Supplement for Sparse Tensor Decomposition for
Haplotype Assembly of Diploids and Polyploids

Abolfazl Hashemi\textsuperscript{1*}, Banghua Zhu\textsuperscript{2} and Haris Vikalo\textsuperscript{1}

\textbf{Derivation of the proposed step size}

The proof relies on the results of Theorem 1 in [1]. For the proposed algorithm to converge, it must hold that

\[ f(U_{t+1}, V_{t+1}) \leq f(U_t, V_t), \quad (1) \]

\forall t. First, by noting (9) it holds that \( f(U_{t+1}, V_t) \leq f(U_t, V_t) \). It now remains to show that \( f(U_{t+1}, V_{t+1}) \leq f(U_{t+1}, \hat{V}_t) \). First, for the sake of notations and clarity, define

\[ \text{Tr}_\Omega(A^T B) = \text{Tr} \left( P_\Omega(A^T) P_\Omega(B) \right) = \sum_{(i,j) \in \Omega} A_{ij} B_{ij}. \quad (2) \]

Recall that \( \nabla f(\hat{V}_t) = -(P_\Omega(\hat{R} - U_{t+1}V_t^\top))V_{t+1} \) and \( V_{t+1} = P_\Omega(V_{t+1}) \) where \( V_{t+1} = V_t - \alpha \nabla f(V_t) \). Since \( P_\Omega \) is a projection onto a convex set of constraints, \( f(U_{t+1}, V_{t+1}) \leq f(U_{t+1}, \hat{V}_t) \). Thus, it remains to show \( f(U_{t+1}, \hat{V}_{t+1}) \leq f(U_{t+1}, \hat{V}_t) \). Given that,

\[ f(U_{t+1}, \hat{V}_{t+1}) - f(U_{t+1}, \hat{V}_t) \]

\begin{align*}
&= \frac{1}{2} \left\| P_\Omega(\hat{R} - U_{t+1}V_t^\top) + \alpha P_\Omega(U_{t+1} \nabla f(V_t)^\top) \right\|_F^2 \\
&\quad - \frac{1}{2} \left\| P_\Omega(\hat{R} - U_{t+1}V_t^\top) \right\|_F^2 \\
&= \alpha \text{Tr} \left( P_\Omega(\hat{R} - U_{t+1}V_t^\top) V_{t+1} \right) + P_\Omega(U_{t+1} \nabla f(V_t)^\top) \\
&\quad + \frac{1}{2} \left\| P_\Omega(U_{t+1} \nabla f(V_t)^\top) \right\|_F^2. \quad (3)
\end{align*}

Now, consider the first term in the last line of (19). Following straightforward linear algebra we obtain

\begin{align*}
- \text{Tr} \left( P_\Omega(\hat{R} - U_{t+1}V_t^\top) V_{t+1} \right) &\mathbf{T} P_\Omega(U_{t+1} \nabla f(V_t)^\top) \\
= - \text{Tr}(\hat{R} - U_{t+1}V_t^\top) \text{Tr}(U_{t+1} \nabla f(V_t)^\top) \\
= \text{Tr}(\hat{R} - U_{t+1}V_t^\top) \text{Tr}(U_{t+1} \nabla f(V_t)^\top) \\
= \text{Tr}(\hat{R} - U_{t+1}V_t^\top) \text{Tr}(U_{t+1} \nabla f(V_t)^\top) \\
&= \| V_{t+1} \|^2_F - \| V_t \|^2_F.
\end{align*}

Therefore,

\[ f(U_{t+1}, \hat{V}_{t+1}) - f(U_{t+1}, \hat{V}_t) \]

\begin{align*}
&= \frac{\alpha^2}{2} \left\| P_\Omega(U_{t+1} \nabla f(V_t)^\top) \right\|_F^2 - \alpha \| \nabla f(V_t)^\top \|_F^2 \\
&= \left( \frac{C^2}{2} - C \right) \frac{\| \nabla f(V_t)^\top \|_F^2}{\| P_\Omega(U_{t+1} \nabla f(V_t)^\top) \|_F^2}. \quad (5)
\end{align*}

where the last equality follows according to the step size in (11). Clearly if \( C \in (0, 2) \) it must hold that \( f(U_{t+1}, \hat{V}_{t+1}) \leq f(U_{t+1}, \hat{V}_t) \), which in turn implies convergence.

\textbf{Derivation of the MEC and CPR bounds}

Recall that under conditions of Theorem 4.1, with probability \( 1 - \frac{1}{m^2} \) it holds that \( \| M - U^* V^\top \|_F^2 \leq \frac{C_1 \kappa^2 \rho \kappa m}{2 \epsilon_{\text{avg}}} \). Once the stationary point \( \hat{M} = U^* V^\top \) is found, Althap performs a decoding (rounding) step in order to obtain the binary solution \( M = U^* V^\top \). In this rounding procedure Althap first normalizes all unfolded fibers such that sum of entries of each fiber equals 1. Then Althap sets the largest entry of each unfolded fiber to 1 and the remaining three entries to 0. Eventually, Althap reshapes the solution \( \hat{M} \) to the tensor \( M \). Note that this normalization is not required and we only consider it for the analysis purposes. Therefore, it is required to establish a bound on \( \| \hat{M} - M \|_F^2 \). Let \( \mathcal{L} \) be the set of mismatching fibers, i.e., \( \forall f \in \mathcal{L}, \| \hat{M}_f - M_f \|_2^2 \neq 0 \). First, notice that \( \forall f \in \mathcal{L}, \| \hat{M}_f - M_f \|_2^2 \geq \| M - \hat{M} \|_F^2 = 0 \). In addition, consider a fiber \( \forall f \in \mathcal{L} \). The minimum value of \( \| M_f - \hat{M}_f \|_2^2 \) occurs when two entries in \( f \) are both equal to 0.5 and the remaining two entries are both 0. Hence, it becomes clear that \( \| M_f - \hat{M}_f \|_2^2 \geq 0.5 \). Thus, with probability \( 1 - \frac{1}{m^2} \) it holds that

\[ \frac{1}{2} \| \hat{M} - M \|_F^2 \leq \| M - U^* V^\top \|_F^2 \leq \frac{C_1 \kappa^2 \rho \kappa m}{2 \epsilon_{\text{avg}}} \quad (6) \]

which is the desired relation. We now establish the MEC bound. Using the linearity of expectation and the above
discussion, we obtain
\[
\frac{1}{2} \mathbb{E}\{\|P_Ω(\mathbf{R} - \mathbf{M})\|_F^2\}
\]
\[
= \frac{1}{2} \mathbb{E}\{\|P_Ω(\mathbf{R} - \hat{\mathbf{M}}) + P_Ω(\hat{\mathbf{M}} - \mathbf{M})\|_F^2\}
\]
\[
\leq \mathbb{E}\{\|P_Ω(\mathbf{R} - \hat{\mathbf{M}})\|_F^2\} + \mathbb{E}\{\|P_Ω(\hat{\mathbf{M}} - \mathbf{M})\|_F^2\}
\]
\[
= \mathbb{E}\{\|P_Ω(\mathbf{N})\|_F^2\} + p\|\hat{\mathbf{M}} - \mathbf{M}\|_F^2
\]
\[
\leq \mathbb{E}\{\|P_Ω(\mathbf{N})\|_F^2\} + 2C_1\kappa^4p_kk
\]
\[
= 2p_emC_{seq} + 2C_1\kappa^4p_kk.
\]
Thus, \(\mathbb{E}\{\text{MEC}\} \leq 2p_e(C_{seq}m + \kappa^4C_1k)\). We now establish the CPR bound. The following is an equivalent definition of CPR computed using unfolded tensors of the true and the reconstructed haplotype sequences,
\[
\text{CPR} = 1 - \frac{1}{2mk} \min_{\mathbf{M}} \mathbb{E}\{\|\mathbf{V} - \mathbf{M}(\hat{\mathbf{V}})\|_F^2\},
\]
where \(\mathbf{M}\) is a one-to-one mapping from the corresponding entries of the lateral slices of \(\hat{\mathbf{V}}\) to those of \(\mathbf{V}\). Assuming that sequencing reads uniformly sample haplotype sequences, on average, the mismatches between \(\hat{\mathbf{V}}\) and \(\mathbf{V}\) contribute equally to the number of mismatches between \(\hat{\mathbf{M}}\) and \(\mathbf{M}\). That is, \(\frac{1}{2} \mathbb{E}\{\|\mathbf{M} - \hat{\mathbf{M}}\|_F^2\} = \frac{4}{2n} \mathbb{E}\{\|\mathbf{V} - \hat{\mathbf{V}}\|_F^2\}\). Therefore,
\[
\mathbb{E}\min_{\mathbf{M}} \|\mathbf{V} - \mathbf{M}(\hat{\mathbf{V}})\|_F^2 \leq \mathbb{E}\{\|\mathbf{V} - \hat{\mathbf{V}}\|_F^2\}
\]
\[
= \frac{k}{n} \mathbb{E}\|\mathbf{M} - \hat{\mathbf{M}}\|_F^2
\]
\[
\leq \frac{2C_1\kappa^4p_kk^2m}{nC_{snp}}.
\]
Thus, \(\mathbb{E}\{\text{CPR}\} \geq 1 - \frac{C_1\kappa^4p_kk}{nC_{snp}}\) which is the desired bound.

Author details
1 Department of ECE, University of Texas at Austin, Austin, Texas, USA. 2 EE Department, Tsinghua University, Beijing, China.

References