0.1 Quantitative Results

Overall, we see that our models are able to learn to estimate colors based on sequences of terms. The CNN and SOWE models perform almost as well as the direct methods. With the SOWE taking the lead for distribution estimation, and the CNN for point estimation. We believe the reason for this is the SOWE is an easier to learn model from a gradient descent perspective: is very shallow model with only one true hidden layer. The RNN did not perform well at this task, it may be hard to capture the significant relationship between terms in the sequence. However, as will be shown in the examples in the next section it did learn generally acceptable colors.

The performance of SOWE on the order tasks (Table 2 and Table 5) is surprising. SOWE can not take into account word order, its good results suggest that the word-order is not very significant for color names. While word order matters, colors with the same terms in there name but in different order are similar enough that it still performs well.

It can be seen that smoothing has very little effect on the performance of any of the neural network based distribution estimation models. All 3 term based models (SOWE, CNN, RNN) all perform very similarly whether or note the training data is smoothed. The direct method which bypasses learning the functioning of terms is very substantially effected by the smoothing. This is because without smoothing it results in eastimating the probability based on bins unfilled by any observation (Which is zero due to the cap on the minimum value. See **??**). This is particularly notable in the case of the direct, unsmoothed nonextrapolating result reported in Table 3. As these were some of the rarest terms in the training set, they were thus more likely to not coincide with any terms from the testing set. Conversely, on this dataset the term models do quiet well, with or without smoothing. The SOWE model out out even performs the the Direct method. As the network can effectively learn the smoothness, not just from the observations of one color but from all of the observations. It learns that increasing the value of one bin should increase adjacent values. As such it does not need the smoothing applied to the training data.

We see both for point estimation (**??**) and for distribution estimation (Table 3), when the network is forced to extrapolate to new combinations of color names, the SOWE and CNN can do so with a reasonable degree of correctness. The RNN results continue to perform badly when extrapolating, doing much worse than the already poor results on the non-extrapolating task. However, the CNN and SOWE do well; while as expected the results are worse than for non-extrapolating they are not much worse. For the Distribution extrapolation task, they are actually still below the overall full datasets best result for perplexity.

In the point estimation results we find that using the networks trained specifically for point estimation (As described by **??**) performs consistently

Method	$\frac{PP}{256^3}$
Direct	0.164
Direct-smoothed	0.071
SOWE	**0.075**
SOWE-smoothed	**0.075**
CNN	0.078
CNN-smoothed	0.079
RNN	0.089
RNN-smoothed	0.088

Table 1: The results for the **full distribution estimation task**. Lower perplexity (PP) is better.

Method	$\frac{PP}{256^3}$
Direct	0.244
Direct-smoothed	0.053
SOWE	**0.055**
SOWE-smoothed	**0.055**
CNN	0.057
CNN-smoothed	0.058
RNN	0.124
RNN-smoothed	0.122

Table 2: The results for the **order distribution estimation task**. Lower perplexity (PP) is better. This is a subset of the full test set containing only tests where the order of the words matters.

better than taking the weighted mean of the distributions output by the distribution estimation methods. We attribute this primarily to the effective additional training data. The distribution estimation method combines all training observation points in to one training datum per color description. Where as the point estimation uses the points directly..We note that for the direct method, the distributions mean is almost identical to the true mean of points, as is expected.

0.2 Point Estimation

0.2.1 Full Task

0.2.2 Order Task

Method	Nonextrapolating $\frac{PP}{256^3}$	Extrapolating $\frac{PP}{256^3}$
Direct	175.883	–
Direct-smoothed	0.050	–
SOWE	**0.050**	**0.055**
SOWE-smoothed	**0.050**	0.056
CNN	0.052	0.065
CNN-smoothed	0.053	0.063
RNN	0.117	0.182
RNN-smoothed	0.112	0.183

Table 3: The results for the **extrapolation distribution estimation task**. Lower perplexity (PP) is better. This uses the extrapolation subset of the full test set. In the extrapolating results certain rare word combinations were removed from the training and development sets. In the non-extrapolating results the full training and development stet was used.

Method	MSE
Direct	0.066
SOWE	**0.067**
CNN	**0.067**
RNN	0.071
Distribution Mean Direct	0.066
Distribution Mean Direct-smoothed	0.066
Distribution Mean SOWE	0.068
Distribution Mean SOWE-smoothed	0.068
Distribution Mean CNN	0.069
Distribution Mean CNN-smoothed	0.069
Distribution Mean RNN	0.077
Distribution Mean RNN-smoothed	0.077

Table 4: The results for the **full point estimation task**. Lower mean squared error (MSE) is better.

Method	MSE
Direct	0.065
SOWE	**0.066**
CNN	**0.066**
RNN	0.096
Distribution Mean Direct	0.065
Distribution Mean Direct-smoothed	0.065
Distribution Mean SOWE	**0.066**
Distribution Mean SOWE-smoothed	**0.066**
Distribution Mean CNN	**0.066**
Distribution Mean CNN-smoothed	**0.066**
Distribution Mean RNN	0.095
Distribution Mean RNN-smoothed	0.097

Table 5: The results for the **order point estimation task**. Lower mean squared error (MSE) is better. This is a subset of the full test set containing only tests where the order of the words matters.