# Dark Matter Detection: Finding a Halo in a Haystack

Paul Covington, Dan Frank, Alex Ioannidis

## 1 Introduction

The predictive modeling competition platform Kaggle $^{\mathrm{TM}}$ recently posed the Observing Dark Worlds prize, challenging researchers to design a supervised learning algorthim that predicts the position of dark matter halos in images of the night sky. Although dark matter neither emits nor absorbs light, its mass bends the path of a light beam passing near it. Regions with high densities of dark matter–termed dark matter halos–exhibit this so-called 'gravitational lensing,' warping light from galaxies behind them. When a telescope captures an image of the sky, such galaxies appear elongated tangential to the dark matter halo in the image. That is, each galaxy appears stretched along a direction perpendicular to the radial vector from the dark matter to that galaxy. Aggregating this phenomenon across a sky filled with otherwise randomly oriented elliptical galaxies, one observes more galactic elongation (ellipticity) tangential to the dark matter than one would otherwise expect (Fig. 1 and 2). The challenge is to use this statistical bias in observed ellipticity to predict the position of the unseen dark matter halos (Fig. 1 and 2, left).

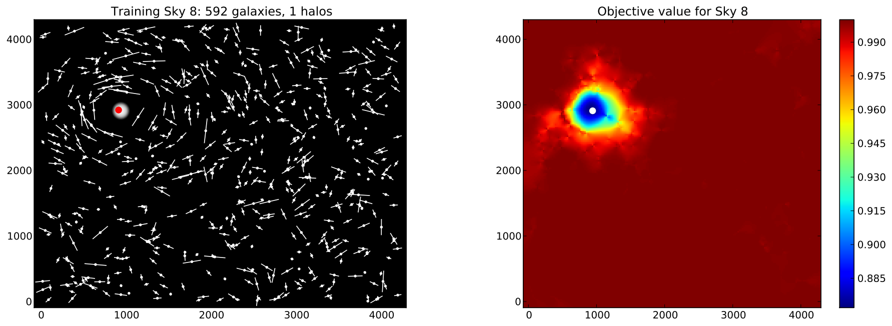

Figure 1: Left: A sky with one halo (white ball at upper left). Galactic ellipticities are notably skewed by the halo. Right: Plot of an objective function (described below) that we designed to have its minimum at the most likely location for the dark matter halo. The position of this minimum is marked by a red dot in the plot at left.

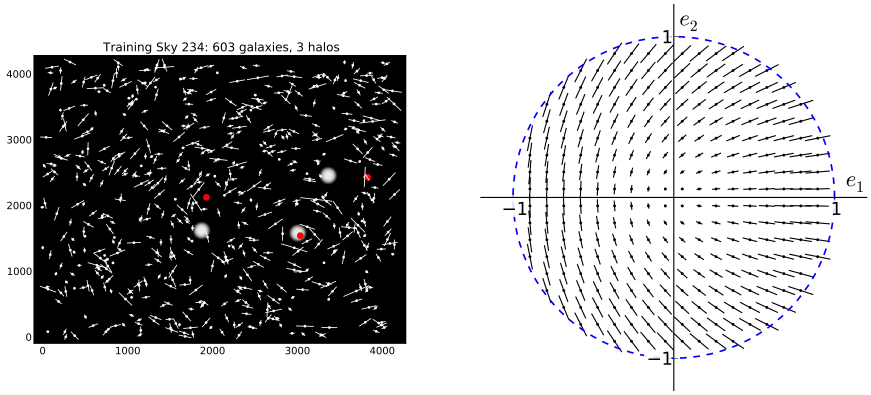

Figure 2: Left: A sky with three halos (white balls) with our method's predictions indicated by red dots. Due to interactions of halos with differing strengths on each galaxy it is no longer easy to discern the position of the halos. The now 6-d objective (2 dimensions in the position of each halo) cannot be visualized. Right: Ellipticity, the elongation and orientation of an ellipse, can be described by components $e_1$ and $e_2$ such that $\|(e_1, e_2)\|^2 \leq 1$.

# 2 Methods

## 2.1 Datasets and Competition Metric

The training data consists of 300 skies, each with between 300 and 720 galaxies. For each galaxy $i$, a coordinate $(x^{(i)}, y^{(i)})$ is given ranging from 0 to 4200 pixels along with an ellipticity $(e_1^{(i)}, e_2^{(i)})$. Here $e_1^{(i)}$ represents stretching along the $x$-axis (positive for elongation along x and negative for elongation along y), and $e_2^{(i)}$ represents elongation along the line $45\,^\circ$ to the x-axis (Fig. 2, right). Each sky also contains one to three halos, whose coordinates are given. Predicting these halo coordinates in test set skies is the challenge. In these test skies only galaxy coordinates, ellipticities, and the total number of halos is provided.

The algorithm's performance on a test set of skies is evaluated by Kaggle according to the formula $m = F/1000 + G$ where $F$ represents the average distance of a predicted halo from the actual halo and $G$ is an angular term that penalizes algorithms with a preferred location for the halo in the sky (positional bias).

## 2.2 Objective Function

Physically, the distortion induced by a halo on a galaxy's ellipticity should be related to the distance of closest approach to the halo of a light beam traveling from the galaxy to the observer. We explored the functional form of this radial influence on distortion by plotting distance of a galaxy from the halo on the horizontal and on the vertical either $e_{\text{tangential}} = -(e_1 \cos(2\phi) + e_2 \sin(2\phi))$ (the elongation along the tangential direction) or $e_{radial}$ (the complementary elongation along a line 45 degrees to the tangential direction) for each galaxy in all single halo training skies (Fig.3). Two functional forms were proposed to fit the dependence of tangential ellipticity on radius: $K_{\underline{\theta}}(r) = e_{\text{tangential}} \propto \exp(-(r/a)^2)$ (Gaussian) and the more general $K_{\underline{\theta}}(r) = e_{\text{tangential}} \propto \exp(-(r/a)^d)$ (learned exponential). The parameter vector, $\underline{\theta} = a$ or $\underline{\theta} = (a, d)$ respectively, is learned from the training skies (see Parameter Fitting section below).

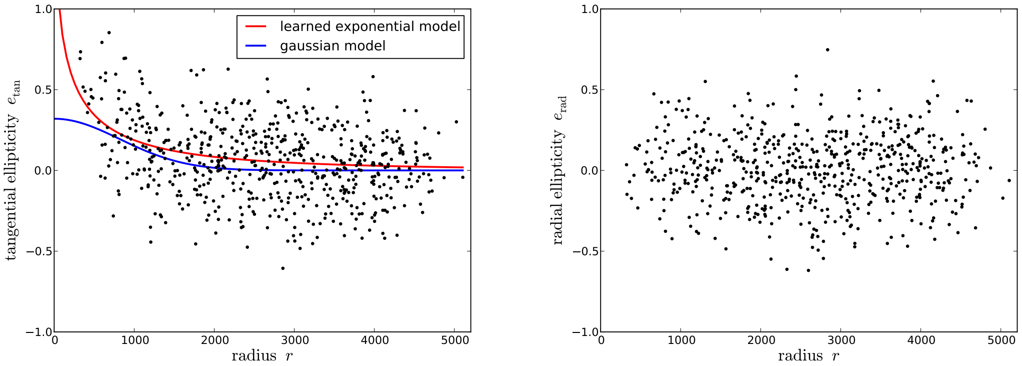

Figure 3: Left: Tangential ellipticity (left) for galaxies in a particular sky varies with distance from the dark matter halo, while radial ellipticity (right) does not.

For a one halo sky the ellipticity for a particular galaxy is then modeled as,

$$\begin{aligned} \hat{e}_1(x, y, \alpha) &= -\alpha K_\theta(r) \cos(2\phi) + \epsilon_1 \\ \hat{e}_2(x, y, \alpha) &= -\alpha K_\theta(r) \sin(2\phi) + \epsilon_2 \end{aligned}$$

where $\epsilon_1$ and $\epsilon_2$ represent the random components of the galaxy's ellipticity, and $\alpha$ is a parameter associated with each halo that represents its strength, and is determined from the observed galaxy data (see Optimization section below). For a multiple halo sky the influence of each of the halos on a galaxy's ellipticity is assumed to form a linear superposition. The predicted ellipticity for a galaxy $i$ is thus,

$$\hat{e}_1^{(i)}(\{(x,y)\},\{\alpha\}) = -\sum_{j=1}^{N_h} \alpha_j K_\theta(r_j^{(i)}) \cos(2\phi_j^{(i)}) + \epsilon_1^{(i)}$$
$$\hat{e}_2^{(i)}(\{(x,y)\},\{\alpha\}) = -\sum_{j=1}^{N_h} \alpha_j K_\theta(r_j^{(i)}) \sin(2\phi_j^{(i)}) + \epsilon_2^{(i)}$$

The objective function is formed by summing the squared errors between the model predicted ellipticity components ($\hat{e}_1^{(i)}$ and $\hat{e}_2^{(i)}$) and the observed ellipticity components ($e_1^{(i)}$ and $e_2^{(i)}$) over all galaxies in a given sky.

$$E(\{(x,y)\},\{\alpha\}) = \sum_i \left[ e_1^{(i)} + \sum_j \alpha_j K_\theta(r_j^{(i)}) \cos(2\phi_j^{(i)}) \right]^2 + \left[ e_2^{(i)} + \sum_j \alpha_j K_\theta(r_j^{(i)}) \sin(2\phi_j^{(i)}) \right]^2$$

Finally, predicted halo locations $\{(x^*,y^*)\}$ and strengths $\{\alpha^*\}$ are found by minimizing the objective for a particular sky,

$$(\{(x^*,y^*)\},\{\alpha^*\}) = \underset{(x,y,\alpha)}{\operatorname{argmin}} \; E(\{(x,y)\},\{\alpha\})$$

This objective is a squared error loss function for the deviation of the observed galaxy ellipticity components from their maximum likelihood predictions given the halo positions. Assuming that the deviations ($\epsilon_1$ and $\epsilon_2$) of each galaxy's ellipticity components from their model predictions follows a bivariate Gaussian with spherical covariance, the optimum of this squared error loss function objective also is the maximum likelihood estimate for the halo positions. Moreover, assuming a uniform prior probability for the halos' positions in the sky, this maximum likelihood estimator for the halo positions is also the Bayesian maximum *a posteriori* estimator for the halo positions. Thus, under these assumptions the highest probability positioning for the halos given the observed galaxy ellipticities is the argmin of this objective.

## 2.3 Optimization

The optimal halo strength parameters $\alpha_i$ for each galaxy can be found analytically, since the objective is quadratic in the $\alpha_i$. Differentiating the objective with respect to each $\alpha_i$ and setting equal to zero yields a linear system for the $\alpha_i^*$,

$$A\alpha^* = b \text{ with } A_{j,k} = \sum_i K_\theta(r_j^{(i)}) K_\theta(r_k^{(i)}) \cos(2(\phi_j^{(i)} - \phi_k^{(i)})) \text{ and } b_j = \sum_i K_\theta(r_j^{(i)}) e_{\text{tangential}}^{(i)}(x_j, y_j)$$

With the $\alpha_i^*$ determined explicitly, and assuming the model parameters are already fit (see Parameter Fitting section below), we need to optimize the objective only over the space of possible halo positions. Each halo may be positioned anywhere in the 2-dimensional sky image, so we have a 2-d, 4-d, or 6-d search space for the one, two, and three halo problems respectively. This optimization is performed using a downhill simplex algorithm, employing random starts to overcome local minima.

## 2.4 Parameter Fitting

The radial influence models $e_{\text{tangential}} \propto \exp(-(r/a)^2)$ (Gaussian) and $e_{\text{tangential}} \propto \exp(-(r/a)^d)$ (learned exponential) require fitting values to the parameter vector $\underline{\theta}$, where $\underline{\theta} = a$ or $\underline{\theta} = (a,d)$ respectively. Since we consider these models to be approximations to a universal physical law, we require these parameters be constant for all halos across all skies. Fixing $\underline{\theta}$ both prevents overfitting to the training data and allows us to arrive at a more accurate estimation of it. When fitting the more general learned exponential model these benefits are crucial, since, in constast to $a$, which just alters the scaling, the parameter $d$ alters the functional form of the model, introducing significant flexibility and implying very high dimensionality. The parametric estimation of $\underline{\theta}$ was performed similarly to the non-parametric estimation of halo positions previously described. Specifically, the sum of squared errors between modeled and observed galaxy ellipticities for all single-halo training skies was minimized with respect to $\underline{\theta}$.

# 3 Discussion

To determine the efficacy of our optimization algorithm, we compared the value of the objective at the true halo locations in the training data to the value returned by the optimization routine (see Fig. 4). For modest number of random starts (corresponding to a few minutes per sky on a desktop machine), the optimization algorithm plateaus consistently finding a minimum below that of the true solution. This diagnostic suggests the optimization is finding the objective function's global minumum; but this objective function minimum does not correspond to the true halo positions. We concluded that we should focus on improving our objective function rather than pursuing improvements to our optimization algorithm. We decided to improve our objective by refining our model of $K_{\underline{\theta}}(r)$ from the Gaussian form to the learned exponential form. Another question addresses the validity of our initial assumption of Gaussian distributed deviations ($\epsilon_1^{(i)}$ and $\epsilon_2^{(i)}$) for the galaxy ellipticity components ($e_1^{(i)}$ and $e_2^{(i)}$) from their predicted values ($\hat{e}_1^{(i)}$ and $\hat{e}_2^{(i)}$). This assumption was crucial, since it allowed us to use a squared error loss objective to find our maximum likelihood (and maximum *a posteriori*) estimators. The quadratic form of this objective further allowed us to solve explicitly for the halo strength parameters $\alpha_i^*$, reducing the dimensionality of our final optimization space. Unfortunately the assumption is clearly not true. The support of $\epsilon_1^{(i)}$ and $\epsilon_2^{(i)}$ is the unit cirlce (Fig. 2), so their deviations must come from a pdf with a similarly limited support not a Gaussian with infinite support. However this theoretical objection is not a severe problem in practice. Figure 5 shows that for moderate $e_{tangential}$ values (less than .4) the ellipticity pdfs are close to bivariate Gaussians with spherical covariance. Moreover, so little probability density is near the unit circle boundary that approximating the ellipticity deviations with a Gaussian pdf is tolerable. This self-consistently justifies our original use of a minimum squared error objective. For regions with predicted $e_{\text{tangential}}$ (plotted as $\tilde{e}_1$) larger that .4, approximating the ellipticity deviations with a bivariate Gaussian of spherical covariance becomes less tenable, see Fig. 5 bottom plot. Fortunately, very few galaxies lie so close to a dark matter halo as to be inside such a high predicted $e_{\text{tangential}}$ region. (One might incorporate the departure from spherical covariance for galaxies in high predicted

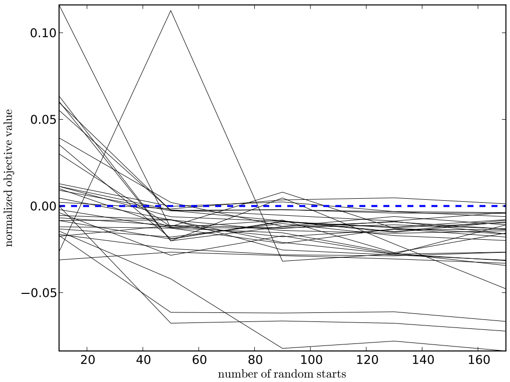

Figure 4: A line is plotted for each sky showing the minimal normalized objective value (given by $\frac{E(x^*,y^*)-E(x_{\text{true}},y_{\text{true}})}{E(x_{\text{true}},y_{\text{true}})}$) returned for increasing numbers of random starts. (The predicted values were derived with our Gaussian model.)

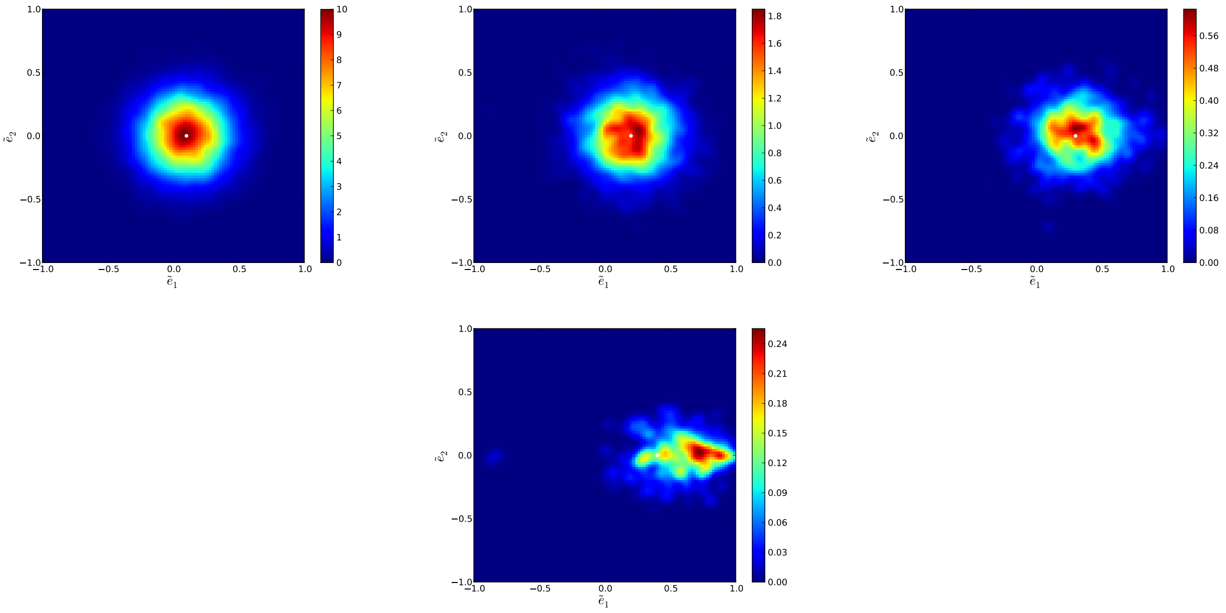

Figure 5: These four plots look across all training skies at regions of those skies where tangential ellipticity is predicted (by the learned exponential model) to be .1, .2, .3, and .4 respectively. The actual tangential ellipticity observed for each galaxy in such regions is labeled $\tilde{e}_1$, and its complementary ellipticity component is labeled $\tilde{e}_2$. By aggregating these ellipticity values across the many training galaxies found in each region, and using kernel density estimation, the corresponding ellipticity pdf can be found. These are the pdfs plotted above.

$e_{\text{tangential}}$ regions by adjusting the weights of the squared errors of $e_1$ and $e_2$ to vary with the magnitude of the predicted $e_{\text{tangential}}$ for the galaxy. This would still not alleviate the departures from even spherical Gaussian shape that occur for extreme $e_{\text{tangential}}$ values.)

# 4 Results

Our dark matter halo position prediction algorithm performed well on both the training set and test set and compared favorably with the other Kaggle ™ prize entrants, see Table 1 below. Here we see a comparison of the performance of our initial Gaussian model and our later more general learned exponential model to the performance of Kaggle ™ provided benchmarks. Against all benchmarks we showed superior results. Indeed, our algorithm outperformed even an astrophysical standard, the 50,000 line *Lenstool Maximum Likelihood* code.

| | 1 halo skies distance error | 2 halo skies distance error | 3 halo skies distance error | all skies distance error | Kaggle metric |
|---|---|---|---|---|---|
| gridded signal (kaggle) | 1645 | 1767 | 1483 | 1605 | 1.77 |
| maximum likelihood (kaggle) | 632 | - | - | - | - |
| gaussian model | 177 | 804 | 1007 | 801 | .955 |
| learned exponential | 133 | 704 | 934 | 723 | .856 |

Table 1: The distance metric gives the average distance of predicted halos from actual halos on the training set. The Kaggle metric was described in Methods (lower is better) and describes error on the test set.

Kaggle Ranking (team Skynet): $55^{th}$ out of 337 competitors with score of 0.97831