The Quest for Dark Matter
CS229 Project Report, Fall 2012

Abhishek Sheshadri*, Pranav Pai†, Sreenath Krishnan‡

*Aeronautics and Astronautics, Stanford University, Email: abisheks@stanford.edu
†Electrical Engineering, Stanford University, Email: pmpai@stanford.edu
‡Mechanical Engineering, Stanford University, Email: ksree@stanford.edu

***Abstract*—The main objective of this project is to devise an algorithm that is able to predict the locations of multiple dark matter halos in a given image of galaxy distributions precisely and without directional bias. It is based on the active Kaggle competition Observing Dark Worlds, and uses the data, rules and accuracy measures prescribed by the competition.**

I. INTRODUCTION

Dark matter halos are not directly observable, and this continues to be a fundamental problem in its detection, and therefore in our understanding of its behavior. However, these large clusters tend to have observable effects on the surrounding visible neighbors and on light due to their ability to bend the fabric of space-time via gravitational effects. This leads to the phenomenon of gravitational lensing, whereby light from galaxies behind these dark matter clusters is distorted along the path to our observation point causing an observable distortion (tangential shearing) of these nearby (in the 2D sky sense) galaxies.

Understanding dark matter requires accurate estimates of the positions of these clusters. Till now, physicists have not been able to predict dark matter locations accurately without detailed red-shift and x-ray based analysis of particular small regions in the sky. This process is expensive and can only be used to verify halos once a reasonable prediction of the halo locations are made.

II. TRAINING DATA AND OBSERVATIONS

We were provided with 300 training skies, each with 300-700 galaxies and 1-3 halos. The locations of the galaxy centers and the ellipticities (e_1 and e_2) of the galaxies as well as the halo positions have been provided. The final algorithm will be tested on a different set of skies. In order to evaluate a prediction, the competition has set a 'metric' described as follows:

$$m = \frac{F}{1000} + G \tag{1}$$

where m is the metric, F is the average radial distance from the prediction to the true position of the halo and

$$G = \sqrt{(\frac{1}{N}\sum_{i=1}^{N}\cos\phi_i)^2 + (\frac{1}{N}\sum_{i=1}^{N}\sin\phi_i)^2} \tag{2}$$

where N is the total number of halos across all the skies and ϕ is a measure of directional bias.

For a completely random prediction of the position of halos, the metric comes out to be about 1.94. An open source software called lenstool is the most widely used and accepted tool in this field and gets a metric of about 1.01 on the test skies. Instead of just trying to improve on lenstool, we started off from scratch so that we can have a rich learning experience and understand and use a variety of algorithms rather than just trying to improve a fixed algorithm.

If the galaxies were actually circular, they would have just appeared to be ellipses tangentially aligned to the halo. But since the galaxies themselves are elliptical (in the 2D sky) and have random orientations to begin with, the shearing effect of the galaxies rotates and/or magnifies them towards having a higher tangential ellipticity with respect to the halo. It was observed that the distributions of e_1 and e_2 were $\sim \mathcal{N}(0, 0.04)$.

III. HALO MODEL

The effect of halo on a galaxy can be modeled as a tangential shear. The magnitude of tangential shear should go down with distance(r). Assuming the halo is spherically symmetric the shear can be modeled as:

$$\tau_{tangential} = \frac{s}{r^n} \tag{3}$$

$$\tau_{normal} = \nu\tau_{tangential} \tag{4}$$

This model has 3 parameters [s,n,ν]. The parameter 's' can be interpreted as the strength of the halo and 'ν' can be interpreted as an equivalent to Poissons ratio that is encountered in mechanics of materials. The positive integer 'n' controls the rate of decay with distance. As a further simplification, we assume that ν=0. It was found that n=1 fits well with the data provided.

In order to provide an intuitive picture of the model, let us consider what effect a halo following equations (3, 4) has on a circular galaxy with radius 'R'. The galaxy becomes elliptical with major axis a = R(1+$\frac{s}{r}$) and minor axis b = R. Also the major axis will be perpendicular to the line joining the galaxy and halo. It can be proved that the application of halo is not completely reversible, in the sense there are multiple initial configurations of galaxy that can result in the same final configuration.

IV. Methods for single halo skies

A. Net Tangential Ellipticity (NTE)

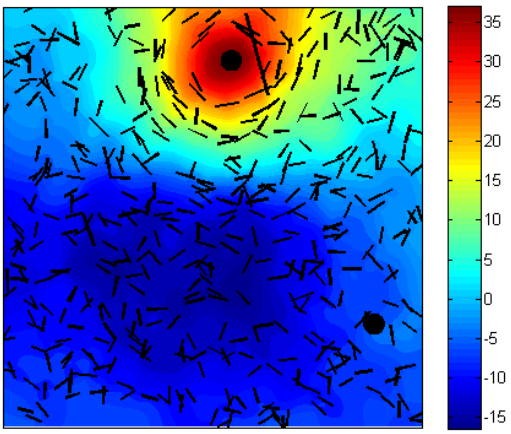

Fig. 1. A sample 2 halo case. The lines are the major axes of the galaxies and the 2 black circles are the halo locations. We can see the strong halo is in a region of high NTE

We observed that the effect of the halo is to increase the effective ellipticity of a galaxy in a direction tangential to the line joining the halo and the galaxy. So we calculated the net tangential ellipticity at each point in the sky due to all the galaxies and plotted it as a contour plot. In the single halo case, the point of maximum tangential ellipticity is in general close to the actual halo location. But in the case of multiple halos, usually only the strong halo can be identified in this manner as shown in Fig. 1.

$$NTE = \sum_{i=1}^{n} -(e_1^{(i)} \cos\phi^{(i)} + e_2^{(i)} \sin\phi^{(i)}) \tag{5}$$

B. Aligned galaxies as an indicator for outliers

Though the NTE is a good indicator of the location of the strongest halo for most cases, there are a few cases for which this does not work and we referred to these cases as 'outliers'. An example of such a sky is shown in Fig. 2.

In such cases a brief look at the galaxies aligned almost tangentially to the halo gave us the idea of using the aligned galaxies alone for calculating the maximum NTE and this sometimes improves the accuracy of the prediction. Fig. 3 shows all galaxies aligned almost tangentially to the halo. So, at every point we consider all galaxies almost tangentially aligned to that point and compute the NTE using these galaxies only.

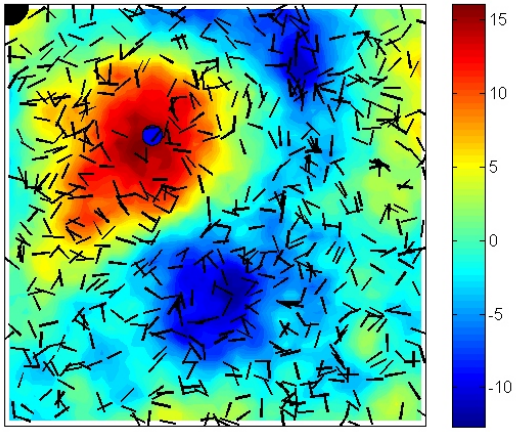

Fig. 2. An example of a case wherein the halo is not close to the maximum NTE point. The blue circle shows the predicted halo position and the black circle shows the actual halo

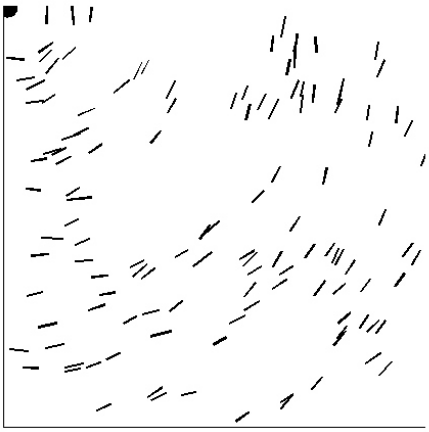

Fig. 3. The no. of galaxies aligned almost tangential to the halo is high

C. Logistic Regression

The data as such is not in a form where one can apply logistic regression directly. This is because we have a single output (which is the halo position) as a function of a set of inputs (which are the properties of galaxies). Not only does the input set has varying length across different training skies, the corresponding input sets of two different skies are not related. However this problem can be converted to a standard logistic regression problem using the following tricks:

- Divide each sky into a fixed number of bins. Each bin is assumed to be an observation.
- Instead of considering features of each galaxy, one can now set features for each bin which will be an average of the feature of galaxies.
- The target variable can now be set as whether a halo is present in the given bin or not

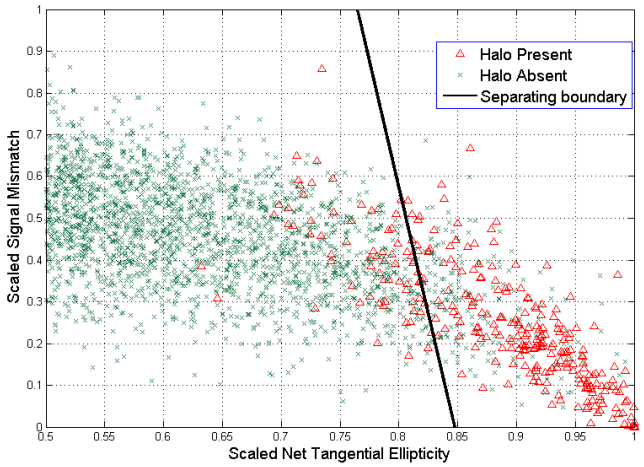

Fig. 4. Logistic Regression Results

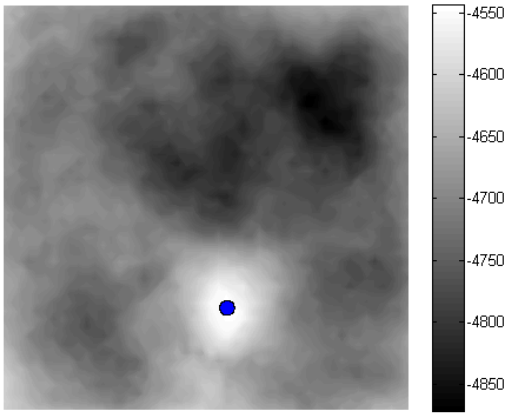

Fig. 5. Maximum Likelihood contours based on Generative Learning Results

1) Feature Selection: It has been shown that a high value of NTE and a low value of signal mismatch (Section: VI-C) are both good indicators of the presence of halos. In addition we also saw there are significant numbers of galaxies aligned tangential to the halo. Hence it makes sense to use the following as the input features

- NTE at the center of the bin
- Signal mismatch w.r.t the center of the bin
- The number of galaxies aligned tangentially with the center of the bin

Since NTE and signal mismatch had widely varying limits across skies, we decided to scale it between 0 and 1 before performing the regression.

2) Results: It was found that the third feature is a relatively weak one. We present the results of logistic regression for the single halo case in Figure [4]. Such a classification can also accurately predict the position of strongest halo in multiple halo cases. The performance metric was as low as 0.2387 for single halo predictions.

V. Generative Learning

Let $G^{(i)}$ denote the features of galaxy 'i' and H denote the halo features (position and strength) and S = $\{G^{(i)}\}$, which denotes the entire sky. The given problem is to find H that maximizes P(H|S). Using Bayes rule, assuming the galaxy observations are conditionally independent given H,

$$P(H|S) = \frac{P(S|H)P(H)}{P(S)} = \prod_{i=1}^{n} \frac{P(G^{(i)}|H)P(H)}{P(G^{(i)})} \quad (6)$$

$P(G^{(i)})$ is constant for a given sky. Since it's equally likely that the halo is present anywhere, the prior P(H) is just a uniform distribution. Since we have already assumed a model for the halo we can find a discrete estimate for $P(G^{(i)}|H)$. Even though the expression involves position of galaxies and halos as well as the halo strength, all these can be characterized by a particular value of tangential shear. In other words, if we have a distribution $P(e_1, e_2|\tau)$ where e_1 is in the tangential direction of the halo, we can fully characterize $P(G^{(i)}|H)$. A discrete approximation of $P(e_1, e_2|\tau)$ can be obtained by the following sampling technique:

- Divide the possible values of e_1, e_2 and τ into a finite number of bins. Let the set of such points be $\{e_1^i, e_2^j, \tau^k\}$ and the bin size be {m x m x p}
- Generate 'q' (q>>m) pairs of random samples drawn from $\mathcal{N}$(0,0.04) as candidates for halo absent sky.
- Apply the tangential shear to the samples to find the resulting $e_1^{(i,j)}$, $e_2^{(i,j)}$ for all the 'p' tangential shear values. Here 'i' is an index for the sample and 'j' is an index for the tangential shear.
- The resulting values are discretized to the respective bins
- The distribution can be approximated as:

$$P(e_1^i, e_2^j|\tau^k) = \frac{\sum_{r=1}^{q} 1\{e_1^{(r,k)} = e_1^i, e_2^{(r,k)} = e_2^j\} + 1/m}{m+q} \quad (7)$$

We have used laplace smoothing to avoid 0/0 cases when using this distribution. Hence we can determine the likelihood $P(H|S)$ for a given sky at any number of points using equation 6. The maximum likelihood point is chosen as the position of halo. The likelihood contour of a typical sky is shown in Figure [5] This approach is also successful in finding the strongest halo (metric of 0.554 for single halo case), which further justifies the assumptions in the model. This model can be extended to the case with multiple halos by appropriately calculating the probability distribution.

VI. Methods for multiple halo skies

Since we are able to identify the strongest halo by means of one of the methods described above, most of our methods to find the location of the second and third halo assume the knowledge of the first halo.

A. Iterative method based on alignment of galaxies

So far we have seen methods to capture the strongest or the last-applied halo. In order to get the second and third halo location, one of the first methods we tried was to remove the galaxies which are aligned almost tangentially to the first halo. These are the galaxies which contribute the most to the high NTE values around the strong halo. So we removed all galaxies which are almost tangential (within $\pm 10^\circ$ of tangential direction) and rebuilt the NTE contour and considered the point with the maximum NTE to be the halo location. We do this repeatedly in case of more than 2 halos. This method gave us a metric of 0.9499 across the training skies.

B. Halo Removal/Elimination Strategy

Given that we have effective methods to determine the position of the strongest or last-applied halo even in multiple-halo skies, we now try to unapply the identified halo so as to capture the effect of the other halo. From physics literature, it can be found that the shearing strength of a halo decreases with distance from the galaxy. We already mentioned this in the Halo Model section and we also noted that such a model is non-invertible. However we can find an approximate inverse by specifying additional constraints. In the present case, given a final configuration we selected the initial configuration that has the minimum ellipticity. This makes the model invertible i.e. we can apply a negative strength of the same magnitude and approximately recover the original sky. However, we need to have an estimate for the strength of the halo before unapplying the halo. To find an estimate for the strength:

1. We assume the position of the strong halo known (from one of the single-halo methods).

2. We apply a halo onto the ideal sky at this location with varying halo strengths and calculate minimum-mismatch value (described in the next section) of the sky.

3. We consider the strength which causes the least min-mismatch as the strength of the halo. Lets call this strength S_o.

Now we can use the halo model to obtain e_1 and e_2 of sky without the halo. We can then try one of our single-halo capturing methods to get the second halo. Fig. 3 shows the unapplication/removal of the halo. Here the strength of the halo removed is varied and the resultant NTE maps shown.

This method works well for 2-halo and 3-halo cases in identifying the first 2 halos but does not work very well when going from second to third halo. This method gives us a overall metric of 0.94 across all the training skies and about 1.1 on the test skies.

C. Minimum mismatch methods

The mean of the ellipticities e_1 and e_2 are both zero. This tells us that many of the galaxies are close to circular before halo application. We also experimented with the order of application of the halo and based on our observations that the last applied halo has a high probability of being the strongest halo in terms of the NTE, we decided to assume that the strongest halo according to the NTE is also the last applied. In this method we do the following:

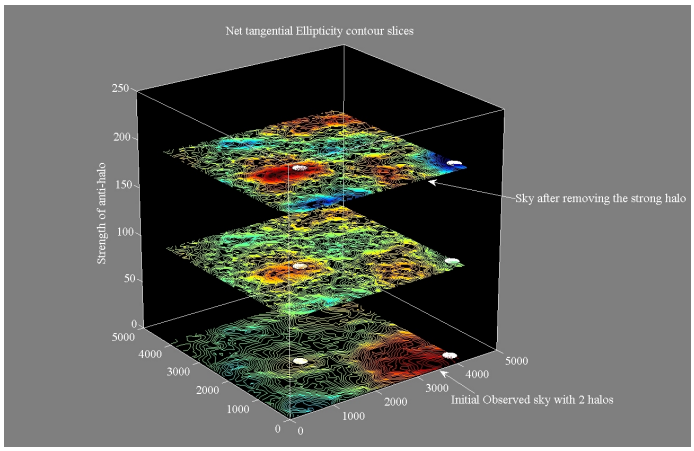

Fig. 6. Effect of halo removal. The strong halo at the bottom right corner is removed gradually and it can be seen that the red patch moves near the second halo indicating its position

1. Discretize the sky into bins or basically create sampling points across the sky.

2. Create an ideal sky i.e a sky with the same galaxy locations but zero ellipticities (basically circular galaxies).

3. At each point (or bin center), use our halo model and a guessed strength and apply the halo on the ideal sky placing it at this sampling point.

4. Now apply the first known strong halo onto this above sky using our model and the 'optimal strength' for this halo (described in the previous section). Let the resulting sky have ellipticities e_{1o} and e_{2o}.

5. Calculate the "mismatch" between this resulting sky and the actual given sky. We have used various definitions of mismatch like $\sqrt{\|e_1 - e_{1o}\|_2^2 + \|e_2 - e_{2o}\|_2^2}$ and $\|e_{tan} - etan_o\|_2$ etc.

5. We find the point in the sky which has the least mismatch as the halo prediction and call this mismatch value as the minimum mismatch or min-mismatch.

6. We repeat steps 1-5 for various strengths and take the result which had the least overall min-mismatch value as the location of the second halo.

This method of using min-mismatch described above can be used in the single halo case also where step 4 will not be required. We used this as an alternate to NTE-based approached for the single halo cases and obtained slightly better results overall.

This method again works well for 2 halo cases but not so well for 3 halo cases because in the 3 halo case, even after assuming the location and strength of 1 halo (the strongest one), we have 2 halo locations and strengths to vary and test the mismatch. It becomes both inefficient and ineffective when trying to vary all 4 parameters and trying to find the least mismatch. The overall metric is about 1.01 on the training

skies.

VII. Fixed Point Method

It was shown how NTE calculations fail for some single halo cases. The root cause of the problem is that the halo is weak to cause any significant change to the galaxies. However, this would have been possible if we had information about the initial sky(the sky before the application of halos). It is natural to ask if there is a method to recover the initial sky. Let E_i and E_o denote the ellipticities of initial sky and observed sky respectively. These two skies can be related using the halo model. Let F denote the halo function of the given sky that takes galaxy features as input, applies the tangential shear and return the modified galaxy. Then we have:

$$E_o = F(E_i) \tag{8}$$

F(0) will then denote the ellipticities had the galaxies been circular to start with. One key observation that we made is depicted in Figure [8]. The figure is generated based on the model. The top left figure shows the NTE distribution of the observed sky for a weak halo case. Clearly a prediction using maxima of NTE will result in a significant error. The reason for that is clear in the top right figure, which shows the NTE distribution of the initial sky. The effect of the halo is weak enough not to shift the maxima to its position. However if we look at the difference(bottom left), we find that the maxima is exactly at the halo. This means that if we apply maxima of NTE criteria over the difference, it would not fail even for weak halos. Another remarkable observation is that the difference is very similar to the NTE distribution of sky formed by applying the halo onto a circular set of galaxies. Since equation 5 can be represented as NTE = A E, where E denotes the ellipticities, the above observation can be written as

$$AF(0) \approx AE_o - AE_i \tag{9}$$

$$\Rightarrow AF(0) \approx AF(E_i) - AE_i \tag{10}$$

Assuming g(x) = AF(x) - AF(0), the equation simplifies to:

$$g(E_i) \approx E_i \tag{11}$$

This is clearly finding a fixed point of the function g different from zero. The fixed point gives us the initial sky. This is slightly more complicated than a standard fixed point problem because the function g(x) is specified only if one specifies both halo position and strength.

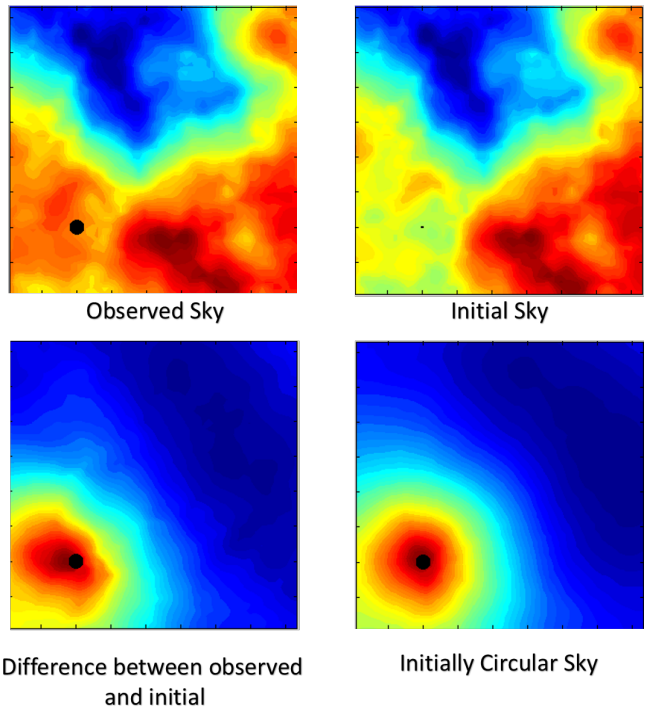

Fig. 7. NTE Contours to explain the fixed point iteration

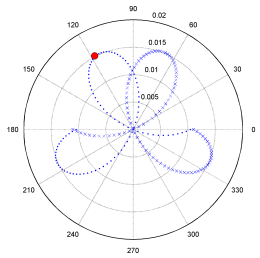

Fig. 8. Direction of maximum expectation. The red dot indicate the angle between halos

VIII. Other ideas explored

1) Direction of maximum expectation: This is a concept useful in case of two halos. Here we assume that the expected value of ellipticity will be maximum in the direction of line joining the two halos. The assumption holds fairly well if the strength of both halos are relatively close.(Figure [8])
2) k-means over intersection points: This is again an extension of the aligned galaxies concept. The assumption that there will be a significant number of galaxies aligned with the halo, would also mean that the intersection of minor axis of galaxies are expected to cluster around the halo. We calculated all the intersection points and applied k-means clustering over those points. However, this method works only for strong halos.
3) Gravitational lensing approximations: The average ellipticity for an unlensed sky (a sky without a halo) is zero. According to the weak lens approximation, the average ellipticity in a lensed sky can be used as an approximation to the average shear with lensing. We used this idea to get an estimate of the strength of the halo in order to un-apply it. However this did not work, because in a multiple halo case this average is an indicator of the net strength rather than the strength of the strongest halo and probably also because the simulated data is not made to conform to this.