1 09/15/2015: Atlanta Talk

Over the past week, I have

- Continued work on presentation for the talk in October.
- Considered, geometrically, the effect of “greasing”in a trough-like section of the error function. The geometric interpretation indicates that in such a section, “greasing”will make faster progress than will ALS, possibly after a period of movement in an incorrect direction. ALS would not move ”upstream”, as it were.
- Continued to attempt a proof that for a degenerate local minimum x^*, there exists an open set U such that $B(x^*, \epsilon) \cap \{x : f^*(x) < \delta\} \subset U$, where f^* is the infimum of f over rescaling of the summands, and such that the Łojasiewicz inequality holds. If proven, previous results could be adapted to provide a guarantee of convergence toward degenerate local minima, and provide greater understanding of the rate of approach toward such minima.

2 09/22/2015: Greasing and Second-order Taylor Approximations

- Using advice received, the presentation has been much improved, primarily through re-ordering of information.
- We now have an example target for which “greasing”significantly exacerbates a swamp. Asymptotic behavior still shows improvement in rate of convergence.
- Under the assumption that the error function $f(\mathbf{x}) \approx \mathbf{b}^*\mathbf{x} + \mathbf{x}^*A\mathbf{x} + k$, for some vector $\mathbf{b}$ and Hermitian matrix A, we have an explicit formula for the result of ALS, optionally with “greasing”:

$$\mathbf{x}_n = \mathbf{x}_0 + \left(\sum_{j=1}^{n} \frac{(-\theta)^j}{2} \sum_{1 \le k_j < \cdots < k_1 \le n} B_{k_1} A_{k_1}^{-1} B_{k_1}^* \prod_{i=2}^{j} A B_{k_i} A_{k_i}^{-1} B_{k_i}^* \right) (\mathbf{b} + 2A\mathbf{x}_0)$$

  Here A_i is the block of A corresponding to the i-th dimensional update, B_i is a linear map from a dimensional block to parameter space. Results not yet complete include

  - Analysis of best values of θ for minimizing distance to (approximate) target at a given iteration, if $A \succ 0$.
  - “Correct” direction if A is not positive-definite, and values of θ maximizing motion in that direction.

3 09/29/2015: Greasing and Second-order Taylor Approximations, continued.

I continued last week’s study of second-order Taylor expansions as they apply to (“greased”) ALS.

- The expression created above shows that ALS is well-approximated by a discrete affine dynamical system, provided that the parameter-space iterations do not move far.

  - This will occur near local minima of the objective function. In this case, however, Newton's method might provide quadratic convergence rather than the probable linear convergence of ALS.
  - This should occur in swamps.

- When in a swamp, the effectiveness of "greasing" (as measured by squared distance to target, or total progress after a full round of ALS) should be well-approximated by a polynomial of degree $2d$. This suggests a justifiable algorithm for choosing λ_{grease}.
  - Numerical experiments indicate that a better estimate of progress is needed, as moving into a swamp reduces effectiveness.
  - Some other aspects of "greased" behavior are not exploited by this algorithm. In asymptotic cases, the algorithm performs similarly to a known "good" constant, but underperforms compared to a more carefully chosen case-specific constant.

- Examination of asymptotic behavior when iterations converge in parameter space is not a primary concern, as Newton's method should provide quadratic convergence in that case. The computational complexity of Newton's method may be unreasonable.

4 10/06/2015: Topology and continued second-order Taylor approximation.

- I have found necessary and sufficient conditions for path-connectedness of unit real or complex tensors, provided that their factor spaces are equipped with an inner product.

- As noted last week, the behavior of ("greased") ALS does not correspond to that expected of an afine dynamical system.
  - Observed behavior indicates that sufficiently "greased" ALS tends to alternate between two linear rates, asymptotically. The mechanism causing this is not yet known.

5 10/13/2015: Algorithms.

This week, I continued examining the second-order Taylor expansion under the assumption that it is "near" the target. Progress remains slow.

- The update matrix for the affine approximation of ALS does not have any apparent exploitable structure (eg. block triangular or block diagonal). Computing its eigenvalues remains impractical.
- In the proposed algorithm for choosing λ_{grease}, replacing the tensor-space norm with a parameter-space norm seems to give correct choices of λ_{grease}.
  - Knowing a factorization of the target renders the approximation problem redundant, though numerical experiments showed error reaching $\epsilon_{\text{machine}} = 10^{-5}$ in 87 iterations. (15442 for pure ALS)
  - Creating an approximate factorization of the target is possible, but such an approximation would necessarily be created by a different algorithm, such as Gauss-Newton. Numerical experiments indicate that λ_{grease} does not need to be updated often, so this method can be used to gain some of the advantages of more expensive algorithms, such as Newton's method, without incurring those expenses at every iteration.
  - Estimated complexity of more sophisticated approximate factorization is $O(r^3d^6) + O(r^2d^4n)$.
- Dr. Mohlenkamp and I discussed, briefly, what constitutes an "impractical" matrix operation. If we assume 32-bit floating point matrix entries and typical modern hardware and operating systems,
  - 4×4 matrices fit in a cache line. (fastest)
  - 32×32 matrices fit in a page.
  - 32768×32768 matrices fit in 4GB of RAM.
  - 524288×524288 matrices fit on a 1TB hard disk. (slow)

6 10/20/2015: Parameter-space and behavior near singularities.

- Convergence behavior when ALS, MBI, or RBCD are near a unique local minimizer of the usual objective function is well studied. Convergence behavior when norms of summands diverge is practically unknown. I've been considering the behavior of a different objective function which ignores the lengths of the separable summands:

  $$E : \prod_{i=1}^{r} S(1, \{V_i\}_{i=1}^{d}) \to \mathbb{R} \qquad \begin{bmatrix} \mathcal{T}^1 \\ \vdots \\ \mathcal{T}^r \end{bmatrix} \mapsto \frac{\langle LC^* L^* \mathcal{T}_{\text{target}}, \mathcal{T}_{\text{target}} \rangle}{|L^* L|}$$

  Where L is a linear map dependent on the summands $\mathcal{T}^l$:

  $$L : \mathbb{R}^r \to S(\infty, \{V_i\}_{i=1}^{d}) \qquad \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_r \end{bmatrix} \mapsto \sum_{l=1}^{r} \mathbf{x}_l \mathcal{T}^l$$

  And C is the matrix of cofactors of $L^* L$.

  - E is bounded. All output is in the interval $[0, \|\mathcal{T}_{\text{target}}\|^2]$.
  - If each $V_i = \mathbb{R}^{n_i}$, and each $\mathcal{T}^l = \bigotimes_{i=1}^{d} \mathbf{x}_i^l$, then E is a multivariate rational function on the parameter space $\mathbb{R}^{r\left(\sum_{i=1}^{d} n_i\right)}$.
  - Restricting parameters to any line creates a bounded univariate rational function, which has a continuous extension.
  - In an attempt to show that limits of sequences linearly approaching a singularity are continuously dependent on the direction of approach (hopefully almost-everywhere), I have been working on derivatives of the determinant $|L^* L|$. I have a result for 2 summands, but a more general result eludes me.
  - My goal is a result similar to the Łojasiewicz inequality, but usable near singularities.

- Graphing all pairs of parameters for a $2 \times 2 \times 2$ tensor with a "wobbly" error graph shows (stretched) spirals in at least 2 parameter pairs. 3 parameter pairs remain constant. Note that "wobbly" error has only been observed in "greased" ALS.
  - For approximately correlated pairs, behavior is not well-approximated by a 1-dimensional linear discrete dynamical system.
  - Spirals are slow, so any complex eigenvalues have a small imaginary component.

7 10/27/2015: Attending SIAM Conference.

This Tuesday, I will have given the 20-minute presentation which I practiced earlier. Additionally:

- I have improved the visualization process in my tensor code. Output is now directly compatible with LaTeX.
- I was able to reduce the problem of homotopies of general paths in $S(r, \{V_i\}_{i=1}^d)$ to that of paths with piecewise-continuous summands.
  - The proof is constructive.
  - It follows that the set in question is simply-connected if and only if all paths which form loops in tensor space and have continuous (but not necessarily looping) summands are homotopic to a point.
    - It is my understanding that this statement was previously only known to be true in the "only if" direction.

8 11/03/2015: SIAM LA15 postmortem, small update on "shape" of error landscape

- Work continues, slowly, on a proof of a Łojasiewicz-like inequality near singularities of the error landscape. I have found that an open set around a generic linear approach will have nice (continuous on direction) limiting error. I'm now working on bounds for the higher-order terms of the Taylor expansion.
- The talk I gave on Tuesday went well. I have the following observations:
  - There is interest, in the community, in the topic of linear extrapolation. Though it has been studied in terms of matrices, it seems our work is the first time it has been applied to ALS.
  - The first thought of one of the audience members on the subject of swamps was that they may be caused by saddle points.
  - I should have included a slide listing the costs of the deregularization and greasing algorithms.
    - Greasing may cause swan-dives into swamps.
    - Deregularization increases the condition number of the matrices used.
- From the conference in general, I have the following observations:
  - "Sketching", most relevantly described as the matrix version of low-rank separated representation, is a hot topic.
    - The typical method for creating these "sketches" (pick several corresponding rows/columns) is related to choosing the best subset of separable summands from a separable representation.
    - Randomization may be employed when creating "sketches".
  - There is much investigation into less-studied matrix forms, such as Toeplitz, Hankel, etc. New decomposition algorithms, such as BATMAN2, are also being created. It was indicated that a stable algorithm to create Toeplitz decompositions in less than $O(n^3)$ time is an open problem.
  - Tensors are being applied to graph theory.
  - Tensor Train and Hierarchical Tucker formats are hot topics in tensor approximation.

  - Parallel programming, particularly on the unforgiving GPUs, is a hot topic. If you see a good way to break the problem down into easily-handled chunks, this is a potential application.
  - Many presentations included notes on preconditioners.
  - For an iterative algorithm, being able to re-use data from previous iterations can greatly benefit the algorithm. If computational complexity grows at each iteration, look for a way to use the results of previous iterations.
  - Regularization using approximations of Hamming distance is the currently preferred approach to encouraging low-rank decompositions.

9 11/10/2015: Little progress

- I have made some progress on the Łojasiewicz-like inequality.
  - The terms of the Taylor expansion of the continuous extension of a linear restriction of the usual objective function are analytic with respect to the direction of the linear restriction, on a dense, open, full-measure set. (more precisely, the complement of the set has Lebesque measure 0)

10 11/17/2015: Proposal

- This week, I concentrated primarily on writing my proposal, and have greatly improved the rate at which it is being written. For those intending to write a proposal, I have the following advice: Write the introduction first. If you attempt to write a different chapter before the introduction, you may find your writing better suited to the introduction; this can result in slow writing, due to the conflict between what should be written, and what is written.
- Dr. Mohlenkamp suggests that the "badness" measurement $\frac{|f(\mathbf{x})|}{\|\nabla f(\mathbf{x})\|^2}$ may be useful for choosing the best value of θ_{grease}.
  - To test this, I will generate tuples of n, d, rank (target), rank (approximation), badness, "optimal" value for theta (as measured by difference of parameters), and condition number. From these, I intend to graph pairs of values, hopefully revealing some connection.
  - At this time, I am unable to generate these, as my compiler is giving trouble that I have not yet been able to explain or fix.
- Work on the Łojasiewicz-like inequality has seen little progress this week. I tried a potentially elegant approach to solving one of my problems, but couldn't make it work. I will attempt a less-elegant approach.