Matrix Completion and Performance Guarantees for Single Individual Haplotyping

Somsubhra Barik and Haris Vikalo, *Senior Member, IEEE*

***Abstract*—Single individual haplotyping is an NP-hard problem that emerges when attempting to reconstruct an organism's inherited genetic variations using data typically generated by high-throughput DNA sequencing platforms. Genomes of diploid organisms, including humans, are organized into homologous pairs of chromosomes that differ from each other in a relatively small number of variant positions. Haplotypes are ordered sequences of the nucleotides in the variant positions of the chromosomes in a homologous pair; for diploids, haplotypes associated with a pair of chromosomes may be conveniently represented by means of complementary binary sequences. In this paper, we consider a binary matrix factorization formulation of the single individual haplotyping problem and efficiently solve it by means of alternating minimization. We analyze the convergence properties of the alternating minimization algorithm and establish theoretical bounds for the achievable haplotype reconstruction error. The proposed technique is shown to outperform existing methods when applied to synthetic as well as real-world Fosmid-based HapMap NA12878 datasets.**

***Index Terms*—matrix completion, single individual haplotyping, chromosomes, sparsity, alternating minimization.**

I. Introduction

DNA of diploid organisms, including humans, is organized into pairs of homologous chromosomes. The two chromosomes in a pair differ from each other due to point mutations, i.e., they contain so-called single nucleotide polymorphisms (SNPs) in a fraction of locations. SNPs are relatively rare; for humans, the SNP rate between two homologous chromosomes is roughly 1 in 300 base-pairs [1]. The ordered sequence of SNPs located on a chromosome in a homologous pair is referred to as a *haplotype*. Haplotype information is of critical importance for personalized medical applications, including the discovery of an individual's susceptibility to diseases [2], whole genome association studies [3], gene detection under positive selection and discovery of recombination patterns [4]. High-throughput DNA sequencing platforms rely on so-called shotgun sequencing strategy to randomly oversample the pairs of chromosomes and generate a library of overlapping paired-end reads (fragments). Parts of the reads that do not cover variant positions are typically discarded; the remaining data is conveniently organized in a read-fragment matrix where the rows correspond to reads and columns correspond to SNPs. Since the SNPs are relatively rare and reads are relatively short, the read-fragment matrix is typically very sparse. If the reads were free of sequencing errors, haplotype assembly would be straightforward and would require partitioning the reads into two clusters, one for each chromosome in a pair.

S. Barik and H. Vikalo are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, 78712 USA e-mail: sbarik@utexas.edu.

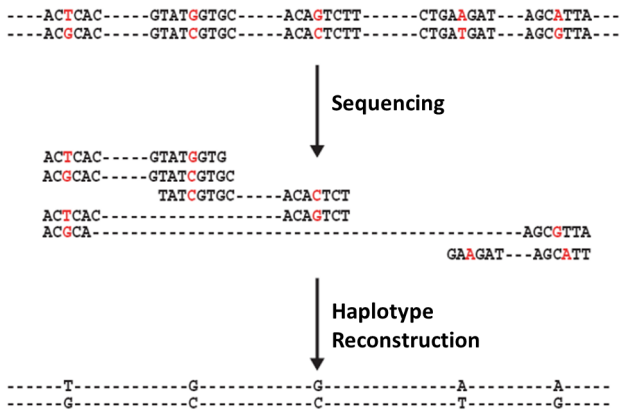

Fig. 1: An illustration of the single individual haplotyping problem. The nucleotides marked in red denote the SNPs on a pair of chromosomes. During sequencing, paired-end reads containing the SNPs are generated from multiple copies of the chromosomal sequences; one can think of each read as being obtained by sampling (with replacement) one of the chromosomes. We assume that the relative ordering of reads can be determined by mapping them to a reference genome. The goal of single individual haplotyping is to determine the order of SNPs associated with each chromosome in the pair.

However, presence of sequencing errors (of the order 10^{-3} to 10^{-2}) gives rise to ambiguities about the origin of reads and renders the single individual haplotyping (SIH) problem computationally very challenging.

Approaches to SIH attempt to perform optimization of various criteria including minimum fragment removal, minimum SNP removal and most widely used minimum error correction (MEC) objectives [5]. Finding the optimal MEC solution to the SIH problem is known to be NP-hard [5], [6]. A branch-and-bound approach in [7] solves the problem optimally but the complexity of the scheme grows exponentially with the haplotype length. Similar approach was adopted in [8] where statistical information about sequencing errors was exploited to solve the MEC problem using sphere decoding. However, the complexity of this scheme grows exponentially with haplotype length and quickly becomes prohibitive. Suboptimal yet efficient methods for SIH include greedy approach [9], max-cut based solution [10], Bayesian methods based on MCMC [11], greedy cut based [12] and flow-graph based approaches [13]. More recent heuristic haplotype assembly approaches include a convex optimization program for minimizing the MEC score [14], a communication-theoretic approach solved using belief propagation [15], dynamic programming based approach using graphical models [16] and probabilistic mixture model based approach [17]. Generally, these heuristic methods come without performance guarantees.

Motivated by recent developments in the research on matrix completion (overviewed in Section II-A), in this paper we formulate SIH as a rank-one matrix completion problem and propose a binary-constrained variant of alternating minimization algorithm to solve it. We analyze the performance and convergence properties of the proposed algorithm, and provide theoretical guarantees for haplotype reconstruction expressed in the form of an upper bound on the MEC score. Furthermore, we determine the sample complexity (essentially, sequencing coverage) that is sufficient for the algorithm to converge. Experiments performed on both synthetic and HapMap sample NA12878 datasets demonstrate superiority of the proposed framework over competing methods. Note that a matrix factorization framework was previously leveraged to solve SIH via gradient descent in [18]; however, [18] does not provide theoretical performance guarantees that are established for the alternating minimization algorithm proposed in the current paper. Preliminary version of our work was presented in [19].

A. Notation

Matrices are represented by uppercase bold letters and vectors by lowercase bold letters. For a matrix $\mathbf{M}$, $\mathbf{M}^{(i)}$ and $\mathbf{M}_i$ represent its i^{th} row and i^{th} column, respectively. M_{ij} denotes the $(i,j)^{th}$ entry of matrix $\mathbf{M}$ and u_i denotes the i^{th} entry of vector $\mathbf{u}$. $\mathbf{M}^\dagger, \|\mathbf{M}\|_2$, $\|\mathbf{M}\|_F$ and $\|\mathbf{M}\|_1$ represent respectively the transpose, the spectral norm (or 2-norm), the Frobenius norm and entry-wise ℓ_1 norm (i.e., $\sum_{ij}|M_{ij}|$) of the matrix $\mathbf{M}$, whereas the 2-norm of a vector $\mathbf{u} \in \mathbb{R}^m$ is denoted by $\|\mathbf{u}\|_2 = \left(\sum_{i=1}^m |u_i|^2\right)^{1/2}$. Each vector is assumed to be a column vector unless otherwise specified. A range of integers from 1 to m is denoted by $[m] = \{1, 2, \ldots, m\}$. $\mathbb{I}$ stands for the identity matrix of an appropriate dimension. Sign of an entry u_i is $\text{sign}(u_i) = 1$ if $u_i \geq 0$, -1 otherwise, and $\text{sign}(\mathbf{u})$ is the vector of entry-wise signs of $\mathbf{u}$. A standard basis vector with 1 in the i^{th} entry and 0 everywhere else is denoted by $\mathbf{e}_i$. The singular value decomposition (SVD) of a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ of rank k is given by $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{m\times k}$ and $\mathbf{V} \in \mathbb{R}^{n\times k}$ are matrices of the left and right singular vectors, respectively, of $\mathbf{M}$, with $\mathbf{U}^T\mathbf{U} = \mathbb{I}$ and $\mathbf{V}^T\mathbf{V} = \mathbb{I}$, and $\Sigma \in \mathbb{R}^{k\times k}$ is a diagonal matrix whose entries are $\{\sigma_1, \sigma_2, \ldots, \sigma_k\}$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$ are the singular values of $\mathbf{M}$. Projection of a matrix $\mathbf{M}$ on the subspace spanned by the columns of another matrix $\mathbf{U}$ is denoted by $\mathbb{P}_{\mathbf{U}}(\mathbf{M}) = \|\mathbf{U}\mathbf{U}^T\mathbf{M}\|_2$ and the projection to the orthogonal subspace is denoted by $\mathbb{P}_{\mathbf{U}^\perp}(\mathbf{M}) = \|(\mathbb{I} - \mathbf{U}\mathbf{U}^T)\mathbf{M}\|_2$. Subspace spanned by vectors $\mathbf{u}_i$ is denoted by $\text{span}\{\mathbf{u}_i\}$. Lastly, $\mathbf{1}_{\mathcal{A}}$ denotes the indicator function for the event $\mathcal{A}$, i.e.,$\mathbf{1}_{\mathcal{A}}$ = 1, if $\mathcal{A}$ is true, 0 otherwise.

II. Matrix Completion Formulation of Single Individual Haplotyping

A. Brief background on matrix completion

Matrix completion is concerned with finding a low rank approximation to a partially observed matrix, and has been an active area of research in recent years. Finding a rank-k approximation $\mathbf{M} \in \mathbb{R}^{m\times n}$, $k < \min\{m, n\}$, to a partially observed matrix is often reduced to the search for factors $\mathbf{U} \in \mathbb{R}^{m\times k}$ and $\mathbf{V} \in \mathbb{R}^{n\times k}$ such that $\mathbf{M} = \mathbf{U}\mathbf{V}^T$ [20]–[25]. Formally, the low rank matrix completion problem for $\mathbf{M}$ with noisy entries over a partial set $\Omega \in [m] \times [n]$ is stated as

$$(\hat{\mathbf{U}}, \hat{\mathbf{V}}) \;=\; \underset{\substack{\mathbf{U}\in\mathbb{R}^{m\times k}\\ \mathbf{V}\in\mathbb{R}^{n\times k}}}{\arg\min} \sum_{(i,j)\in\Omega} (\mathbf{R}_{ij} - \mathbf{U}^{(i)}\mathbf{V}_j)^2, \qquad (1)$$

where $\mathbf{R}$ is the partially observed noisy version of $\mathbf{M}$. The task of inferring missing entries of $\mathbf{M}$ by the above factorization is generally ill-posed unless additional assumptions are made about the structure of $\mathbf{M}$ [22], e.g., $\mathbf{M}$ satisfies the incoherence property (see definition (II.1)) and the entries of Ω are sampled uniformly at random.

Definition II.1. [22] A rank-k matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ with SVD given by $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^T$ is said to be incoherent with parameter μ if

$$\|\mathbb{P}_{\mathbf{U}}(\mathbf{e}_i)\|_2 \leq \frac{\mu\sqrt{k}}{\sqrt{m}} \;\forall\; i \in [m], \quad \text{and}$$

$$\|\mathbb{P}_{\mathbf{V}}(\mathbf{e}_j)\|_2 \leq \frac{\mu\sqrt{k}}{\sqrt{n}} \;\forall\; j \in [n].$$

The optimization in (1) is NP-hard [26]; a commonly used heuristic for approximately solving (1) is the alternating minimization approach that keeps one of $\mathbf{U}$ and $\mathbf{V}$ fixed and optimizes over the other factor, and then switches and repeats the process [24], [27]. Each of the alternating optimization steps is convex and can be solved efficiently. One of the few works that provide a theoretical understanding of the convergence properties of alternating minimization based matrix completion methods is [24], where it was shown that for a sufficiently large sampling probability of Ω, the reconstruction error can be minimized to an arbitrary accuracy. The original noiseless analysis was extended to a noisy case in [28].

In a host of applications, factors $\mathbf{U}$, $\mathbf{V}$ or both may exhibit structural properties such as sparsity, non-negativity or discreteness. Such applications include blind source separation [29], gene network inference [30], and clustering with overlapping clusters [31], to name a few. In this work, we consider the rank-one decomposition of a binary matrix $\mathbf{M} \in \{1, -1\}^{m\times n}$ from its partial observations that are perturbed by bit-flipping noise. This formulation belongs to a broader category of non-negative matrix factorization [21] or, more specifically, binary matrix factorization [32]–[35]. Related prior works include [32], [33], which consider decomposition of a binary matrix $\mathbf{M}$ in terms of non-binary $\mathbf{U}$ and $\mathbf{V}$, while [34] explores a Bayesian approach to factorizing matrices having binary components. The approach in [35] constrains $\mathbf{M}$, $\mathbf{U}$ and $\mathbf{V}$ to all be binary; however, it requires a fully observed input matrix $\mathbf{M}$. On the other hand, [36] considers a factorization of a non-binary $\mathbf{M}$ into a binary and a non-binary factor, with the latter having "soft" clustering constraints imposed. As opposed to these works, we aim for approximate factorization in the scenario where all of $\mathbf{M}$, $\mathbf{U}$ and $\mathbf{V}$ are binary, having only limited and noisy access to the entries of $\mathbf{M}$.

Next, we define the notion of distance between two vectors, which will be used throughout the rest of this paper.

Definition II.2. [37] Given two vectors $\tilde{\mathbf{u}} \in \mathbb{R}^m$ and $\tilde{\mathbf{w}} \in$

$\mathbb{R}^m$, the principal angle distance between $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{w}}$ is defined as

$$\text{dist}(\tilde{\mathbf{u}}, \tilde{\mathbf{w}}) = \|\mathbb{P}_{\mathbf{u}^\perp}(\mathbf{w})\|_2 = \|(\mathbb{I} - \mathbf{u}\mathbf{u}^T)\mathbf{w}\|_2 = \sqrt{1 - (\langle \mathbf{u}, \mathbf{w}\rangle)^2},$$

where $\mathbf{u}$ and $\mathbf{w}$ are normalized[1] forms of $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{w}}$.

B. System model

The first steps in the haplotype assembly data processing pipeline include mapping sequencing reads onto a reference followed by SNP and genotype calling. The former is typically done using short read mapping tools such as the Burrows-Wheeler Aligner (BWA) [38] while the latter is performed by relying on software packages such as FreeBayes [39] and SAMtools [40]. Following prior work on single individual haplotyping ([5] - [18]), we keep only parts of the reads that provide haplotype information, i.e., cover SNP positions along the genome. Let the number of reads carrying information about the haplotypes (after discarding reads which cover no more than one SNP) be n. If m denotes the haplotype length ($m \leq n$), then the reads can be organized into an $m \times n$ SNP fragment matrix $\mathbf{R}$, whose i^{th} column $\mathbf{R}_i$ contains information carried by the i^{th} read and whose j^{th} row $\mathbf{R}^{(j)}$ contains information about the j^{th} SNP position on the chromosomes in a pair. Since diploid organisms typically have bi-allelic chromosomes (i.e., only 2 possible nucleotides at each SNP position), binary labels $+1$ or -1 can be ascribed to the informative entries of $\mathbf{R}$, where the mapping between nucleotides and the binary labels follows arbitrary convention. Let Ω be the set of entries of $\mathbf{R}$ that carry information about the SNPs; the number of informative entries in each column of $\mathbf{R}$ is much smaller than m, reflecting the fact that the reads are much shorter than chromosomes. Let us define the sample probability as $p = \frac{|\Omega|}{mn}$. Furthermore, let us define the operator $P_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ as

$$[P_\Omega(\mathbf{R})]_{ij} = \begin{cases} R_{ij}, & (i,j) \in \Omega \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

R_{ij} represents the information about the i^{th} SNP site provided by the j^{th} read. Adopting the convention that 0's in column j correspond to SNP positions not covered by the j^{th} read, $P_\Omega(\mathbf{R})$ becomes a matrix with entries in $\{-1, 0, 1\}$. Let $\mathcal{H} = \{h_1, h_{-1}\}$ be the set of haplotype sequences of the diploid organism under consideration, with $h_k \in \{-1, 1\}^m, k = 1, -1$. Note that the binary encoding of SNPs along the haplotypes implies that the haplotypes are binary complements of each other, i.e., $h_1 = -h_{-1}$.

$P_\Omega(\mathbf{R})$ can be thought of as obtained by sampling, with errors, a *rank one* matrix $\mathbf{M}$ with entries from $\{1, -1\}$, given by

$$\mathbf{M} = \hat{\mathbf{u}}^\star(\hat{\mathbf{v}}^\star)^\dagger = \sigma^\star \mathbf{u}^\star(\mathbf{v}^\star)^\dagger \tag{3}$$

where $\hat{\mathbf{u}}^\star$ and $\hat{\mathbf{v}}^\star$ are vectors of lengths m and n respectively, with entries from $\{1, -1\}$, $\mathbf{u}^\star$ and $\mathbf{v}^\star$ are normalized versions of $\hat{\mathbf{u}}^\star$ and $\hat{\mathbf{v}}^\star$, and $\sigma^\star > 0$ is the singular value of $\mathbf{M}$. $\hat{\mathbf{u}}^\star$ represents the haplotype h_1 or h_{-1} (the choice can be arbitrary) and $\hat{v}_j^\star$ denotes the membership of j^{th} read, i.e., $\hat{v}_j^\star = k$ if and only if the j^{th} read is sampled from $h_k, k = 1, -1$. If $\mathbf{N}$ denotes the sequencing error noise matrix, then the erroneous SNP fragment matrix is given by

$$\mathbf{R} = \mathbf{M} + \mathbf{N}, \quad \text{or}$$
$$P_\Omega(\mathbf{R}) = P_\Omega(\mathbf{M}) + P_\Omega(\mathbf{N}). \tag{4}$$

The objective of SIH is to infer $\hat{\mathbf{u}}^\star$ (and $\hat{\mathbf{v}}^\star$) from the data matrix $P_\Omega(\mathbf{R})$ which is both sparse as well as noisy.

C. Noise model

Let p_e denote the sequencing error probability. The noise matrix $\mathbf{N}$ capturing the sequencing errors can be modeled as an $m \times n$ matrix with entries in $\{-N_{\max}, 0, N_{\max}\}$,[2] where each entry is given by

$$N_{ij} = \begin{cases} 0, & \text{w. p. } (1 - p_e) \\ -2M_{ij}, & \text{w. p. } p_e. \end{cases} \tag{5}$$

$\mathbf{N}$ has full rank since the errors occur independently across the reads and SNPs. The SVD of $\mathbf{N}$ is given by $\mathbf{N} = \mathbf{U}_N \Sigma_N (\mathbf{V}_N)^\dagger$, where $\mathbf{U}_N \in \mathbb{R}^{m\times m}, \mathbf{V}_N \in \mathbb{R}^{n\times m}, \Sigma_N = \text{diag}(\sigma_1^N, \sigma_2^N, \ldots, \sigma_m^N)$.

An important observation about the noise model defined in (5) is that it fits naturally into the worst case noise model considered in [25], [28]. Under this model, the entries of $\mathbf{N}$ are assumed to be distributed arbitrarily, with the only restriction that there exists an entry-wise uniform upper bound on the absolute value i.e., $|N_{ij}| \leq C$, where C is a constant. This is trivially true for the above formulation of SIH, where $C = N_{\max}$, leading to $\|N\|_F \leq \sqrt{mn} N_{\max}$. With the entries of $\mathbf{N}$ modeled as Bernoulli variables with probability p_e, the following lemma provides a bound on the spectral norm of the partially observed noise matrix $P_\Omega(\mathbf{N})$ and is proved in the Appendix A.

Lemma 1. *Let $\mathbf{N}$ be an $m \times n$ sequencing error matrix as defined in (5). Let Ω be the sample set of observed entries and let p be the observation probability. If p_e denotes the sequencing error probability, then with high probability we have*

$$\frac{\|P_\Omega(\mathbf{N})\|_2}{p} \leq 2N_{\max} p_e m\sqrt{n}.$$

III. Single Individual Haplotyping via Alternating Minimization

As seen in Section II-B, SIH can be formulated as the problem of low-rank factorization of the underlying SNP-fragment matrix $\mathbf{M}$,

$$\mathbf{M} = \mathbf{u}\mathbf{v}^T. \tag{6}$$

To perform the factorization, we optimize the loss function given by

$$f(\mathbf{u}, \mathbf{v}) = \|P_\Omega(\mathbf{R} - \mathbf{u}\mathbf{v}^T)\|_0 = \sum_{(i,j)\in\Omega} \mathbf{1}_{R_{ij} \neq u_i v_j}, \tag{7}$$

which is identical to the MEC score associated with the factorization in (6). However, ℓ_0-norm optimization problems

[1] Normalized version of any vector $\mathbf{x}$ is given by $\mathbf{x}/\|\mathbf{x}\|_2$.

[2] For the labeling scheme adopted in this paper, $N_{\max} = 2$.

are non-convex and computationally hard; instead, we use a relaxed ℓ_2-norm loss function

$$f(\mathbf{u},\mathbf{v}) = \|P_\Omega(\mathbf{R}-\mathbf{u}\mathbf{v}^T)\|_F^2 = \sum_{(i,j)\in\Omega}(R_{ij}-u_iv_j)^2.$$

Then, the optimization problem can be rewritten as finding $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ such that

$$(\hat{\mathbf{u}},\hat{\mathbf{v}}) = \underset{\substack{\mathbf{u}\in\{1,-1\}^m\\ \mathbf{v}\in\{1,-1\}^n}}{\arg\min} f(\mathbf{u},\mathbf{v}) = \underset{\substack{\mathbf{u}\in\{1,-1\}^m\\ \mathbf{v}\in\{1,-1\}^n}}{\arg\min} \sum_{(i,j)\in\Omega}(R_{ij}-u_iv_j)^2.$$

The above optimization problem can be further reduced to a continuous and simpler version by relaxing the binary constraints on $\mathbf{u}$ and $\mathbf{v}$,

$$(\hat{\mathbf{u}},\hat{\mathbf{v}}) = \underset{\mathbf{u}\in\mathbb{R}^m,\mathbf{v}\in\mathbb{R}^n}{\arg\min} \sum_{(i,j)\in\Omega}(R_{ij}-u_iv_j)^2. \tag{8}$$

A. Basic alternating minimization for SIH

The minimization (8) is a non-convex problem and often eludes globally optimal solutions. However, (8) can be solved in a computationally efficient manner by using heuristics such as alternating minimization, which essentially alternates between least-squares solution to $\mathbf{u}$ or $\mathbf{v}$. In other words, the minimization problem boils down to an ordinary least-squares update at each step of an iterative procedure, summarized as

$$\hat{\mathbf{v}} \leftarrow \arg\min_{\mathbf{v}\in\mathbb{R}^n} \sum_{(i,j)\in\Omega}(R_{ij}-\hat{u}_iv_j)^2, \quad \text{and} \tag{9}$$

$$\hat{\mathbf{u}} \leftarrow \arg\min_{\mathbf{u}\in\mathbb{R}^m} \sum_{(i,j)\in\Omega}(R_{ij}-u_i\hat{v}_j)^2. \tag{10}$$

The procedure (9)-(10) is initialized by using power iterations to generate the top singular vector of $P_\Omega(\mathbf{R})/\hat{p}$, where $\hat{p}$ denotes estimated sample probability; the entries of the vector greater than $\frac{2}{\sqrt{m}}$ are set to zero, followed by normalization to obtain $\hat{\mathbf{u}}^{(0)}$. Once a termination condition is met for the iterative steps (9)-(10), entries of $\hat{\mathbf{u}}$ are rounded off to ±1 to satisfy binary constraint on the solution.

Note that the performance of the basic alternating minimization procedure (9)-(10) depends on the choice of the initial vector $\hat{\mathbf{u}}^{(0)}$. The singular vector corresponding to the topmost singular value of the noisy and partially observed matrix $P_\Omega(\mathbf{R})$ serves as a reasonable starting point since, as shown later in Section IV, this vector has a small distance[3] to $\mathbf{u}^\star$. However, performing singular value decomposition requires $\mathcal{O}(mn^2)$ operations; therefore, it is computationally prohibitive for large-scale problems typically associated with haplotyping tasks. In practice, the power method is employed to find the topmost singular vector of the appropriately scaled matrix $P_\Omega(\mathbf{R})$ by iteratively computing vectors $\mathbf{x}^{(j)}$ and $\mathbf{y}^{(j)}$ as

$$\mathbf{x}^{(j)} = P_\Omega(\mathbf{R})\mathbf{y}^{(j-1)},\ \mathbf{y}^{(j)} = [P_\Omega(\mathbf{R})]^T\mathbf{x}^{(j)},\ \forall j = 0,1,\ldots \tag{11}$$

with the initial $\mathbf{y}^{(0)}$ chosen to be a random vector. Let us assume that the singular values of $P_\Omega(\mathbf{R})$ are $\sigma_1' \geq \sigma_2' \geq \ldots 0$. The power method is guaranteed to converge to the singular vector (say, $\hat{\mathbf{u}}^{(0)}$) corresponding to σ_1', provided $\sigma_1' > \sigma_2'$ holds strictly. The convergence is geometric with a ratio $(\sigma_2'/\sigma_1')^2$. Through successive iterations, the iterate $\mathbf{x}^{(j)}$ gets closer to the true singular vector; specifically,

$$\frac{\text{dist}(\mathbf{x}^{(j)},\hat{\mathbf{u}}^0)}{\|\mathbb{P}_{\hat{\mathbf{u}}^0}(\mathbf{x}^{(j)})\|_2} \leq \left(\frac{\sigma_2'}{\sigma_1'}\right)^{2j} \frac{\text{dist}(\mathbf{x}^{(0)},\hat{\mathbf{u}}^0)}{\|\mathbb{P}_{\hat{\mathbf{u}}^0}(\mathbf{x}^{(0)})\|_2}, \tag{12}$$

with per iteration complexity of $\mathcal{O}(mn)$ [18].

It has been shown in [24] that the convergence guarantees for the procedure described in this section can be established, provided the incoherence of the iterates $\hat{\mathbf{u}}^{(t)}$ and $\hat{\mathbf{v}}^{(t)}$ is maintained for iterations $t \geq 0$ (see Definition II.1). To ensure incoherence at the initial step, one needs to threshold or "clip" the absolute values of the entries of $\hat{\mathbf{u}}^{(0)}$ that exceed $2/\sqrt{m}$. Although the singular vector obtained by power iterations minimizes the distance from the true singular vector, it is the clipping step that makes sure that the information contained in $\hat{\mathbf{u}}^{(0)}$ is spread across every dimension instead of being concentrated in only few, much like the true vector $\mathbf{u}^\star$ (see Lemma 2).

B. Binary-constrained alternating minimization

Updates at each iteration of the procedure (9)-(10) ignore the fact that the underlying true factors, namely $\mathbf{u}$ and $\mathbf{v}$, have discrete $\{1,-1\}$ entries; instead, the procedure imposes binary constraints on $\mathbf{u}$ and $\mathbf{v}$ at the final step only. This may adversely impact the convergence of alternating minimization. Note that when $\hat{\mathbf{v}}$ is updated according to (9), its j^{th} entry is found as

$$\hat{v}_j^{(t+1)} = \arg\min_{v\in\mathbb{R}} \sum_{i|(i,j)\in\Omega}(R_{ij}-\hat{u}_i^{(t)}v)^2 = \frac{\sum_{i|(i,j)\in\Omega} R_{ij}\hat{u}_i^{(t)}}{\sum_{i|(i,j)\in\Omega}(\hat{u}_i^{(t)})^2}. \tag{13}$$

Clearly, if the absolute value of $\hat{u}_j^{(t)}$ is very large (or very small) compared to 1 at a given iteration t, then, given that $|R_{ij}| = 1$ for $(i,j) \in \Omega$, we see from (13) that the absolute value of $\hat{v}_j^{(t+1)}$ at iteration $t+1$ becomes close to 0 (or much bigger than 1). We empirically observe that as the iterations progress, the value of $\hat{v}_j^{(t+1)}$ becomes increasingly bounded away from ±1, which leads to potential incoherence of the iterates in subsequent iterations. To maintain incoherence, it is desirable that the entries of $\hat{\mathbf{u}}^{(t)}$ and $\hat{\mathbf{v}}^{(t)}$ remain close to ±1. It is therefore of interest to explore if we can do better by restricting the update steps in the discrete domain; in other words, enforce the discreteness condition in each step, rather than using it at the final step only.

One way of enforcing discreteness is to project the solution of each update onto the set $\{1,-1\}$, i.e., impose the inherent binary structure of $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ in (10) and (9). This leads us to the updates

$$\hat{\mathbf{v}} \leftarrow \arg\min_{\mathbf{v}\in\{1,-1\}^n} \sum_{(i,j)\in\Omega}(R_{ij}-\hat{u}_iv_j)^2, \quad \text{and} \tag{14}$$

$$\hat{\mathbf{u}} \leftarrow \arg\min_{\mathbf{u}\in\{1,-1\}^m} \sum_{(i,j)\in\Omega}(R_{ij}-u_i\hat{v}_j)^2. \tag{15}$$

Replacing $\mathbf{u}$ and $\mathbf{v}$ updates (9)-(10) by (14)-(15) leads to a discretized version of the alternating minimization algorithm for single individual haplotyping, given as Algorithm 1. Clearly, rounding-off of the final iterate is no longer required since the

[3]Please refer to Section II-A for a definition of distance measure.

individual iterates are constrained to be binary at each step of the algorithm.

Algorithm 1 SIH via discrete alternating minimization

Require: SNP-fragment matrix $\mathbf{R}$, observed set Ω, estimated sample probability $\hat{p}$.
Power Iteration: Use power iteration to generate the top singular vector of $P_\Omega(\mathbf{R})/\hat{p}$ and denote it by $\mathbf{u}^{(0)}$
Clipping: Set entries of $\mathbf{u}^{(0)}$ greater than $\frac{2}{\sqrt{m}}$ to zero, and then normalize to get $\hat{\mathbf{u}}^{(0)}$.
for $t = 0, 1, 2, \ldots, T-1$ **do**
$\hat{\mathbf{v}}^{(t+1)} \leftarrow \arg \min_{\mathbf{v}\in\{1,-1\}^n} \sum_{(i,j)\in\Omega} (R_{ij} - \hat{u}_i^{(t)} v_j)^2$
$\hat{\mathbf{u}}^{(t+1)} \leftarrow \arg \min_{\mathbf{u}\in\{1,-1\}^m} \sum_{(i,j)\in\Omega} (R_{ij} - u_i^t \hat{v}_j^{(t+1)})^2$
end for
Output: $\hat{\mathbf{u}}^{(T)}$ is the estimate $\hat{\mathbf{u}}$ of the haplotype vector

A closer look at the iterative update of $\hat{\mathbf{v}}$ in Algorithm 1 reveals that the update can be written as

$$\hat{v}_j^{(t+1)} = \begin{cases} 1 & \sum_{i|(i,j)\in\Omega} R_{ij}\hat{u}_i^{(t)} \geq 0 \\ -1 & \text{otherwise}, \quad \forall\ j \in [n]. \end{cases} \tag{16}$$

Similar update can be stated for $\hat{\mathbf{u}}$. The non-differentiability of the update (16), however, makes the analysis of convergence of Algorithm 1 intractable. In order to remedy this problem, the "hard" update in (16) is approximated by a "soft" update using a logistic function $f(x) = (e^x - 1)/(e^x + 1)$, thus replacing the $\hat{\mathbf{v}}$ and $\hat{\mathbf{u}}$ updates at iteration t in Algorithm 1 by

$$\hat{v}_j^{(t+1)} = f\left(\frac{1}{m}\sum_{i|(i,j)\in\Omega} R_{ij}u_i^{(t)}\right), \quad \forall\ j\ \in\ [n], \tag{17}$$

and

$$\hat{u}_i^{(t+1)} = f\left(\frac{1}{n}\sum_{j|(i,j)\in\Omega} R_{ij}v_j^{(t+1)}\right), \quad \forall\ i\ \in\ [m], \tag{18}$$

where $\mathbf{u}^t$ and $\mathbf{v}^t$ are vectors representing normalized $\hat{\mathbf{u}}^t$ and $\hat{\mathbf{v}}^t$. Note that the update steps (17) and (18) can be represented in terms of the normalized vectors since it holds that sign $\left(\sum_{i|(i,j)\in\Omega} R_{ij}u_i^t\right)$ = sign $\left(\sum_{i|(i,j)\in\Omega} R_{ij}\hat{u}_i^t\right)$. Updates (17) and (18) relax integer constraints on $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ while ensuring that their values remain in the interval $[-1, 1]$; this relaxation allows us to derive an upper bound on the MEC score through an analysis of convergence of the algorithm in Section IV. Alternating minimization procedure that relies on soft update steps given by (17) and (18) is formalized as Algorithm 2. In experiments that we conducted on real Fosmid dataset [12], Algorithm 1 and Algorithm 2 achieve similar MECs (comparison of results omitted for brevity).

It is worth pointing out the main differences between the approach considered here and the method in [18]. In the latter, the authors propose an alternating minimization based haplotype assembly method by imposing structural constraints on only one of the two factors, namely, the read membership factor. However, for the diploid case considered in this work, use of binary labels allow us to impose similar constraints on both $\mathbf{u}$ and $\mathbf{v}$, thereby leading to computationally efficient yet accurate (as demonstrated in the results section) method formalized as Algorithm 1. Furthermore, the alternating minimization algorithm in [18] is not amenable to performance analysis. Our aim is to recover (up to noise terms) the underlying true factors and analytically explore relation between the recovery error and the number of iterations required.

Algorithm 2 SIH via discrete alternating minimization with soft updates

Require: SNP-fragment matrix $\mathbf{R}$, observed set Ω, estimated sample probability $\hat{p}$.
Power Iterations: Use power iterations to generate the top singular vector of $P_\Omega(\mathbf{R})/\hat{p}$ and denote it by $\mathbf{u}^{(0)}$.
Clipping: Set entries of $\mathbf{u}^{(0)}$ greater than $\frac{2}{\sqrt{m}}$ to zero, and then normalize to get $\hat{\mathbf{u}}^{(0)}$.
for $t = 0, 1, 2, \ldots, T-1$ **do**
$\hat{v}_j^{(t+1)} \leftarrow \frac{\exp\left(\frac{1}{m}\sum_{i|(i,j)\in\Omega} R_{ij}u_i^{(t)}\right)-1}{\exp(\frac{1}{m}\sum_{i|(i,j)\in\Omega} R_{ij}u_i^{(t)})+1}, \quad \forall\ j = 1, \ldots, n,$
$\mathbf{v}^{(t+1)} \leftarrow \hat{\mathbf{v}}^{(t+1)}/\|\hat{\mathbf{v}}^{(t+1)}\|_2$
$\hat{u}_i^{(t+1)} \leftarrow \frac{\exp\left(\frac{1}{n}\sum_{j|(i,j)\in\Omega} R_{ij}v_j^{(t+1)}\right)-1}{\exp\left(\frac{1}{n}\sum_{j|(i,j)\in\Omega} R_{ij}v_j^{(t+1)}\right)+1}, \quad \forall\ i = 1, \ldots, m,$
$\mathbf{u}^{(t+1)} \leftarrow \hat{\mathbf{u}}^{(t+1)}/\|\hat{\mathbf{u}}^{(t+1)}\|_2$
end for
Output: Round-off entries of $\hat{\mathbf{u}}^{(T)}$ to ± 1 to get estimate $\hat{\mathbf{u}}$ of the haplotype vector.

IV. Analysis of Performance

We begin this section by presenting our main result on the convergence of Algorithm 2. The following theorem provides a sufficient condition for the convergence of this algorithm.

Theorem 1. *Let $\hat{\mathbf{u}}^\star \in \{1,-1\}^m$ and $\hat{\mathbf{v}}^\star \in \{1,-1\}^n$ denote the haplotype and read membership vectors, respectively, and let $\mathbf{R} = \mathbf{M} + \mathbf{N}$ denote the observed SNP-fragment matrix where $\mathbf{M} = \hat{\mathbf{u}}^\star(\hat{\mathbf{v}}^\star)^T = \mathbf{u}^\star\sigma^\star(\mathbf{v}^\star)^T$, $\mathbf{N}$ is the noise matrix with $N_{\max}$ and p_e as defined in (5), $\mathbf{u}^\star$ and $\mathbf{v}^\star$ are normalized versions of $\hat{\mathbf{u}}^\star$ and $\hat{\mathbf{v}}^\star$ respectively, and $\sigma^\star$ is the singular value of $\mathbf{M}$. Let $\alpha = n/m \geq 1$ and $\epsilon > 0$ be the desired accuracy of reconstruction. Assume that each entry of $\mathbf{M}$ is observed uniformly randomly with probability*

$$p > C\frac{\sqrt{\alpha}}{m\delta_2^2}\log n \log\left(\frac{\|\mathbf{M}\|_F}{\epsilon}\right)\left(p_e + \frac{64}{3}\delta_2\right), \tag{19}$$

where $\delta_2 \in \left[0, \frac{1}{21}(3.93 - C'N_{\max}p_e)\right]$ and $C, C' > 0$ are global constants. Then, after $T = \mathcal{O}(\log(\|\mathbf{M}\|_F/\epsilon))$ iterations of Algorithm 2, the estimate $\hat{\mathbf{M}}^{(T)} = \hat{\mathbf{u}}^{(T)}[\hat{\mathbf{v}}^{(T)}]^T$ with high probability satisfies

$$\|\mathbf{M} - \hat{\mathbf{M}}^{(T)}\|_F \leq \epsilon + 16\frac{p_e\sigma^\star}{3\delta_2}(2 + (2 + 3N_{\max})\delta_2). \tag{20}$$

The following corollary follows directly from Theorem 1.

Corollary 1. *Define $\tilde{\mathbf{M}}^{(T)} = sign\left(\hat{\mathbf{M}}^{(T)}\right)$. Under the conditions of Theorem 1, the normalized minimum error correction*

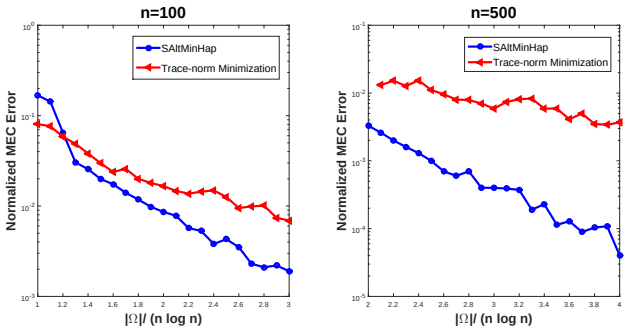

Fig. 2: A comparison of the normalized minimum error correction score of Algorithm 2 and the trace-norm minimization (SVT) method for matrices of dimensions $n = 100$ and $n = 500$, plotted as a function of sample size. α is 2 and error probability p_e is set to 5%. The values shown are averaged over 100 simulation runs.

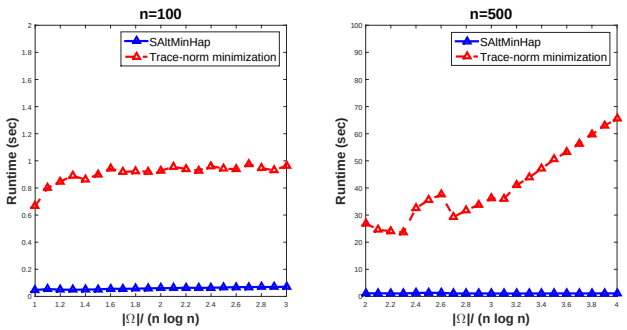

Fig. 3: A comparison of the runtime of Algorithm 2 and the trace-norm minimization (SVT) method for matrices of dimensions $n = 100$ and $n = 500$, plotted as a function of sample size. α is 2 and error probability p_e is set to 5%. The values shown are averaged over 100 simulation runs.

score with respect to $\mathbf{R}$*, defined as* $\tilde{MEC} = \frac{1}{mn}\|P_\Omega(\mathbf{R} - \tilde{\mathbf{M}}^{(T)})\|_0$*, satisfies*

$$\tilde{MEC}(\tilde{\mathbf{M}}^{(T)}) \leq \frac{\epsilon}{\sqrt{mn}} + \frac{16p_e}{3\delta_2}(2 + (2 + 3N_{\max})\delta_2) + \frac{1}{\sqrt{mn}}\|P_\Omega(\mathbf{N})\|_F. \quad (21)$$

Theorem 1 and Corollary 1 are proved in Appendix A. These two results imply that if the sample probability p satisfies the condition (19) for a given sequencing error probability p_e, then Algorithm 2 can minimize the MEC score up to certain noise factors in $\mathcal{O}(\log(\|\mathbf{M}\|_F/\epsilon))$ iterations. The corresponding sample complexity, i.e., the number of entries of $\mathbf{R}$ needed for the recovery of $\mathbf{M}$ is $|\Omega| = \mathcal{O}\left(\frac{\sqrt{\alpha}}{\delta_2^2} n \log n \log\left(\frac{\|\mathbf{M}\|_F}{\epsilon}\right)\left(p_e + \frac{64}{3}\delta_2\right)\right)$. Note that compared to (20), expression (21) has an additional noise term. This is due to the fact that unlike the loss function $\|\mathbf{M} - \hat{\mathbf{M}}^{(T)}\|_F$ in (20), the MEC score of $\tilde{\mathbf{M}}^{(T)}$ is calculated with respect to the observed matrix $P_\Omega(\mathbf{R})$.

Factor $\log(\|\mathbf{M}\|_F/\epsilon)$ in the expression for sample complexity (19) is due to using independent Ω samples at each of $T = \mathcal{O}(\log\|\mathbf{M}\|_F/\epsilon)$ iterations [24]. This circumvents potentially complex dependencies between successive iterates which are typically hard to analyze [41]. We implicitly assume independent samples of Ω in each iteration of Algorithm 2 for the sake of analysis, and consider fixed sample set in our experiments. As pointed out in [41], practitioners typically utilize alternating minimization to process the entire sample set Ω at each iteration, rather than the samples thereof.

An interesting observation in this context is that sequencing coverage, defined as the number of reads that cover a given base in the genome, can conveniently be represented as the product of the sample probability p and the number of reads. Then, (19) implies that the required sequencing coverage for convergence is $\mathcal{O}\left(\frac{\alpha\sqrt{\alpha}}{\delta_2^2}\log n \log\left(\frac{\|\mathbf{M}\|_F}{\epsilon}\right)\left(p_e + \frac{64}{3}\delta_2\right)\right)$, which is roughly logarithmic in n.

In Fig. 2 we compare the MEC error rate performance of Algorithm 2 (denoted as *SAltMinHap* in the figure) with another matrix completion approach, namely singular value thresholding (SVT) [42]; SVT is a widely used trace-norm minimization based method. We compare their performance on randomly generated binary rank-one matrices of size 50×100 and 250×500, and flip the entries in a uniformly randomly chosen sample set Ω with probability 0.05. Both methods are run 100 times for each chosen sample size and error rates are averaged over those runs. Results of the trace-norm minimization method are rounded off in the final iterations. Fig. 2 suggests that alternating minimization based matrix completion approach performs better than the trace-norm minimization for both problem dimensions, and the performance gap is wider for $n = 500$ compared to $n = 100$. Fig. 3 plots the runtime of the two methods for the same problem instances. Trace-norm minimization based method, that was reported in literature as the most accurate version of SVT, is much slower of the two, primarily due to the computationally expensive SVD operation at each iteration step.

Note that our analysis of convergence assumes that the samples of $\mathbf{M}$ are observed uniformly at random. This implies that each read contains SNPs located independently and uniformly at random along the haplotype. In practice, this assumption does not hold due to physical properties of reads (short length, limited insert size). However, note that the MEC score – the metric that the proposed algorithms attempt to minimize – is invariant to permutations of the rows of $\mathbf{M}$ (SNP locations on each read). This, combined with the fact that alternating minimization does not require specific structure of Ω, suggests the average performance of our algorithm may not be affected by permutations of the rows of $\mathbf{M}$. This can be readily verified empirically, which we have done on the experimental Fosmid data discussed in Section V. Random permutations of the rows of $\mathbf{R}$ whiten its columns and result in an input matrix with informative entries that are distributed uniformly, thus conforming to the assumptions under which we performed convergence analysis.

The remainder of this section presents the necessary components for establishing the main result of the paper (Theorem 1, Corollary 1). We start with a discussion of a suitable initialization for Algorithm 2 in Section IV-A. In Section IV-B, we present the essential building blocks for the convergence analysis, namely, characterization of the decay of distance of iterates from the true vectors (Theorem 2 and Corollary 2) and the incoherence of iterates (Lemma 3).

A. Initialization Analysis

In order for Algorithm 2 to converge, a suitable initial starting point close to the ground truth is necessary. In addition to the power iteration step, which gives a singular vector that is close to the true vector $\mathbf{u}^\star$, the subsequent clipping step helps retain the incoherence of the same, without sacrificing the closeness property. The following lemma uses Theorem 3 to establish the incoherence of $\hat{\mathbf{u}}^{(0)}$, and that $\hat{\mathbf{u}}^{(0)}$ remains close to the true left singular vector $\mathbf{u}^\star$.

Lemma 2. *Let $\mathbf{u}^{(0)}$ be obtained after normalizing the output of the power iteration step in Algorithm 2. Let $\mathbf{u}^C$ be the vector obtained after setting entries of $\mathbf{u}^{(0)}$ greater than $2\frac{\mu}{\sqrt{m}}$ to zero. If $\tilde{\mathbf{u}}$ is the normalized $\mathbf{u}^C$, then, with high probability, we have dist$(\tilde{\mathbf{u}}, \mathbf{u}^\star) \leq 1/2$, and $\tilde{\mathbf{u}}$ is incoherent with parameter 4μ, where μ is the incoherence parameter of $\mathbf{u}^\star$.*

Proof. The proof follows directly from Lemma C.2 from [24] and Lemma 2 from [28] after suitably using conditions from Lemma 1 and Theorem 3 (see Appendix A). □

B. Convergence Analysis

Let us denote $\lambda_j = \frac{1}{m} \sum_{i|(i,j)\in\Omega} R_{ij} u_i^{(t)}$. The Taylor series expansion of $\hat{v}_j^{(t+1)}$ from (17) is given by

$$\hat{v}_j^{(t+1)} = \frac{\lambda_j}{2} - \frac{{\lambda_j}^3}{24} + \frac{{\lambda_j}^5}{240} - \ldots, \quad \forall\ j = 1, \ldots, n. \tag{22}$$

Now, we have

$$|\lambda_j| = \left| \frac{1}{m} \sum_{i|(i,j)\in\Omega} R_{ij} u_i^{(t)} \right| \leq \frac{1}{m} \sum_{i|(i,j)\in\Omega} |R_{ij}||u_i^{(t)}| \leq \frac{1}{m} \sum_{i|(i,j)\in\Omega} |R_{ij}|.$$

Clearly, $\sum_{i|(i,j)\in\Omega} |R_{ij}| \sim \text{Bin}(m, p)$ for a given j, implying that the absolute value of λ_j is close to the entry-wise observation probability $p < 1$. This, in turn, implies that in (22) all terms with higher powers of λ_j are much smaller than the dominant linear term, and the Taylor's series expansion can be written as $\hat{v}_j^{(t+1)} \approx \frac{\lambda_j}{2} + \epsilon(\lambda_j)$, where the error term $\epsilon(\lambda_j)$ can be bounded as $|\epsilon(\lambda_j)| \leq |\lambda_j|^3/3! \ll 1$ using the Lagrange error bound. Therefore, we approximate the update in (17) as

$$\begin{aligned}
\hat{v}_j^{(t+1)} &= \frac{1}{2m} \sum_{i|(i,j)\in\Omega} R_{ij} u_i^{(t)} \\
&= \frac{1}{2m} \sum_{i|(i,j)\in\Omega} (M_{ij} + N_{ij}) u_i^{(t)} \\
&= \frac{1}{2m} \sum_{i|(i,j)\in\Omega} \left(\sigma^\star u_i^\star v_j^\star + [\mathbf{U}_N^{(i)}]^T \Sigma_N \mathbf{V}_N^{(j)} \right) u_i^{(t)} \\
&= \frac{1}{2m} \Bigg(\sigma^\star \langle \mathbf{u}^{(t)}, \mathbf{u}^\star \rangle v_j^\star - \Big[\sigma^\star \langle \mathbf{u}^{(t)}, \mathbf{u}^\star \rangle v_j^\star \\
&\quad - \sigma^\star v_j^\star \sum_{i|(i,j)\in\Omega} u_i^{(t)} u_i^\star \Big] + \sum_{i|(i,j)\in\Omega} u_i^{(t)} [\mathbf{U}_N^{(i)}]^T \Sigma_N \mathbf{V}_N^{(j)} \Bigg)
\end{aligned} \tag{23}$$

for $j = 1, \ldots, n$.

Let us introduce an error vector $\mathbf{F} \in \mathbb{R}^n$ as $\mathbf{F} = \sigma^\star \mathbf{B}^{-1} \left(\langle \mathbf{u}^{(t)}, \mathbf{u}^\star \rangle \mathbf{B} - \mathbf{C} \right) \mathbf{v}^\star$, where $\mathbf{B} = \frac{1}{p} \mathbb{I}_n$ and $\mathbf{C} \in \mathbb{R}^{n\times n}$ is diagonal with $C_{jj} = \frac{1}{p} \sum_{i|(i,j)\in\Omega} u_i^{(t)} u_i^\star$, $\forall j = 1, \ldots, n$. Furthermore, let us define a noise vector $N_{\text{res}} \in \mathbb{R}^{n\times 1}$ as $N_{\text{res}} = \mathbf{B}^{-1} \mathbf{C}^N \mathbf{S}^N \mathbf{v}^N$, where the quantities are as follows:

- $\mathbf{C}^N = [\mathbf{C}_1^N\ \mathbf{C}_2^N \cdots \mathbf{C}_m^N] \in \mathbb{R}^{n\times nm}$ where $(\mathbf{C}_q^N)_{jj} = \frac{1}{p} \sum_{i|(i,j)\in\Omega} u_i^{(t)} \mathbf{U}_{iq}^N, \quad \forall q \in [m], \forall j \in [n]$;
- $\mathbf{S}^N \in \mathbb{R}^{nm\times nm}$ is a diagonal matrix given by $\mathbf{S}^N = \text{diag}\left(\sigma_1^N \mathbb{I}_n, \ldots, \sigma_m^N \mathbb{I}_n\right)$;
- $\mathbf{v}^N = \left[(\mathbf{v}_1^N)^T\ (\mathbf{v}_2^N)^T\ \ldots\ (\mathbf{v}_m^N)^T \right]^T \in \mathbb{R}^{nm\times 1}$ where $\mathbf{v}_j^N \in \mathbb{R}^n$ is the j^{th} column of $\mathbf{V}_N$.

We also define $C_{jj}^N = \left[(\mathbf{C}_1^N)_{jj}\ (\mathbf{C}_2^N)_{jj}\ \ldots\ (\mathbf{C}_m^N)_{jj} \right]$ for any given $j \in [n]$. Using the above definitions, (23) can be written as

$$\hat{v}_j^{(t+1)} = \frac{1}{2m} \left[\sigma^\star \langle \mathbf{u}^{(t)}, \mathbf{u}^\star \rangle v_j^\star - F_j + \frac{1}{B_{jj}} C_{jj}^N \Sigma_N \mathbf{V}_N^{(j)} \right]. \tag{24}$$

Therefore, using vector-matrix notation, the update of $\hat{\mathbf{v}}$ can be written as

$$\begin{aligned}
\hat{\mathbf{v}}^{(t+1)} &= \frac{1}{2m} \left[\sigma^\star \langle \mathbf{u}^{(t)}, \mathbf{u}^\star \rangle \mathbf{v}^\star - \mathbf{F} + N_{\text{res}} \right] \\
&= \frac{1}{2m} \left[\mathbf{M}^T \mathbf{u}^{(t)} - \mathbf{F} + N_{\text{res}} \right].
\end{aligned} \tag{25}$$

Recalling (11), one can identify that $\mathbf{M}^T \mathbf{u}^{(t)}$ in the above expression is the update term in the power iteration applied to the true matrix $\mathbf{M}$. Therefore, the update described in (25) is essentially power iteration of $\mathbf{M}$ that would have led to the true singular vector $\mathbf{u}^\star$ except that it is perturbed by error terms due to incomplete observations and sequencing noise ($\mathbf{F}$ and N_{res} in the above expression, respectively). This observation leads to the analysis approach where appropriate upper bounds to the aforementioned errors terms are derived. The proof of convergence presented in this work follows the framework in [24], [28] and consists of an inductive analysis meant to establish guarantees that, given an incoherent $\mathbf{u}^{(t)}$ that is close in terms of principal angle to $\mathbf{u}^\star$, the subsequent iterate $\mathbf{v}^{(t+1)}$ is also incoherent with identical parameter and is closer to $\mathbf{v}^\star$ by at least a constant factor. This statement is formally expressed using Theorem 2 and Lemma 3 for the distance and incoherence conditions, respectively. Note that Theorem 2, which is a starting point for proving Theorem 1, demonstrates geometric decay of the distance between the subspaces spanned by $\hat{\mathbf{u}}^{(t)}$, $\hat{\mathbf{v}}^{(t)}$ and $\mathbf{u}^\star$, $\mathbf{v}^\star$, respectively; furthermore, Lemma 3 establishes the incoherence condition satisfied by the iterates. Proofs of both are deferred to Appendix A.

Theorem 2. *Under the assumptions of Theorem 1, the $(t+1)^{th}$ iterates $\hat{\mathbf{u}}^{(t+1)}$ and $\hat{\mathbf{v}}^{(t+1)}$ of Algorithm 2 satisfy with high probability*

$$\begin{aligned}
&dist(\hat{\mathbf{v}}^{(t+1)}, \mathbf{v}^\star) \leq \frac{1}{4} dist(\hat{\mathbf{u}}^{(t)}, \mathbf{u}^\star) + \frac{\mu_1 p_e}{\delta_2}, \quad \textit{and} \\
&dist(\hat{\mathbf{u}}^{(t+1)}, \mathbf{u}^\star) \leq \frac{1}{4} dist(\hat{\mathbf{v}}^{(t+1)}, \mathbf{v}^\star) + \frac{\mu_1 p_e}{\delta_2}, \ 0 \leq t \leq T-1,
\end{aligned}$$

where μ_1 is the incoherence parameter of the intermediate iterates $\hat{\mathbf{u}}^{(t)}$ and $\hat{\mathbf{v}}^{(t)}$.

The following corollary is based on the findings of Theorem 2 and is used to complete the proof of Theorem 1 in Appendix A.

Corollary 2. *Under the assumptions of Theorem 2, at the end of $T = \mathcal{O}(\log \frac{\|\mathbf{M}\|_F}{\epsilon})$ iterations, it holds that*

$$dist(\hat{\mathbf{v}}^{(T+1)}, \mathbf{v}^\star) \leq \frac{1}{2}\frac{\epsilon}{\|\mathbf{M}\|_F} + \frac{4\mu_1 p_e}{3\delta_2}, \quad \text{and}$$
$$dist(\hat{\mathbf{u}}^{(T+1)}, \mathbf{u}^\star) \leq \frac{1}{2}\frac{\epsilon}{\|\mathbf{M}\|_F} + \frac{4\mu_1 p_e}{3\delta_2}.$$

Lemma 3. *Let $\mathbf{M}, \mathbf{N}$, p, Ω be defined as in Theorem 1. Let $\mathbf{u}^{(t)}$ be the unit vector obtained at the t^{th} iteration of Algorithm 2 with incoherence parameter μ_1. Then, with probability greater than $1 - 1/n^3$, the next iterate $\mathbf{v}^{(t+1)}$ is also μ_1 incoherent.*

In our analysis of the matrix factorization approach to single individual haplotyping, we adopted techniques proposed by [24], [28]; note, however, that the scope of analysis in the current paper goes beyond the prior works. In particular, the authors of [24] considered a noiseless case of matrix factorization and did not impose structural constraints on the iterates. Their work was extended to a noisy case in [28] for a similar unconstrained setting. Additionally, [28] did not exploit statistical properties of the noise, except the entry-wise upper bound, whereas the present work uses the bit-flipping model as discussed in Section II-C, allowing us to characterize dependence of the performance on the sequencing error probability p_e.

V. Results and Discussions

We begin this section by stating the metrics for evaluating performance of our single individual hapolotyping algorithm.

A. Performance Metrics

A widely used metric for characterizing the quality of single individual haplotyping is the minimum error correction (MEC) score. This metric captures the smallest number of entries of $P_\Omega(\mathbf{R})$ which need to be changed from 1 to -1 and vice versa so that $P_\Omega(\mathbf{R})$ can be interpreted as a noiseless version of $P_\Omega(\mathbf{M})$. Essentially, it is the most likely number of sequencing errors, defined for diploids as

$$\text{MEC} = \sum_{i=1}^{n} \min\left(D(r_i, \hat{h}_1), D(r_i, \hat{h}_{-1})\right), \tag{26}$$

where $D(r_i, \hat{h}_j)$ denotes the generalized Hamming distance between read r_i (regarded as m length vector in $\{1, -1, \text{`-'}\}$) and the estimated parent haplotypes $\hat{h}_k,\ k = 1, -1$. This, in turn, is defined as $D(r_i, \hat{h}_j) = \sum_{j=1}^{m} d(r_{i,j}, \hat{h}_{k,j})$, where

$$d(x, y) = \begin{cases} 1, & \text{if } x \neq \text{`-'} \text{ and } y \neq \text{`-'} \text{ and } x \neq y \\ 0, & \text{otherwise,} \end{cases} \tag{27}$$

and $r_{i,j}$ and $\hat{h}_{k,j}$ denote the j^{th} entries of r_i and $\hat{h}_k$, respectively.

The MEC score is a relevant and most commonly studied performance metric for single individual haplotyping [43], and critically important for experimental data where the ground truth is not known in advance. It is also a proxy for the most meaningful haplotype assembly metric referred to as *reconstruction rate*. Recall that $\mathcal{H} = \{h_1, h_{-1}\}$ denotes the set of true haplotypes. Then the reconstruction rate of $\hat{\mathcal{H}} = \{\hat{h}_1, \hat{h}_{-1}\}$ with respect to $\mathcal{H}$ is defined as [44]

$$\mathcal{R}_{\mathcal{H},\hat{\mathcal{H}}} = 1 - \frac{\min\{D(h_1, \hat{h}_1 + D(h_{-1}, \hat{h}_{-1}), D(h_1, \hat{h}_{-1}) + D(h_{-1}, \hat{h}_1)\}}{2m}, \tag{28}$$

where $D(h_i, h_j)$ denotes the generalized Hamming distance between the haplotype pair h_i and h_j.

B. Experiments

In this section, for convenience we refer to our algorithm for single individual haplotyping as *SAltMinHap*. All of the methods described here were run on a Linux OS desktop with 3.07 GHz CPU and 8 GB RAM on an Intel Core i7 880 Processor.

We first tested our algorithm on the experimental dataset containing Fosmid pool-based NGS data for HapMap trio child $NA12878$ [12]. The Fosmid dataset is characterized by very long fragments, high SNP to read ratio, and low sequencing coverage of about 3X, consisting of around $1,342,091$ SNPs spread across 22 chromosomes. We compare the performance of SAltMinHap with the structurally-constrained gradient descent (SCGD) algorithm of [18] and one more recent SIH software ProbHap [16], which was shown to be superior to several prior methods on this dataset [10], [12], [17]. Table I shows the MEC rate (average number of mismatches per SNP position across the reads) and runtimes for all 22 chromosomes. As seen there, our SAltMinHap outperforms other methods for majority of the chromosomes shown; it is second best in terms of runtime (behind SCGD).

Next, we focus on the evaluation of performance on simulated dataset using reconstruction rate metric. For this purpose, we use a widely popular standard benchmarking dataset from [44] which also provides the true haplotypes used to generate the read data. The dataset contains reads at a sequencing error rate with values in the set $\{0.0, 0.1, 0.2, 0.3\}$, and a depth of coverage in the set $\{3X, 5X, 8X, 10X\}$. Reconstruction rate of our algorithm is compared with that of SCGD [18] and two more recent SIH methods known as HGHap [45] and MixSIH [17]. In particular, [45] is chosen for performance comparison since it has been shown to outperform a number of existing SIH methods such as [9], [10], [46]–[48]. A comparison with ProbHap is not shown for this data since it reconstructed haplotypes with a large fraction of SNPs missing (and therefore has inferior performance compared to the methods used in the comparisons). The results, shown in Table II, are obtained by averaging over 100 simulation runs for each combination of sequencing error rate and sequencing coverage, for a haplotype length of 700 base pairs and pairwise hamming distance 0.7. As evident from the results, our method

TABLE I: MEC rates and runtimes on HapMap sample NA12878 dataset.

Chr	SAltMinHap		SCGD		ProbHap	
	MEC	time(s)	MEC	time(s)	MEC	time(s)
1	0.034	65.0	0.04	44.2	0.058	87.7
2	0.035	71.6	0.035	49.5	0.055	88.9
3	0.034	61.1	0.036	41.5	0.057	84.3
4	0.029	60.7	0.034	41.8	0.053	67.1
5	0.032	52.9	0.036	39.9	0.054	64.6
6	0.038	34.7	0.037	27.9	0.050	53.4
7	0.038	26.4	0.035	25.05	0.055	40.8
8	0.033	24.3	0.034	23.9	0.05	42.8
9	0.036	21.9	0.037	17.6	0.052	45.2
10	0.036	24.7	0.037	21.0	0.053	44.4
11	0.034	24.7	0.038	20.8	0.055	39.5
12	0.037	23.5	0.037	20.2	0.057	38.9
13	0.039	14.6	0.035	15.6	0.053	26.4
14	0.035	16.6	0.039	13.7	0.055	27.4
15	0.038	14.1	0.041	11.9	0.056	26.5
16	0.046	20.3	0.0405	12.2	0.051	36.5
17	0.048	15.3	0.046	11.1	0.061	27.4
18	0.033	12.2	0.037	11.8	0.053	24.4
19	0.052	12.8	0.046	9.0	0.063	19.8
20	0.044	18.1	0.044	13.0	0.055	30.9
21	0.035	11.5	0.041	8.5	0.051	15.6
22	0.054	11.7	0.055	8.6	0.061	31.4

TABLE II: Reconstruction rate comparison on simulated data. Boldface values indicate best performance.

Error Rate	Cov.	SAlt-MinHap	SCGD	HGHap	MixSIH
0.0	3X	**1**	0.983	0.934	0.776
0.0	5X	**1**	0.976	0.989	0.923
0.0	8X	**1**	0.999	0.994	0.995
0.0	10X	**1**	0.999	0.999	**1**
0.1	3X	**0.935**	0.869	0.934	0.775
0.1	5X	0.979	0.951	**0.990**	0.942
0.1	8X	**0.996**	**0.996**	0.987	0.972
0.1	10X	**0.999**	**0.999**	0.997	0.993
0.2	3X	**0.735**	0.677	0.677	0.68
0.2	5X	0.864	0.785	**0.91**	0.774
0.2	8X	**0.943**	0.899	0.884	0.932
0.2	10X	0.966	0.934	0.894	**0.969**
0.3	3X	0.555	0.527	0.592	**0.65**
0.3	5X	0.595	0.524	0.621	**0.667**
0.3	8X	0.68	0.518	0.646	**0.714**
0.3	10X	0.723	0.58	0.696	**0.751**

is either the best or the second best in all of the scenarios. SAltMinHap performs particularly well in the more realistic error range of $0.0 - 0.2$ and is marginally inferior to MixSIH only for higher sequencing error values.

To illustrate the performance guarantees in Section IV, we synthesized a dataset with 300bp-long reads that sample randomly generated reference genomes of length 2000bp, 2500bp, 3000bp and 3500bp, each with an average SNP rate of 10^{-2}. Sequencing error rate and coverage are set to 0.1% and $20X$, respectively. Fig. 4 shows the MEC rate and the MEC bound stated in Corollary 1. As the genome length increases, dimension of the SNP fragment matrix grows as well; since the bound scales with problem dimension slower than empirical MEC, the bound becomes less tight for longer genomes.

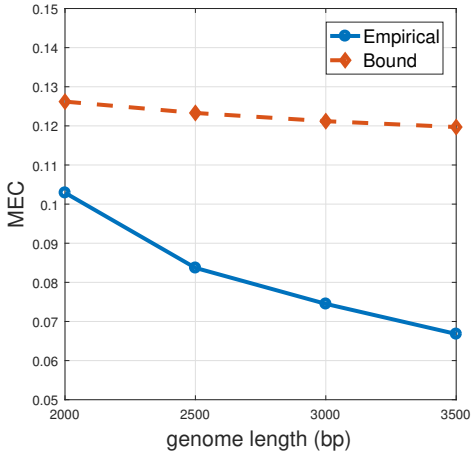

Fig. 4: *A comparison of the MEC bound in Corollary 1 and empirical MEC rates (averaged over* 50 *simulation runs) for a short-read dataset with* 300*bp-long reads at* $20X$ *coverage with sequencing error as* 0.1% *and SNP rate of* $1/100$*. Parameters used to compute the bound are found via grid search.*

VI. Conclusion

Motivated by the single individual haplotyping problem from computational biology, we proposed and analyzed a binary-constrained variant of the alternating minimization algorithm for solving the rank-one matrix factorization problem. We provided theoretical guarantees on the performance of the algorithm and analyzed its required sample probability; the latter has important implications on experimental specifications, namely, sequencing coverage. Performance of haplotype reconstruction is often expressed in terms of the minimum error correction score; we establish theoretical guarantees on the achievable MEC score for the proposed binary-constrained alternating minimization. Experiments with a HapMap sample NA12878 dataset as well as those with a widely used benchmarking simulated dataset demonstrated efficacy of our algorithm.

Appendix A
Proofs of Lemmas and Theorems

A. Preliminaries

The following property and lemma are well-known classical results that will be useful in the forthcoming proofs.

Property 1. *For a given matrix* $\mathbf{M} \in \mathbb{R}^{m\times n}$ *of rank* k*, the following relations hold between the 2-norm, the Frobenius norm and the entry-wise* ℓ_1 *norm of* $\mathbf{M}$*:*

- $\|\mathbf{M}\|_2 \leq \|\mathbf{M}\|_F \leq \sqrt{k}\|\mathbf{M}\|_2$
- $\|\mathbf{M}\|_F \leq \|\mathbf{M}\|_1 \leq \sqrt{mn}\|\mathbf{M}\|_F$

Lemma 4 (Bernstein's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables. Also, let $|X_i| \leq L \in \mathbb{R}\ \forall i$ w.p. 1. Then it holds that*

$$Pr\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mathbb{E}[X_i]\right| > t\right) \leq 2\exp\left(-\frac{t^2/2}{Lt/3 + \sum_{i=1}^{n} Var(X_i)}\right).$$

The following theorem from [25] provides an upper bound on the error between the true matrix $\mathbf{M}$ and the best rank-k approximation of the noisy and partially observed version of $\mathbf{M}$, and is used in the proof of Lemma 2.

Theorem 3. *[25] Let $\mathbf{R} = \mathbf{M} + \mathbf{N}$, where $\mathbf{M}$ is an $m \times n$ μ-incoherent matrix with rank k ($m \leq n$) and the indices in the sampling set $\Omega \in [m] \times [n]$ are chosen uniformly at random. Let $\alpha = n/m$, $|M_{ij}| \leq M_{\max}$ and p be the sampling probability. Furthermore, from the SVD of $\frac{1}{p}P_\Omega(\mathbf{R})$, we get a rank-$k$ approximation given by $[P_\Omega(\mathbf{R})]_k = \mathbf{U}^0\Sigma^0(\mathbf{V}^0)^T$. Then there exists numerical constants C and C' such that, with probability greater than $1 - \frac{1}{n^3}$, we have*

$$\frac{1}{\sqrt{mn}}\|\mathbf{M} - [P_\Omega(\mathbf{R})]_k\|_2 \leq CM_{\max}\left(\frac{m\alpha^{3/2}}{|\Omega|}\right)^{1/2} + \frac{C'm\sqrt{\alpha}}{|\Omega|}\|P_\Omega(\mathbf{N})\|_2.$$

B. Induction Proofs

Lemma 5. *Let $\mathbf{M}, \mathbf{N}, \Omega$ and $\mathbf{u}^{(t)}$ be defined as in Algorithm 2. Then, with high probability we have*

$$\|\mathbf{F}\|_2 \leq \sigma^\star\delta_2\sqrt{1 - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle^2}.$$

Proof. From the definition of $\mathbf{F}$ stated in Section IV-B, we have

$$\|\mathbf{F}\|_2 \leq \sigma^\star\|\mathbf{B}^{-1}\|_2\|\left(\mathbf{C} - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\mathbf{B}\right)\mathbf{v}^\star\|_2. \tag{29}$$

Since $\mathbf{B}$ is a diagonal matrix, $\|\mathbf{B}^{-1}\|_2 = \frac{1}{\min_i B_{ii}} = p \leq 1$. Let $\mathbf{x} \in \mathbb{R}^n$ be such that $\|\mathbf{x}\|_2 = 1$. Then, for all such $\mathbf{x}$,

$$\begin{aligned}
&\mathbf{x}^T\left(\mathbf{C} - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\mathbb{I}\right)\mathbf{v}^\star \\
&= \frac{1}{p}\sum_j x_j v_j^\star\left(\sum_{i|ij\in\Omega} u_i^{(t)}u_i^\star - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\right) \\
&\leq \frac{1}{p}\sum_{ij\in\Omega} x_j v_j^\star\left(u_i^{(t)}u_i^\star - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle(u_i^{(t)})^2\right) \\
&\overset{\zeta_1}{\leq} \frac{C}{p}\sqrt{np}\sqrt{\sum_j x_j^2(v_j^\star)^2}\sqrt{\sum_i\left(u_i^{(t)}u_i^\star - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle(u_i^{(t)})^2\right)^2} \\
&\overset{\zeta_2}{\leq} \frac{C}{p}\sqrt{np}\sqrt{\frac{\mu_1^2}{n}\sum_j x_j^2}\sqrt{\sum_i (u_i^{(t)})^2\left(u_i^\star - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle u_i^{(t)}\right)^2} \\
&\overset{\zeta_3}{\leq} \frac{C}{p}\frac{\sqrt{np}\mu_1^2}{\sqrt{mn}}\sqrt{1 - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle^2} \\
&\leq \delta_2\sqrt{1 - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle^2}, \qquad \text{if } \ p \geq C'\frac{\mu_1^4}{m\delta_2^2},
\end{aligned}$$

where $C' = C^2 > 0$ is a global constant and ζ_1 follows from Lemma 6 (which imposes the condition $p \geq C\frac{\log n}{m}$) and ζ_2 follows from the incoherence of $\mathbf{v}^\star$, and ζ_3 from that of $\mathbf{u}^{(t)}$. Then,

$$\begin{aligned}
\|\left(\mathbf{C} - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\mathbf{B}\right)\mathbf{v}^\star\|_2 &= \max_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\left(\mathbf{C} - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\mathbf{B}\right)\mathbf{v}^\star \\
&\leq \delta_2\sqrt{1 - \langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle^2}.
\end{aligned}$$

Hence, the lemma follows from (29) if $p \geq C'\frac{\mu_1^4\log n}{m\delta_2^2}$. □

Lemma 6. *([24]) Let $\Omega \in [m] \times [n]$ be a set of indices sampled uniformly at random with sampling probability p that satisfies $p \geq C\frac{\log n}{m}$. Then with probability $\geq 1 - \frac{1}{n^3}$ $\forall \mathbf{x} \in \mathbb{R}^m, \forall \mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{x}$ satisfies $\sum_i x_i = 0$, it holds that $\sum_{ij\in\Omega} x_i y_j \leq C\sqrt{\sqrt{mn}p}\|\mathbf{x}\|_2\|\mathbf{y}\|_2$, where $C > 0$ is a global constant.*

Lemma 7. *Let $\mathbf{M}, \mathbf{N}, \Omega$ and p_e be defined as before. Then with high probability it holds that*

$$\|N_{res}\|_2 \leq 2N_{\max}\mu_1 p_e\sqrt{mn}.$$

Proof. The proof follows from Lemma B.3, [28] for the case $k = 1$, and by noting the fact that $\|\mathbf{B}^{-1}\|_2 \leq 1$, and from the observation that $\frac{\|\mathcal{P}_\Omega(\mathbf{N})\|_2}{p} \leq 2N_{\max}p_e\sqrt{mn}$, with high probability (see Lemma 1). □

Lemma 8. *Let $\mathbf{F}$, N_{res} and $\mathbf{u}^{(t)}$ be defined as in (25). Then we have*

$$\|\hat{\mathbf{v}}^{(t+1)}\|_2 \geq \frac{1}{2m}\left(\sigma^\star\sqrt{1 - dist^2(\mathbf{u}^\star, \mathbf{u}^{(t)})} - \|\mathbf{F}\|_2 - \|N_{res}\|_2\right).$$

Proof.

$$\begin{aligned}
2m\|\hat{\mathbf{v}}^{(t+1)}\|_2 &= \|\sigma^\star\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\mathbf{v}^\star - \mathbf{F} + N_{\text{res}}\|_2 \\
&\overset{\zeta_1}{\geq} \|\sigma^\star\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\mathbf{v}^\star\|_2 - \|\mathbf{F} - N_{\text{res}}\|_2 \\
&\overset{\zeta_2}{\geq} \|\sigma^\star\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\mathbf{v}^\star\|_2 - \|\mathbf{F}\|_2 - \|N_{\text{res}}\|_2 \\
&= \sigma^\star\sqrt{1 - \text{dist}^2(\mathbf{u}^\star, \mathbf{u}^{(t)})^2} - \|\mathbf{F}\|_2 - \|N_{\text{res}}\|_2,
\end{aligned}$$

where both ζ_1 and ζ_2 follow from the reverse triangle inequality for vectors. □

Lemma 9. *Under the conditions of Theorem 1, with probability greater than $1 - 1/n^3$ it holds that for a given $j \in [n]$,*

$$\left|\frac{1}{B_{jj}}C_{jj}^N\Sigma_N\mathbf{V}_N^{(j)}\right| \leq N_{\max}\mu_1\sqrt{m}(p_e + \delta_2).$$

Proof. Let a Bernoulli random variable δ_{ij} characterize the event that the (i, j) entry in $\mathbf{R}$ is observed and is in error. Therefore, $\delta_{ij} = 1$ w.p. pp_e and 0 otherwise. Also, let us define $Z_i = \frac{1}{p}\delta_{ij}u_i^{(t)}N_{\max}$ and $Z = \sum_{i=1}^m Z_i$. Then,

$$\mathbb{E}[Z] = \mathbb{E}\left[\sum_{i=1}^m Z_i\right] = p_e\sum_{i=1}^m u_i^{(t)}N_{\max} \overset{\zeta_1}{\leq} N_{\max}p_e\mu_1\sqrt{m},$$

where ζ_1 follows from the incoherence of $u_i^{(t)}$. Moreover,

$$\begin{aligned}
\text{Var}(Z) &= \frac{p_e}{p}(1 - pp_e)N_{\max}^2\sum_{i=1}^m |u_i^{(t)}|^2 \\
&\leq N_{\max}^2\frac{p_e}{p}(1 - pp_e) \leq N_{\max}^2\frac{p_e}{p},
\end{aligned}$$

and $\max_i |Z_i| = \frac{1}{p}\max_i |u_i^{(t)} N_{ij}| \leq \frac{\mu_1 N_{\max}}{p\sqrt{m}}$. Using Bernstein's inequality, we have

$$\begin{aligned} &Pr\left(Z - \mathbb{E}[Z] > N_{\max}\mu_1\sqrt{m}\delta_2\right) \\ &\quad\leq \exp\left(-\frac{N_{\max}^2\mu_1^2 m\delta_2^2/2}{N_{\max}^2\frac{p_e}{p} + \frac{N_{\max}\mu_1}{3p\sqrt{m}}N_{\max}\mu_1\sqrt{m}\delta_2}\right) \\ &\quad= \exp\left(-\frac{p\mu_1^2 m\delta_2^2/2}{p_e + \frac{\mu_1^2\delta_2}{3}}\right) \overset{\zeta_2}{\leq} \exp\left(-3\log n\right) = \frac{1}{n^3}, \end{aligned}$$

where ζ_2 follows by using the condition from Theorem 1 that $p > \frac{6\log n}{\mu_1^2 m\delta_2^2}(p_e + \frac{\mu_1^2\delta_2}{3})$.

Therefore, using the definition from (24), with probability greater than $1 - 1/n^3$ it holds that

$$\begin{aligned} \left|\frac{1}{B_{jj}} C_{jj}^N \Sigma_N \mathbf{V}_N^{(j)}\right| &= \left|\frac{1}{p}\sum_{i|ij\in\Omega} u_i^{(t)} [\mathbf{U}_N^{(i)}]^T \Sigma_N \mathbf{V}_N^{(j)}\right| \\ &= \left|\frac{1}{p}\sum_{i|ij\in\Omega} u_i^{(t)} N_{ij}\right| \leq N_{\max}\mu_1\sqrt{m}(p_e + \delta_2). \end{aligned}$$

□

Proof of Lemma 3. We bound the largest magnitude of the entries of $\hat{\mathbf{v}}^{(t+1)}$ as follows. For every $j \in [n]$, using (24) we have

$$\begin{aligned} 2m\left|\hat{v}_j^{(t+1)}\right| &\leq |\sigma^\star\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle v_j^\star| \\ &+ \left|\frac{\sigma^\star}{B_{jj}}(\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle B_{jj} - C_{jj})v_j^\star\right| + \left|\frac{1}{B_{jj}} C_{jj}^N \Sigma_N \mathbf{V}_N^{(j)}\right| \\ &\overset{\zeta_1}{\leq} \sigma^\star\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\frac{\mu}{\sqrt{n}} + \sigma^\star\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\frac{\mu}{\sqrt{n}} \\ &+ \sigma^\star(\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle + \delta_2)\frac{\mu}{\sqrt{n}} + N_{\max}\mu_1\sqrt{m}(p_e + \delta_2) \\ &\overset{\zeta_2}{\leq} \sigma^\star(3 + \delta_2)\frac{\mu}{\sqrt{n}} + N_{\max}\sigma^\star\frac{\mu_1}{\sqrt{n}}(p_e + \delta_2) \\ &\overset{\zeta_3}{\leq} \sigma^\star\frac{\mu}{\sqrt{n}}(3 + \delta_2 + 8N_{\max}(p_e + \delta_2)) \end{aligned} \tag{30}$$

where ζ_1 follows from the fact that $|C_{jj}| \leq (|\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle| + \delta_2)$ (Lemma C.3, [24]), and Lemma 9, ζ_2 follows from $\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle \leq 1$, and ζ_3 follows by setting $\mu_1 = 8\mu$.

Furthermore, from Lemma 8 and using Lemma 5 and Lemma 7, we have

$$\begin{aligned} 2m\|\hat{\mathbf{v}}^{(t+1)}\| &\geq \sigma^\star\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle - \sigma^\star\delta_2\sqrt{1 - (\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle)^2} - \|N_{\text{res}}\|_2 \\ &\overset{\zeta_1}{\geq} \sigma^\star\langle\mathbf{u}^0, \mathbf{u}^\star\rangle - \sigma^\star\delta_2\sqrt{1 - (\langle\mathbf{u}^0, \mathbf{u}^\star\rangle)^2} \\ &- 2N_{\max}\mu_1 p_e\sqrt{mn} \\ &\overset{\zeta_2}{\geq} \sigma^\star\left(\frac{\sqrt{3}}{2} - \frac{\delta_2}{2} - 2N_{\max}\mu_1 p_e\right), \end{aligned} \tag{31}$$

where ζ_1 follows from $\text{dist}(\mathbf{u}^{(t)}, \mathbf{u}^\star) \leq \text{dist}(\mathbf{u}^0, \mathbf{u}^\star)$ and Lemma 7, ζ_2 follows from $\text{dist}(\mathbf{u}^0, \mathbf{u}^\star) \leq 1/2$ and $\sigma^\star = \sqrt{mn}$.

Using the two inequalities from (30) and (31), we have

$$\|\mathbf{v}^{(t+1)}\|_\infty = \frac{\|\hat{\mathbf{v}}^{(t+1)}\|_\infty}{\|\hat{\mathbf{v}}^{(t+1)}\|_2} = \frac{\sigma^\star\frac{\mu}{\sqrt{n}}(3 + \delta_2 + 8N_{\max}(p_e + \delta_2))}{\sigma^\star\left(\frac{\sqrt{3}}{2} - \frac{\delta_2}{2} - 2N_{\max}\mu_1 p_e\right)}.$$

From the condition on δ_2 as specified in Theorem 1, and since $\mu_1 = 8\mu$, we simplify the above equation as

$$\|\mathbf{v}^{(t+1)}\|_\infty \leq \frac{8\mu}{\sqrt{n}} = \frac{\mu_1}{\sqrt{n}}.$$

□

C. MEC proofs

Lemma 10. *Let $\mathbf{M}, \mathbf{N}, \mathbf{R}$ and Ω be defined as in Theorem 1. Let $\hat{\mathbf{M}}^{(T)}$ denote the estimate of matrix $\mathbf{M}$ after T iterations of Algorithm 2; furthermore, let $\tilde{\mathbf{M}}^{(T)} = sign(\hat{\mathbf{M}}^{(T)})$. Then $\tilde{\mathbf{M}}^{(T)}$ satisfies*

$$\frac{1}{\sqrt{mn}}\|P_\Omega(\mathbf{R} - \tilde{\mathbf{M}}^{(T)})\|_0 \leq \|P_\Omega(\mathbf{N})\|_F + \|P_\Omega(\mathbf{M} - \hat{\mathbf{M}}^{(T)})\|_F.$$

Proof. Clearly, $\forall i, j \in \Omega \subseteq [m] \times [n]$ it holds that

$$R_{ij}\begin{cases} = \tilde{M}_{ij}^{(T)} & \text{if } |R_{ij} - \hat{M}_{ij}^{(T)}| \leq 1, \\ \neq \tilde{M}_{ij}^{(T)} & \text{otherwise.} \end{cases}$$

Now, $\|P_\Omega(\mathbf{R} - \tilde{\mathbf{M}}^{(T)})\|_0$ denotes the number of non-zero entries among the observed entries of the difference matrix $\mathbf{R} - \tilde{\mathbf{M}}^{(T)}$. In other words,

$$\begin{aligned} \|P_\Omega(\mathbf{R} - \tilde{\mathbf{M}}^{(T)})\|_0 &= \left|\left\{i, j \in \Omega \text{ s.t. } R_{ij} \neq \tilde{M}_{ij}^{(T)}\right\}\right| \\ &\leq \left|\left\{i, j \in \Omega \text{ s.t. } |R_{ij} - \hat{M}_{ij}^{(T)}| > 1\right\}\right| \\ &\leq \sum_{ij\in\Omega}\left|R_{ij} - \hat{M}_{ij}^{(T)}\right|, \end{aligned}$$

where the last quantity is the entry-wise ℓ_1-norm of the matrix $P_\Omega(\mathbf{R} - \hat{M}^{(T)})$, denoted by $\|P_\Omega(\mathbf{R} - \hat{M}^{(T)}\|_1$. Therefore,

$$\begin{aligned} &\frac{1}{\sqrt{mn}}\|P_\Omega(\mathbf{R} - \tilde{\mathbf{M}}^{(T)})\|_0 \\ &\quad\leq \frac{1}{\sqrt{mn}}\|P_\Omega(\mathbf{R} - \mathbf{M} + \mathbf{M} - \hat{\mathbf{M}}^{(T)})\|_1 \\ &\quad\leq \frac{1}{\sqrt{mn}}\left(\|P_\Omega(\mathbf{R} - \mathbf{M})\|_1 + \|P_\Omega(\mathbf{M} - \hat{\mathbf{M}}^{(T)})\|_1\right) \\ &\quad\leq \|P_\Omega(\mathbf{N})\|_F + \|P_\Omega(\mathbf{M} - \hat{\mathbf{M}}^{(T)})\|_F. \end{aligned}$$

□

Proof of Theorem 2.

$$\begin{aligned} &\text{dist}(\mathbf{v}^{(t+1)}, \mathbf{v}^\star) \\ &= \|\mathbb{P}_{\mathbf{v}_\perp^\star}(\hat{\mathbf{v}}^{t+1})\|_2/\|\hat{\mathbf{v}}^{(t+1)}\|_2 \\ &= \frac{1}{2m}\|\mathbb{P}_{\mathbf{v}_\perp^\star}(\sigma^\star\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle\mathbf{v}^\star - \mathbf{F} + N_{\text{res}})\|_2/\|\hat{\mathbf{v}}^{(t+1)}\|_2 \\ &\overset{\zeta_1}{=} \frac{1}{2m}\|\mathbb{P}_{\mathbf{v}_\perp^\star}(-\mathbf{F} + N_{\text{res}})\|_2/\|\hat{\mathbf{v}}^{(t+1)}\|_2 \\ &\overset{\zeta_2}{\leq} \frac{1}{2m}\|\mathbf{F} + N_{\text{res}}\|_2/\|\hat{\mathbf{v}}^{(t+1)}\|_2 \\ &\overset{\zeta_3}{\leq} \frac{1}{2m}\left(\|\mathbf{F}\|_2 + \|N_{\text{res}}\|_2\right)/\|\hat{\mathbf{v}}^{(t+1)}\|_2 \\ &\overset{\zeta_4}{\leq} \frac{\sigma^\star\delta_2\sqrt{1 - (\langle\mathbf{u}^{(t)}, \mathbf{u}^\star\rangle)^2} + 2N_{\max}p_e\mu_1\sqrt{mn}}{\sigma^\star\sqrt{1 - \text{dist}^2(\mathbf{u}^{(t)}, \mathbf{u}^\star)} - \delta_2\text{dist}(\mathbf{u}^{(t)}, \mathbf{u}^\star) - 2N_{\max}p_e\mu_1\sqrt{mn}} \end{aligned}$$

$$\overset{\zeta_5}{\leq} \frac{\sigma^\star\left(\delta_2 \text{dist}(\mathbf{u}^{(t)},\mathbf{u}^\star)+2N_{\max}p_e\mu_1\right)}{\sigma^\star\left(\frac{\sqrt{3}}{2}-\frac{\delta_2}{2}-2N_{\max}p_e\mu_1\right)}$$
$$\overset{\zeta_6}{\leq} \frac{1}{4}\text{dist}(\mathbf{u}^{(t)},\mathbf{u}^\star)+\frac{\mu_1 p_e}{\delta_2},$$

where ζ_1 follows since the first term in the numerator is orthogonal to $\mathbf{v}^\star$, ζ_2 follows since $\forall \mathbf{x},\mathbf{y}, \|\mathbb{P}_{\mathbf{y}}(\mathbf{x})\|_2 \leq \|\mathbf{x}\|_2$, ζ_3 follows from the triangle inequality, ζ_4 follows by using Lemma 5 and Lemma 7, ζ_5 follows from the fact that $\sigma^\star = \sqrt{mn}$ and $\text{dist}(\mathbf{u}^{(t)},\mathbf{u}^\star) \leq \text{dist}(\mathbf{u}^0,\mathbf{u}^\star) \leq 1/2$, and finally ζ_6 follows by using the condition on δ_2 from Theorem 1. Using similar arguments, we show that $\text{dist}(\mathbf{u}^{(t+1)},\mathbf{u}^\star) \leq \frac{1}{4}\text{dist}(\mathbf{v}^{(t+1)},\mathbf{v}^\star)+\frac{\mu_1 p_e}{\delta_2}$. □

Proof of Corollary 2. It follows from Theorem 2 that, after T iterations,

$$\begin{aligned}
\text{dist}(\hat{\mathbf{u}}^{(T)},\mathbf{u}^\star) &\leq \frac{1}{4}\left(\frac{1}{4}\text{dist}(\hat{\mathbf{u}}^{(T-1)},\mathbf{u}^\star)+\frac{\mu_1 p_e}{\delta_2}\right)+\frac{\mu_1 p_e}{\delta_2}\\
&= \frac{1}{16}\text{dist}(\hat{\mathbf{u}}^{(T-1)},\mathbf{u}^\star)+\frac{5\mu_1 p_e}{4\delta_2}\\
&\vdots\\
&\leq \frac{1}{16^T}\text{dist}(\hat{\mathbf{u}}^{(0)},\mathbf{u}^\star)+\frac{5\mu_1 p_e}{4\delta_2}\left(1+\frac{1}{16}\frac{1}{16^2}+\right.\\
&\quad \left.+\cdots(\text{T terms})\right)\\
&\overset{\zeta_1}{\leq} \frac{1}{2}\frac{1}{16^T}+\frac{4\mu_1 p_e}{3\delta_2}\\
&\leq \frac{1}{2}\frac{\epsilon}{\|\mathbf{M}\|_F}+\frac{4\mu_1 p_e}{3\delta_2}, \text{ as } T=\mathcal{O}\left(\log\frac{\|\mathbf{M}\|_F}{\epsilon}\right)
\end{aligned}$$

where ζ_1 follows from the fact that $\text{dist}(\hat{\mathbf{u}}^{(0)},\mathbf{u}^\star) \leq 1/2$ (see Lemma 2). □

Proof of Theorem 1. In order to prove this theorem, firstly we need to bound the error between the true matrix $\mathbf{M}$ and the output of Algorithm 2 prior to the rounding step. Let us denote the latter as the scaled estimate $\hat{\mathbf{M}}^{(T)}$, where we have $\tilde{\mathbf{M}}^{(T)} = \text{sign}(\hat{\mathbf{M}}^{(T)})$. By using (25), the difference between $\mathbf{M}$ and (appropriately scaled) $\hat{\mathbf{M}}^{(T)}$ is

$$\begin{aligned}
\mathbf{M}-\hat{\mathbf{M}}^{(T)} &= \mathbf{M}-\mathbf{u}^{(T)}\left[\sigma^\star\langle\mathbf{u}^{(T)},\mathbf{u}^\star\rangle\mathbf{v}^\star-\mathbf{F}+N_{\text{res}}\right]^T\\
&= \mathbf{M}-\mathbf{u}^{(T)}(\mathbf{u}^{(T)})^T\mathbf{u}^\star\sigma^\star(\mathbf{v}^\star)^T+\mathbf{u}^{(T)}\mathbf{F}^T\\
&\quad -\mathbf{u}^{(T)}N_{\text{res}}^T\\
&= \left(\mathbb{I}-\mathbf{u}^{(T)}(\mathbf{u}^{(T)})^T\right)\mathbf{M}+\mathbf{u}^{(T)}\mathbf{F}^T-\mathbf{u}^{(T)}N_{\text{res}}^T.
\end{aligned}$$

Using the fact that $\|\mathbf{v}^\star\|_2 = 1$, $\|\mathbf{u}^{(T)}\|_2 = 1$, and using Lemma 5, Lemma 7, and Corollary 2, we have

$$\begin{aligned}
&\|\mathbf{M}-\hat{\mathbf{M}}^{(T)}\|_F\\
&\leq \|\left(\mathbb{I}-\mathbf{u}^{(T)}(\mathbf{u}^{(T)})^T\right)\mathbf{u}^\star\sigma^\star\|_2+\|\mathbf{F}\|_2+\|N_{\text{res}}\|_2\\
&\leq \sigma^\star\text{dist}(\mathbf{u}^{(T)},\mathbf{u}^\star)+\sigma^\star\delta_2\text{dist}(\mathbf{u}^{(T)},\mathbf{u}^\star)+2\sigma^\star N_{\max}p_e\mu_1\\
&= \sigma^\star(1+\delta_2)\left(\frac{\epsilon}{2\|\mathbf{M}\|_F}+\frac{4\mu_1 p_e}{3\delta_2}\right)+2\sigma^\star N_{\max}p_e\mu_1\\
&\leq \frac{\epsilon(1+\delta_2)}{2}+2\sigma^\star p_e\mu_1\left(\frac{(3N_{\max}+2)\delta_2+2}{3\delta_2}\right).
\end{aligned}$$

The theorem then follows by setting $\epsilon' = \frac{\epsilon(1+\delta_2)}{2}$ and substituting the value of μ_1. □

Proof of Corollary 1. Using Lemma 10 and Theorem 1, and noting that $\sigma^\star = \sqrt{mn}$, the normalized minimum error correction score $\tilde{\text{MEC}} = \frac{1}{mn}\|P_\Omega(\mathbf{R}-\tilde{\mathbf{M}}^{(T)})\|_0$ can be bounded as

$$\begin{aligned}
\tilde{\text{MEC}} &\leq \frac{1}{\sqrt{mn}}\left(\|P_\Omega(\mathbf{M}-\hat{\mathbf{M}}^{(T)})\|_F+\|P_\Omega(\mathbf{N})\|_F\right)\\
&\leq \epsilon'+\frac{2p_e\mu_1\sigma^\star}{3\delta_2\sqrt{mn}}(2+(2+3N_{\max})\delta_2)+\frac{1}{\sqrt{mn}}\|P_\Omega(\mathbf{N})\|_F\\
&= \epsilon'+\frac{2p_e\mu_1}{3\delta_2}(2+(2+3N_{\max})\delta_2)+\frac{1}{\sqrt{mn}}\|P_\Omega(\mathbf{N})\|_F.
\end{aligned}$$

□

D. Characterizing the noise matrix

Proof of Lemma 1. Noise matrix $\mathbf{N}$ clearly follows the worst case noise model as described in [25] since $\forall i,j \in [m]\times[n], |N_{ij}| \leq N_{\max}$. Let $\Omega' \subseteq \Omega$ be the set of indices where a sequencing error has occurred, i.e., $\forall (i,j)\in\Omega', N_{ij}\neq 0$. Define δ_{ij} to be a random variable indicating the membership of the index (i,j) in Ω', i.e., $\delta_{ij}=1$ if $(i,j)\in\Omega'$, 0 otherwise. Since sampling and error occur independently, the probability that $\delta_{ij}=1$ is pp_e. Therefore, $|\Omega'| \approx \mathbb{E}\left[\sum_{ij}\delta_{ij}\right] = mnpp_e$ w.h.p. Using Theorem 4 below, we conclude that

$$\frac{\|P_\Omega(\mathbf{N})\|_2}{p} \leq 2\frac{|\Omega'|\sqrt{m}}{pm\sqrt{n}}N_{\max} = 2N_{\max}p_e\sqrt{mn}.$$

□

Theorem 4 ([25]). *If $\mathbf{N}\in\mathbb{R}^{m\times n}$ $(m\leq n)$ is a matrix with entries chosen from the worst case model, i.e., $|N_{ij}|\leq N_{\max}$ $\forall(i,j)$ for some constant $N_{\max}$, then for a sample set Ω drawn uniformly at random, it holds that*

$$\|P_\Omega(\mathbf{N})\|_2 \leq \frac{2|\Omega|\sqrt{m}}{m\sqrt{n}}N_{\max}.$$

ACKNOWLEDGMENT

This work was funded in part by the NSF grant CCF 1618427.

REFERENCES

[1] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth *et al.*, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," in *Nature*, vol. 409(6822), 2001, pp. 928–933.

[2] A. G. Clark, "The role of haplotypes in candidate gene studies," in *Genetic epidemiology*, vol. 27(4), 2004, pp. 321–333.

[3] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, and P. K. H. Tam, "The international hapmap project," in *Nature*, vol. 426(6968), 2003, pp. 789–796.

[4] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner *et al.*, "Detecting recent positive selection in the human genome from haplotype structure," in *Nature*, vol. 419(6909), 2002, pp. 832–837.

[5] G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz, "Snps problems, complexity, and algorithms," in *AlgorithmsESA*. Springer Berlin Heidelberg, 2001, pp. 182–193.

[6] R. Cilibrasi, L. Van Iersel, S. Kelk, , and J. Tromp, "On the complexity of several haplotyping problems," in *Algorithms in bioinformatics*. Springer Berlin Heidelberg, 2005, pp. 128–139.

[7] R. S. Wang, L. Y. Wu, Z. P. Li, and X. S. Zhang, "Haplotype reconstruction from snp fragments by minimum error correction," in *Bioinformatics*, vol. 21(10), 2005, pp. 2456–2462.

[8] S. Das and H. Vikalo, "Optimal haplotype assembly via a branch-and-bound algorithm," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2016.

[9] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov *et al.*, "The diploid genome sequence of an individual human," *PLoS Biol*, vol. 5, no. 10, p. e254, 2007.

[10] V. Bansal and V. Bafna, "Hapcut: an efficient and accurate algorithm for the haplotype assembly problem," *Bioinformatics*, vol. 24, no. 16, pp. i153–i159, 2008.

[11] V. Bansal, A. L. Halpern, N. Axelrod, and V. Bafna, "An mcmc algorithm for haplotype assembly from whole-genome sequence data," in *Genome research*, vol. 18(8), 2008, pp. 1336–1346.

[12] J. Duitama, G. K. McEwen, T. Huebsch, S. Palczewski, S. Schulz, K. Verstrepen *et al.*, "Fosmid-based whole genome haplotyping of a hapmap trio child: evaluation of single individual haplotyping techniques," in *Nucleic acids research*, 2011, p. gkr1042.

[13] D. Aguiar and S. Istrail, "Hapcompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data," *Journal of Computational Biology*, vol. 19, no. 6, pp. 577–590, 2012.

[14] S. Das and H. Vikalo, "Sdhap: haplotype assembly for diploids and polyploids via semi-definite programming," *BMC genomics*, vol. 16, no. 1, p. 260, 2015.

[15] Z. Puljiz and H. Vikalo, "Decoding genetic variations: Communications-inspired haplotype assembly," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 13, no. 3, pp. 518–530, 2016.

[16] V. Kuleshov, "Probabilistic single-individual haplotyping," in *Bioinformatics*, vol. 30(17), 2014, pp. i379–i385.

[17] H. Matsumoto and H. Kiryu, "Mixsih: a mixture model for single individual haplotyping." in *BMC Genomics*, vol. 14(Suppl 2) S5, 2013.

[18] C. Cai, S. Sanghavi, and H. Vikalo, "Structured low-rank matrix factorization for haplotype assembly," in *Journal of Selected Topics in Signal Processing*, vol. 10.4. IEEE, 2016, pp. 647–657.

[19] S. Barik and H. Vikalo, "Binary matrix completion with performance guarantees for single individual haplotyping," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2716–2720.

[20] Netflix, "Netflix prize," in *http://www.netflixprize.com/*. [Onine].

[21] H. Kim and H. Park, "Nnonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," in *SIAM J. Matrix Anal. Appl.*, vol. 30(2), 2008, pp. 713–730.

[22] E. Candès and B. Recht, "Exact matrix completion via convex optimization," in *Foundations of Computational mathematics*, vol. 9.6, 2009, pp. 717–772.

[23] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," in *SIAM Review*, vol. 52(3), 2010, pp. 471–501.

[24] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *ArXiv e-prints*, 2012.

[25] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," in *Journal of Machine Learning Research*, vol. 11, 2010, pp. 2057–2078.

[26] R. Meka, P. Jain, C. Caramanis, and I. S. Dhillon, "Rank minimization via online learning," in *Proceedings of the 25th International Conference on Machine learning*. ACM, 2008, pp. 656–663.

[27] M. Hardt, "Understanding alternating minimization for matrix completion," in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 2014, pp. 651–660.

[28] S. Gunasekar, A. Acharya, N. Gaur, and J. Ghosh, "Noisy matrix completion using alternating minimization," in *ECML PKDD*. Springer, 2013, pp. 194–209.

[29] A. J. v. d. Veen, "Analytical method for blind binary signal separation," in *IEEE Signal Processing*, vol. 45, 1997, pp. 1078–1082.

[30] R. Liao, J. C.and Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury, "Network component analysis: reconstruction of regulatory signals in biological systems," in *Proceedings of the National Academy of Sciences*, 2003, pp. 15 522–15 527.

[31] A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. J. Mooney, "Model-based overlapping clustering," in *SIGKDD*. ACM, 2005, pp. 532–537.

[32] A. I. Schein, L. K. Saul, and L. H. Ungar, "A generalized linear model for principal component analysis of binary data." in *AISTATS*, vol. 3(9), 2003, p. 10.

[33] A. Kabán and E. Bingham, "Factorisation and denoising of 0–1 data: a variational approach." in *Neurocomputing*, vol. 71(10), 2008, pp. 2291–2308.

[34] E. Meeds, Z. Ghahramani, R. M. Neal, and S. T. Roweis, "Modeling dyadic data with binary latent factors." in *NIPS*, vol. 3(9), 2006, pp. 977–984.

[35] Z. Zhang, T. Li, C. Ding, and X. Zhang, "Binary matrix factorization with applications." in *ICDM*. IEEE, 2007, pp. 391–400.

[36] M. Slawski, M. Hein, and P. Lutsik, "Matrix factorization with binary components," in *Advances in Neural Information Processing Systems*, 2013, pp. 3210–3218.

[37] G. H. Golub and C. Van Loan, "Matrix computations," vol. 3. JHU Press, 2012.

[38] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[39] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," *arXiv preprint arXiv:1207.3907*, 2012.

[40] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[41] P. Jain and P. Netrapalli, "Fast exact matrix completion with finite samples." in *COLT*, 2015, pp. 1007–1034.

[42] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[43] R. Schwartz *et al.*, "Theory and algorithms for the haplotype assembly problem," *Communications in Information & Systems*, vol. 10, no. 1, pp. 23–38, 2010.

[44] F. Geraci, "A comparison of several algorithms for the single individual snp haplotyping reconstruction problem," in *Bioinformatics*, vol. 26(18), 2010, pp. 2217–2225.

[45] X. Chen, Q. Peng, L. Han, T. Zhong, and T. Xu, "An effective haplotype assembly algorithm based on hypergraph partitioning," in *Journal of theoretical biology*, vol. 358, 2014, pp. 85–92.

[46] Z. Chen, B. Fu, R. Schweller, B. Yang, Z. Zhao, and B. Zhu, "Linear time probabilistic algorithms for the singular haplotype reconstruction problem from snp fragments," *Journal of Computational Biology*, vol. 15, no. 5, pp. 535–546, 2008.

[47] L. M. Genovese, F. Geraci, and M. Pellegrini, "A fast and accurate heuristic for the single individual snp haplotyping problem with many gaps, high reading error rate and low coverage," in *International Workshop on Algorithms in Bioinformatics*. Springer, 2007, pp. 49–60.

[48] A. Panconesi and M. Sozio, "Fast hare: A fast heuristic for single individual snp haplotype reconstruction," in *International Workshop on Algorithms in Bioinformatics*. Springer, 2004, pp. 266–277.