Alvaro Aguilar
Sebastian Echegaray

Observing Dark Worlds

Project Overview

In this project, we analyze images showing galaxies in space to predict the location of Dark Matter Halos (DMH), which by definition cannot be observed directly.

Galaxies are elliptical in nature, but we assume this property to be inherently random across all space, such that the total ellipticity with respect to a given point averages out to zero. A DMH between us and the observed galaxies will bend space-time, forcing the galaxies to appear to have a different ellipticity than they actually do. This phenomenon is known as gravitational lensing, and was first introduced in the 1920's.

The stated problem is part of a competition between 204 teams organized by Winton Capital and Kaggle [1].

Data

Our training and testing data consists of 300 simulated skies of 4200x4200 pixels. Each simulated sky contains anywhere from 300 to 720 galaxies. Each galaxy is represented by an (x,y) coordinate (0 to 4200 pixels) and a measure of ellipticity, which tells us the shape and angle of the galaxy with respect to a given point in the sky. Each image can contain from 1 to 3 DMH. Additionally, the DMH are represented as a point source by their (x,y) coordinates, giving the ground truth for each simulated sky image. The data is provided by Winton Capital. In Figure 1 we can observe three generated images showing the position of the DMHs as a red circle and each galaxy as a white ellipse. The ellipses maintain the original ellipticity of the simulated galaxies (after being affected by the DMH).

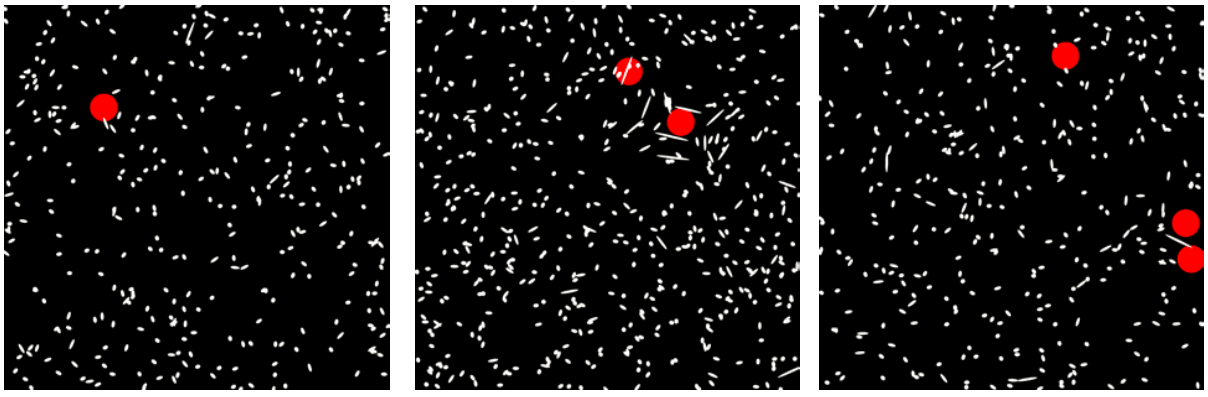

Figure 1. Rendered images from our training dataset. DMH are shown as red circles with arbitrary radius and galaxies as white ellipses

A limiting factor is that the above data is simulated by the contest organizers, therefore subject to random noise, which is added to the underlying physics euqations describing the gravitational lensing

effect. A key assumption for the two and three-halo skies is that the effect of each DMH can be added linearly to obtain the final effect on total ellipticity.

Since each image consist of 4200 x 4200 pixels, our approach was to divide each image into bins of 100 x 100, with each bin consisting of 42 pixels. This division was necessary due to the computing complexity of evaluating each pixel in a 4200x4200 grid. By lowering the number of possible locations we make the problem easier to tract in memory an in processing power.

Feature Extraction

We initially focused on extracting two features (Signal Map and Maximum Likelihood) for each of the bins described above which helped us find the centers of the DMH. These features were inspired by the tutorial in [2].

Our first feature, called Signal Map, calculates the interaction tangential ellipticity from all the galaxies in the sky to a given point in the image. To do this, the algorithm assumes there is a DMH on a particular bin, and calculates the angle of each galaxy in the sky with respect to that bin. Using the angular information, it then adds the tangential ellipticity from every galaxy in the sky with respect to that bin. The algorithm assigns the bin a score proportional to the amount of ellipticity it can account. The process is repeated until all bins have been assigned a score. As was stated in the introduction, in the absence of DMHs, we would expect the average tangential ellipticity of any given point in the sky to be zero. Using this information, we know that the bin which records the highest tangential ellipticity is the bin where the halo is most likely to be. A heat map used for visualization of this data is shown in Figure 2.

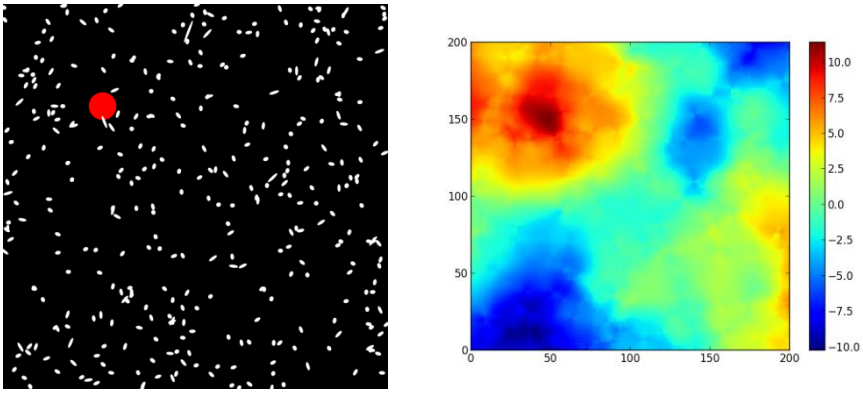

Figure 2. Signal map showing most probably location of a DMH (note for this rendering, the sky was divided into 200x200 bins, rather than 100x100)

Our second feature, called Maximum Likelihood, assumes that the distortion caused by the DMH has an effect that acts only on galaxies that are close to it. This is commonly described as a having a $1/r^a$ drop off, where r is the radial distance between the center of the proposed DMH and a given galaxy, and a is a factor which allows control of the expected drop off. The algorithm for calculating this feature is as follows: assume there is a DMH in a particular bin, calculate what the total ellipticity of the sky would look like given a $1/r^a$ model, and finally try to fit the predicted ellipticity of the sky to the actual ellipticity using a chi squared distribution. Sweeping the exponent on the r term allows us to vary how

local the gravitational lensing caused by the DMH is. The chi-square distribution is used because the random variable for total ellipticity is described by the sum of squares of two independent random variables: $\left(e_{1(predicted)} - e_1\right)^2 + \left(e_{2(predicted)} - e_2\right)^2$. A heat map for the same simulated sky as Figure 2, but created using the maximum likelihood approach is shown in Figure 3 below.

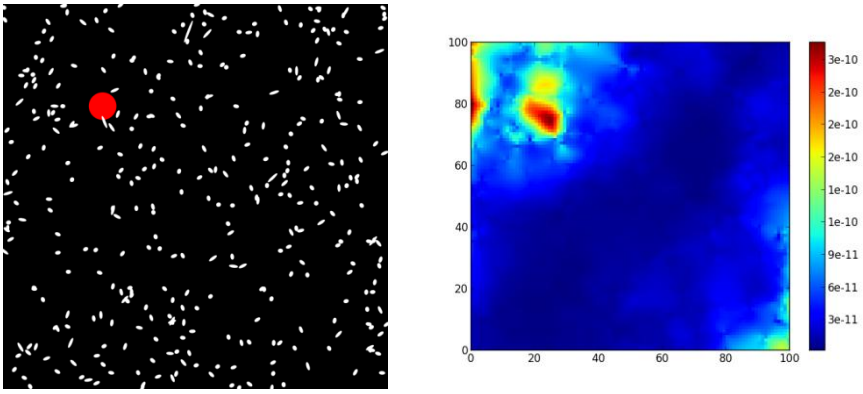

Figure 3. Heat map generated using the maximum likelihood approach with a $1/r^a$ dropoff factor, with $a = 0.2$

Inspired by the better performance of this $1/r^a$ model, we modified our initial Signal Map algorithm to include this factor in its calculations and therefore generate a third set of features. The idea is that the total ellipticity calculated can be multiplied by a $1/r^a$ factor in order to assign different weights to ellipticities of galaxies according to distance. By analyzing how well our algorithm predicted the location of DMH's in our training dataset we empirically found that $a = 0.2$ provides the best compromise between local and global information. However, we also included heat maps using this technique for $a = 0.1$, $a = 0.2$ and $a = 0.5$ in our feature set.

As a fourth approach, which we have called Directional Bias, we take advantage on the fact that ellipticities of all galaxies should sum to zero if there is no distortion. For this feature we create a multi-scale grid and calculate the ellipticity bias (i.e. the average ellipticity) in each of the cells. Our hypothesis is that for each galaxy: $\varepsilon_{observed} = \varepsilon_{galaxy} + \varepsilon_{distortion}$ where $\varepsilon_{observed}$ is the ellipticity that we can read from the dataset, ε_{galaxy} is the original galaxy ellipticity, and $\varepsilon_{distortion}$ is the distortion caused by the warping effect of the DMH. Then:

$$\sum \varepsilon_{observed} = \sum (\varepsilon_{galaxy} + \varepsilon_{distortion})$$

$$= \sum \varepsilon_{galaxy} + \sum \varepsilon_{distortion}$$

$$= 0 + \sum \varepsilon_{distortion}$$

By taking the mean of a group of galaxies close to each other, we can approximate the ellipticity caused by the distortion of that cluster. Using this feature, we can try to find patterns which show the location of the DMH across different skies. The idea is drawn in Figure 4, where we can observe how the biases (shown as blue lines) organize themselves tangentially to the DMH (shown as a red circle).

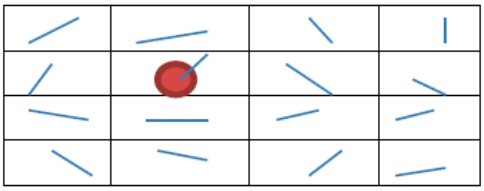

Figure 4. Calculated biased directional ellipticity in a grid. The ellipticities (shown in blue) aligned tangentially to the DMH center (shown in red)

Feature fitting

Using the first three features (Signal Map, Maximum Likelihood, and Local Signal Map), we created features for each bin in the 300 training skies (recall each sky has 100x100 bins, each measuring 42x42 pixels). These features were evaluated by themselves and in group in their ability to find the DMH centers. To combine the features we used an ensemble of trees classifier (the TreeBagger algorithm provided by Mathworks) [4]. The results are shown in Table 1.

	Distance from Closest Halo		
Feature used	**Average (bins)**	**Median (bins)**	**Standard Dev (bins)**
Global Signal Map	4.90	2.00	12.46
Max Likelihood (a=0.2)	7.49	2.00	17.94
Local Signal Map (a=0.1)	4.88	2.00	12.41
Local Signal Map (a=0.2)	4.24	2.00	10.72
Local Signal Map (a=0.5)	5.45	2.23	11.29

Table 1. Results after using an ensemble of trees classifier for 300 simulated skies

For testing, we used five-fold cross-validation by sky. That is, we separated the skies randomly into five groups of 60 skies each. These groups were in turn composed by 20 skies with one halo, 20 skies with two halos and 20 skies with three halos. Then we trained the classifier in four of the five groups and used this to test the fifth group. We repeated this procedure for each of the five groups. This allowed us to maximize the amount of data that we could extract from our training set as we tested against the samples that were not trained (and that we had ground truth).

Conclusions

The project was a very educational experience in two separate fronts. First came the understanding of the underlying physics explaining the gravitational lensing effect. Using this as our basis, we were then able to extract useful features for our classifier. Finally, we ended up with a classifier that over-fit our particular training set and didn't perform as well when presented with new data. As of right now we are creating new ground truth by convolving a Gaussian with the original position of the DMH. This will give us a probability map of each location; we will try to use regression to improve our results and submit to the competition, which finalizes on Sunday December 16, 2012.

Future Improvements

There is room for improvement in our initial feature extracting algorithms when it comes to dealing with multiple DMH in a given sky. Given DMH effects add linearly for each halo present in the sky, two approaches for dealing with multiple halos can be implemented

Separately detect multiple halos and add their effects

The proposed feature extraction algorithms all perform reasonably well when extracting the location of one DMH, even when there are multiple DMHs in an image. Under two- or three-DMH skyes, the algorithms could be modified to compute the effect that the first DMH had on the total ellipticity in the sky, and attempt to subtract it in order to find the effect that the second halo will have. This could get overly complicated for skies containing three DMHs. Presumably, on the two-DMH sky, knowing the location and the exact effect of the second halo will somehow affect our initial choice for the first halo. This means that after finding the second halo, we should go back and modify the location and effect of the first halo, and then try to find the second one again. It is easy to see that having three halos would become computationally expensive.

Simultaneously detect multiple halos

Given knowledge of the number of halos in each sky, the maximum likelihood algorithm could potentially sweep the sky and calculate the effect two or three halos at different positions would have on the ellipticity. Since the maximum likelihood algorithm allows us to fit the predicted ellipticity to the overall ellipticity given by the sky, modification to include multiple halos comes naturally. By trying to fit the predicted ellipticity to the known ellipticity for all DMH at once, the problem of assigning different weights to the effect each halo contributes to the overall ellipticity can be avoided. A possible algorithm would proceed as follows: assign DMHs to three separate bins, predict the ellipticity cause by these halos assuming a $1/r^a$ model, and fit the predicted ellipticity to the actual ellipticity using a chi-square distribution (now using the sum of squares of six independent variables, namely $e1$ and $e2$ contributed by each of the three DMHs). The algorithm would then change the location of one DMH at a time until all halos have been assigned to each bin in the image. Finally, the combination that produces the highest likelihood gives the correct location for each of the three DMHs.

References

[1] Kaggle, "Observing Dark Worlds," Kaggle, [Online]. Available: http://www.kaggle.com/c/DarkWorlds. [Accessed 18 11 2012].

[2] D. Harvey, “Obersving Dark Worlds: A Beginners Guide to Dark Matter & How to Find It”, Kaggle, 12 10 2012. [Online]. Available: http://blog.kaggle.com/2012/10/12/observing-dark-worlds-a-beginners-guide-to-dark-matter-how-to-find-it/

[3] Kaggle, “Observing Dark Worlds – Evaluation Criteria”, 12 10 2012 [Online]. Available: https://www.kaggle.com/c/DarkWorlds/details/evaluation

[4] Mathworks, 12 14 2012 [Online]. Available: http://www.mathworks.com/help/stats/treebagger.html