Model A

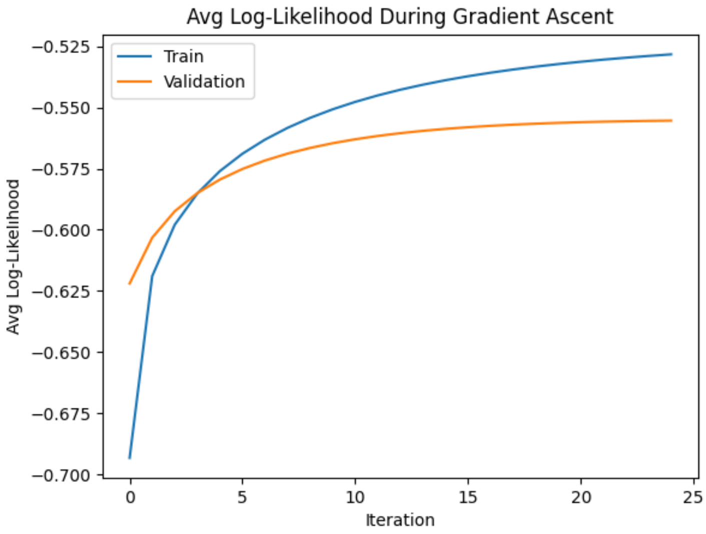
We have that: $logp(C|\theta, \beta) = log(\prod_{i,j}(I(C_{ij} = 1)(exp(\theta_i - \beta_j))/(1 + exp(\theta_i - \beta_j))$
$+ I(C_{ij} = 0)(1 - exp(\theta_i - \beta_j)/(1 + exp(\theta_i - \beta_j)))))$

$= \sum_{i,j}(I(C_{ij} = 1)log(exp(\theta_i - \beta_j)/(1 + exp(\theta_i - \beta_j)))$
$+ I(C_{ij} = 0)log(1 - exp(\theta_i - \beta_j)/(1 + exp(\theta_i - \beta_j))))$

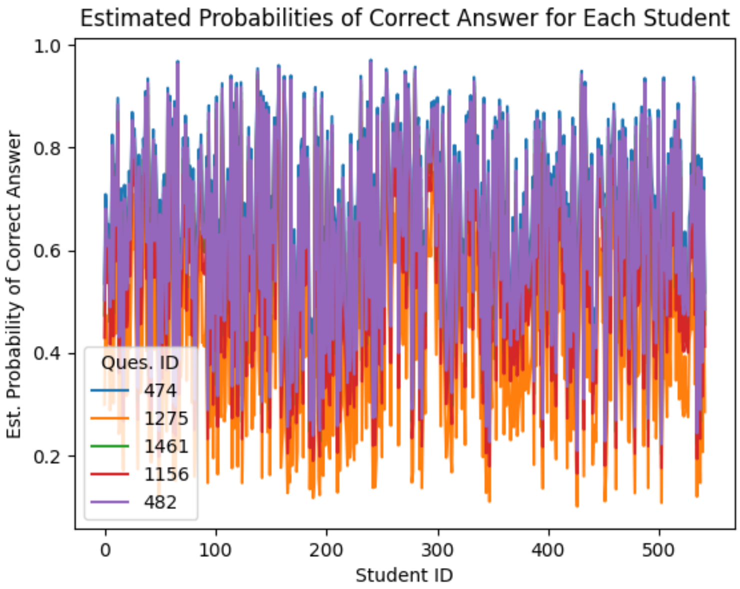
$= \sum_{i,j}(I(C_{ij} = 1)((\theta_i - \beta_j) - log(1 + exp(\theta_i - \beta_j)))$
$+ I(C_{ij} = 0)log(1 - exp(\theta_i - \beta_j)/(1 + exp(\theta_i - \beta_j))))$

Then, $\frac{\partial logp(C|\theta,\beta)}{\partial\theta_i} = \sum_j(I(C_{ij} = 1)(exp(\beta_j)/(exp(\theta_i) + exp(\beta_j)))$
$+ I(C_{ij} = 0)(- exp(\theta_i)/(exp(\theta_i) + exp(\beta_j))))$ and

$\frac{\partial logp(C|\theta,\beta)}{\partial\beta_j} = \sum_i(I(C_{ij} = 1)(- exp(\beta_j)/(exp(\beta_j) + exp(\theta_i)))$
$+ I(C_{ij} = 0)(exp(\theta_i)/(exp(\theta_i) + exp(\beta_j))))$



Learning Rate: 0.01
Number of iterations: 25
Final Validation Accuracy: 0.7067456957380751
Final Test Accuracy: 0.7056167090036692



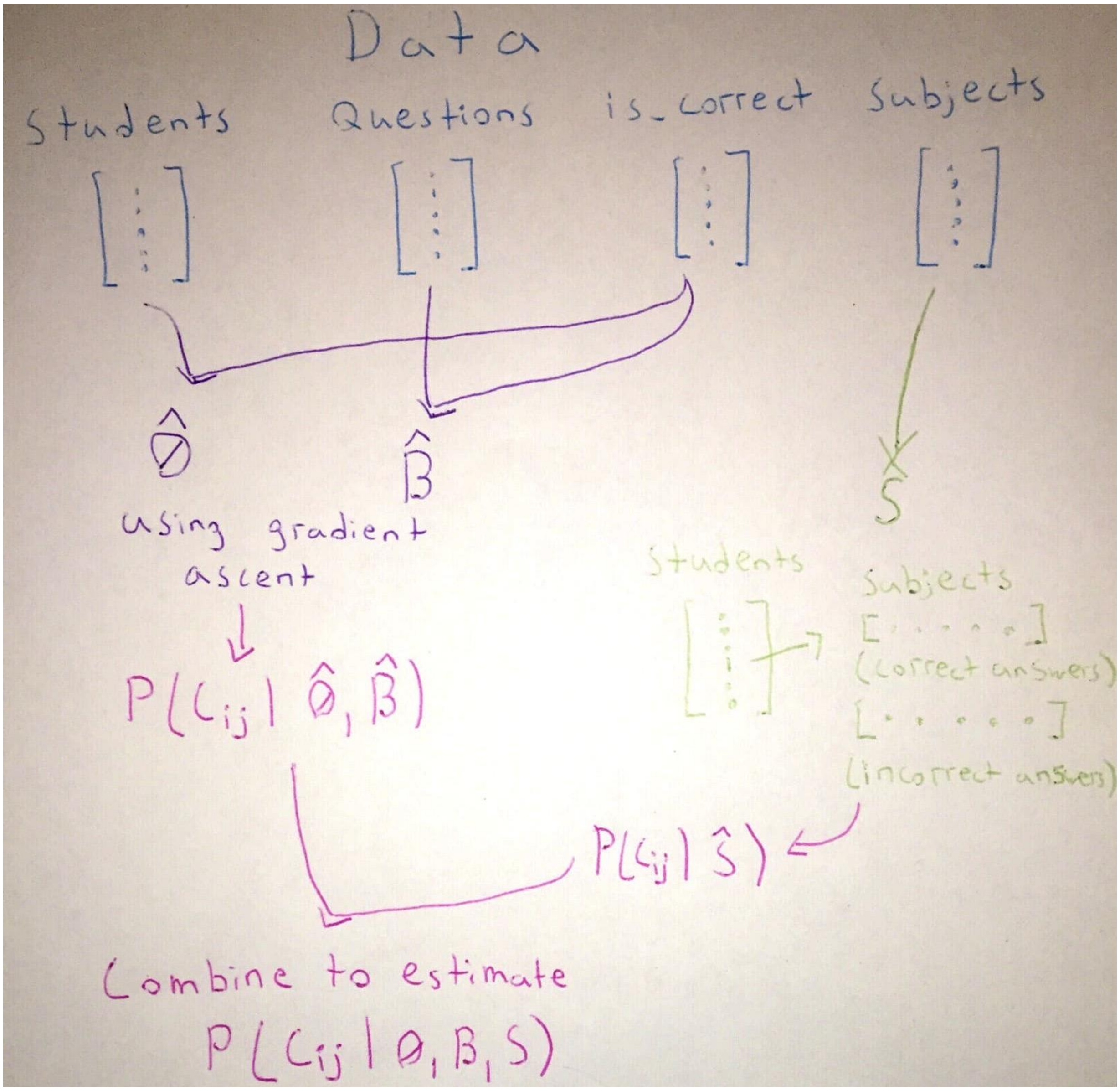
We can see that each of the curves follows the same pattern in that they seem to oscillate up and down in the same places. This is to be expected because if a certain student's perceived ability is high, the model will predict a higher probability that the student will get a question right than it will for a student with a lower perceived ability, regardless of the question. This explains why the peaks and dips occur in the same places, but why each curve may be shifted up or down depending on the difficulty of the question.

Model B

Utilizing item response theory in Model A, we attempted to model the probability that a given student correctly answered some particular question using an estimate of the student's ability and an estimate of the question's difficulty based on the proportion of questions the student answered correctly and the proportion of students who answered that question correctly, respectively. However, this seemed overly simplistic in that the proficiency of students usually varies across subject area. In fact, it is often the case that there are certain areas one is proficient in and others where one has trouble. For this reason, we assume the item response theory model is underfitting by not taking into account enough of the available data.

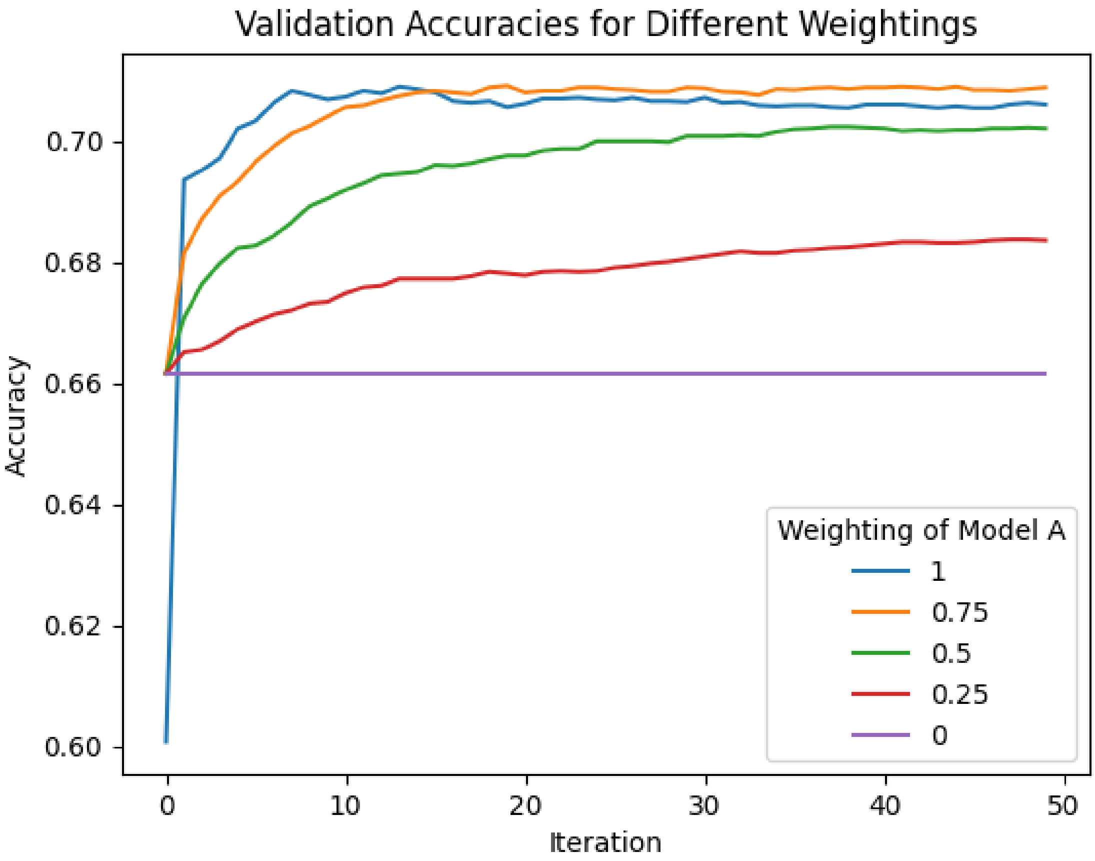
We use the metadata in question_meta.csv to model $P(C_{ij} = 1 \mid \theta, \beta, S)$ for a particular student i and question j, where C is the sparse matrix and S is a parameter that denotes the proficiency of each student within each subject. We expect the bias to decrease in this latter model by taking into account the parameter S.

We do not need to convert to a generative model because we can already calculate $P(C_{ij} = 1 \mid \theta, \beta, S)$ without needing to explicitly find $P(S \mid \theta, \beta, C_{ij} = 1)$. So for the new model, we take the question_meta.csv file and turn it into a dictionary with question IDs as keys and tuples of the subject IDs as values. We then use this in conjunction with the training data to create a new dictionary where the keys are student IDs and the values are arrays that keep track of how many questions the corresponding student answered correctly and incorrectly for each subject. We train the θ and β parameters the same way we did in Part A, but now we have a new parameter S which we can use to aid in our predictions. Given student i and a new question j we are trying to determine C_{ij} for, we average their answer accuracy for each subject that question belongs to, which gives us a new prediction. We will try combining the prediction of our old Model A with the prediction given by Model S for different weights to see if we can arrive at a new Model B with improved accuracy.



Comparison/Demonstration

We see from the figure that Model A outperforms the mixed model for every set of weights tested, except when the weights on Model A and Model S are 0.75 and 0.25, respectively. This model, which we will adopt as Model B, has a higher accuracy than Model A past 16 iterations of gradient ascent. The final accuracy of Model B after 50 iterations is 0.7088625458650861 which is marginally greater than the accuracy of 0.7060400790290714 of Model A after 50 iterations.



By testing out different weights, we can begin to see just what kind of effect Model S has on Model B as the weighting of the former increases. This way, we are able to assess if our hypothesis that Model B would decrease underfitting was correct.

We see that when the weighting on Model S is 1, the accuracy is constant with respect to the number of iterations. This is to be expected because in this case, the parameters $\theta\ and\ \beta$ are not used, so the gradient ascent has no effect on the model predictions. Our experiment tells us that we were wrong in our assumptions that Model B would have an increased accuracy and decrease overfitting because it seems as though the less we take into account Model S, the greater our accuracy until we reach our Model B, at which point the difference is negligible. Thus, we cannot conclude that Model B is a better model than Model A.

Limitations

Our modified model failed to improve upon our original one, as we couldn't find a set of weights that had more than a negligible effect on improving accuracy. Thus, this was a setting in which all our existing models failed to improve upon our first one. Our results from the previous section suggest that a question's subject matter is relatively not relevant to predicting whether a student will answer it correctly. This was initially surprising, but we must consider the subject categories more closely. All the data came from Eedi, who only offer mathematical diagnostic questions. It seems fairly reasonable that mathematical ability can be well-generalized across mathematical sub-fields in that one would expect a student proficient with fractions to also be proficient at algebra, especially more so than a student proficient at fractions to also be proficient with reading comprehension. Thus, it should be expected that proficiency in any particular subject would be highly correlated with a measure of a student's proficiency in general, θ and the addition of parameter estimate $\hat{S}$ would not yield any useful information given that we already have an estimate for θ. Even if this were not the case, we notice that for many of the students, our $\hat{S}$ stored a fairly sparse array representing the proficiencies of the student. This means we have many subject categories, but perhaps not enough questions answered by students in each category to make the parameter estimate useful. This is why we combine the models rather than solely use Model S because we arrive at the problem of our Model S predictions being far too variable and sensitive to the effect of individual questions due to data sparsity and the large spread of possible subjects. Thus, one way we could address this problem is simply to collect more data. Additionally, if in the future, we were given the same task, but with questions across the range of academic disciplines, perhaps our method would be more useful.