Introduction

Let Γ represent the knowledge of the table. Γ is composed of $\gamma_1, \ldots, \gamma_n$, corresponding to the objects on the table, and $\gamma_{\text{wall}_1}, \gamma_{\text{wall}_2}, \gamma_{\text{wall}_3}, \gamma_{\text{wall}_4}$, corresponding to the four walls (edges) of the table. Each γ contains position and dimension information about the object in question.

Let λ be the command the user gives, consisting of a distance λ_d a direction (relation) λ_r and a reference object λ_f. For the sake of testing this algorithm, we will assume perfect ability to parse the command and extract this information. Therefore we assume there is a perfect correspondence between λ_f and γ_{ref}, where $\gamma_{\text{ref}} \in \{\gamma_1, \ldots, \gamma_n\}$. In addition, we assume that the direction is in the set {left, right, in front, behind} and map λ_r to the corresponding direction vectors $\{[1,0], [-1,0], [0,1], [0,-1]\}$. Underlying this assumption is the larger assumption that all commands map to a single direction vector - discussion of how we plan to improve this can be found later in this paper. We assume that the distance λ_d of a command is independent of the direction λ_r. Finally, we only deal with two dimensions, so don't consider commands that involve the z-axis.

As an example, if the command is 'five inches to the right of the bowl', then we have $\lambda_d = 5$, $\lambda_r = [1,0]$, $\lambda_f =$ the bowl.

Our goal then is to estimate a function $\mathbb{P}(x, y|\lambda, \Gamma)$ which gives the probability that a user was referring to the point (x, y) given the command λ and world Γ.

From the command and reference, we calculate a naive mean, which is the point you would select if you went exactly the distance specified by the command, in the direction specified by the command. From this we calculate four features for a log-linear model.

Feature Calculation

Calculating $\hat{\mu}$

The naive mean, $\hat{\mu}$, is what is obtained by going exactly the distance specified in the command from the edge of the object. This is calculated as follows:

$$\hat{\mu} = \begin{cases} \gamma_{\text{ref}}.center + \frac{1}{2}\gamma_{\text{ref}}.height + \lambda_d, & \lambda_r = [0,-1] \\ \gamma_{\text{ref}}.center - \frac{1}{2}\gamma_{\text{ref}}.height - \lambda_d, & \lambda_r = [0,1] \\ \gamma_{\text{ref}}.center + \frac{1}{2}\gamma_{\text{ref}}.width + \lambda_d, & \lambda_r = [1,0] \\ \gamma_{\text{ref}}.center - \frac{1}{2}\gamma_{\text{ref}}.width - \lambda_d, & \lambda_r = [-1,0] \end{cases}$$

Calculating T_1 and T_2

T_1 and T_2 create a gaussian-like distribution around the naive mean $\hat{\mu}$. T_1 penalizes points based on squared distance from $\hat{\mu}$ *in the direction of the command* (e.g. if the direction is 'left', then T_1 penalizes points based on squared distance from $\hat{\mu}$ in the x-axis). In addition, T_1 penalizes less if the command is a longer distance (e.g. 1 foot) than if it is a shorter distance (e.g. 1 inch). To experimentally verify this assumption, we fit a line to the empirical Gaussian variances for all commands in experiment 1 (Figure 1). The weighting for this variance, however, was computed using the MLE method described below. T_2 penalizes points based on squared distance from $\hat{\mu}$ in the orthogonal direction (e.g. if the direction is 'left', then T_2 penalizes points based on squared distance from $\hat{\mu}$ in the y-axis).

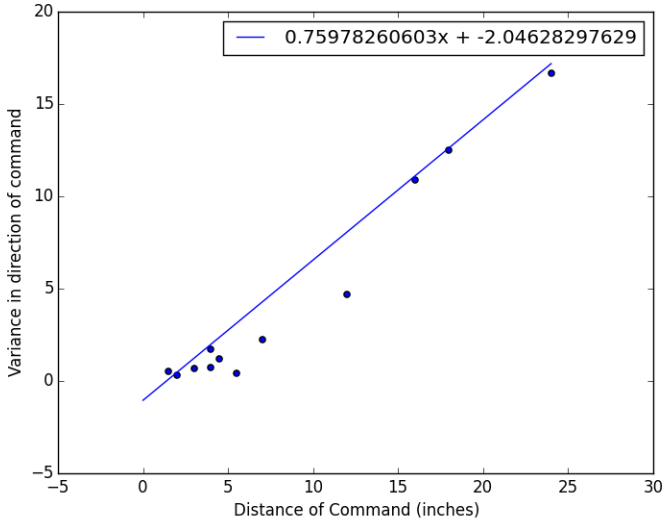


Figure 1: Plot of variance versus command distance

To arrive at these features, we make three assumptions about the data. First, that the data are distributed in a gaussian manner. Second, that the variance in the direction of the command (i.e. in the x direction for 'left' or 'right' and the y direction for 'in front' and 'behind') is independent of the variance in the orthogonal direction. Third, that variance in the direction of the command scales linearly with the distance of the command, while variance in the orthogonal direction is constant.

From this, our goal is to generate features that tell us about the probability of a point (x, y) given λ, Γ. Let $v = (x, y) - \hat{\mu}$. A gaussian version of this probability incorporating

the above assumptions would be:

$$\frac{1}{Z}\exp\left(\frac{\langle v,\lambda_r\rangle^2}{k_1\lambda_d}\right)\exp\left(\frac{[v-\langle v,\lambda_r\rangle\lambda_r]^2}{k_2}\right)$$

Turning these into features in a log-linear distribution, we then get:

$$T_1(x,y|\lambda,\Gamma)=\frac{1}{\lambda_d}\langle v,\lambda_r\rangle^2$$
$$T_2(x,y|\lambda,\Gamma)=[v-\langle v,\lambda_r\rangle\lambda_r]^2$$

with $v=(x,y)-\hat{\mu}$ as before.

Calculating T_3

T_3 is a hinge loss, designed to penalize points that are closer to an object other than the reference object (the one used in the command). The value should be zero for points where the reference object is the closest object, and increase linearly as (x,y) get closer to some other object. Let $p_{(x,y),\gamma_i}$ be the point on γ_i closest to (x,y). Then,

$$T_3(x,y|\lambda,\Gamma)=||(x,y)-p_{(x,y),\gamma_{\text{ref}}}||-\min_{\gamma_i\in\Gamma}||(x,y)-p_{(x,y),\gamma_i}||$$

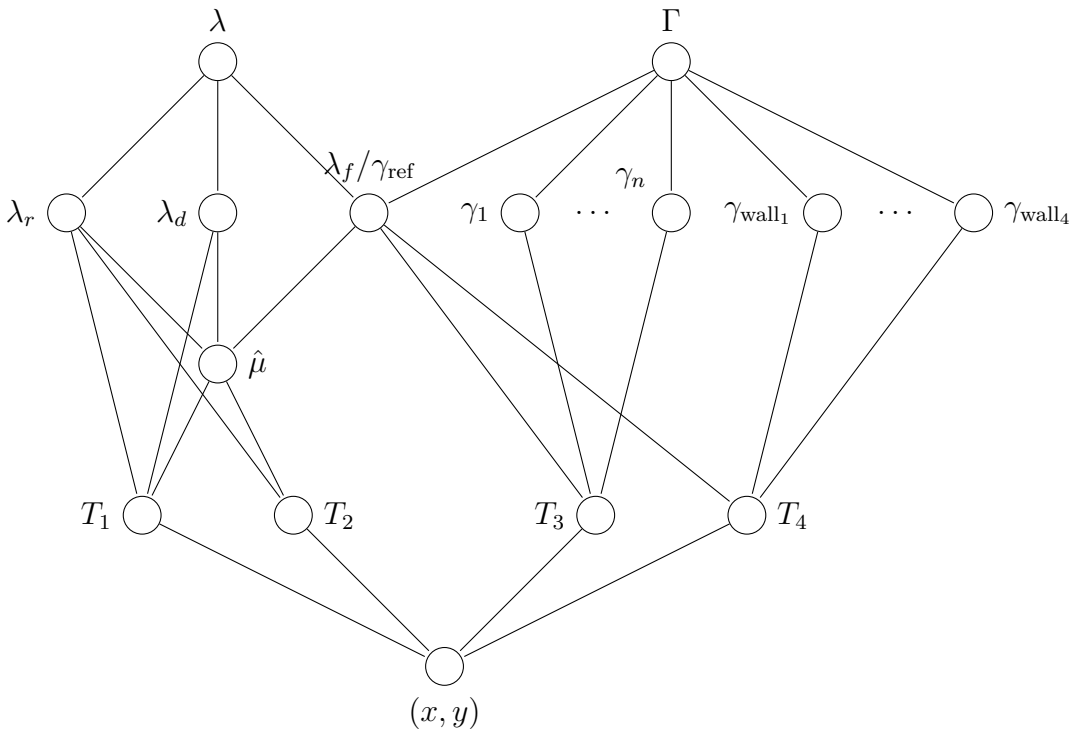
Calculating T_4

T_4 is a hinge loss, designed to penalize points that are closer to a wall than to the reference object. The value should be zero for points where the reference object is closer than any wall, and increase linearly as (x,y) get closer to some other object. Let $p_{(x,y),\gamma_i}$ be the point on γ_i closest to (x,y). Then,

$$T_4(x,y|\lambda,\Gamma)=\max\left(0,||(x,y)-p_{(x,y),\gamma_{\text{ref}}}||-\min_{\gamma_{\text{wall}_i}\in\Gamma}||(x,y)-p_{(x,y),\gamma_{\text{wall}_i}}||\right)$$

Putting the Model Together

All the features described above are related to each other via the following graphical model:



Given these features, we obtain the log-linear model

$$\mathbb{P}_w(x, y|\lambda, \Gamma) = \frac{1}{z_w(\lambda, \Gamma)} \exp\left(\sum_{c=1}^{4} w_c T_c(x, y|\lambda, \Gamma)\right)$$

where w is a vector of weights. We then learn the weights using MLE estimation.

MLE Estimation for Learning Weights

We trained the model using the data from scene #1. We assume each data point is an iid sample from $\mathbb{P}_w(x, y|\lambda, \Gamma)$. Let the 12 commands be specified by $\lambda^{(1)}, \ldots, \lambda^{(12)}$. For each command we collected 10 data points. Let X represent all the data and let $X_j^{(i)}$ be the j-th data point for the i-th command. Then the joint probability of the data points can be expressed as

$$\begin{aligned}
\mathbb{P}_w(X|\lambda^{(1)}, \ldots, \lambda^{(12)}, \Gamma) &= \prod_{i=1}^{12}\prod_{j=1}^{10} \frac{1}{z_w(\lambda^{(i)}, \Gamma)} \exp\left(\sum_{c=1}^{4} w_c T_c(X_j^{(i)}|\lambda^{(i)}, \Gamma)\right) \\
&= \left(\prod_{i=1}^{12}\prod_{j=1}^{10} \frac{1}{z_w(\lambda^{(i)}, \Gamma)}\right)\left(\prod_{i=1}^{12}\prod_{j=1}^{10} \exp\left(\sum_{c=1}^{4} w_c T_c(X_j^{(i)}|\lambda^{(i)}, \Gamma)\right)\right) \\
&= \left(\prod_{i=1}^{12}\prod_{j=1}^{10} \frac{1}{z_w(\lambda^{(i)}, \Gamma)}\right) \exp\left(\sum_{c=1}^{4} w_c \sum_{i=1}^{12}\sum_{j=1}^{10} T_c(X_j^{(i)}|\lambda^{(i)}, \Gamma)\right)
\end{aligned}$$

Our goal then is to find the argmax over w of this function. This is the same as finding the argmin of the negative log, so we have

$$\begin{aligned} w^* &= \underset{w}{\text{argmin}} - \log \mathbb{P}_w(X|\lambda^{(1)}, \ldots, \lambda^{(12)}, \Gamma) \\ &= \underset{w}{\text{argmin}} \sum_{i=1}^{12} \sum_{j=1}^{10} \log(z_w(\lambda^{(i)}, \Gamma)) - \sum_{c=1}^{4} w_c \sum_{i=1}^{12} \sum_{j=1}^{10} T_c(X_j^{(i)}|\lambda^{(i)}, \Gamma) \end{aligned}$$

This is the sum of exponential families, one for each i, j. Therefore the derivative with respect to w_c of the log partition function is the expected value of the c-th feature. Therefore,

$$\frac{\partial}{\partial w_c} - \log \mathbb{P}_w(X|\lambda^{(1)}, \ldots, \lambda^{(12)}, \Gamma) = \sum_{i=1}^{12} \sum_{j=1}^{10} \mathbb{E}_w[T_c(x, y|\lambda^{(i)}, \Gamma)] - \sum_{i=1}^{12} \sum_{j=1}^{10} T_c(X_j^{(i)}|\lambda^{(i)}, \Gamma)$$

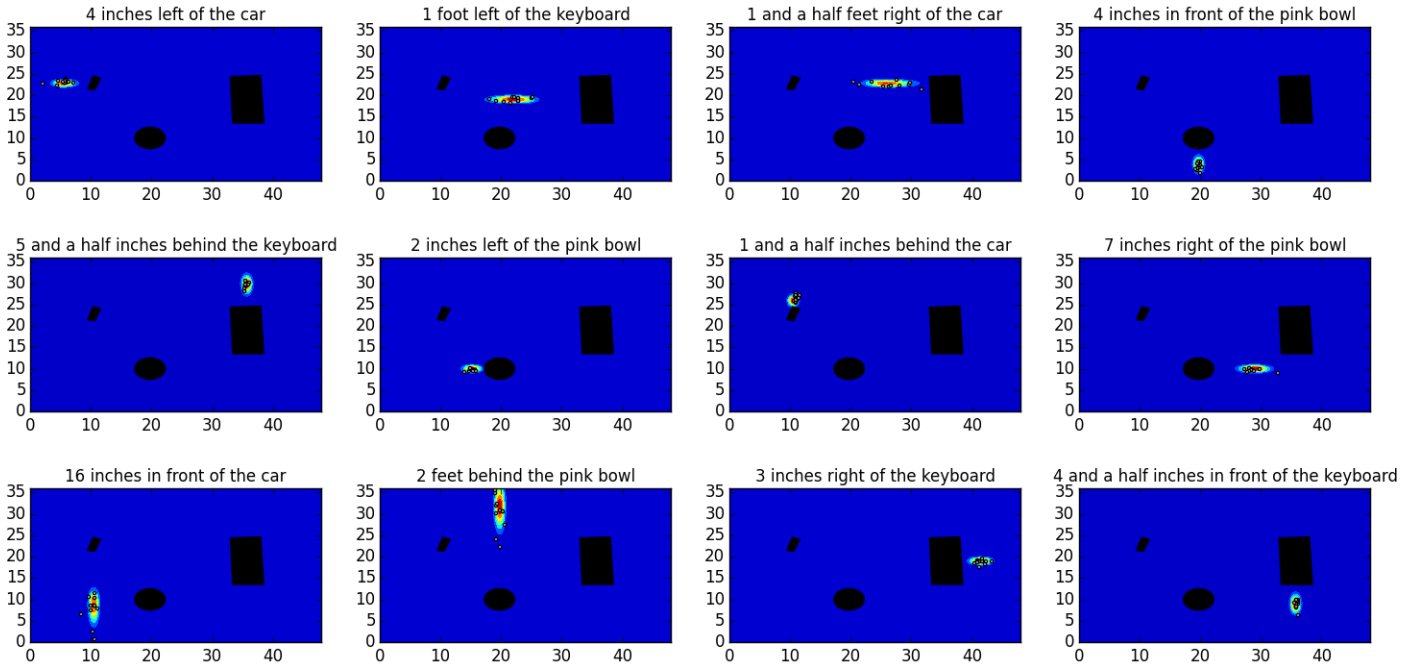
We use this to perform gradient descent. In order to calculate the probabilities and the partition function, we discretize the table with a grid of step size 0.1 inches. Once we do this, we have all the components necessary to calculate the gradient, and so we perform gradient descent until the gradient falls below a threshold level.

Modeling Results

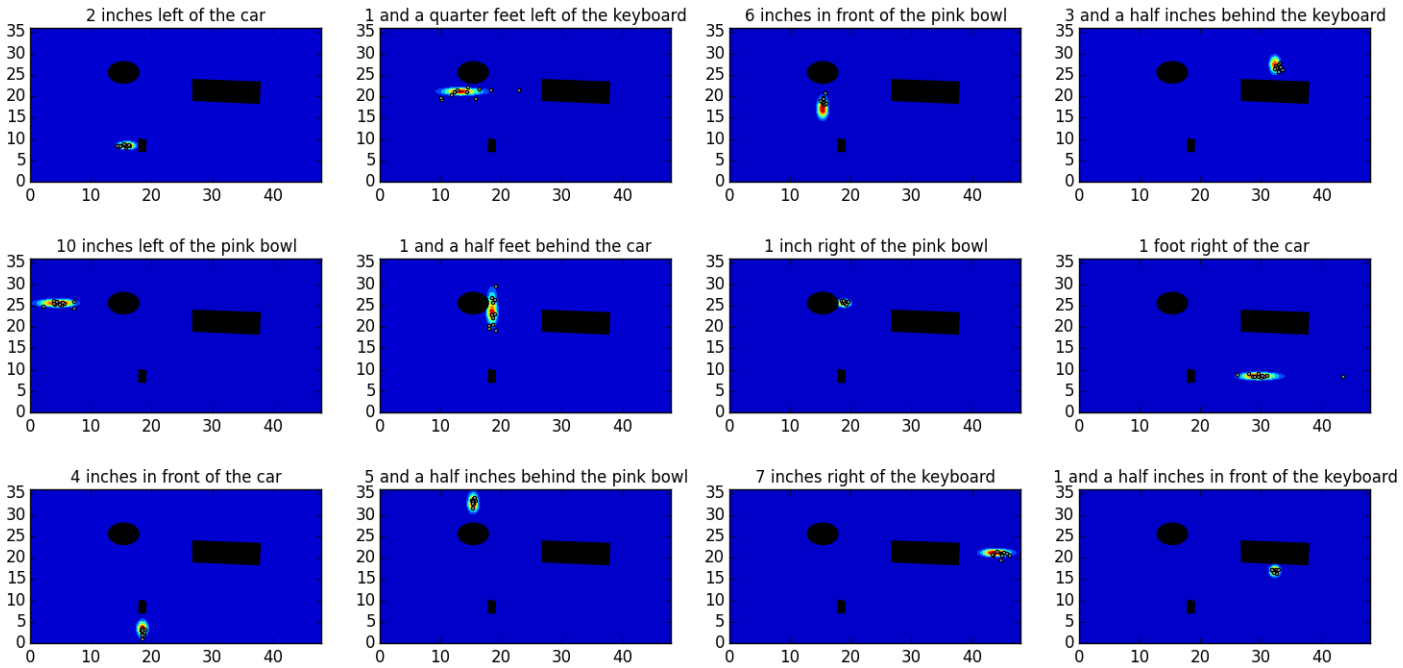
We have successfully implemented this model in MATLAB and Python, and learned values of w from the training data gathered in Experiment 1. Once trained using the MLE code implemented in MATLAB, our algorithm can parse a natural language command and, combined with information about the world Γ, produce a distribution over (x, y) coordinates. Plotted below are distributions generated by the log-linear model with all four features enabled:

Algorithm Evaluation

To evaluate the performance of our algorithm, we compared its performance to a baseline algorithm that fits a gaussian distribution around $\hat{\mu}$, the naive mean generated above. In addition, we compared it to log-linear algorithms trained and tested on subsets of our features, to determine the relative performance of different feature selections. All of these algorithms were compared to an empirical Gaussian distribution generated from the data in question - i.e. fitting a Gaussian to data for the specific λ, Γ pair. In the following evaluation metrics, "empirical mean" and "empirical standard deviation" refer to the mean and standard deviation of this distribution. This is distinct from the baseline algorithm, where a Gaussian distribution is generated for a λ, Γ using weights learned from the entire data set. We used three different performance metrics to evaluate our system: the distance between predicted and empirical means, and the combined



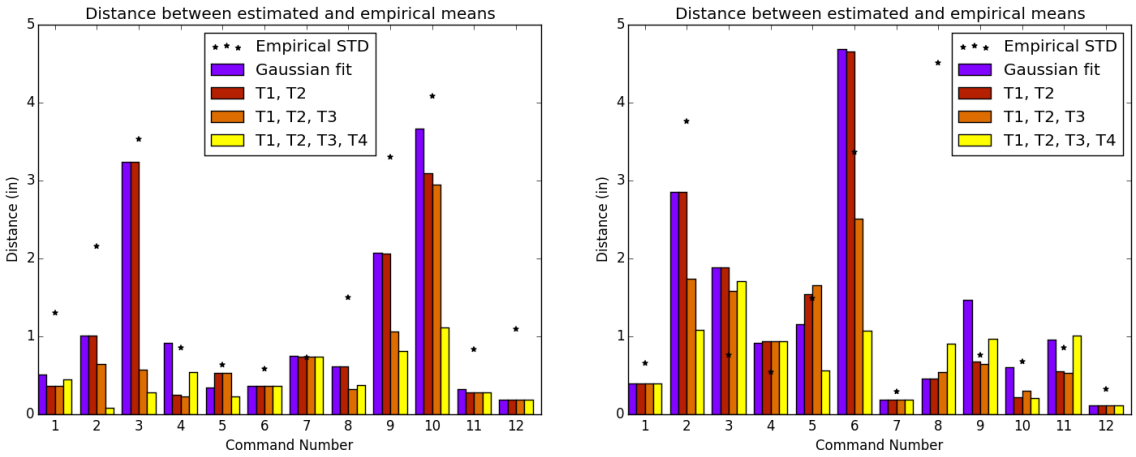
(a) Distributions on training data



(b) Distributions on test data

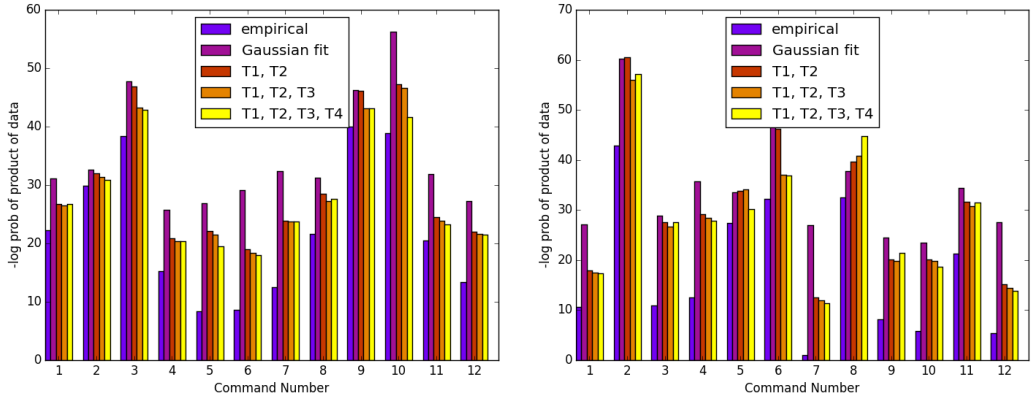
Figure 2: Distributions from natural language commands and world information. Higher probability regions are in red.

log-probability of the data given our distributions, mean-squared error (MSE) of the algorithms on our data sets. Plots of these metrics are displayed below, followed by a discussion of their meaning.



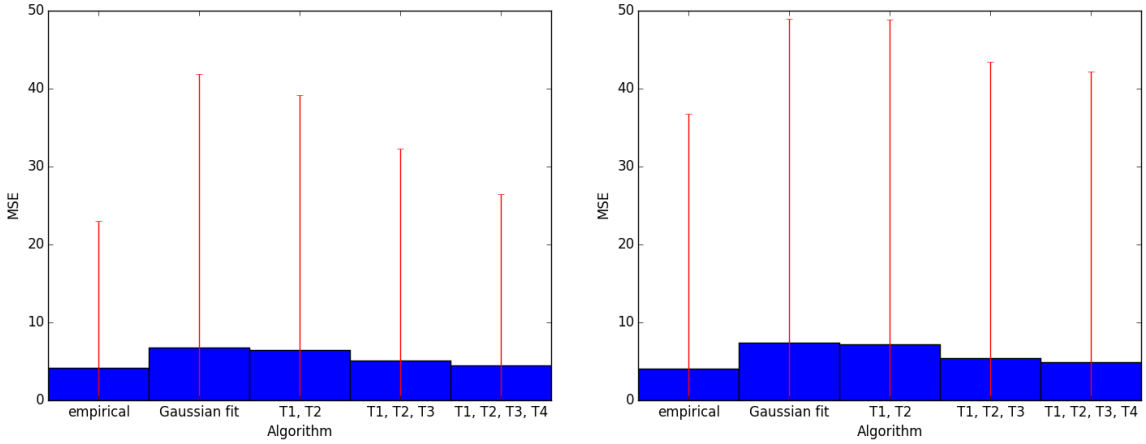
(a) Training data (b) Test data

Figure 3: Distance between empirical and predicted mean for all NL phrases.



(a) Training data (b) Test data

Figure 4: Negative log probability of algorithms for all NL phrases



(a) MSE across all training data

(b) MSE across all test data

Figure 5: Mean-squared error by feature set. Confidence interval is twice the standard deviation of the squared error.

Discussion of Results

These evaluation metrics lead to two primary conclusions about the performance of the algorithm and possible avenues for future research. The first, and most obvious, conclusion can be derived from the command-wise plots: the performance of different feature sets, as well as the relative performance between algorithms using different feature sets, varies across commands. This is entirely logical and consistent with the algorithms' design - in cases where objects and walls in Γ have minimal effect, the algorithms including T3 and T4 hinge losses would be expected to perform similarly to the gaussian fit and algorithms only using the first two features. As well, in commands with short distances, the distance term in feature T1 has little effect on the variance of the distributions. However, in commands with larger distances, the non-Gaussian log-linear models have a superior performance. Overall, the larger feature sets maximize log-probability of the data, although the distance between the empirical and predicted mean varies across the commands. For nearly every command, the predicted mean is within one standard deviation of the empirical mean. A notable exception is command 4 in the test data. In that case, participants expressed confusion about reference points on the object, due to duct tape used to secure the object to the whiteboard.

The second conclusion derives from the MSE evaluation. As shown in Figure 5, increasing the size of the feature sets lowers the mean-squared error across all distributions. However, the confidence interval with p=0.95 shows that the MSE of every feature set is statistically indistinguishable. Given that we only have 10-12 data points per λ, Γ pair, a large confidence interval is to be expected. In particular, the magnitude of the standard deviation is similar to that of the empirical distribution, demonstrating that

the algorithms perform similarly in that regard. Even so, the small data sets do not allow a definitive conclusion on the relative performance of each feature set. While the complete feature set does look promising in regards to log-probability, distance between predicted and empirical means, and MSE, we cannot definitively say that the performance of the log-linear model with all features is superior to Gaussian or log-linear models with smaller feature sets. Discussion of how we plan to resolve this issue, and expand the capabilities of our model, is included below.

Future Research

The limitations of our current model suggest multiple avenues for future research. The first is dramatically increasing the size of our data set, both in terms of λ, Γ pairs and number of data points gathered per pair. This will reduce statistical variation in the training and test sets, and provide a larger set of training data so more accurate weight vectors can be gathered. With a sufficiently large data set, we could use cross-validation to train our model instead of merely using a training and test set. A larger data set will also provide insight into other factors affecting human behavior, allowing us to add more features or parameters to the model if necessary. We plan to investigate the use of Amazon Mechanical Turk to rapidly gather large data sets, if the tradeoffs (especially the lack of spatial grounding in relation to performing in-person experiments) are not too great.

One such factor is the use of multiple reference frames and reference points on objects with off-axis rotations. Humans possess multiple reference frames in which they can refer to objects, such as the relative frame, which contains directions relative to the speaker's orientation, and the intrinsic frame, where directions are defined based on inherent properties of the object. (Levinson 2001) For instance, a car might be placed in any position relative to the speaker, but the speaker has some knowledge of what the "front" of the car should be regardless of its orientation. As well, we observed in experiments that objects with off-axis rotations cause human subjects to vary which reference point on they object they use, even if they operate within the relative reference frame. The data points for these objects appears to separate into distinct clusters, suggesting that humans make a decision between discrete reference points and reference frames before choosing a point. We plan to use a multi-peaked log-linear distribution to more accurately model this situation, by modifying features T1 and T2 to accept arbitrary direction vectors and reference points. With a knowledge of the distribution of reference points ρ and direction vectors $\vec{v}$, we can then marginalize over the joint probability $\mathbb{P}(x, y|\lambda, \Gamma, \rho, \vec{v})\mathbb{P}(\rho, \vec{v}|\lambda, \Gamma)$ to produce a multi-peaked distribution.

$$\sum_{\rho,\vec{v}} \mathbb{P}(x, y|\lambda, \Gamma, \rho, \vec{v})\mathbb{P}(\rho, \vec{v}|\lambda, \Gamma) = \mathbb{P}(x, y|\lambda, \Gamma)$$

However, we do not possess enough data about the relationships between objects, ref-

erence frames, and reference points to accurately model the distribution $\mathbb{P}(\rho, \vec{v}|\lambda, \Gamma)$. A large part of the data collection described above will be gathering enough data to accurately model this relationship.

We also plan to pursue a better model of the relationship between space and language by developing more advanced parsing capabilities. To simplify our parser, we restricted the language used to a standardized format when conducting our experiments. However, humans can use a variety of natural language phrases when describing points in space. We plan to perform "inverted" versions of our experiments, where humans are given a point in space and asked to describe it using natural language. From there, in combination with linguistic and psychological research on this subject, we can begin to develop a parser that more accurately captures a mapping between natural language and spatial relationships.

Conclusion

In this project we developed a log-linear model, which relates natural language commands to distributions of points in 2D space. We created an experimental procedure to determine how humans interpret these natural language commands. Our model takes in a subset of our data, and via maximum likelihood estimation, finds the feature weights. The model obtains promising results in terms of MSE and log probability of the data, but ultimately we have too little data to be conclusive either way. There are some phenomena that the model does not support, such as rotated objects. We hope to address this using a multi-peaked model built from our current one. In addition, our limited amount of data is a major constraining factor. We hope to address this either by performing more experiments or by moving to a platform such as Mechanical Turk.