks with increasing lightcone radii from  $R_L = 1$  to  $R_L = 8$  arcminutes, all plotted in purple with various line styles. These are compared to the same ray-traced correlation from Figure 11a. The predicted correlation seems to converge to the ray-traced value with larger  $R_L$  as more halo contributions are accounted for.

FIG. 11



(a) The galaxy-mass correlation function of 1440 sources at separation distances from 0.1 to 2 arcminutes and a lightcone radius of  $R_L = 6$  arcminutes. The halo model predicts the statistic well on scales above 0.3 arcminutes, but systematically overpredicts correlation on scales below 0.2 arcminutes. This is likely due to strong lensing effects which are not currently accounted for in Pangloss.



(b) The galaxy-mass correlation functions of 1440 sources for a series of halo model frameworks with increasing lightcone radii from  $R_L = 1$  to  $R_L = 8$  arcminutes, all plotted in purple with various line styles. These are compared to the same ray-traced correlation from Figure 12a. The predicted correlation seems to converge to the ray-traced value with larger  $R_L$  as more halo contributions are accounted for.

FIG. 12

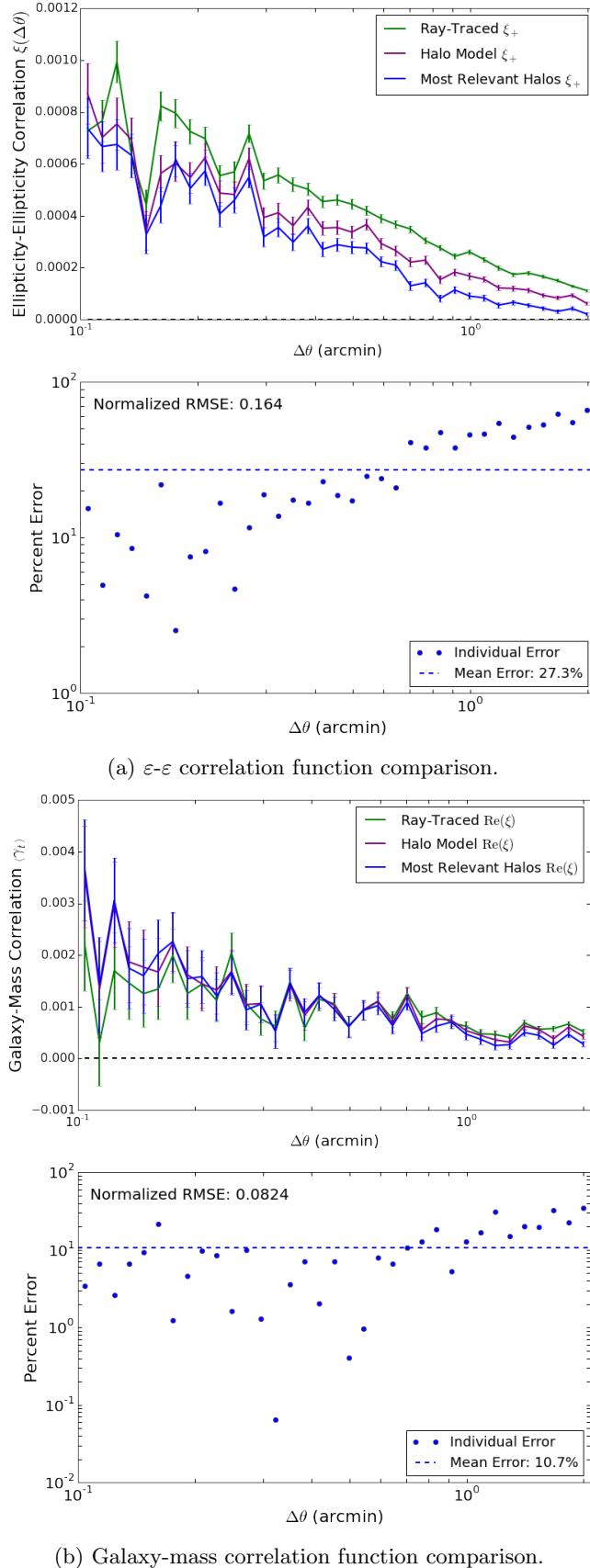


FIG. 13: Comparing the two correlation functions using all halos (purple) and only the  $66 \pm 10$  most relevant halos resulting from a relevance limit of  $10^{-5} M_\odot/\text{Mpc}^3$  (blue). The percent error of each point as well as the mean error and NRMSE is plotted below each figure.

but then the calculation speed depends quadratically on the number of halos and logarithmically on the relevance limit. The runtime saved by using only the most relevant halos can only counteract the increase in runtime due to larger lightcones for so long before the desired correlation function NRMSE becomes too large, so accurate analysis with radii above  $R_L = 8$  arcminutes may only be feasible with parallelization.

The distribution of total CPU time in significant parts of the lensing calculation for run 6 are as follows: 24.0% of total time calculating individual halo convergence contributions where 14.4% is spent in lensing table lookup functions, 65.1% of CPU time was spent combining all convergence contributions where 62.1% is spent on the smooth-component correction, and the majority of the remaining time spent on writing the output `Astropy` columns. However, time spent across all components of the lensing calculation was dominated by `Astropy` column use with nearly 55% of total time spent initializing, copying, or summing column elements.

## 5 FEASIBILITY OF PARAMETER INFERENCE USING PANGLOSS

One of the long-term goals of `Pangloss` is to be able to use all available sky survey data (sky positions, stellar masses, magnitude, photometric redshift, ellipticities, etc.) to make a hierarchical inference of the Universe’s cosmological and galaxy population model hyperparameters such as the density parameters  $\Omega_m$  and  $\Omega_\Lambda$ , the Hubble constant  $H_0$ , and stellar mass - halo mass relation [14]. This is a large undertaking and will require considerable testing and calibration on large N-body simulations such as the Millennium Simulation. As a first step towards this goal, I attempted to gauge the feasibility of implementing two methods of parameter inference, maximum likelihood estimation and approximate Bayesian computation, using the current framework. The results of this feasibility test, particularly the CPU time per likelihood estimation, will help guide future work on `Pangloss`. For a detailed description of a probabilistic modeling approach to gravitational lensing inference for large sky surveys, see [33]. Exploratory notes of implementing such an inference in the framework is available on `Pangloss`’s public GitHub repository<sup>4</sup>.

### 5.1 Likelihood Estimation

Perhaps the most common method of parameter estimation of statistical models is called maximum-likelihood

<sup>4</sup> <https://github.com/drphilmarshall/Pangloss/issues/23>

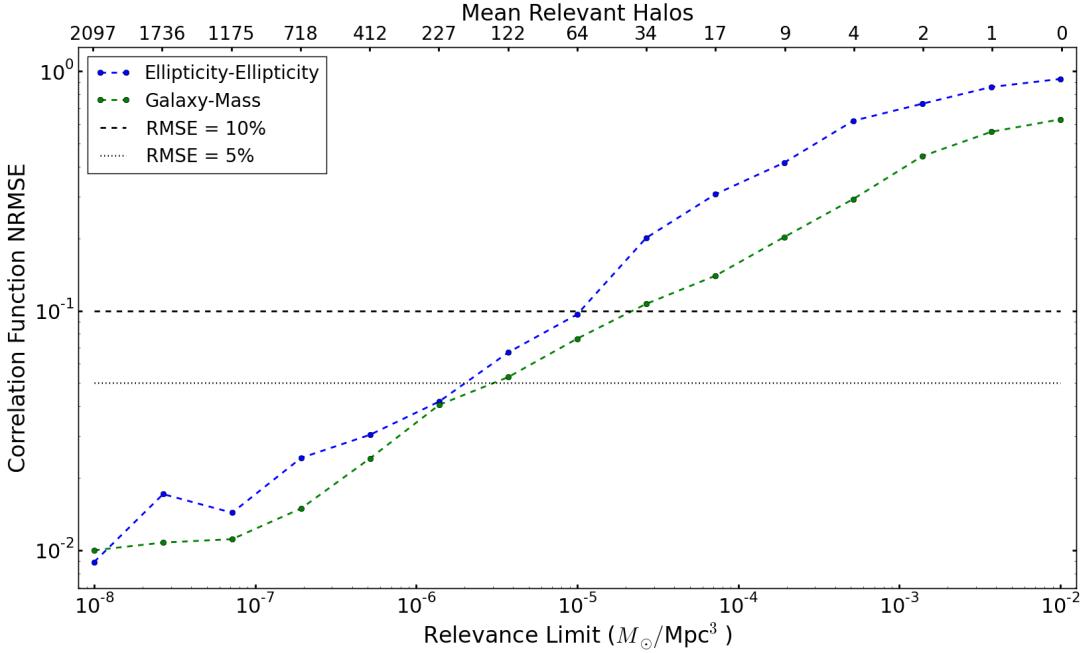


FIG. 14: The NRMSE of the  $\varepsilon$ - $\varepsilon$  and galaxy-mass correlation function, plotted in blue and green respectively, as a function of relevance limit / number of relevant halos used for 1440 sources with  $R_L = 6$  arcminutes. The lines corresponding to an NRMSE of 5% and 10% are plotted for convenience. This suggests that, on average, either best-case summary statistic can be approximated using only the 70 most relevant halos for a NRMSE below 10% and the 300 most relevant halos for a NRMSE below 5%. This figure can be used to determine what relevance limit should be used to recover the statistics to a desired accuracy.

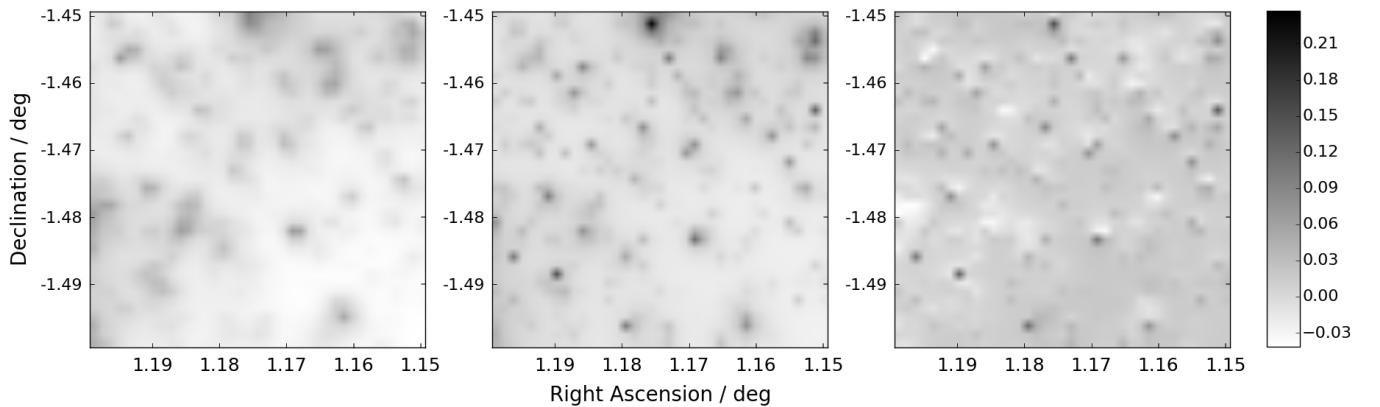


FIG. 15: The projected convergence map of a 9 arcmin<sup>2</sup> field of the Millennium Simulation using (left) the ray-traced convergence, (center) the Pangloss predicted halo model convergence, and (right) the ray-traced convergence subtracted from the halo model convergence. All maps are plotted with the same intensity scaling. The halo model predicts individual foreground objects, but systematically overpredicts the convergence throughout the map. This could be a result of incorrect halo profiles, a too-simplistic modeling of large-scale mass structures, and/or a significant amount of dark matter that is not correlated with galaxies.

estimation (MLE). The likelihood is a function of the parameters of a statistical model that measures the probability of a data set or outcomes given the model parameters [34]. It is the mathematical answer to the question, “Given a particular set of parameters, what is the probability that this data set could have occurred?” The form of the likelihood depends on the properties and circumstances of the data, but generally has the form of the product of the probability density function  $f$  of each in-

ability that this data set could have occurred?” The form of the likelihood depends on the properties and circumstances of the data, but generally has the form of the product of the probability density function  $f$  of each in-

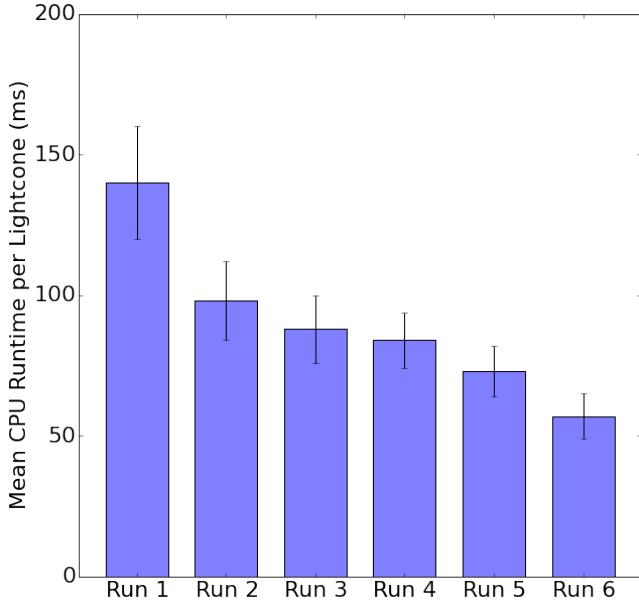


FIG. 16: The mean CPU runtime per lightcone with radius  $R_L = 6$  arcminutes for various runs with different (and cumulative) code optimizations described in Section 4.5. The cumulative changes account for a 59% decrease in the mean CPU time and a halving of the initial variation. On average, the lensing prediction for a single lightcone can now be made in  $57 \pm 8$  ms.

individual observation  $x_i$  given model parameters  $\vec{\theta}$ :

$$\mathcal{L}(\vec{\theta}, x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i | \vec{\theta}). \quad (36)$$

A much more detailed introduction to likelihood and MLE is given in [34]. For this work, the likelihood is given by [35]:

$$\mathcal{L}(\Sigma | \varepsilon_{i,j}, g_{i,j}) = \frac{1}{Z_L} \exp\left(\frac{-\chi^2}{2}\right), \quad (37)$$

where  $\varepsilon_{i,j}$  is the  $j$ -component of the  $i$ -th predicted lensed galaxy ellipticity,  $g_{i,j}$  is the  $j$ -th component of the ‘true’ reduced shear corresponding to  $\varepsilon_i$ ,  $\chi^2$  is the typical test statistic

$$\chi^2 = \sum_{i=1}^N \sum_{j=1}^2 \frac{(\varepsilon_{i,j} - g_{i,j})^2}{\sigma^2}, \quad (38)$$

the normalization factor  $Z_L$  is

$$Z_L = (2\pi\sigma^2)^{(2N)/2}, \quad (39)$$

where  $N$  is the number of background galaxies (the  $2N$  exponent is a result of the two ellipticity components for each of the  $N$  galaxies), and the uncertainty in galaxy ellipticity estimation  $\sigma$  is the sum in quadrature of the

dispersions of intrinsic and ‘observed’ (i.e. Pangloss-predicted) ellipticities:

$$\sigma = \sqrt{\sigma_{\text{int}}^2 + \sigma_{\text{obs}}^2}. \quad (40)$$

While this works in principle, computationally it is much easier to work with the natural logarithm of the likelihood function called the log-likelihood. As the logarithm is a monotonically increasing function, the likelihood and log-likelihood achieve their maximum at the same points and so can be used interchangeably in maximum likelihood estimation. From Equations (37)-(40) it follows that the log-likelihood is given by

$$\begin{aligned} \ln \mathcal{L} &= -\ln Z_L - \frac{1}{2} \chi^2 \\ &= -N \ln(2\pi\sigma) - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^2 \frac{(\varepsilon_{i,j} - g_{i,j})^2}{\sigma^2}. \end{aligned} \quad (41)$$

To use Pangloss to infer hyperparameters using MLE, a set of trial parameters would be sampled from appropriate prior distributions and used to calculate the lensed ellipticities of a mock background catalog using the methodology shown in Section 3. The log-likelihood of this set of parameters would be calculated using Equation (41), and the process would be repeated for many parameter combinations until a sufficient approximation of the likelihood is made. The most likely hyperparameters would then simply be the set of parameters that led to the maximum likelihood.

However, exploring the parameter space efficiently is a significant problem in MLE, especially in a dataset such as this which there are potentially millions or even billions of parameters (i.e. halo masses). For this thesis, I only compute the log-likelihood for a small handful of parameter sets to gauge how feasible such an inference would be in future work with far more computational power.

## 5.2 Approximate Bayesian Computation

MLE is not the only common method of parameter inference. As discussed in the previous section, the likelihood is computationally expensive to calculate and so MLE may not lend itself well to inference when large number of background galaxies are used. Approximate Bayesian computation (ABC) is a class of computational methods that estimate the posterior distribution without calculating the likelihood function. This can be vital for complex models in which the likelihood function does not have an analytic form or is computationally costly to evaluate. While fortunately Equation (41) is a relatively simple likelihood function, the calculation may become too costly for very large mock catalogs of millions or billions of source galaxies. However, ABC does come at a cost; as the name implies, there are various assumptions

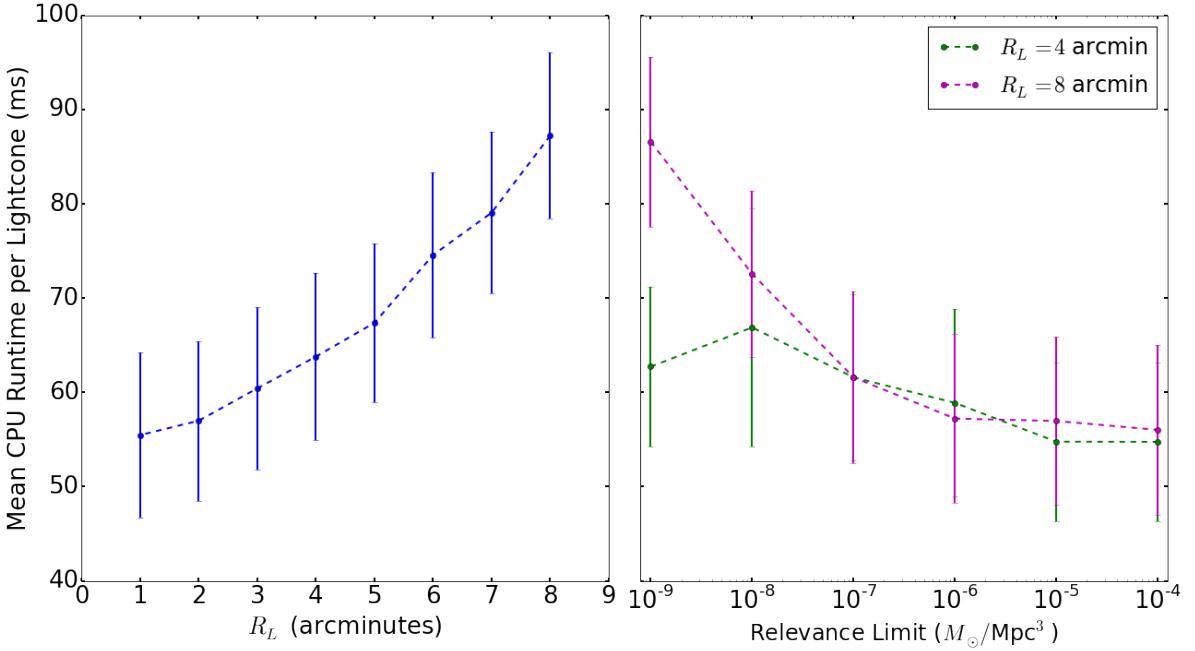


FIG. 17: The mean CPU runtime required for a single lensing prediction as a function of lightcone radius  $R_L$  (left) and relevance limit at constant radius (right). The mean runtime scales quadratically with lightcone radius, and therefore halo count, with an overhead of about  $55 \pm 10$  ms. Increasing the relevance limit to  $10^{-6} M_\odot/\text{Mpc}^3$ , or about 300 relevant halos, decreases the mean runtime to near the minimum overhead.

and approximations made in the process that can potentially lead to an inaccurate posterior prediction. See [36] for an excellent introduction to the topic. I summarize a common ABC implementation and its application to **Pangloss** below.

All ABC methods approximate the likelihood function by simulating data using parameters sampled from the prior distribution. If the simulated data are very ‘similar’ to the observed data, then the parameter set is accepted; if not, the parameter set is rejected. The posterior is then approximated by a histogram of the accepted parameters after a large number of simulations have been run. These methods are approximations as simulated data that is ‘similar enough’ to observed data is accepted rather than requiring exact matches. Practically, determining what metric of ‘similar’ to use and what threshold value  $\delta$  should be chosen for data rejection becomes a compromise of computation speed and posterior accuracy.

ABC using a simple distance metric like Euclidean distance or  $\chi^2$  is highly inefficient for data with large dimensionality as the probability of the ‘distance’ being small quickly vanishes. Instead, the similarity of observed and simulated data is often characterized by summary statistics such as mean and variance for simple data or correlation functions for more advanced data. Thus an ABC rejection algorithm reduces to predicting the summary statistic for the simulated data, comparing to the corresponding summary statistic of the observed data, and basing rejection on a chosen threshold of difference in

summary statistics.

A possible implementation of the above for **Pangloss** would be as follows: Choose the ray-traced lensed ellipticities as the ‘observed’ data and compute its  $\varepsilon$ - $\varepsilon$  correlation function to be used as the desired summary statistic. Let the distance metric to be the NRMSE between the predicted and ray-traced  $\varepsilon$ - $\varepsilon$  correlation function. Compute the best-case  $\varepsilon$ - $\varepsilon$  correlation function using all halo data as shown in Section 3 and make the threshold value  $\delta$  be some multiple of the best-case NRMSE (e.g.  $\delta = 2 \cdot \text{NRMSE}_{\text{best-case}}$ ). Create mock data by repeating the methodology in 3 except use the stellar mass - halo mass relation, dependent on parameters we sample from a prior, to predict halo masses based upon each galaxy’s stellar mass. Compute the simulated  $\varepsilon$ - $\varepsilon$  correlation function and determine whether to accept or reject the data by determining if the NRMSE is below the threshold  $\delta$ . The posterior will be estimated after a large number of iterations through the rejection algorithm.

### 5.3 Results

I implemented the methodology in section 3 to make weakly lensed ellipticity predictions for 20 mock catalogs of  $N$  background sources with lightcone radii of  $R_L = 4$  arcminutes for  $N$  ranging from  $N = 10^2$  to  $N = 5 \cdot 10^6$ . The mean CPU runtimes for a single likelihood calculation and ABC rejection decision as a function of  $N$

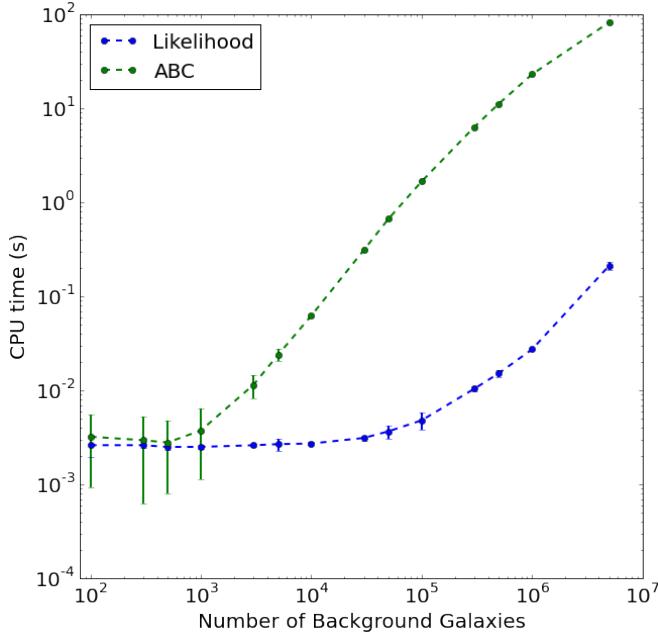


FIG. 18: The mean CPU runtime of a single likelihood evaluation and ABC decision, based upon the  $\varepsilon\text{-}\varepsilon$  correlation function calculation, as a function of the number of background sources  $N$ . The mean runtime was determined using 20 mock catalogs for each  $N$ . Both sampling methods require an overhead of about 3 ms and increase runtime with  $N$ , but a single ABC decision takes three orders of magnitude more runtime for  $N \geq 10^5$  galaxies.

is shown in Figure 18. The two methods take approximately the same amount of time up to  $N = 10^3$  sources, but the correlation function quickly becomes far too computationally inefficient at  $N$  values above  $10^4$ . While further analysis is needed to determine whether more or less ABC samples are required than posterior samples from MLE, this result, as well as the limitation that ABC is fundamentally an approximation of the posterior, suggests that any future inference work will likely use MLE instead of ABC. However, the time spent for either inference method is quite small in comparison to the CPU time required to create lightcones and make the LOS lensing predictions. This is an encouraging result which deemphasizes the choice of sampling method and suggests that further work on the framework should be on additional optimization of the lensing prediction code rather than posterior estimation.

## 6 DISCUSSION

### 6.1 Issues with the Halo Model

Both the  $\varepsilon\text{-}\varepsilon$  and galaxy-mass correlation functions show clear issues with using the halo model to predict the lensing of background sources with `Pangloss`, even in the best-case scenario of perfect knowledge of the halo masses and their spectroscopic redshifts. The systematic underprediction of correlation on scales larger than 0.2 arcminutes in the  $\varepsilon\text{-}\varepsilon$  function indicates that larger structures like filaments and voids are not being fully captured by the halo model. This is not entirely surprising as the only structure beyond halos accounted for in the presented model was the smooth-component correction which only adds sheets of uniform mass; clearly a better model of filaments and voids is needed to make accurate correlation predictions on scales larger than  $\sim 1$  arcminute. The progression of model predicted ellipticity-ellipticity correlation functions at larger radii converging to the ray-traced correlation in Figure 11b supports this analysis as the prediction was progressively better with larger number of foreground halos and thus more mass and structure.

A possible solution, briefly discussed in Section 2.2, would be to model filaments as ‘cylinders’ of dark matter attached to the most massive halos with densities, length distributions, and frequencies matching simulated data. This implementation might lead to better correlation predictions on large scales, but the details of the implementation, as well as analyzing how filaments affect LOS lensing measurements, still needs to be explored. Stellar masses could also be incorporated into the `Pangloss` prediction but any effect would likely be dominated by the missing features described above. Resolving the over-prediction of correlation on scales below 0.2 arcminutes in the galaxy-mass function is likely more straightforward, as sources this close to the center of foreground masses are often in the strong lensing regime. `Pangloss` currently removes any sources with total reduced shear  $g$  greater than 0.5, but this potentially leaves sources that are not entirely described by weak lensing. Correctly modeling LOS strong lensing will require analyzing the convergence and shear contributions of BMO halos in the strong lensing regime and at the interface between strong and weak lensing.

### 6.2 Most Halos are not Relevant

From Section 4.3 and Figure 14, we clearly do not need to use all halos in the field to predict the best-case scenario correlation functions within a NRMSE below 10% or even 5%. This will be of essential importance when making lensing predictions in fields of many square degrees that otherwise would be computationally infeasible. Figure 17 suggests that setting the relevance limit

to  $10^{-6} M_\odot/\text{Mpc}^3$ , or 300 relevant halos per lightcone, will decrease the mean CPU time per lightcone prediction to roughly the minimum overhead runtime of about 60 ms. According to Figure 14, this relevance limit will on average keep the  $\varepsilon$ - $\varepsilon$  and galaxy-mass correlation function's NRMSE at or below 5%. These two figures can be used in this way to decide what relevance limit is appropriate for the desired tradeoff between mean runtime speed and correlation function accuracy.

### 6.3 Scaling Up Pangloss

While the current **Pangloss** framework can handle the lensing predictions of 1440 sources in a  $144 \text{ arcmin}^2$  field comfortably on a single processor, the goal is to scale up the framework to make predictions for background sources of the same number density in a  $100 \text{ deg}^2$  field; this will require 3.6 million lightcones, each containing at least 300 relevant foreground halos for an estimated  $\varepsilon$ - $\varepsilon$  NRMSE of less than 5%. Luckily, the prediction is trivially parallelizable as the shear and convergence calculation of each lightcone is independent of all other lightcones. This makes GPUs an attractive candidate for future work, as it would only take 360 GPUs with 10,000 threads each to carry out the prediction. Additionally, the  $\sim 5 \text{ MB}$  of RAM per lightcone is sufficiently small to fit 10,000 lightcones on a GPU.

However, the **Pangloss** code needs to be optimized before the parallelization will be efficient or even feasible. The breakdown of CPU time spent on various components of the lensing calculation in Section 4.5 indicates that (1) the smooth-component correction needs to be implemented *far* more quickly as it is responsible for on average 62.1% of the total runtime and (2) the current use of **Astropy** tables for data storage is not practical for fields much larger than  $1,000 \text{ arcmin}^2$  or with lightcone radii larger than 8 arcminutes. Possible solutions to (1) are to experiment with faster numerical integrators or use a lookup table for appropriate parts of the calculation, while the second issue suggests that **Pangloss** should store catalog data in databases rather than **Astropy** tables as databases are designed for rapid retrieval, modification, and deletion of data [37]. While this will require significant reworking of the framework to implement, the CPU distribution indicates that such a change is necessary for even a moderate scale up.

### 6.4 Future Inference Work

The results of Section 5 demonstrated that neither the likelihood nor correlation function computation present a significant computational issue in comparison to the lensing calculation, so either are feasible approaches for a hierarchical inference of the stellar mass - halo mass relation. As a single ABC rejection algorithm performed many orders of magnitude *worse* than a single likeli-

hood calculation for  $10^2 \leq N \leq 5 \cdot 10^6$  sources, and ABC fundamentally an approximation, it appears that the method is not suitable for this work unless ABC ends up requiring far less samples than the number of needed posterior samples in MLE. However as discussed above and in Section 5.3, there is still much more work to the **Pangloss** framework that needs to be done, both in modeling features and computational infrastructure, before an inference using a large number of background sources comparable to modern sky surveys can be achieved.

## 7 CONCLUSION

In this thesis, a simple halo model was used to reconstruct lensing mass along lines of sight in the Millennium Simulation to make predictions of how foreground mass weakly lensed the ellipticities of 1440 generated background sources across a field of  $144 \text{ arcmin}^2$ . The lensed ellipticities were characterized globally using the  $\varepsilon$ - $\varepsilon$  and galaxy-mass correlation functions and then compared with the same statistics from the ray-traced data calculated in [19]. I then explored what fraction of foreground halos were truly relevant to the lensing prediction, the computational performance of the used framework, and how feasible an inference of hyperparameters like the stellar mass - halo mass relation will be in future work using either MLE or ABC. My conclusions to the questions posed in the introduction are as follows:

- There was a systematic-underprediction of correlation on scales larger than 0.2 arcminutes in the  $\varepsilon$ - $\varepsilon$  correlation function suggesting missing features of large-scale structure in the halo model. The galaxy-mass correlation function predicted by the halo model was in relative agreement with the ray-traced correlation function on scales from 0.2 to 2 arcminutes, but significantly overpredicted the correlation on scales below 0.2 arcminutes due to unaccounted for strong lensing effects. Figures 11b and 12b show that the predicted summary statistics converge towards the ‘true’ statistics at large lightcone radii, although at quickly prohibitive computational cost. The model was also used to predict convergence mass maps in small  $9 \text{ arcmin}^2$  regions that visually demonstrated the prediction of excess mass both in regions of foreground mass and the voids between them. This result suggested that a more complex model of filaments and voids is needed to accurately predict lensing mass and the correlation from large-scale mass structures.

- Figure 14 describes the fraction of relevant halos needed to reproduce the best-case correlation functions to a desired accuracy using the relevance metric in Equation (33) derived by McCully et al. in [18]. It suggests that, on average, the best-case correlation functions can be approximated to

a NRMSE of less than 5% or 10% using a relevance limit of  $10^{-6} M_{\odot}/\text{Mpc}^3$ , or  $\sim 300$  halos, and  $10^{-5} M_{\odot}/\text{Mpc}^3$ , or  $\sim 70$  halos, respectively. This is an encouraging result as Figure 17 shows that the lensing prediction scales quadratically with  $R_L$  and thus halo number. Instead of using the many thousands of halos contained in each lightcone for radii above 4 arcminutes, only a few hundred are needed to reproduce the best-case summary statistics within a NRMSE of below 5%. This will be essential in future scale-ups where the lightcone radius and field of view will both be much larger.

- The mean CPU time per likelihood evaluation and ABC decision for catalogs of background sources ranging from  $10^2 \leq N \leq 5 \cdot 10^6$  is shown in Figure 18. While the likelihood calculation outperformed the summary statistic calculation for all values of  $N$ , the main result is that neither computation took a significant amount of time relative to the lensing calculation for a given  $N$ . This indicates that inference feasibility rests upon optimizing the lensing prediction code to sufficient speed as discussed in Sections 4.5 and 6.3 rather than the chosen sampling method.

Now that a proof of concept has been demonstrated, work can be done to scale up the **Pangloss** framework to make lensing predictions of sources across  $100 \text{ deg}^2$  after implementing the suggested computational and model improvements. While the results are encouraging for a toy model, many strong and unrealistic assumptions were

made in the methodology that limit its use in observational surveys. In the future, a similar analysis should be made using simulated photometric redshifts and inferred halo masses to see how different the predictions by the halo model are from this best-case scenario.

## 8 ACKNOWLEDGEMENTS

I would like to thank the Department of Energy for funding this research and giving me the wonderful opportunity to meet so many talented peers from across the country. This work was greatly aided by numerous conversations with Risa Wechsler from Stanford University and Matthew Becker from the Kavli Institute for Particle Astrophysics and Cosmology (KIPAC), as well as Jesus Pando and Anuj Sarma from DePaul University. Additionally, I am sincerely appreciative of the many scientists, faculty, and graduate students at SLAC National Laboratory and Stanford University who continually welcomed a unknown undergraduate student who wanted to learn a little bit more about the universe. Finally, I am incredibly grateful to Phil Marshall of SLAC National Laboratory for his ceaseless support, guidance, enthusiasm, and friendship throughout this research and thesis. I look forward to continued collaboration in the future!

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTD) under the Science Undergraduate Laboratory Internships Program (SULI).

## REFERENCES

- 
- [1] Planck Collaboration, R. Adam, P. A. R. Ade, N. Aghanim, Y. Akrami, M. I. R. Alves, M. Arnaud, F. Arroja, J. Aumont, C. Baccigalupi, and et al. Planck 2015 results. I. Overview of products and scientific results. *ArXiv e-prints*, February 2015.
  - [2] V. C. Rubin, W. K. J. Ford, and N. . Thonnard. Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 /R = 4kpc/ to UGC 2885 /R = 122 kpc/. *Astrophys. J.*, 238:471–487, June 1980.
  - [3] F. Zwicky. On the Masses of Nebulae and of Clusters of Nebulae. *Astrophys. J.*, 86:217, October 1937.
  - [4] D. J. Eisenstein, I. Zehavi, D. W. Hogg, R. Scoccimarro, M. R. Blanton, R. C. Nichol, R. Scranton, H.-J. Seo, M. Tegmark, Z. Zheng, S. F. Anderson, J. Annis, N. Bahcall, J. Brinkmann, S. Burles, F. J. Castander, A. Connolly, I. Csabai, M. Doi, M. Fukugita, J. A. Frieman, K. Glazebrook, J. E. Gunn, J. S. Hendry, G. Hennessy, Z. Ivezić, S. Kent, G. R. Knapp, H. Lin, Y.-S. Loh, R. H. Lupton, B. Margon, T. A. McKay, A. Meiksin, J. A. Munn, A. Pope, M. W. Richmond, D. Schlegel, D. P. Schneider, K. Shimasaku, C. Stoughton, M. A. Strauss, M. SubbaRao, A. S. Szalay, I. Szapudi, D. L. Tucker, B. Yanny, and D. G. York. Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. *Astrophys. J.*, 633:560–574, November 2005.
  - [5] D. Clowe, A. Gonzalez, and M. Markevitch. Weak-Lensing Mass Reconstruction of the Interacting Cluster 1E 0657-558: Direct Evidence for the Existence of Dark Matter. *Astrophys. J.*, 604:596–603, April 2004.
  - [6] Y. Wang, X. Yang, H. J. Mo, C. Li, F. C. van den Bosch, Z. Fan, and X. Chen. Probing the intrinsic shape and alignment of dark matter haloes using SDSS galaxy groups. *MNRAS*, 385:1511–1522, April 2008.
  - [7] J. A. Tyson, F. Valdes, J. F. Jarvis, and A. P. Mills, Jr. Galaxy mass distribution from gravitational light deflection. *ApJ*, 281:L59–L62, June 1984.
  - [8] J. A. Tyson, F. Valdes, and R. A. Wenk. Detection of systematic gravitational lens galaxy image alignments - Mapping dark matter in galaxy clusters. *ApJ*, 349:L1–L4, January 1990.

- [9] J. F. Navarro, C. S. Frenk, and S. D. M. White. The Structure of Cold Dark Matter Halos. *Astrophys. J.*, 462:563, May 1996.
- [10] J. E. Gunn and J. R. Gott, III. On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution. *Astrophys. J.*, 176:1, August 1972.
- [11] J. F. Navarro, C. S. Frenk, and S. D. M. White. A Universal Density Profile from Hierarchical Clustering. *Astrophys. J.*, 490:493–508, December 1997.
- [12] E. A. Baltz, P. Marshall, and M. Oguri. Analytic models of plausible gravitational lens potentials. *JCAP*, 1:015, January 2009.
- [13] J. Einasto. Kinematics and dynamics of stellar systems. *Astrofiz.*, 5, 1965.
- [14] A. Kravtsov, A. Vikhlinin, and A. Meshcheryakov. Stellar mass – halo mass relation and star formation efficiency in high-mass halos. *ArXiv e-prints*, January 2014.
- [15] R. M. Reddick, R. H. Wechsler, J. L. Tinker, and P. S. Behroozi. The Connection between Galaxies and Dark Matter Structures in the Local Universe. *Astrophys. J.*, 771:30, July 2013.
- [16] R. B. Metcalf and P. Madau. Compound Gravitational Lensing as a Probe of Dark Matter Substructure within Galaxy Halos. *Astrophys. J.*, 563:9–20, December 2001.
- [17] T. E. Collett, P. J. Marshall, M. W. Auger, S. Hilbert, S. H. Suyu, Z. Greene, T. Treu, C. D. Fassnacht, L. V. E. Koopmans, M. Bradač, and R. D. Blandford. Reconstructing the lensing mass in the Universe from photometric catalogue data. *MNRAS*, 432:679–692, June 2013.
- [18] C. McCully, C. R. Keeton, K. C. Wong, and A. I. Zabludoff. Quantifying Environmental and Line-of-Sight Effects in Models of Strong Gravitational Lens Systems. *ArXiv e-prints*, January 2016.
- [19] S. Hilbert, J. Hartlap, S. D. M. White, and P. Schneider. Ray-tracing through the Millennium Simulation: Born corrections and lens-lens coupling in cosmic shear and galaxy-galaxy lensing. *A&A*, 499:31–43, May 2009.
- [20] V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, and F. Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature (London)*, 435:629–636, June 2005.
- [21] G. Battaglia, A. Helmi, H. Morrison, P. Harding, E. W. Olszewski, M. Mateo, K. C. Freeman, J. Norris, and S. A. Shectman. The radial velocity dispersion profile of the Galactic halo: constraining the density profile of the dark halo of the Milky Way. *MNRAS*, 364:433–442, December 2005.
- [22] S. Dodelson. *Modern Cosmology*. Academic Press. Academic Press, 2003.
- [23] V. Belokurov, N. W. Evans, A. Moiseev, L. J. King, P. C. Hewett, M. Pettini, L. Wyrzykowski, R. G. McMahon, M. C. Smith, G. Gilmore, S. F. Sanchez, A. Udalski, S. Koposov, D. B. Zucker, and C. J. Walcher. The Cosmic Horseshoe: Discovery of an Einstein Ring around a Giant Luminous Red Galaxy. *ApJ*, 671:L9–L12, December 2007.
- [24] G. Meylan, P. Jetzer, P. North, P. Schneider, C. S. Kochanek, and J. Wambsganss, editors. *Gravitational Lensing: Strong, Weak and Micro*, 2006.
- [25] P. Schneider and D. Sluse. Mass-sheet degeneracy, power-law models and external convergence: Impact on the determination of the Hubble constant from gravitational lensing. *A&A*, 559:A37, November 2013.
- [26] E. S. Cypriano, A. Amara, L. M. Voigt, S. L. Bridle, F. B. Abdalla, A. Réfrégier, M. Seiffert, and J. Rhodes. Cosmic shear requirements on the wavelength dependence of telescope point spread functions. *MNRAS*, 405:494–502, June 2010.
- [27] E. S. Sheldon, D. E. Johnston, J. A. Frieman, R. Scranton, T. A. McKay, A. J. Connolly, T. Budavári, I. Zehavi, N. A. Bahcall, J. Brinkmann, and M. Fukugita. The Galaxy-Mass Correlation Function Measured from Weak Lensing in the Sloan Digital Sky Survey. *AJ*, 127:2544–2564, May 2004.
- [28] C. Oaxaca Wright and T. G. Brainerd. Gravitational Lensing by NFW Halos. *ArXiv Astrophysics e-prints*, August 1999.
- [29] V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, and F. Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature (London)*, 435:629–636, June 2005.
- [30] D. N. Spergel, L. Verde, H. V. Peiris, E. Komatsu, M. R. Nolta, C. L. Bennett, M. Halpern, G. Hinshaw, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, L. Page, G. S. Tucker, J. L. Weiland, E. Wollack, and E. L. Wright. First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters. *ApJS*, 148:175–194, September 2003.
- [31] M. Jarvis. TreeCorr: Two-point correlation functions. *Astrophysics Source Code Library*, August 2015.
- [32] F. Pace, M. Manera, D. J. Bacon, R. Crittenden, and W. J. Percival. The importance of the cosmic web and halo substructure for power spectra. *MNRAS*, 454:708–723, November 2015.
- [33] M. D. Schneider, D. W. Hogg, P. J. Marshall, W. A. Dawson, J. Meyers, D. J. Bard, and D. Lang. Hierarchical Probabilistic Inference of Cosmic Shear. *Astrophys. J.*, 807:87, July 2015.
- [34] R. Trotta. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49:71–104, March 2008.
- [35] P.J. Marshall. Bayesian Analysis of Clusters of Galaxies. *PhD thesis*, September 2003.
- [36] M. Sunnker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate Bayesian Computation. *PloS Computational Biology*, 9, January 2013.
- [37] L. Lyons and M.K. Ünel. *Statistical Problems in Particle Physics, Astrophysics and Cosmology: Proceedings of PHYS-TAT05, Oxford, UK, 12-15 September 2005*. Imperial College Press, 2006.
- [38] J. Miralda-Escude. The correlation function of galaxy ellipticities produced by gravitational lensing. *Astrophys. J.*, 380:1–8, October 1991.

## 9 APPENDIX

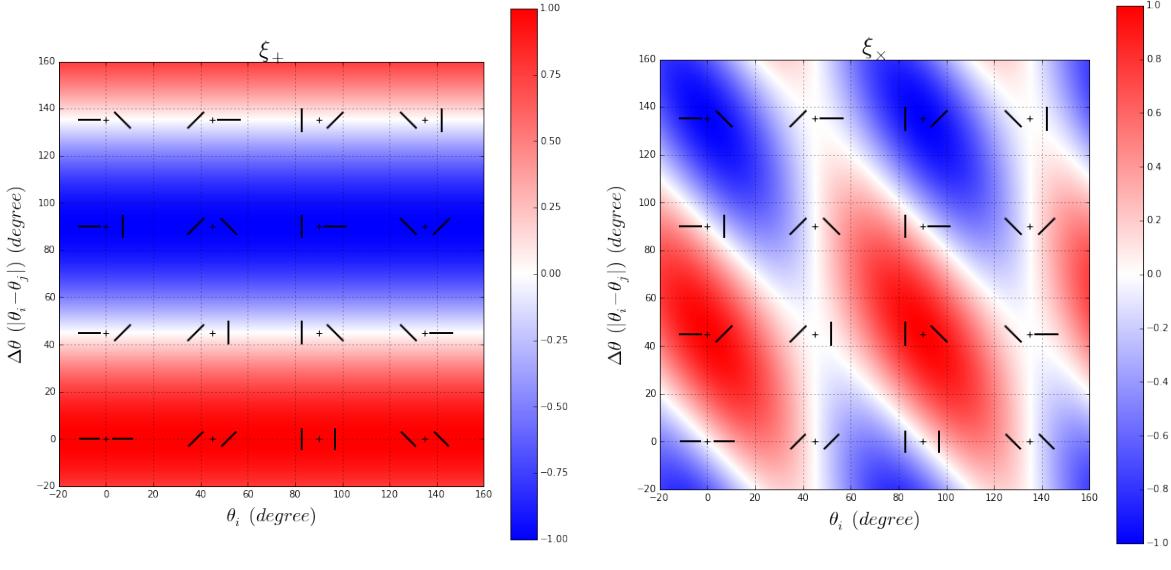
### 9.1 Visualizing Correlation Functions

At first encounter, the use of correlation functions to analyze data can be difficult to comprehend, let alone visualize. While by no means a thorough or mathematical analysis (see [38] for a more formal introduction), I have created some explanatory plots to aid in visualizing how weakly lensed ellipticities can be measured with correlation functions and place the main results of this paper in more context.

The main statistic used in this work is the ellipticity-ellipticity ( $\varepsilon\text{-}\varepsilon$ ) correlation function, which has two tangential components  $\xi_{\pm}$  and a cross component  $\xi_{\times}$ . The components most interesting for gravitational lensing are the positive tangential component  $\xi_+$ , expected to be positive for lensed galaxies, and the cross component  $\xi_{\times}$ , which is expected to be zero. To get an intuition for what these quantities measure, first consider the correlation values for individual pairs of galaxies as shown in Figure 19. In each plot, the value of the correlation between two galaxies as a function of their relative orientation is plotted as a colormap with a few sampled galaxy orientations superimposed. For  $\xi_+$ , the correlation is positive for parallel galaxies, negative for perpendicular galaxies, and zero for galaxies offset by  $45^\circ$ . In contrast,  $\xi_{\times}$  is zero for all parallel and perpendicular orientations while nonzero for most  $45^\circ$  offsets. As gravitational lensing shears source images tangentially around a foreground mass, the tangential component is a measure of how similarly two galaxies are lensed, while the cross component is an estimate of the bias (as it should be zero).

Now consider a region of space around a foreground lens populated with many background sources *before lensing*, as in Figure 20a. In this plot, source ellipticities are plotted as sticks and colored by their distance from the lens. Calculating the  $\xi_+$  component correlation pairs for all galaxies in the same color bin and making a scatter plot as a function of the separation distance between the source pairs, we find the distribution in Figure 20b. The  $\xi_+$  component of the  $\varepsilon\text{-}\varepsilon$  correlation function is simply the average of this scatter plot binned in separation distances along the  $x$ -axis. As the scatter is randomly distributed around zero, the  $\varepsilon\text{-}\varepsilon$  correlation function for unlensed sources should be zero on all scales.

The same plots for the source galaxies *after lensing* are shown in Figure 20c. The source ellipticities have clearly aligned tangentially around the foreground lens in Figure 20c, and distinct patterns have appeared in the correlation pair distribution for each color bin in Figure 20d. First consider a single color bin, such as the blue. Galaxies next to one another are lensed in approximately the same way and thus nearly parallel, leading to a positive correlation at



(a) The positive tangential component of the  $\varepsilon\text{-}\varepsilon$  correlation as a function of galaxy orientation plotted as a colormap. The correlation component is positive for parallel galaxies, negative for perpendicular galaxies, and zero for galaxies offset by  $45^\circ$ .

(b) The cross component of the  $\varepsilon\text{-}\varepsilon$  correlation as a function of galaxy orientation plotted as a colormap. While the pattern is more complicated than that of the tangential component, note that the correlation is zero for all preferred orientations after lensing.

FIG. 19: Correlation pair colormaps for  $\xi_+$  and  $\xi_{\times}$ .

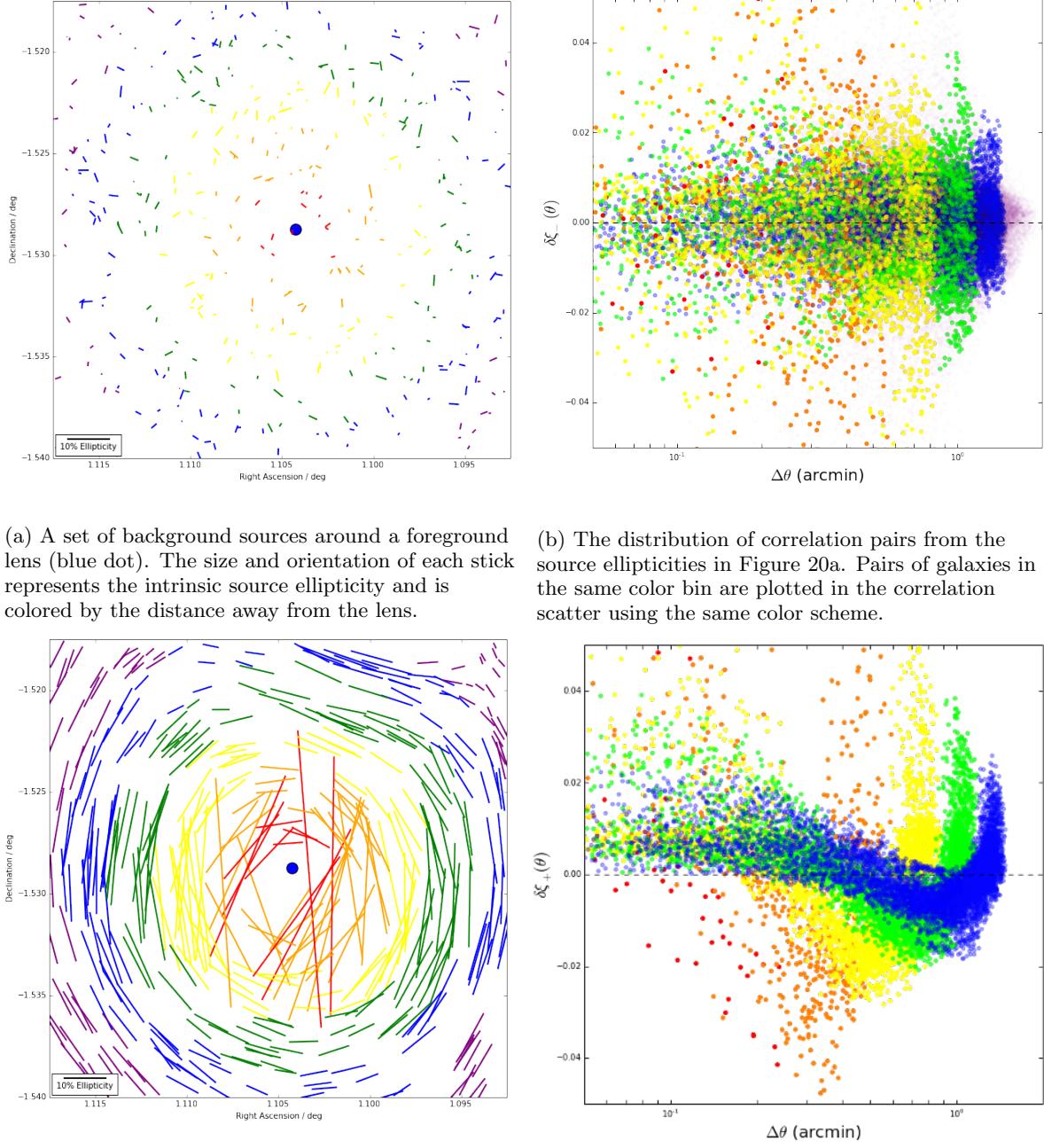


FIG. 20:  $\varepsilon$ - $\varepsilon$  correlation function visualization.

small separation distances in Figure 20d. Moving a quarter of the way around the blue circle of galaxies, the lensed ellipticities become perpendicular and have negative correlation. This can be seen as the dip in correlation at ‘middle’ separation distances in 20d. Finally for galaxies on opposite sides of the lens, their relative orientation is again parallel and the correlation has returned to positive values. This spike in correlation is clearly visible in the distribution and is shifted to the right for colors further from the lens as the radius of the color bin increases. Averaging all scatter points in discrete separation distance bins, the  $\varepsilon$ - $\varepsilon$  correlation function will no longer be zero and a detectable signal will remain.

## 9.2 Deriving the Standard Correlation Function Definition from TreeCorr's Definition

The definition for the shear-shear (or equivalently, in the case of weak lensing, ellipticity-ellipticity) correlation function components in Mike Jarvis's `TreeCorr` package is different from the common definition given in Equation (31) from Schneider [24]. For the aid of those who wish to use this very useful package but still use the conventional correlation function definition, I have outlined the connection between the two definitions and the outputs of `TreeCorr` below:

From Jarvis [31] (Page 3):

$$\xi_+ = \langle \gamma_i \gamma_j^* \rangle = \mathbf{xip} + i(\mathbf{xip_im}) \quad (42)$$

$$\xi_- = \langle \gamma_i \gamma_j e^{-4i\alpha} \rangle = \mathbf{xim} + i(\mathbf{xim_im}) \quad (43)$$

where  $\alpha$  is the angle between the two objects  $i, j$  and each  $\gamma$  is given by  $\gamma_n = |\gamma_n|e^{2i\theta_n}$  in polar form.

From Schneider [24] (Page 92):

$$\xi_{\pm} = \langle \gamma_{i_t} \gamma_{j_t} \rangle \pm \langle \gamma_{i_x} \gamma_{j_x} \rangle \quad (44)$$

$$\xi_x = \langle \gamma_{i_t} \gamma_{j_x} \rangle \quad (45)$$

where  $\gamma_{n_t} = -\operatorname{Re}(\gamma_n e^{-2i\alpha})$  and  $\gamma_{n_x} = -\operatorname{Im}(\gamma_n e^{-2i\alpha})$ .

### SCHNEIDER'S $\xi_+$ TO JARVIS'S XIP

Starting with Schneider's definition, observe that

$$\begin{aligned} \xi_+ &= \langle \gamma_{i_t} \gamma_{j_t} \rangle + \langle \gamma_{i_x} \gamma_{j_x} \rangle \\ &= \langle \operatorname{Re}(\gamma_i e^{-2i\alpha}) \cdot \operatorname{Re}(\gamma_j e^{-2i\alpha}) \rangle + \langle \operatorname{Im}(\gamma_i e^{-2i\alpha}) \cdot \operatorname{Im}(\gamma_j e^{-2i\alpha}) \rangle \\ &= \left\langle \operatorname{Re}\left(|\gamma_i| e^{2i(\theta_i - \alpha)}\right) \cdot \operatorname{Re}\left(|\gamma_j| e^{2i(\theta_j - \alpha)}\right) \right\rangle + \left\langle \operatorname{Im}\left(|\gamma_i| e^{2i(\theta_i - \alpha)}\right) \cdot \operatorname{Im}\left(|\gamma_j| e^{2i(\theta_j - \alpha)}\right) \right\rangle \\ &= \langle |\gamma_i| \cdot |\gamma_j| \cos(2(\theta_i - \alpha)) \cos(2(\theta_j - \alpha)) \rangle + \langle |\gamma_i| \cdot |\gamma_j| \sin(2(\theta_i - \alpha)) \sin(2(\theta_j - \alpha)) \rangle. \end{aligned}$$

Using the trig identities

$$\cos(u) \cos(v) = \frac{1}{2} [\cos(u - v) + \cos(u + v)], \quad (46)$$

$$\sin(u) \sin(v) = \frac{1}{2} [\cos(u - v) - \cos(u + v)], \quad (47)$$

and the linearity of the expectation operator, it is straightforward to show that

$$\xi_+ = \langle |\gamma_i| \cdot |\gamma_j| \cos(2(\theta_i - \theta_j)) \rangle.$$

Now using the identity

$$\cos(u \pm v) = \cos(u) \cos(v) \mp \sin(u) \sin(v), \quad (48)$$

the previous equation can be written as

$$\begin{aligned} \xi_+ &= \langle |\gamma_i| \cdot |\gamma_j| \cdot [\cos(2\theta_i) \cos(2\theta_j) + \sin(2\theta_i) \sin(2\theta_j)] \rangle \\ &= \langle \operatorname{Re}(\gamma_i) \cdot \operatorname{Re}(\gamma_j) + \operatorname{Im}(\gamma_i) \cdot \operatorname{Im}(\gamma_j) \rangle. \end{aligned}$$

Using the complex number identities

$$\operatorname{Re}(z) = \frac{1}{2} [z + z^*], \quad (49)$$

$$\operatorname{Im}(z) = \frac{1}{2i} [z - z^*], \quad (50)$$

it follows that

$$\begin{aligned} \xi_+ &= \left\langle \frac{1}{4} [(\gamma_i + \gamma_i^*)(\gamma_j + \gamma_j^*) - (\gamma_i - \gamma_i^*)(\gamma_j - \gamma_j^*)] \right\rangle \\ &= \left\langle \frac{1}{2} [\gamma_i \gamma_j^* + (\gamma_i \gamma_j^*)^*] \right\rangle \\ &= \langle \operatorname{Re}(\gamma_i \gamma_j^*) \rangle = \text{xip}. \end{aligned} \quad (51)$$

### SCHNEIDER'S $\xi_-$ TO JARVIS'S XIM

The first few steps are identical to the previous section except for the minus sign in the definition of  $\xi_-$ , giving

$$\begin{aligned} \xi_- &= \langle \gamma_{i_t} \gamma_{j_t} \rangle - \langle \gamma_{i_x} \gamma_{j_x} \rangle \\ &= \langle |\gamma_i| \cdot |\gamma_j| \cos(2(\theta_i - \alpha)) \cos(2(\theta_j - \alpha)) \rangle - \langle |\gamma_i| \cdot |\gamma_j| \sin(2(\theta_i - \alpha)) \sin(2(\theta_j - \alpha)) \rangle \\ &= \langle |\gamma_i| \cdot |\gamma_j| \cos(2(\theta_i + \theta_j - 2\alpha)) \rangle. \end{aligned}$$

Now using Equation (48) twice, first letting  $u = 2(\theta_i + \theta_j)$  and  $v = -4\alpha$ , and then letting  $u = 2\theta_i$  and  $v = 2\theta_j$ , this equation becomes

$$\begin{aligned} \xi_- &= \left\langle |\gamma_i| \cdot |\gamma_j| \left( [\cos(2\theta_i) \cos(2\theta_j) - \sin(2\theta_i) \sin(2\theta_j)] \cos(4\alpha) + [\sin(2\theta_i) \cos(2\theta_j) + \cos(2\theta_i) \sin(2\theta_j)] \sin(4\alpha) \right) \right\rangle \\ &= \left\langle [\operatorname{Re}(\gamma_i) \cdot \operatorname{Re}(\gamma_j) - \operatorname{Im}(\gamma_i) \cdot \operatorname{Im}(\gamma_j)] \cos(4\alpha) + [\operatorname{Im}(\gamma_i) \cdot \operatorname{Re}(\gamma_j) + \operatorname{Re}(\gamma_i) \cdot \operatorname{Im}(\gamma_j)] \sin(4\alpha) \right\rangle. \end{aligned}$$

Using the identities (49) and (50), and simplifying the leftover terms, this equation can be shown to equal

$$\xi_- = \langle \operatorname{Re}(\gamma_i \gamma_j) \cos(4\alpha) + \operatorname{Im}(\gamma_i \gamma_j) \sin(4\alpha) \rangle.$$

Now observe that for two complex numbers  $a$  and  $b$ , it is true that

$$\operatorname{Re}(a \cdot b^*) = \operatorname{Re}(a) \cdot \operatorname{Re}(b) + \operatorname{Im}(a) \cdot \operatorname{Im}(b). \quad (52)$$

Then setting  $a = \gamma_i \gamma_j$  and  $b = e^{4i\alpha}$ , it must be true that

$$\operatorname{Re}(\gamma_i \gamma_j e^{-4i\alpha}) = \operatorname{Re}(\gamma_i \gamma_j) \cos(4\alpha) + \operatorname{Im}(\gamma_i \gamma_j) \sin(4\alpha).$$

Combining this with our previous result, this means that

$$\xi_- = \langle \operatorname{Re}(\gamma_i \gamma_j) \cos(4\alpha) + \operatorname{Im}(\gamma_i \gamma_j) \sin(4\alpha) \rangle = \langle \operatorname{Re}(\gamma_i \gamma_j e^{-4i\alpha}) \rangle = \text{xim}. \quad (53)$$

### SCHNEIDER'S $\xi_\times$ TO JARVIS'S $\frac{1}{2}(\text{xim\_im} - \text{xip\_im})$

Starting from Schneider's definition of  $\xi_\times$ ,

$$\begin{aligned}
\xi_{\times} &= \langle \gamma_{i_t} \gamma_{j_{\times}} \rangle \\
&= \left\langle \operatorname{Re} \left( |\gamma_i| e^{2i(\theta_i - \alpha)} \right) \cdot \operatorname{Im} \left( |\gamma_j| e^{2i(\theta_j - \alpha)} \right) \right\rangle \\
&= \langle |\gamma_i| \cdot |\gamma_j| \cos(2(\theta_i - \alpha)) \sin(2(\theta_j - \alpha)) \rangle.
\end{aligned}$$

Using the trig identity

$$\cos(u) \sin(v) = \frac{1}{2} [\sin(u+v) - \sin(u-v)], \quad (54)$$

the previous equation becomes

$$\xi_{\times} = \left\langle |\gamma_i| \cdot |\gamma_j| \cdot \frac{1}{2} [\sin(2(\theta_i + \theta_j - 2\alpha)) - \sin(2(\theta_i - \theta_j))] \right\rangle.$$

Next, applying the identities

$$\sin(u \pm v) = \sin(u) \cos(v) \pm \cos(u) \sin(v) \quad (55)$$

along with (48) iteratively until each trig function only has a single term, the equation becomes

$$\begin{aligned}
\xi_{\times} &= \left\langle |\gamma_i| \cdot |\gamma_j| \cdot \frac{1}{2} \left( [\sin(2\theta_i) \cos(2\theta_j) + \cos(2\theta_i) \sin(2\theta_j)] \cos(4\alpha) \right. \right. \\
&\quad \left. \left. - [\cos(2\theta_i) \cos(2\theta_j) - \sin(2\theta_i) \sin(2\theta_j)] \sin(4\alpha) \right. \right. \\
&\quad \left. \left. - \sin(2\theta_i) \cos(2\theta_j) + \cos(2\theta_i) \sin(2\theta_j) \right) \right\rangle \\
&= \left\langle \frac{1}{2} \left( [\operatorname{Im}(\gamma_i) \cdot \operatorname{Re}(\gamma_j) + \operatorname{Re}(\gamma_i) \cdot \operatorname{Im}(\gamma_j)] \cos(4\alpha) \right. \right. \\
&\quad \left. \left. - [\operatorname{Re}(\gamma_i) \cdot \operatorname{Re}(\gamma_j) - \operatorname{Im}(\gamma_i) \cdot \operatorname{Im}(\gamma_j)] \sin(4\alpha) \right. \right. \\
&\quad \left. \left. - \operatorname{Im}(\gamma_i) \operatorname{Re}(\gamma_j) + \operatorname{Re}(\gamma_i) \cdot \operatorname{Im}(\gamma_j) \right) \right\rangle.
\end{aligned}$$

Using the identities (49) and (50), this can be simplified to

$$\xi_{\times} = \left\langle \frac{1}{2} [\operatorname{Im}(\gamma_i \gamma_j) \cos(4\alpha) - \operatorname{Re}(\gamma_i \gamma_j) \sin(4\alpha) - \operatorname{Im}(\gamma_i \gamma_j^*)] \right\rangle.$$

Consider again two complex numbers  $a$  and  $b$ . Observe that

$$\operatorname{Im}(a \cdot b^*) = \operatorname{Im}(a) \cdot \operatorname{Re}(b) - \operatorname{Re}(a) \cdot \operatorname{Im}(b). \quad (56)$$

Then setting  $a = \gamma_i \gamma_j$  and  $b = e^{4i\alpha}$ , it must be true that

$$\operatorname{Im}(\gamma_i \gamma_j e^{-4i\alpha}) = \operatorname{Im}(\gamma_i \gamma_j) \cos(4\alpha) - \operatorname{Re}(\gamma_i \gamma_j) \sin(4\alpha).$$

Combining these results give

$$\begin{aligned}
\xi_{\times} &= \left\langle \frac{1}{2} [\operatorname{Im}(\gamma_i \gamma_j) \cos(4\alpha) - \operatorname{Re}(\gamma_i \gamma_j) \sin(4\alpha) - \operatorname{Im}(\gamma_i \gamma_j^*)] \right\rangle \\
&= \left\langle \frac{1}{2} [\operatorname{Im}(\gamma_i \gamma_j e^{-4i\alpha}) - \operatorname{Im}(\gamma_i \gamma_j^*)] \right\rangle \\
&= \frac{1}{2} [\operatorname{xim_im} - \operatorname{xip_im}].
\end{aligned} \quad (57)$$

as desired.