1 01/29/2014: Geometric Intuition and Miscellaneous

Over the break, I was able to prove that $S_1(1, d, V)$ is simply-connected if the sets of unit vectors in the component vector spaces are simply-connected.

Intuitive results have been acheived for geometric obstacles near local minima of the error function. In particular every local minimum will either be the target, or be bordered by a geometric obstacle witha minimum curvature determined by the distance to the target. For a given target T, every local minimum x has some $\epsilon > 0$ such that $B(T, \|T - x\|) \cap B(x, \epsilon)$ contains no tensor of the approximation rank.

Local minima of error function for symmetric rank-2 approximations of Laplacian target demonstrate different behavior when symmetry condition is removed. Next step in analysis is to view error landscapes for asymmetric tensors near local minimum.

The following ideas have been considered and shelved due to low probability of success or usefulness:

- Find bounds on minimum/maximum spans of geometric obstacles, use to decide whether in swamp. Similar problems are NP-hard.
- Find good fast separable approximation of arbitrary tensor, use for rank increment. Fairly well-studied problem, doesn't aid intuition.
- Find low-complexity formula for best separable approximation of rank-2 tensor. Use for rank decrement. Unlikely to be a significant improvement over dropping smallest summand for cases where rank-decrease is wise, unless significant cancellation occurs.

2 02/05/2014: Condition number

The previous definition used for rank-r tensor approximation condition number does not behave well. Instead define the rank-r condition number of a tensor T by

$$\kappa_r(T) = \inf \left\{ \frac{\sum_{i=1}^r \|T_i\|}{\|T\|} \middle| T_i \text{ separable tensors}, T = \sum_{i=1}^r T_i \right\}$$

This definition behaves well, and is related to the previously used definition by the inequality $\kappa_{\text{old}}(T) \leq \kappa_r(T) \leq \sqrt{r}\kappa_{\text{old}}(T)$, so the new definition is penalized by regularization, though less directly than the old one was. Note: Error induced by representing a rank-r tensor T using r separable summands on a floating-point machine is not $\epsilon_{\text{machine}}\kappa_r(T)$, but rather $((1+\epsilon_{\text{machine}})^d - 1)\kappa_r(T)$. This is still $O(\epsilon_{\text{machine}}\kappa_r(T))$.

This definition also leads to a far simpler bound on perturbations of condition number from perturbations of summands. Despite this, it still does not significantly aid intuition, as interpretation of the bound has indicated continuity of κ, rather than a deep property of ALS. Further interpretation may be possible.

Error landscapes for asymmetric approximations of Laplacian target indicate that the local minimum of the symmetric approximation is a local minimum of the asymmetric approximation as well. Because this does not match the observed behavior of ALS near this point, further investigation is still required.

Shelved ideas:

- Approximate using complex first, then approximate the approximation using real summands. Complex approximation is also unlikely to be close to target.

3 02/12/2014: Laplacian Target

I have calculated the limiting error as $\alpha \to \beta$ with regularization parameter λ for the Laplacian target. This provides an example of critical points independent of λ.

Some further thought has been given to the "approximate using complex first" idea, but with no result.

Literature review indicated that examining the curvature of the graph of the error may lead to earlier detection of swamps. It also highlighted the inadequacy of current definitions of the term "swamp".

4 02/19/2014: Corrections, Confirmation, and Generalization

As Dr. Mohlenkamp expected, $\lambda = 0$ gives unique behavior as $\alpha \to \beta$. Further, and also as he expected, the local minima of the asymptotic ($\lambda \neq 0$) case correspond to the local minima of the rank-1 error, for the Laplacian target.

The rank-2 symmetric (regularized) ALS update function has been created for the Laplacian target, but it is too complicated to provide immediate insight. Hopefully, this function will provide insight into the "best λ" problem.

I have reformulated the problem of real approximation of complex as a problem of approximation on subspaces of the factor spaces. It's not perfect, as we lose some facets of complex multiplication, but if solved in a useful manner will be more broadly applicable. Current thoughts: Project each factor onto the subspace.

5 02/26/2014: Slices and angles

The problem of approximation on a subspace is solved for separable tensors. The naive approach, projection of each factor onto the subspace, is optimal. To complete the problem of finding real approximations of complex tensors, we need only choose scalar multiples of the factors to maximize the norms of their projections. This and the rank-r case are not high priority.

The slice of $\mathbb{R}^{2\times2\times2}$ described by Paatero is not symmetric, and thus does not fit well the error landscapes we have been examining. I am searching for a different, symmetric, slice with similar properties.

6 03/12/2014: Slices and (De)regularization

I have found several interesting symmetric slices of $\mathbb{R}^{2\times2\times2}$, one of which has relatively nice image in the angle representation. One slice appears to give an example of a (partially) convex set of rank-2 tensors in a neighborhood of a local minimum of approximation error, rather than the concave sets previously observed.

We now have a proof that $(A^*A + \lambda I)\mathbf{x} = A^*\mathbf{b}$ minimizes $\|A\mathbf{x} - \mathbf{b}\|^2 + \lambda\|\mathbf{x}\|^2$ for $\lambda \in (-\sigma_{\min}, \infty)$. Previously, we only had the result for $\lambda \geq 0$. This result allows error landscapes to be created for deregularized ALS, and permits intuition on the behavior of deregularized ALS as a hill-climbing algorithm.

7 03/19/2014: Deregularization, damping, local minima

Linear combinations of $\|A_i\mathbf{x} - \mathbf{b}_i\|^2$ terms have a simple formula for optimization, though some care must be exercised in the choice of constants. I am investigating several options for using this:

- Reward/penalize movement (in factor space): $\lambda\|\mathbf{x}_{n+1} - \mathbf{x}_n\|^2$
- Reward/penalize movement (in tensor space): $\lambda\|A_{n+1}\mathbf{x}_{n+1} - A_n\mathbf{x}_n\|^2$
- Penalize proximity to previously determined local minima (in tensor space): $\lambda\|A_{n+1}\mathbf{x}_{n+1} - T_{\min}\|^2$, with $\lambda < 0$.

Each has some potential application to resolving difficulties with the ALS algorithm; the first may help escape swamps, and the second may aid in finding globally optimal solutions by systematic elimination of locally optimal solutions. In each of these uses, useful values of the parameter λ are negative.

8 03/26/2014: "Greasing"

I have been analyzing the behavior of modified ALS, where each iteration minimizes an objective function which includes a term rewarding progress, conceptually similar to successive over-relaxation. The resulting algorithm typically provides faster convergence than ALS, but otherwise has similar behavior. I have (unproven) explanations for all observed behavior so far.

To explain the algorithm, it is best to define it in terms of ALS. Let $T : \times_{i=1}^{r} V_d \to S(r, d, \{V_i\})$ be a linear map from factor space to tensor space defined by

$$T(\mathbf{x}_1, \ldots, \mathbf{x}_r) = \sum_{i=1}^{r} (T_i \otimes \mathbf{x}_i)$$

for tensors $T_i \in S(1, d-1, \{V_i\})$. This captures the notion of holding fixed all but one dimension of a sum of separable tensors, which is the approach of ALS.
Iterations of the modified algorithm solve the optimization problem

$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_r} \|T(\mathbf{x}_1, \ldots, \mathbf{x}_r) - X_{\text{target}}\|^2 + \lambda \|T(\mathbf{x}_1, \ldots, \mathbf{x}_r) - X_{\text{previous}}\|^2$$

With target X_{target}, previous approximation X_{previous}, and tuning parameter λ. Choice of λ alters the behavior of the algorithm, as does choice of X_{previous}, typically either the tensor from before the d dimensions were altered or the most recent approximation.

- $\lambda = 0$, pure ALS.
- $\lambda > 0$, "damping", causes iterations to remain close to previous iterations, when compared with pure ALS.
- $\lambda < 0$, "greasing", causes iterations to become further from previous iterations, when compared with pure ALS.

Observed behaviors in test problems, dependent on choice of λ and X_{previous}:

- Rapid convergence (exact rate not yet determined).
- Divergence, apparently the result of an unstable dynamical system.
- "Spinning", where each iteration is distant from the previous, yet makes little progress toward the target. In tensor space, this typically appears as a spiraling motion.

The problem of choosing an optimal λ, both improving convergence of ALS and avoiding divergence, is not yet solved. We do, however, have evidence that certain poor choices of λ arise from an unstable dynamical system. In particular, choosing $\lambda \leq \frac{-1}{2}$ creates an unstable linear dynamical system in the case $d = 1$. The case $d > 1$ is not linear, but demonstrates similar problems.

The modified algorithm is little more complicated than ALS, as the optimization problem has an exact solution similar to the normal equations, mentioned last week.

Finally, the method of modifying the algorithm is easily extended to include similar algorithms, such as regularized and damped ALS, which should allow many of our results to have broad application.

9 04/02/2014: Convergence of *Greased* ALS

In most papers on ALS-like algorithms, two types of convergence results are given, if they apply:

- An accumulation point of the algorithm is a KKT (KarushKuhnTucker) point, thus satisfying necessary conditions for being a local minimum.
- Error does not increase.

The rate of convergence is rarely studied, outside of numerical examples, due to the difficulty of studying it symbolically.

We now have a result stating that, if an error function may be induced by an inner product and is minimized on a linear subspace by a vector $\mathbf{x}_{\text{new}}$, linear interpolations and extrapolations $\text{lerp}(\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}}, \theta) = (1-\theta)\mathbf{x}_{\text{old}} + \theta\mathbf{x}_{\text{new}}$ do not increase error for $\theta \in [0, 2]$.

This result, given as a linear algebra result, may be applied to *greased* variations of ALS and regularized ALS to show decreasing error.

The term "greasing" is awkward, and not descriptive. Perhaps "extrapolated"?

10 04/09/2014: Convergence of *Greased* ALS

Study of "greased"/"extrapolated" ALS continues. The current investigation simplifies the problem by considering how linear subspaces are altered, rather than requiring that a "best" approximation be found.

To understand this simplification, Let $T^1, \ldots, T^r \in S(1, d-1, \{V_i\}_{i=1}^{d-1})$ be separable tensors. Then

$$\left\{ \sum_{l=1}^{r} T^l \otimes v^l : v^l \in V_d \right\} \subset S(r, d, \{V_i\}_{i=1}^{d})$$

is a linear subspace of $S(\infty, d, \{V_i\}_{i=1}^{d})$. Updating one dimension while keeping the others fixed, as in the ALS algorithm, is equivalent to choosing a vector in this subspace.

Linearly interpolating two subspaces may be defined in several ways. The first, naive, approach can be applied to any two linear subspaces, but is not helpful for our research. Given a vector space W over a subfield $\mathbb{F} \subset \mathbb{C}$, consider the set-valued function

$$\begin{aligned} &\text{lerp} : 2^W \times 2^W \times \mathbb{F} \to 2^W \\ &\text{lerp}(U, V, \theta) = \{(1-\theta)\mathbf{u} + \theta\mathbf{v} : \mathbf{u} \in U, \mathbf{v} \in V\} \end{aligned}$$

Unfortunately, $\theta \notin \{0, 1\}$ implies

$$\text{lerp}(U, V, \theta) = \{\mathbf{u} + \mathbf{v} : \mathbf{u} \in U, \mathbf{v} \in V\}$$

So the function is useless for analysis of *greasing*.

A span of r linearly interpolated vectors does not share this problem, but is not as easy to study, as linearly interpolating between a vector in the first subspace and a vector in the second subspace need not give a vector in this span.

11 04/16/2014: Convergence of *Greased* ALS - Continued

I have continued to study how linearly extrapolating affects ALS. The current avenue of investigation is, as it was last week, examining the effect of extrapolation on subspaces.

Finding an optimal choice of extrapolation constant θ is expected to be difficult, so I am simplifying the problem by evaluating the "goodness" of a subspace using a single point within or near the subspace, rather than an exact solution.

Methods of choosing the point vary in both ease and expected utility:

- Choosing points arbitrarily on extrapolated subspaces is unlikely to give useful results.
- Extrapolation of the best points on two subspaces is expected to give a good approximation of the best point on an extrapolated subspace. It is not expected, however, to lie on the extrapolated subspace. Determining how well this method approximates the "best" θ has been vexing.
- A combination of multiple linear extrapolations (reminiscent of a cubic Bezier curve) can be made to lie in the extrapolated subspace, and may prove to be ideal, if initially more difficult to study.

Dr. Mohlenkamp has pointed out that previous work has studied optimization over two dimensions simultaneously. Such optimization should be expected to provide results at least as good as those of extrapolated ALS (though this has not been proven), but also to be far more difficult to study.

This is no longer expected to be the case, as it does not take into account the effect on subsequent dimensions.

12 04/23/2014: Numerical experiments with greasing

I have run several experiments with extrapolated ALS, testing estimation of a best extrapolation parameter.

Current methods of estimation attempt to extrapolate in such a way as to minimize error in the next calculated dimension. As such, it may be considered a subproblem of minimizing error with two dimensions free. Note that ALS minimizes error over only one free dimension.

Numerical experiments with greased ALS indicate the following:

- Estimation of best linear extrapolation using a single point gives results that beat ALS, and require only occasional updates to the extrapolation parameter.
- Applying multipliers to extrapolated vectors (to give a better estimate of extrapolation parameter) does not significantly improve results.
- Though estimated "best" parameters beat ALS, in no experiment did they beat iteration with fixed extrapolation parameter $\lambda = -0.487$. This indicates that studying only two dimensions is insufficient. The Laplacian test case may shed further light on this.