STRUCTURALLY-CONSTRAINED GRADIENT DESCENT FOR MATRIX FACTORIZATION IN HAPLOTYPE ASSEMBLY PROBLEMS

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Changxiao Cai
Department of Electrical Engineering, Tsinghua University

Sujay Sanghavi and Haris Vikalo
Department of Electrical and Computer Engineering, The University of Texas at Austin

Motivation

- Finding a low-rank approximation to a partially observed matrix is a frequently encountered problem
- The bi-linearity of the objective function renders the problem non-convex and computationally challenging
- Our focus: the problem of *haplotype assembly*, i.e., reconstructing single individual haplotypes from high-throughput sequencing data
- We formulate haplotype assembly as a structured low-rank matrix factorization problem
- Contributions: an efficient algorithm, analysis of its convergence, and experimental verification

Mathematical Model

- The set of n reads bearing information relevant for haplotype assembly are organized into an $n \times m$ SNP fragment matrix $\mathbf{R}$, where m is the haplotype length.
- In humans (and other diploids), SNP sites are bi-allelic, i.e. only two out of four possible nucleotides A, C, G or T are possible at any position.
- Bases in SNP positions are labeled by binary symbols {1;-1}, where the mapping between letters and binary symbols follows arbitrary convention. Entries in r_i that do not provide any SNP information are labeled by 0.
- Introduce a projector operator $P_\Omega(\cdot)$ defined as

$$P_\Omega(\mathbf{M}) = \begin{cases} M_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{otherwise,} \end{cases}$$

where Ω denotes the set of indices (i, j) such that $\mathbf{R}_{ij} \neq 0$.

- Matrix $\mathbf{R}$ can be thought of as being obtained by sampling, with errors, a low-rank $n \times m$ unobservable matrix $\mathbf{M}$

$$\mathbf{M} = \mathbf{U}\mathbf{V}^{\mathbf{T}}.$$

- Here $\mathbf{U}$ and $\mathbf{V}$ are $n \times k$ and $m \times k$ matrices, respectively, and k denotes the ploidy (the number of haplotypes).
- The i^{th} row of $\mathbf{U}$, $\mathbf{u}_i$, is the indicator of the origin of the i^{th} read. The rows of $\mathbf{U}$ are the k-dimensional standard unit vectors consisting of all 0's except for one entry which is equal to 1.
- DNA sequencing is erroneous and $P_\Omega(\mathbf{M}) \neq P_\Omega(\mathbf{R})$. We assume the model where the entries in $\mathbf{R}$ are perturbed versions of the corresponding entries in $\mathbf{M}$, i.e., the (i, j) entry in $\mathbf{R}$, $\mathbf{R}_{ij}$, is obtained as

$$R_{ij} = \begin{cases} M_{ij}, & \text{w.p.} \quad 1-p, \\ -M_{ij}, & \text{w.p.} \quad p, \end{cases}$$

where p denotes the sequencing/genotyping error rate

Illustration

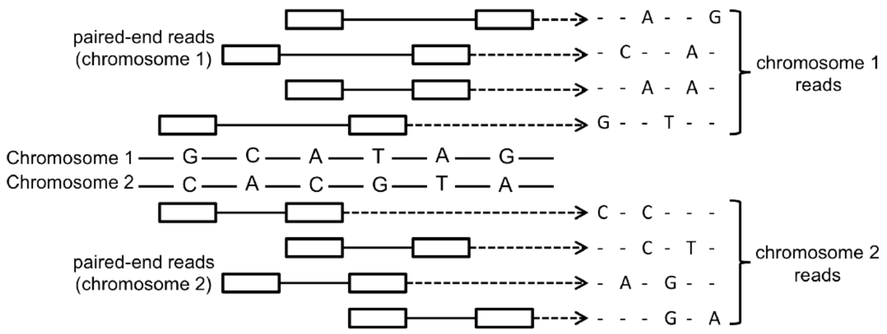

Fig. 1: The reads sample chromosomes/haplotypes but their origin (the chromosome from which they originate) is unknown and needs to be inferred.

Structurally-Constrained Gradient Descent

- We phrase haplotype assembly as the problem of low-rank matrix factorization $\mathbf{M} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ of an unobservable matrix $\mathbf{M}$ from its noisy sample with missing entries, $\mathbf{R}$.
- This can be expressed as minimization of the objective function

$$f(\mathbf{U}, \mathbf{V}) = \|P_\Omega(\mathbf{R} - \mathbf{U}\mathbf{V}^T)\|_F^2,$$

where $||\cdot||_F$ denotes the Frobenius norm of its argument.

- Imposing the special structure of matrix $\mathbf{U}$, i.e., the rows of $\mathbf{U}$ are standard unit vectors, we perform iterations

$$\mathbf{V}_{t+1} = \mathbf{V}_t - \alpha \nabla f(\mathbf{V}_t)$$

$$\mathbf{U}_{t+1} = \arg\min_{\mathbf{u}_i \in \Phi} f(\mathbf{U}, \mathbf{V}_{t+1})$$

where the gradient of f($\mathbf{U}$, $\mathbf{V}$) with respect to $\mathbf{V}$ is computed as

$$\nabla f(\mathbf{V}) = -2(P_\Omega(\mathbf{R} - \mathbf{U}\mathbf{V}^T))^T\mathbf{U}.$$

- The optimization over $\mathbf{V}$ is done by conventional gradient descent. The optimization over $\mathbf{U}$ is done by exhaustively searching over k vectors in $\boldsymbol{\Phi} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k\} = \{(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \ldots, 1)\}$ to find the most likely $\mathbf{U}_{t+1}$.
- After the termination criterion is met, entries of the most recent iteration $\mathbf{V}_{t^{max}}$ are quantized to generate an estimate of the haplotype matrix $\mathbf{V}$.
- *Theorem:* Let the step size α be selected as

$$\alpha = C\frac{\|\nabla f(V_t)^T\|_F^2}{\|P_\Omega(U_t \nabla f(V_t)^T)\|_F^2},$$

where $C \in (0, 1)$ is a constant. Then the solution ($\mathbf{U}^*$, $\mathbf{V}^*$) found by the structurally constrained gradient search algorithm is a stationary point of the objective function. Moreover, if a fresh set Γ of uniformly distributed test samples is available, then executing one iteration of the algorithm from ($\mathbf{U}^*$, $\mathbf{V}^*$) will reveal whether or not f($\mathbf{U}^*$, $\mathbf{V}^*$) is a global minimum.

Results

- Our structurally-constrained gradient descent is compared with HapCUT, HapCompass, Belief Propagation and several other state-of-the-art haplotype assembly methods on both synthetic and experimental data.
- The accuracy is measured by the reconstruction rate, the minimum error correction (MEC) score and the switch error rate (SWER).
- **Synthetic data:** the data sets emulating haplotype assembly of diploids

TABLE I: *Reconstruction rates for several haplotype assembly algorithms on diploid data*

Data error rate	Coverage	SCGD	SPH	FAST	2d	Hap-Cut	BP
0.1	8	**0.9957**	0.9843	0.9852	0.9641	0.9641	0.8722
0.1	10	**0.9978**	0.9836	0.9948	0.9781	0.9781	0.8727
0.2	8	**0.8802**	0.8247	0.8529	0.7912	0.8616	0.8607
0.2	10	**0.9482**	0.8555	0.8774	0.8169	0.8672	0.8672
0.3	8	**0.6656**	0.6294	0.6259	0.6230	0.6206	0.5715
0.3	10	**0.6967**	0.6381	0.6437	0.6340	0.6641	0.5956

- **Real data**: *1000 Genomes Project*, an international study meant to provide a detailed map of human genetic variation.

TABLE II: *The MEC scores and runtimes for the structurally constrained gradient descent, HapCompass and belief propagation when applied to the experimental data generated by the 1000 Genomes Project.*

chr	SCGD (Alg. 2)		HapCompass		BP	
	MEC	time(s)	MEC	time(s)	MEC	time(s)
1	1300	3.35	1496	9468.7	1488	29.2
2	1763	4.84	1938	10971.0	1921	28.7
3	1434	4.27	1627	8878.1	1615	29.0
4	1663	6.74	1863	9859.5	1849	31.4
5	1330	4.37	1505	8623.8	1488	26.2
6	2326	19.21	2771	8969.2	2719	27.5
7	1262	5.60	1423	8076.2	1417	22.4
8	1177	4.01	1261	8613.7	1255	23.6
9	895	2.92	1007	6145.0	1004	17.2

Conclusions

- We proposed the structurally-constrained gradient descent algorithm for factorizing partially observed low-rank matrices.
- The algorithm imposes special structure of the matrices in the sought after decomposition.
- We analyzed the convergence of the proposed algorithm and provided fundamental convergence guarantees.
- We applied the algorithm to the problem of haplotype assembly, testing it on both synthetic and experimental data. The results demonstrate superior performance in terms of both accuracy and speed over competing schemes.