A Truncation Algorithms for the IHT

In this section, we present the algorithms for the truncation step in the IHT algorithm, i.e., Line 6 in Algorithm 1. The truncation algorithms for sparse vector and low-rank matrix recovery are tabulated in Algorithms 2 and 3, respectively.

Algorithm 2 Evaluation of the truncation function Trunc($\Theta, s$) in Line 6 of Algorithm 1 for nonlinear sparse vector recovery

1: **Input:** Truncation level $s > 0$, a vector $\Theta \in \mathbb{R}^d$
2: Sort $\{|\Theta_j|\}_{j=1}^d$ such that $|\Theta_{j_1}| \geq |\Theta_{j_2}| \geq \ldots \geq |\Theta_{j_d}|$
3: $S \leftarrow \{j_1, \ldots, j_s\}$
4: for $j$ in $\{1, \ldots, d\}$ do
5: $\Theta_j \leftarrow 0$ if $j \notin S$
6: end for
7: return $\Theta$

Algorithm 3 Evaluation of the truncation function Trunc($\Theta, s$) in Line 6 of Algorithm 1 for nonlinear low-rank matrix recovery

1: **Input:** Truncation level $s > 0$, a low rank matrix $\Theta \in \mathbb{R}^{m_1 \times m_2}$ with rank($\Theta$) = $r$
2: Perform singular value decomposition $\Theta = U \Lambda V^\top$ where $U \in \mathbb{R}^{m_1 \times r}$, $\Lambda \in \mathbb{R}^{r \times r}$, $V \in \mathbb{R}^{r \times m_2}$. The diagonal elements of $\Lambda$ are in decreasing order
3: for $j$ in $\{1, \ldots, r\}$ do
4: $\Lambda_{jj} \leftarrow 0$ if $j > s$
5: end for
6: $\Theta \leftarrow U \Lambda V^\top$
7: return $\Theta
B Proof of the Main Results

In this section we give a detailed proof of the main results.

B.1 Proof of Theorem 3.5

Proof. We first define \( \text{supp}(V) := \{ j : v_j \neq 0 \} \), for any \( V = (v_1, \cdots, v_d) \top \in \mathbb{R}^d \) as the support of \( V \). When proceeding Algorithm 1 for sparse vector recovery, for \( t \geq 0 \), we denote \( S(t) = \text{supp}(\Theta(t)) \), \( S^* = \text{supp}(\Theta^*) \) and \( F(t) = S(t) \cup S(t+1) \cup S^* \). By definition, the cardinality of \( F(t) \) is no larger than \( 2s + s^* \), i.e., \( \| \Theta_F(t) \|_0 \leq 2s + s^* \).

We denote \( \Theta(t+1) = \Theta(t) - \eta \cdot \nabla F(t)\ell(\Theta(t)) \), then by definition we have \( \Theta(t+1) = \text{Trunc}(\Theta(t+1), s) \). By denoting \( \Delta(t) = \Theta(t) - \Theta^* \), we have

\[
\| \Theta(t+1) - \Theta^* \|_2 = \| \Delta(t) - \eta \cdot \nabla F(t)\ell(\Theta(t)) \|_2. \tag{B.1}
\]

According to the definition of the least-squares loss function in (1.2), \( \nabla \ell(\Theta(t)) \) is represented as:

\[
\nabla \ell(\Theta(t)) = \frac{1}{n} \sum_{i=1}^{n} \left[ f'(\langle X_i, \Theta(t) \rangle) \cdot Y_i - f'(\langle X_i, \Theta^* \rangle) \cdot E_i \right].
\]

(B.2)

where \( E(t) \) and \( G(t) \) are defined as

\[
E(t) := \frac{1}{n} \sum_{i=1}^{n} \left[ f'(\langle X_i, \Theta(t) \rangle) - f'(\langle X_i, \Theta^* \rangle) \right] \cdot E_i \quad \text{and} \quad G(t) := \frac{1}{n} \sum_{i=1}^{n} f'(\langle X_i, \Theta(t) \rangle) \cdot \left[ f(\langle X_i, \Theta(t) \rangle) - f(\langle X_i, \Theta^* \rangle) \right] \cdot X_i.
\]

(B.3)

Triangle inequality yields that

\[
\| \Theta(t+1) - \Theta^* \|_2 \leq \| \Delta(t) - \eta \cdot G(t) \|_2 + \eta \cdot \| E(t) \|_2. \tag{B.5}
\]

Using Mean Value Theorem, \( G(t) \) can be written as

\[
G(t) = \frac{1}{n} \sum_{i=1}^{n} f'(\langle X_i, \Theta(t) \rangle) \cdot f'(\langle X_i, \Theta_1 \rangle) \cdot (X_i, \Delta(t)) \cdot X_i = A(t) \cdot \Delta(t),
\]

(B.6)

where \( \Theta_1 \) lies between \( \Theta(t) \) and \( \Theta^* \) and \( A(t) = \frac{1}{n} \sum_{i=1}^{n} f'(\langle X_i, \Theta(t) \rangle) \cdot f'(\langle X_i, \Theta_1 \rangle) \cdot X_i X_i \top \). Since \( \text{supp}(G(t)) \subseteq F(t) \), we actually have

\[
G(t) = A(t) \cdot \Delta(t).
\]

(B.7)

Therefore, combining (B.5), (B.6), and (B.7), we obtain

\[
\| \Theta(t+1) - \Theta^* \|_2 \leq \| \Delta(t) - \eta \cdot G(t) \|_2 + \eta \cdot \| E(t) \|_2 \leq \| I - \eta \cdot A(t) \|_2 \cdot \| \Delta(t) \|_2 + \eta \cdot \| E(t) \|_2, \tag{B.8}
\]

where \( I \in \mathbb{R}^{|F(t)| \times |F(t)|} \) is the identity matrix. Here the second inequality follows from the definition of operator norm. Since \( |F(t)| \leq 2s + s^* \), by Assumptions 3.1 and 3.2 we have

\[
a^2 [1 - \delta(2s + s^*)] \cdot \| V \|_2 \leq V \top A(t)_{F(t), F(t)} V \leq b^2 [1 + \delta(2s + s^*)] \cdot \| V \|_2
\]

for any \( V \in \mathbb{R}^{|F(t)|} \). Therefore, given \( \eta \) we can bound \( \| I - \eta \cdot A(t)_{F(t), F(t)} \|_2 \) by

\[
\lambda_1(\eta) := \max \{ 1 - \eta a^2 [1 - \delta(2s + s^*)], \eta b^2 [1 + \delta(2s + s^*)] - 1 \}. \tag{B.9}
\]
By choosing $3/7 \cdot \{a^2[1 - \delta(2s + s^*)]\}^{-1} < \eta < 11/7 \cdot \{b^2[1 + \delta(2s + s^*)]\}^{-1}$, we have $\lambda_1(\eta) \in (0, 4/7)$. Note that given $a$ and $b$, such $\eta$ always exists as long as the constant $\delta(2s + s^*)$ is sufficiently small.

Moreover, due to the property of the norms,

$$\|E_{x,ij}^{(t)}\|_2 \leq \sqrt{2s + s^*} \cdot \|E^{(t)}\|_\infty.$$  \hspace{1cm} (B.10)

Therefore, combining (B.8), (B.9), and (B.10), we obtain that

$$\|\tilde{\Theta}^{(t+1)} - \Theta^*\|_2 \leq \lambda_1(\eta) \cdot \|\Delta^{(t)}\|_2 + \eta \sqrt{2s + s^*} \cdot \|E^{(t)}\|_\infty.$$  \hspace{1cm} (B.11)

The following lemma further upper bounds $\|E^{(t)}\|_\infty$ with high probability.

**Lemma B.1.** For $E^{(t)}$ defined in (B.3), if we assume $X_i \in \mathbb{R}^d$ satisfies Assumption 3.4, then it holds with probability at least $1 - d^{-1}$ that

$$\|E^{(t)}\|_\infty \leq 2b\sigma D \cdot \sqrt{\log d/n}, \ \forall t \geq 0.$$  \hspace{1cm} (B.12)

**Proof.** Recall that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. sub-Gaussian random variables. By the definition of $E_j^{(t)}$, for $j \in \{1, \ldots, d\}$, each entry $E_{j,i}^{(t)}$ is a sub-Gaussian random variable with variance proxy given by

$$\tau_t := \frac{\sigma^2}{n^2} \sum_{i=1}^n [f'(\langle X_i, \Theta^{(t)} \rangle) - f'(\langle X_i, \Theta^* \rangle)]^2 X_{ij}^2 \leq 4b^2 \sigma^2 D^2/n, \ \forall t \geq 0,$$

where $X_{ij}$ denotes the $j$-th entry in $X_i$. Here the inequality follows from the assumption that $f'(u) \leq b$ for all $u \in \mathbb{R}$ and Assumption 3.4. By the definition of sub-Gaussian random vectors, for any $A \in \mathbb{R}^d$, we have

$$\max_{t \geq 0} \mathbb{E}[\exp(A^T E^{(t)})] \leq \max_{t \geq 0} \exp(\|A\|_2^2 \cdot \tau_t^2) \leq \exp(\|A\|_2^2 \cdot 4b^2 \sigma^2 D^2/n).$$

By the tail inequality for sub-Gaussian random variables, we conclude that

$$\|E^{(t)}\|_\infty \leq 2b\sigma D \cdot \sqrt{\log d/n}, \ \forall t \geq 0$$

with probability at least $1 - d^{-1}$. Thus, we conclude the proof of the lemma. \qed

Now we use the following lemma from Jain et al. (2014) to characterize the relation between $\|\tilde{\Theta}^{(t+1)} - \Theta^*\|_2$ and $\|\Theta^{(t+1)} - \Theta^*\|_2$.

**Lemma B.2.** For any $\Theta \in \mathbb{R}^k$ and integer $s \leq k$, let $\Theta_t = \text{Trunc}(\Theta, s)$. Then for any $\Theta^* \in \mathbb{R}^k$ such that $\|\Theta^*_0\|_0 \leq s^*$, we have $\|\Theta_t - \Theta^*_0\|_2^2 \leq (k - s)/(k - s^*) \cdot \|\Theta^* - \Theta^*_0\|_2^2$.

**Proof.** See Lemma 1 of Jain et al. (2014) for a detailed proof. \qed

Applying Lemma B.2 with $\Theta = \tilde{\Theta}^{(t+1)}$, we have

$$\|\tilde{\Theta}^{(t+1)} - \Theta^{(t+1)}\|_2^2 \leq \frac{|F^{(t)}| - s}{|F^{(t)}| - s^*} \cdot \|\tilde{\Theta}^{(t+1)} - \Theta^*\|_2^2 \leq \frac{8 + s^*}{2s} \cdot \|\tilde{\Theta}^{(t+1)} - \Theta^*\|_2^2,$$

where the second inequality follows from $|F^{(t)}| \leq 2s + s^*$. From Assumption 3.2 that $s \geq 8s^*$, we further have

$$\|\tilde{\Theta}^{(t+1)} - \Theta^{(t+1)}\|_2^2 \leq 9/16 \cdot \|\tilde{\Theta}^{(t+1)} - \Theta^*\|_2^2.$$

Therefore, we obtain

$$\|\Theta^{(t+1)} - \Theta^*\|_2 \leq \|\Theta^{(t+1)} - \tilde{\Theta}^{(t+1)}\|_2 + \|\tilde{\Theta}^{(t+1)} - \Theta^*\|_2 \leq 7/4 \cdot \|\tilde{\Theta}^{(t+1)} - \Theta^*\|_2.$$  \hspace{1cm} (B.13)

Finally we obtain the main result by combining (B.11), (B.12), and (B.13):

$$\|\Theta^{(t+1)} - \Theta^*\|_2 \leq \mu_1^t \cdot \|\Theta^{(0)} - \Theta^*\|_2 + C_1 \cdot \sqrt{(2s + s^*) \log d/n},$$

where $\mu_1 \in (0, 1)$ and $C_1 > 0$ is an absolute constant. Therefore, we conclude the proof of the theorem. \qed
B.2 Proof of Theorem 3.6

Proof. The proof has the similar procedure as the proof of Theorem 3.5. When Algorithm 1 is adopted for low-rank matrix recovery, for \( t \geq 0 \), we define \( S(t) \) and \( S^* \) as

\[
S(t) := \{ V \in \mathbb{R}^{m_1 \times m_2} : \text{row}(V) \subseteq \text{row}(\Theta(t)), \text{col}(V) \subseteq \text{col}(\Theta(t)) \},
\]

\[
S^* := \{ V \in \mathbb{R}^{m_1 \times m_2} : \text{row}(V) \subseteq \text{row}(\Theta^*), \text{col}(V) \subseteq \text{col}(\Theta^*) \}.
\]

Therefore, any matrix in subspace \( F(t) := S(t) \cup S((t+1)) \cup S^* \) has rank no larger than \( 2s+s^* \), i.e., \( \text{rank}(\Theta_{F(t)}) \leq 2s+s^* \), for any \( \Theta \in \mathbb{R}^{m_1 \times m_2} \). Therefore, by denoting \( \Theta(t+1) = \Theta(t) - \eta \nabla_{F(t)} f(\Theta(t)) \) and \( \Theta(t+1) = \text{Trunc}(\Theta(t+1), s) \), we obtain the same results as (B.1) and (B.2) for low-rank matrix recovery and have the same definition for \( E(t) \) and \( G(t) \) as (B.3) and (B.4), respectively. With triangle inequality for the Frobenius norm, we obtain

\[
\| \Theta^{(t+1)} - \Theta^* \|_F \leq \| \Theta^{(t)} - \eta \cdot G(t) \|_F + \eta \cdot \| E(t) \|_F.
\]

(B.14)

We need to upper bound the two terms on the right hand side.

Using the Mean Value Theorem, \( G(t) \) can be written as

\[
G(t) = \frac{1}{n} \sum_{i=1}^{n} f'(\langle X_i, \Theta(t) \rangle) \cdot f'(\langle X_i, \Theta_1 \rangle) \cdot \langle X_i, \Delta(t) \rangle \cdot X_i = \frac{1}{n} \sum_{i=1}^{n} B_i \cdot \langle X_i, \Delta(t) \rangle \cdot X_i,
\]

where \( \Theta_1 \) lies between \( \Theta(t) \) and \( \Theta^* \) and \( B_i = f'(\langle X_i, \Theta(t) \rangle) \cdot f'(\langle X_i, \Theta_1 \rangle) \). Hence by definition, \( \| \Delta^{(t)} - \eta \cdot G(t) \|_F \) can be written as

\[
\| \Delta^{(t)} - \eta \cdot G(t) \|_F = \sup_{\Delta \in S(t)} \left| \langle \Delta, \Delta(t) \rangle - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} B_i \cdot \langle X_i, \Delta(t) \rangle \langle \Delta, X_i \rangle \right|
\]

(B.15)

where \( S(t) := \{ \Delta \in F(t) : \| \Delta \|_F = 1 \} \). Using triangle inequality, we have

\[
\sup_{\Delta \in S(t)} \left| \langle \Delta, \Delta(t) \rangle - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} B_i \cdot \langle X_i, \Delta(t) \rangle \langle \Delta, X_i \rangle \right|
\]

\[
\leq \sup_{\Delta \in S(t)} \left| \langle \Delta, \Delta(t) \rangle - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} B_i \cdot \langle \Delta, \Delta(t) \rangle \right| + \left| \eta \cdot \frac{1}{n} \sum_{i=1}^{n} B_i \cdot \langle \Delta, \Delta(t) \rangle - \langle X_i, \Delta(t) \rangle \langle \Delta, X_i \rangle \right|
\]

(B.16)

The first term on the right hand side of (B.16) is bounded as

\[
\sup_{\Delta \in S(t)} \left| \langle \Delta, \Delta(t) \rangle - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} B_i \cdot \langle \Delta, \Delta(t) \rangle \right| \leq \max \{ 1 - \eta a^2, \eta b^2 - 1 \} \cdot \| \Delta(t) \|_F,
\]

(B.17)

due to the boundness of the derivative \( f' \).

For the second term on the right hand side of (B.16), we introduce the following lemma from Carpentier and Kim (2015) to bound it.

Lemma B.3. Under Assumption 3.2, for any \( s \leq 8s^* \), we have that

\[
\sup_{A, B \in R(s)} \left| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, A \rangle \langle X_i, B \rangle - \langle A, B \rangle \right| \leq 2\delta(s) \cdot \| A \|_F \| B \|_F,
\]

(B.18)

where we denote \( R(s) := \{ V \in \mathbb{R}^{m_1 \times m_2} : \| \text{row}(V) \| \leq s, \| \text{col}(V) \| \leq s \} \).

Proof. See Lemma 5.1. of Carpentier and Kim (2015) for a detailed proof.
By applying Lemma B.3, we have
\[
\sup_{\Delta \in \mathcal{S}^{(t)}} \left| \eta \cdot \frac{1}{n} \sum_{i=1}^{n} B_i \cdot (\langle \Delta, \Delta^{(t)} \rangle - \langle X_i, \Delta^{(t)} \rangle \langle \hat{\Delta}, X_i \rangle) \right| \leq 2\eta b^2 \delta (2s + s^*) \cdot \| \Delta^{(t)} \|_F. \tag{B.19}
\]
Thus (B.15) can be upper bounded by combining (B.17) and (B.19)
\[
\| \Delta^{(t)} - \eta \cdot \mathbf{G}^{(t)}_{\mathcal{X}(t)} \|_F \leq \lambda_2(\eta) := \max \{ 1 - \eta a^2, \eta b^2 - 1 \} + 2\eta b^2 \delta (2s + s^*). \tag{B.20}
\]
By choosing $b^{-2} < \eta < 11/7 \cdot b^{-2} \cdot [1 + 2\delta (2s + s^*)]^{-1}$, or $3/7 \cdot [a^2 - 2b^2 \delta (2s + s^*)]^{-1} < \eta < a^{-2}$, we have $\lambda_2(\eta) < 4/7$. Note that such $\eta$ always exists as long as the constant $\delta (2s + s^*)$ in the RIP condition is sufficiently small.

For the term $\| \mathbf{E}^{(t)}_{\mathcal{X}(t)} \|_F$, we also have
\[
\| \mathbf{E}^{(t)}_{\mathcal{X}(t)} \|_F \leq \sqrt{2s + s^*} \cdot \| \mathbf{E}^{(t)} \|_2 \tag{B.21}
\]
due to the relation between the Frobenius norm and the operator norm. Moreover, we can further upper bound $\| \mathbf{E}^{(t)} \|_2$ with high probability for low-rank matrix recovery depending on different assumptions on $\mathbf{X}$. For example, under the assumption that $\mathbf{X}_i$ are i.i.d. sampled from the $\Sigma$-ensemble with some positive definite $\Sigma$, we have $\| \mathbf{E}^{(t)} \|_2 = O_p(\sqrt{m_1 + m_2}/\sqrt{n})$ (Negahban and Wainwright, 2011) (See more discussions in Remark 3.7).

To characterize the relation between $\| \check{\Theta}^{(t+1)} - \Theta^* \|_F$ and $\| \Theta^{(t+1)} - \Theta^* \|_F$, we use another lemma in Jain et al. (2014) for matrix recovery.

**Lemma B.4.** For any $\Theta \in \mathbb{R}^{m_1 \times m_2}$ with rank $k$ and integer $s \leq k$, let $\Theta_1 = \text{Trunc}(\Theta, s)$. Then for any $\Theta^* \in \mathbb{R}^{m_1 \times m_2}$ with rank($\Theta^*$) $\leq s^*$, we have $\| \Theta_1 - \Theta \|_F \leq (k - s)/(k - s^*) \cdot \| \Theta^* - \Theta \|_F^2$.

**Proof.** See Lemma 2 of Jain et al. (2014) for a detailed proof. \qed

By applying Lemma B.4, we arrive at the similar result as (B.13),
\[
\| \Theta^{(t+1)} - \Theta^* \|_F \leq \| \Theta^{(t+1)} - \check{\Theta}^{(t+1)} \|_F + \| \check{\Theta}^{(t+1)} - \Theta^* \|_F \leq 7/4 \cdot \| \check{\Theta}^{(t+1)} - \Theta^* \|_F. \tag{B.22}
\]
Finally we obtain the main results for low-rank matrix recovery by combining (B.14), (B.20), (B.21), and (B.22):
\[
\| \Theta^{(t)} - \Theta^* \|_F \leq \mu_2^t \cdot \| \Theta^{(0)} - \Theta^* \|_F + \eta \sqrt{2s + s^*} \cdot \| \mathbf{E}^{(t)} \|_2
\leq \mu_2^t \cdot \| \Theta^{(0)} - \Theta^* \|_F + C_2 \sqrt{2s + s^*} \cdot \left \| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathbf{X}_i \right \|_2,
\]
with $\mu_2 \in (0, 1)$ and $C_2 = 2b\eta > 0$ as an absolute constant. Thus, we conclude the proof of the theorem. \qed