Journal of Ryan Botts for the Fall of 2007

Ryan Botts

1 9/3/07-9/7/07: Journals, Autobiographies, and EMACS

This week was productive in gaining more familiarity with the project, and getting Caity started in LaTeX. I read through all of the other journals from last year to get a feel for which other projects have been worked on and to understand different aspects of them. I read several of the autobiographies to understand what they should include. I then wrote my own autobiography. It contains much of my mathematical history and perhaps more detail than necessary, however, I think all of these things are important in understanding how people get to where they are.

After the autobiography was finished, I proofed my Journal from the summer and completed a final copy which is ready to be published. It was actually very helpful to look through old work and understand how I got to my current understanding of the question. It also reminded me of many details I had not thought about in some time.

Up to this point I have been editing code for the cluster expansions project on my Mac's text editor and uploading it to the server and running it from there. I decided that I should force myself to use XEMACS. So far it seems to provide many features which will be helpful for debugging, however, I do not enjoy that all of its functions are hidden. After taking the learning styles test for the autobiography this makes sense because I am almost entirely a visual learner, so remembering commands which don't seem to have any pattern, would be difficult. I hope that I will become more fluent in this shortly.

For the upcoming week I am going to work on the bcc.py code. Currently it does not function entirely, it can read in from the test files and provides some output, however it throws some exceptions when it begins to process the bcc data. I have not been able to sort this out, but hope that I can soon.

Once this is done I will add new code to compute inner products using the definitions we have.

2 9/8/07-9/14/07: The Joys of Object Oriented Design

This week I worked on modifying the old methods to account for the group symmetries of each data point. I figured out however, that it would be better to make an object which would extend the BccStructure, and which would store all bcc data initially along with its rotations and translations. This seems to be an efficient way to go as every time that a BccStructure is used it is evaluated over it's group rotations. Class inheritance has been very useful in modifying methods for new processes, especially when the procedure is so similar.

I have created a new test class to perform the regression using these new structures and am fully realizing the uses of object oriented design. I am going to need to come back to make the code more efficient as it currently does not account for repeated structures. We may eventually want to find a better way to check to see if a structure is repeated, but the old technique for this relied on the periodicity of the function, and I need to check for structures all having independant periods. I have devised a technique for doing it, but will have to check that it works. I am going to take the small number of unique sites in a bcc structure, subtract the rotated and translated forms and check the difference to see which ones are multiples of the period vectors. There may have been a method similar to this before, but I don't quite follow the logic. This procedure could perhaps be improved upon later, and once a working version is finished, this will be one place to improve run times.

One of the difficulties I have encountered many times is that by using what someone else has already done, I have a difficult time understanding all of their logic. I have learned a lot about the language by debugging something that is already in place.

Next week, I am going to hopefully finish modifying the ALS procedures to account for the GroupBccStructures. If time permits I will hopefully have a working version to perform tests with.

3 9/15/07-9/22/07: Bcc Structure Storage

This week went much more slowly than anticipated. I am not all that much closer to running tests than I was at the beginning of the week, however, I do have a better understanding of some of the code that has been written.

Dr. Mohlenkamp helped me understand how point values are stored and how we checked to see if two structures were the same. I did not have a good understanding of the reciprocal vectors and for the first time I saw how they were used. I did not previously understand that, the point values are simply the component vectors we desired. This bit of insight, will make it much easier to store all of the equivalent BCC structures. I will in fact only store the equivalent point values. I also gained some other valuable insights from Dr. Mohlenkamp.

If we have two locations in a structure, and we wish to check if they are in fact equivalent locations, we can write them as two column vectors a and b. If they are equivalent then the difference between them will be an integer multiple of the period vectors. Making a matrix A, with the columns formed by the period vector we can solve the following system:

$$Ax = b - a, \tag{1}$$

where x is a column vector, all integer valued, if they are in fact equivalent. In order to solve the system we applyA^{-1} to both sides, however A^{-1} is simply a matrix formed from the reciprocal vectors. This is a very clever way to check to see if two points are the same, which should be very useful.

This upcoming week I will still be working on implementing this ideas. Even the progress was not very tangible, what I have now should make it go much faster, at least in that I have a much more clear idea of how I am going to do it.

4 9/23/07-9/30/07: To Inherit or Not to Inherit?

This week I spent more time on the GroupBCCStructure which would be a derived class from BccStructure, and would hold all of the equivalent structures. Every single time a function is evaluated it is evaluated over the equivalent structures, so this will make it easier to compute things.

I almost finished writing this class and getting it functional, and we decided that since we only needed equivalent point values, that we could use the BccStructures and call on a routine to get all equivalent structures and only store a list of all equivalent structures. This work has nearly been completed, however, I must make a few changes in the ALS routine to accomodate for this.

For next week, we should be able to actually test the regression procedure. So the countdown begins. All of the previous time and energy that didn't seem to produce many results, is now very helpful and has made this portion go much more quickly.

5 10/1/07-10/8/07: Running and Debugging

This week, with the help of Dr. Mohlenkamp, the code is excecutable, however it is not behaving as desired. I completed modifying the ALS procedure to evaluate over equivalence classes of point values. It took some time to correct the weighting as the weighting also takes into account repeats, so in general the weights must be computed every time a function is evaluated at a point value, however, in the paper, the weights are only used on the outer most sums. I believe that this has been corrected at all necessary locations, but I will double check it again.

Currently, GroupBCC, reads in the BccStructures successfully, converts them to equivalence classes of point values, creates an instance of a Sum Separable, as a guess, and begins to perform the ALS procedure on this function. The Sum Separable function is normalized and then it does not undergo any more updates, however, the algorithm will run through the ALS. I am in the process of figuring out why the Sum Separable function is not updating.

For next week I will work on correcting these bugs and double checking the weighting. As soon as the procedure functions correctly I will have to figure out how Gnuplot works and why I cannot open new windows on turing.

6 10/9/07-10/16/07: Eureka! It Runs.

This week, the changes have been finished and there is finally a working version. There were originally problems with the error and it appeared that

the mean squared error was not proportional to the Least Squares error. Amidst all of the changes the weighting had not been computed properly. I walked through all of the code and made sure that the weighting was used at the proper times and after this the code worked, and appears to work as planned. I have run it over large and small training data sets and have noticed that the larger the data set for training the better the error on the test data. I have run the procedure using larger numbers of active sites (cube size) and have also observed that this improves the error in our approximations. It is worth mentioning that the ALS procedure converges much more quickly as the size of the cube is increased. Note that all of these tests were performed using a sum separable of rank 1. I performed several tests using a sum separables of rank 2 and 3. These did provide better approximations, but I need to run more tests using these.

At this point I am going to build a test suite for the code. This will provide a way to check the current code for errors, and to provide a tool to check the effects of future modifications. I am working on plotting the data, however this may be put off until later. I will also be merging the writing I have done on this project with Dr. Mohlenkamp's start to a paper.

The investment of time into learning Python is finally beginning to pay off. A thing to note for the future is that there can be many hours of work researching, which do not produce many tangible results, but, which make future work possible.

7 10/17/07-10/23/07: Testing

This week I worked on more thorough testing of the procedure. I ran tests with cubes of size 2,3 and 4. I tried training over all of the data, which provided smaller errors when interpolating, but could not be fit as well. I need to learn more about Gnuplot so that I can plot the data better. Specifically, I need to learn how to include titles and lables.

I set up EquivBcc so that it would run tests using all 3 ranks and cube sizes 2,3,and 4, however this is not completely working. This should be very simple, but when I run it with cube size 3, the ALS procedure does not show any improvement. Next week I will need to find a better way to run these trials.

I will continue working on the test suites and still have to commit all of the changes that I have made.

8 10/24/07-10/30/07: More Testing

This week I learned how to commit things to the repository using svn. I have uploaded final versions of ALS, and EquivBcc. I need to merge Sum Separable and then I can commit it. When I tried it gives me an error that it has the wrong root, which means that the version in the repository has been modified. I made several little changes I should not have made which do not allow it to merge easily. In the future, I should probably not correct typos or eliminate white space I don't like.

I have made some progress on the test functions, but I am not entirely sure how to test the ALS procedure which accounts for the group symmetry. I have run the procedure for Separable functions of rank 1,2 and 3, with cube sizes 1,2,3, and 4 and have stored the outputs in log files and plotted the trained values against the predicted values. I have found a few interesting results have been that when using a cube of size 1, all 3 ranks predict the same value. Cube of size one only has the nearest neighbors, so this number should be some sort of average value of the energies between closest neighbors. All of the functions train well using cube sizes of 1 and 2. With a cube of size 3, rank 1 functions become constant. I am not sure if this is a problem in the code, or the fact that a rank 1 function does not have enough free parameters to fit this data. I have also found that I need to find a good way to check how significant the error is in comparison to the data.

Next week I hope to make many revisions to the paper and to finish the test suite.

9 10/31/07-11/5/07: Fitting Too Well?

This week I worked on my presentation. It was an excellent exercise, and great practice. It was much less stressful than I would imagine a typical seminar talk would be. I learned a little about the LaTeX "slide" class. One of the greatest difficulties was learning to use precise mathematical language for the presentation. The value of a mentor in helping with this is priceless.

After running over several different training sets we discovered the following issues: there were not always improvements in the error as rank increased. This should not be possible, as increasing the rank gives more degrees of freedom. What is happening is that the functions are being initialized to the same values. Hence all three procedures are finding a local minimum over the

span of the sums of separable functions. To guarantee that we see improvement in the errors as the rank is increased, we can this problem by initializing the rank 2 Sum Separable using the previous rank 1 approximation, and then a second part with components initialized to 1. This overcame the problem to some degree. To test the goodness of fit, we computed

$$\%capturederror = 1 - \frac{MSE}{Variance}. \tag{2}$$

The regression functions now demonstrated the desired properties that they predicted the data to a greater degree as rank increased and as the number of active sites increased. One thing to note was that I found that they were predicting 99.99% of the variance in the data, which seemed a little too high. I may want to check that I really am doing what I think I am doing. One can also overcome the previous problem by intializing the Separable function to random values. I may want to run some tests on this as well.

10 11/6/07-11/13/07: Pesky Groups

I have been working on the problem of G/G_x being well-defined. G_x is not normal in G, hence the factor group, G/G_x is not well-defined. However, if we take G/G_x to be the left cosets of G, namely the set $H = \{yG_x\}$. In general the collection of left or right cosets does not form a group unless the subgroup used in defining the cosets is normal. The group structure should be very similar to that of the wallpaper groups, however, I haven't found any properties of these groups that may help us. I have also noticed one mistake in our assumptions, namely, G should be the group of translations, rotations and reflections. However, this will not effect our algorithm. If we cannot find any nicer ways we could define $G = \{rt|r \in R, t \in T\}$. Where R is the group of all rotations and reflections and T is the collection of all translations. In this setting, we can define $T_x = \{t|tx = x\}$ and then replace G/G_x with $RT/T_x = \{rt|r \in R, t \in T_x\}$, which would have the order we used and would satisfy the consistency property.

It turns out that we can make sense of things as we have written them before:

Define G and G_x as before. Let $H_x = \{hG_x\}$ be the collection of left cosets of G_x. We may refer to these cosets by one of their representatives, h. First, note that H_x is a finite partition of G and that $|hG_x| = |G_x|$ for all h.

It is easy to see that f is invariant on each coset, as for all $\gamma \in hG_x$, $\gamma = hg$, for some $g \in G_x$, so $f(\gamma x) = f(hgx) = f(hx)$. Another useful property of these cosets is for all $h \in G$, $gH_x = \{ghG_x\} = \{hG_x\}$, as multiplying by a member of G is merely a permutation of H_x and we can follow the same proof as that in the paper.

The presentation went well this week. I had a few technical difficulties, teaching me a valuable lesson: if possible set everything up on the system you will use for the presentation before hand. Being able to watch Martin's presentation a few days prior was quite valuable. Attending other seminars has been an excellent way to learn how to give a good talk. In the future, I would think it might be useful to open our presentations up to a larger audience, perhaps undergrads, or other graduate students, to practice fielding more questions. I need to learn how to find more concise statements for the slides in the presentation and how to better incorporate these with my spoken words. I felt like I didn't have any additional information to provide other than what was on the slide, and as an audience member I appreciate when a speaker elaborates on the slides and when the spoken words provide a more thorough glimpse into what is on the slides. This will probably be the one are I will modify the most for next time.