1 Major Changes

There have not been any major changes to the project since the proposal. The only change has been more detailed project milestones. In the original proposal, my milestones were monthly, but after understanding the project better, I have updated the milestones to bi-weekly as required.

2 Accomplishments So Far

I have made two main accomplishments so far: background reading and an initial experiment. The background reading consisted of a collection of research papers. The first paper I read [1] described the basic matrix factorization method that does not account for noise in the ratings data. After understanding basic matrix factorization, I read several papers [2] [3] [4] on robust matrix factorization, which describe more complex methods that assume the observed data has been corrupted by noise.

The initial experiment is a simple experiment to motivate the research project by showing that conventional matrix factorization fails on noisy data. The experiment is as follows:

We generated synthetic data that follows the non-biased matrix factorization model. There are $m = 10000$ users, $n = 1000$ items, and $k = 30$ factors. The user-factor matrix $U$ is a $m \times k$ matrix, and the item-factor matrix $V$ is a $n \times k$ matrix; both are randomly generated from a uniform $[0, 1)$ distribution. We then generated the ratings matrix from these two factors: $R = UV^T$. The rating values approximate a Gaussian distribution with $\mu \approx 7.5$ and $\sigma \approx 1.2$.

We split the ratings data into a training set and test set with an 80/20 split. We evaluated the model after training on the original training data, and then we introduced various amounts of noise into the training data and re-evaluated the model. All noisy trails consisted of a uniformly random value being added to a random 10% of the training data. The only variation across trials was the range of the uniform distribution. The results are shown in the table below:

<table>
<thead>
<tr>
<th>Noise range</th>
<th>Test RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.4454</td>
</tr>
<tr>
<td>[-3, 3)</td>
<td>0.4498</td>
</tr>
<tr>
<td>[-5, 5)</td>
<td>0.6004</td>
</tr>
<tr>
<td>[-7.5, 7.5)</td>
<td>0.9099</td>
</tr>
</tbody>
</table>

Table 1: Test RMSE of Matrix Factorization model with various amounts of noise added to training data.

We can see that adding noise in the range $[-3, 3)$ hardly affects the model, but as soon as we increase the range to $[-5, 5)$, we notice a significant hit to performance. Increasing the range to $[-7.5, 7.5)$ delivers an even greater hit to the model’s performance.
This initial experiment shows that even when the data follows our model, an adversary that corrupts a small percentage of the data can dramatically affect the model. The next step is to show this effect on real-world data, and to begin devising a method that is robust to such corruptions.

3 Meeting First Milestone

My initial milestone as described in my proposal was to conduct background reading on robust matrix factorization to gain a better understanding of the problem, and to conduct a simple experiment that shows conventional matrix factorization failing under noise, motivating the research project. I have achieved these two tasks, and so I have met my milestone.

4 Surprises

There have been no surprises to report so far.

5 Revisions to 15-400 Milestones

As mentioned above, when I had initially wrote the proposal, I had made my milestones monthly. Several weeks later, having a stronger grasp of the project, I was able to modify these milestones to bi-weekly. The general timeline remains the same; it just became more specific.

6 Resources Needed

The only resource I will potentially need is GPU compute if I decide to explore deep learning in this project, which I already have access to.

References