Supplementary Material for *Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery*

A. Proofs for Section 2

Proof of Theorem 1. The proof of Theorem 1 follows from a covering argument, which we establish in several steps as below.

Let

$$\mathfrak{S}_{2r} = \{\boldsymbol{\mathcal{D}} \mid \boldsymbol{\mathcal{D}} \in \mathfrak{T}_{2r}, \|\boldsymbol{\mathcal{D}}\|_F = 1\}. \tag{A.1}$$

The following lemma shows that the required number of measurements can be bounded in terms of the exponent of the covering number for $\mathfrak{S}_{2r}$, which can be considered as a proxy for dimensionality:

Lemma 1. *Suppose that the covering number for $\mathfrak{S}_{2r}$ with respect to the Frobenius norm, satisfies*

$$N(\mathfrak{S}_{2r}, \|\cdot\|_F, \varepsilon) \;\leq\; (\beta/\varepsilon)^d, \tag{A.2}$$

for some integer d and scalar β that does not depend on ε. Then if $m \geq d+1$, with probability one $\operatorname{null}(\mathcal{G}) \cap \mathfrak{S}_{2r} = \emptyset$*, which implies that* $\operatorname{null}(\mathcal{G}) \cap \mathfrak{T}_{2r} = \{\mathbf{0}\}$.

Proof. The arguments we used below are primarily adapted from (Eldar et al., 2012), where their interest is to establish the number of Gaussian measurements required to recover a low-rank matrix by rank minimization.

Notice that every $\boldsymbol{\mathcal{D}} \in \mathfrak{S}_{2r}$, and every i, $\langle \boldsymbol{\mathcal{G}}_i, \boldsymbol{\mathcal{D}} \rangle$ is a standard Gaussian random variable, and so

$$\forall\, t > 0, \quad \mathbb{P}\left[\, |\langle \boldsymbol{\mathcal{G}}_i, \boldsymbol{\mathcal{D}} \rangle| < t \,\right] < 2t \cdot \frac{1}{\sqrt{2\pi}} = t\sqrt{\frac{2}{\pi}}. \tag{A.3}$$

Let $\mathfrak{N}$ be an ε-net for $\mathfrak{S}_{2r}$ in terms of $\|\cdot\|_F$. Because the measurements are independent, for any fixed $\bar{\boldsymbol{\mathcal{D}}} \in \mathfrak{S}_{2r}$,

$$\mathbb{P}\left[\, \left\|\mathcal{G}[\bar{\boldsymbol{\mathcal{D}}}]\right\|_\infty < t \,\right] \;<\; \left(t\sqrt{2/\pi}\right)^m. \tag{A.4}$$

Moreover, for any $\boldsymbol{\mathcal{D}} \in \mathfrak{S}_{2r}$, we have

$$\begin{aligned} \|\mathcal{G}[\boldsymbol{\mathcal{D}}]\|_\infty &\geq \max_{\bar{\boldsymbol{\mathcal{D}}} \in \mathfrak{N}} \left\{ \left\|\mathcal{G}[\bar{\boldsymbol{\mathcal{D}}}]\right\|_\infty - \|\mathcal{G}\|_{F\to\infty} \left\|\bar{\boldsymbol{\mathcal{D}}} - \boldsymbol{\mathcal{D}}\right\|_F \right\} && \text{(A.5)} \\ &\geq \min_{\bar{\boldsymbol{\mathcal{D}}} \in \mathfrak{N}} \left\{ \left\|\mathcal{G}[\bar{\boldsymbol{\mathcal{D}}}]\right\|_\infty \right\} - \varepsilon \|\mathcal{G}\|_{F\to\infty}. && \text{(A.6)} \end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left[\inf_{\boldsymbol{\mathcal{D}}\in\mathfrak{S}_{2r}} \|\mathcal{G}[\boldsymbol{\mathcal{D}}]\|_\infty < \varepsilon\log(1/\varepsilon)\right] \\
&\quad\leq\quad \mathbb{P}\left[\min_{\boldsymbol{\mathcal{D}}\in\mathfrak{N}} \|\mathcal{G}[\boldsymbol{\mathcal{D}}]\|_\infty < 2\varepsilon\log(1/\varepsilon)\right] + \mathbb{P}\left[\,\|\mathcal{G}\|_{F\to\infty} > \log(1/\varepsilon)\,\right] \\
&\quad\leq\quad \#\mathfrak{N}\times\left(2\sqrt{2/\pi}\times\varepsilon\log(1/\varepsilon)\right)^m + \mathbb{P}\left[\,\|\mathcal{G}\|_{F\to\infty} > \log(1/\varepsilon)\,\right] \\
&\quad\leq\quad \beta^d(2\sqrt{2/\pi})^m\varepsilon^{m-d}\log(1/\varepsilon)^m + \mathbb{P}\left[\,\|\mathcal{G}\|_{F\to\infty} > \log(1/\varepsilon)\,\right]. \qquad \text{(A.7)}
\end{aligned}$$

Since $m \geq d+1$, (A.7) goes to zero as $\varepsilon \searrow 0$. Hence, taking a sequence of decreasing ε, we can show that $\mathbb{P}\left[\,\inf_{\boldsymbol{\mathcal{D}}\in\mathfrak{S}_{2r}} \|\mathcal{G}[\boldsymbol{\mathcal{D}}]\|_\infty = 0\,\right] \leq t$ for every positive t, establishing the result. □

Following Lemma 1, it just remains to find the covering number of $\mathfrak{S}_{2r}$. We use the following lemma, which uses the triangle inequality to control the effect of perturbations in the factors of the Tucker decomposition

$$[[\mathcal{C}; \mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_K]] := \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \cdots \times_K \mathbf{U}_K, \qquad \text{(A.8)}$$

where the *mode-i (matrix) product* of tensor $\boldsymbol{\mathcal{A}}$ with matrix $\mathbf{B}$ of compatible size, denoted as $\boldsymbol{\mathcal{A}}\times_i\mathbf{B}$, outputs a tensor $\mathcal{C}$ such that $\mathcal{C}_{(i)} = \mathbf{B}\boldsymbol{\mathcal{A}}_{(i)}$.

Lemma 2. *Let $\mathcal{C}, \mathcal{C}' \in \mathbb{R}^{r_1,\ldots,r_K}$, and $\mathbf{U}_1, \mathbf{U}_1' \in \mathbb{R}^{n_1\times r_1}, \ldots, \mathbf{U}_K, \mathbf{U}_K' \in \mathbb{R}^{n_K\times r_K}$ with $\mathbf{U}_i^*\mathbf{U}_i = \mathbf{U}_i'^*\mathbf{U}_i' = \mathbf{I}$, and $\|\mathcal{C}\|_F = \|\mathcal{C}'\|_F = 1$. Then*

$$\left\|[[\mathcal{C}; \mathbf{U}_1, \ldots, \mathbf{U}_K]] - [[\mathcal{C}'; \mathbf{U}_1', \ldots, \mathbf{U}_K']]\right\|_F \;\leq\; \left\|\mathcal{C} - \mathcal{C}'\right\|_F + \sum_{i=1}^K \|\mathbf{U}_i - \mathbf{U}_i'\|. \qquad \text{(A.9)}$$

Proof. This follows from the basic fact that for any tensor $\boldsymbol{\mathcal{X}}$ and matrix $\mathbf{U}$ of compatible size,

$$\|\boldsymbol{\mathcal{X}} \times_k \mathbf{U}\|_F \;\leq\; \|\mathbf{U}\|\,\|\boldsymbol{\mathcal{X}}\|_F, \qquad \text{(A.10)}$$

which can be established by direct calculation. Write

$$\begin{aligned}
&\left\|[[\mathcal{C}; \mathbf{U}_1, \ldots, \mathbf{U}_K]] - [[\mathcal{C}'; \mathbf{U}_1', \ldots, \mathbf{U}_K']]\right\|_F \\
&\quad\leq\quad \left\|[[\mathcal{C}; \mathbf{U}_1, \ldots, \mathbf{U}_K]] - [[\mathcal{C}'; \mathbf{U}_1, \ldots, \mathbf{U}_K]]\right\|_F \\
&\qquad +\quad \left\|\sum_{i=1}^K [[\mathcal{C}'; \mathbf{U}_1', \ldots, \mathbf{U}_i', \mathbf{U}_{i+1}, \ldots \mathbf{U}_k]] - [[\mathcal{C}'; \mathbf{U}_1', \ldots, \mathbf{U}_{i-1}', \mathbf{U}_i, \ldots \mathbf{U}_K]]\right\|_F \\
&\quad\leq\quad \left\|\mathcal{C} - \mathcal{C}'\right\|_F + \sum_{i=1}^K \|\mathbf{U}_i - \mathbf{U}_i'\|,
\end{aligned}$$

where the first inequality follows from triangle inequality and the second inequality follows from the fact that $\|\mathcal{C}\|_F = 1$, $\|\mathbf{U}_j\| = 1$, $\mathbf{U}_i^*\mathbf{U}_i = \mathbf{I}$ and $\mathbf{U}_i'^*\mathbf{U}_i' = \mathbf{I}$. □

Using this result, we construct an ε-net for $\mathfrak{S}_{2r}$ by building $\varepsilon/(K+1)$-nets for each of the $K+1$ factors $\boldsymbol{\mathcal{C}}$ and $\{\mathbf{U}_i\}$. The total size of the resulting ε net is thus bounded by the following lemma:

Lemma 3. $N(\mathfrak{S}_{2r}, \|\cdot\|_F, \varepsilon) \leq (3(K+1)/\varepsilon)^{(2r)^K + 2nrK}$

Proof. The idea of this proof is to construct a net for each component of the Tucker decomposition and then combine these nets to form a *compound* net with the desired cardinality.

Denote $\mathfrak{C} = \{\boldsymbol{\mathcal{C}} \in \mathbb{R}^{2r\times 2r\times\cdots\times 2r} \mid \|\boldsymbol{\mathcal{C}}\|_F = 1\}$ and $\mathcal{O} = \{\mathbf{U} \in \mathbb{R}^{n\times r} \mid \mathbf{U}^*\mathbf{U} = \mathbf{I}\}$. Clearly, for any $\boldsymbol{\mathcal{C}} \in \mathfrak{C}$, $\|\boldsymbol{\mathcal{C}}\|_F = 1$, and for any $\mathbf{U} \in \mathcal{O}$, $\|\mathbf{U}\| = 1$. Thus by Prop. 4 of (Vershynin, 2007) and Lemma 5.2 of (Vershynin, 2010), there exists an $\frac{\varepsilon}{K+1}$-net $\mathfrak{C}'$ covering $\mathfrak{C}$ with respect to the Frobenius norm such that $\#\mathfrak{C}' \leq (\frac{3(K+1)}{\varepsilon})^{(2r)^K}$, and there exists an $\frac{\varepsilon}{K+1}$-net $\mathcal{O}'$ covering $\mathcal{O}$ with respect to the operator norm such that $\#\mathcal{O}' \leq (\frac{3(K+1)}{\varepsilon})^{2nr}$. Construct

$$\mathfrak{S}'_{2r} = \{[[\boldsymbol{\mathcal{C}}'; \mathbf{U}'_1, \ldots, \mathbf{U}'_K]] \mid \boldsymbol{\mathcal{C}}' \in \mathfrak{C}',\ \mathbf{U}'_i \in \mathcal{O}'\}.$$

Clearly $\#\mathfrak{S}'_{2r} \leq \left(\frac{3(K+1)}{\varepsilon}\right)^{(2r)^K + 2nrK}$. The rest is to show that $\mathfrak{S}'_{2r}$ is indeed an ε-net covering $\mathfrak{S}_{2r}$ with respect to the Frobenius norm.

For any fixed $\boldsymbol{\mathcal{D}} = [[\boldsymbol{\mathcal{C}}; \mathbf{U}_1, \cdots, \mathbf{U}_K]] \in \mathfrak{S}_{2r}$ where $\boldsymbol{\mathcal{C}} \in \mathfrak{C}$ and $\mathbf{U}_i \in \mathcal{O}$, by our constructions above, there exist $\boldsymbol{\mathcal{C}}' \in \mathfrak{C}'$ and $\mathbf{U}'_i \in \mathcal{O}'$ such that $\|\boldsymbol{\mathcal{C}} - \boldsymbol{\mathcal{C}}'\|_F \leq \frac{3(K+1)}{\varepsilon}$ and $\|\mathbf{U}_i - \mathbf{U}'_i\| \leq \frac{3(K+1)}{\varepsilon}$. Then $\boldsymbol{\mathcal{D}}' = [[\boldsymbol{\mathcal{C}}'; \mathbf{U}'_1, \cdots, \mathbf{U}'_K]] \in \mathfrak{S}'_{2r}$ is within ε-distance from $\boldsymbol{\mathcal{D}}$, since by the triangle inequality derived in Lemma 2, we have

$$\|\boldsymbol{\mathcal{D}} - \boldsymbol{\mathcal{D}}'\|_F = \|[[\boldsymbol{\mathcal{C}}; \mathbf{U}_1, \ldots, \mathbf{U}_K]] - [[\boldsymbol{\mathcal{C}}'; \mathbf{U}'_1, \ldots, \mathbf{U}'_K]]\|_F \leq \|\boldsymbol{\mathcal{C}} - \boldsymbol{\mathcal{C}}'\|_F + \sum_{i=1}^{K} \|\mathbf{U}_i - \mathbf{U}'_i\| \leq \varepsilon.$$

This completes the proof. □

With these observations in hand, Theorem 1 follows immediately.

B. Proofs for Section 3

Proof of Corollary 4.

Proof. Denote $\lambda = \delta(\mathcal{C}) - m$. Then following Theorem 7.1 of (Amelunxen et al., 2013), we have

$$\begin{aligned}
\mathbb{P}\left[\mathcal{C} \cap \mathrm{null}(\mathcal{G}) = \{\mathbf{0}\}\right] &\leq 4\exp\left(-\frac{\lambda^2/8}{\min\{\delta(\mathcal{C}), \delta(\mathcal{C}^\circ)\} + \lambda}\right) \\
&\leq 4\exp\left(-\frac{\lambda^2/8}{\delta(\mathcal{C}) + \lambda}\right) \\
&\leq 4\exp\left(-\frac{(\delta(\mathcal{C}) - m)^2}{16\delta(\mathcal{C})}\right).
\end{aligned}$$

□

Proof of Lemma 2.

Proof. Denote $\mathrm{circ}(\mathbf{e}_n, \theta)$ as $\mathrm{circ}_n(\theta)$, where $\mathbf{e}_n$ is the nth standard basis for $\mathbb{R}^n$. Since $\delta\big(\mathrm{circ}(\boldsymbol{x}_0, \theta)\big) = \delta\big(\mathrm{circ}(\mathbf{e}_n, \theta)\big)$, it is sufficient to prove $\delta\big(\mathrm{circ}_n(\theta)\big) \leq n \sin^2\theta + 2$.

Let us first consider the case where n is *even*. Define a discrete random variable V supported on $\{0, 1, 2, \cdots, n\}$ with probability mass function $\mathbb{P}[V = k] = v_k$. Here v_k denotes the k-th intrinsic volumes of $\mathrm{circ}_n(\theta)$. As specified in Ex. 4.4.8 of (Amelunxen, 2011), we have

$$v_k = \frac{1}{2}\binom{\frac{1}{2}(n-2)}{\frac{1}{2}(k-1)} \sin^{k-1}(\theta)\cos^{n-k-1}(\theta) \qquad \text{for } k = 1, 2, \cdots, n-1.$$

From Prop. 5.11 of (Amelunxen et al., 2013), we know that

$$\delta\big(\mathrm{circ}_n(\theta)\big) = \mathbb{E}[V] = \sum_{k=1}^{n} \mathbb{P}[V \geq k].$$

Moreover, by the interlacing result from Prop. 5.6 of (Amelunxen et al., 2013) and the fact that $\mathbb{P}[V \geq 2k] = \mathbb{P}[V \geq 2k-1] - \mathbb{P}[V = 2k-1]$, we have

$$\begin{aligned}
\mathbb{P}[V \geq 1] &\leq 2\mathbb{P}[V=1] + 2\mathbb{P}[V=3] + \cdots + 2\mathbb{P}[V=n-1],\\
\mathbb{P}[V \geq 2] &\leq \mathbb{P}[V=1] + 2\mathbb{P}[V=3] + \cdots + 2\mathbb{P}[V=n-1];\\
\\
\mathbb{P}[V \geq 3] &\leq 2\mathbb{P}[V=3] + 2\mathbb{P}[V=5] + \cdots + 2\mathbb{P}[V=n-1],\\
\mathbb{P}[V \geq 4] &\leq \mathbb{P}[V=3] + 2\mathbb{P}[V=5] + \cdots + 2\mathbb{P}[V=n-1];\\
\\
\vdots \quad & \quad \vdots \quad \vdots\\
\\
\mathbb{P}[V \geq n-1] &\leq 2\mathbb{P}[V=n-1],\\
\mathbb{P}[V \geq n] &\leq \mathbb{P}[V=n-1].
\end{aligned}$$

Summing up the above inequalities, we have

$$\begin{aligned}
\mathbb{E}[V] &= \sum_{k=1}^{n} \mathbb{P}[V \geq k]\\
&\leq \sum_{k=1,3,\cdots,n-1} 2(k-1)v_k + \sum_{k=1,3,\cdots,n-1} 3v_k\\
&\leq (n-2)\sin^2\theta + \frac{3}{2}\sum_{k=0}^{n} v_k\\
&\leq (n-2)\sin^2\theta + \frac{3}{2} \;=\; n\sin^2\theta + 2\cos^2\theta - \frac{1}{2},
\end{aligned}$$

where the second last inequality follows from the observations that $\sum_{k=1,3,\cdots,n-1} \frac{k-1}{2} \cdot (2v_k) = \mathbb{E}\left[Bin(\frac{n-2}{2}, \sin^2\theta)\right]$ and $\sum_{k=0}^{n} v_k \geq \sum_{k=1,3,\cdots,n-1} 2v_k$ again by the interlacing result from Prop. 5.6 (Amelunxen et al., 2013).

Suppose n is *odd*. Since the intersection of $\text{circ}_{n+1}(\theta)$ with any n-dimensional linear subspace containing $\mathbf{e}_{n+1}$ is an isometric image of $\text{circ}_n(\theta)$, by Prop. 4.1 of (Amelunxen et al., 2013), we have

$$\delta(\text{circ}_n(\theta)) = \delta(\text{circ}_n(\theta)\times\{\mathbf{0}\}) \le \delta(\text{circ}_{n+1}(\theta)) \le (n+1)\sin^2\theta + 2\cos^2\theta - \frac{1}{2} \le n\sin^2\theta + \cos^2\theta + \frac{1}{2}.$$

Thus, taking both cases (n is even and n is odd) into consideration, we have

$$\delta\big(\text{circ}_n(\theta)\big) \le n\sin^2\theta + \cos^2\theta + \frac{1}{2} < n\sin^2\theta + 2.$$

□

Proof of Theorem 5.

Proof. Notice that for any fixed $m > 0$, the function $f:\ t \to 4\exp\left(-\frac{(t-m)^2}{16t}\right)$ is decreasing for $t \ge m$. Then due to Corollary 4 and the fact that $\delta(\mathcal{C}) \ge \kappa - 2 \ge m$, we have

$$\begin{aligned}
\mathbb{P}\left[\, \boldsymbol{x}_0 \text{ is the unique optimal solution to 3.3} \,\right] &= \mathbb{P}\left[\, C \cap \text{null}(\mathcal{G}) = \{\mathbf{0}\} \,\right] \\
&\le 4\exp\left(-\frac{(\delta(C)-m)^2}{16\delta(C)}\right) \\
&\le 4\exp\left(-\frac{(\kappa-m-2)^2}{16\,(\kappa-2)}\right).
\end{aligned}$$

□

C. Proofs and Comments for Section 4

C.1. Proof of Lemma 3.

Proof. (1) By the definition of $\boldsymbol{\mathcal{X}}_{[j]}$, it is sufficient to prove that the vectorization of the right hand side of (4.3) equals $\text{vec}(\boldsymbol{\mathcal{X}}_{(1)})$.

Since $\boldsymbol{\mathcal{X}} = \sum_{i=1}^r \lambda_i \boldsymbol{a}_i^{(1)} \circ \boldsymbol{a}_i^{(2)} \circ \cdots \circ \boldsymbol{a}_i^{(K)}$, we have

$$\begin{aligned}
\text{vec}(\boldsymbol{\mathcal{X}}_{(1)}) &= \text{vec}\big(\sum_{i=1}^r \lambda_i \boldsymbol{a}_i^{(1)} \circ (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(2)})\big) \\
&= \sum_{i=1}^r \lambda_i \text{vec}\big(\boldsymbol{a}_i^{(1)} \circ (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(2)})\big) \\
&= \sum_{i=1}^r \lambda_i (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(2)} \otimes \boldsymbol{a}_i^{(1)}),
\end{aligned}$$

where the last equality follows from the fact that $\text{vec}(\boldsymbol{a} \circ \boldsymbol{b}) = \boldsymbol{b} \otimes \boldsymbol{a}$. Similarly, we can derive that

the vectorization of the right hand side of (4.3),

$$\begin{aligned} & \text{vec}(\sum_{i=1}^{r} \lambda_i (\boldsymbol{a}_i^{(j)} \otimes \boldsymbol{a}_i^{(j-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(1)}) \circ (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \cdots \otimes \boldsymbol{a}_i^{(j+1)})) \\ = & \sum_{i=1}^{r} \lambda_i \text{vec}\big((\boldsymbol{a}_i^{(j)} \otimes \boldsymbol{a}_i^{(j-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(1)}) \circ (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \cdots \otimes \boldsymbol{a}_i^{(j+1)})\big) \\ = & \sum_{i=1}^{r} \lambda_i (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(2)} \otimes \boldsymbol{a}_i^{(1)}) \\ = & \text{vec}(\boldsymbol{\mathcal{X}}_{(1)}). \end{aligned}$$

Thus, equation (4.3) is valid.

(2) The above argument can be easily adapted to prove the second claim. Since $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \cdots \times_K \mathbf{U}_K$, we have

$$\begin{aligned} \text{vec}(\boldsymbol{\mathcal{X}}_{(1)}) &= \text{vec}\Big(\mathbf{U}_1 \, \boldsymbol{\mathcal{C}}_{(1)} \, (\mathbf{U}_K \otimes \mathbf{U}_{K-1} \otimes \cdots \otimes \mathbf{U}_2)^*\Big) \\ &= (\mathbf{U}_K \otimes \mathbf{U}_{K-1} \otimes \cdots \otimes \mathbf{U}_1) \, \text{vec}(\boldsymbol{\mathcal{C}}_{(1)}), \end{aligned}$$

where the last equality follows from the fact that $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^* \otimes \mathbf{A})\text{vec}(\mathbf{B})$. Similarly, we can derive that the vectorization of the right hand side of (4.4),

$$\begin{aligned} & \text{vec}\Big((\mathbf{U}_j \otimes \mathbf{U}_{j-1} \otimes \cdots \otimes \mathbf{U}_1) \, \boldsymbol{\mathcal{C}}_{[j]} \, (\mathbf{U}_K \otimes \mathbf{U}_{K-1} \otimes \cdots \otimes \mathbf{U}_{j+1})^*\Big) \\ = & (\mathbf{U}_K \otimes \mathbf{U}_{K-1} \otimes \cdots \otimes \mathbf{U}_1) \, \text{vec}(\boldsymbol{\mathcal{C}}_{[j]}) \\ = & (\mathbf{U}_K \otimes \mathbf{U}_{K-1} \otimes \cdots \otimes \mathbf{U}_1) \, \text{vec}(\boldsymbol{\mathcal{C}}_{(1)}) \\ = & \text{vec}(\boldsymbol{\mathcal{X}}_{(1)}). \end{aligned}$$

Thus, equation (4.4) is valid. □

C.2. Comments on General Square Reshaping.

As suggested in the paper, our square reshaping can be generalized to combine any j modes (say modes $i_1, i_2, \cdots, i_j$) together rather than the first j modes. Denote $\mathcal{I} = \{i_1, i_2, \cdots, i_j\} \subseteq [K]$ and $\mathcal{J} = [K] \backslash \mathcal{I} = \{i_{j+1}, i_{j+2}, \cdots, i_K\}$. Then the embedded matrix $\boldsymbol{\mathcal{X}}_{\mathcal{I}} \in \mathbb{R}^{\prod_{k=1}^{j} n_{i_k} \times \prod_{k=j+1}^{K} n_{i_k}}$ can be defined as in (4.2) after relabeling. Specifically for $1 \le k \le K$, we first relabel the kth mode as the original i_kth mode. Denote the relabeled tensor as $\widehat{\boldsymbol{\mathcal{X}}}$. Then we can define

$$\boldsymbol{\mathcal{X}}_{\mathcal{I}} := \widehat{\boldsymbol{\mathcal{X}}}_{[j]} = \text{reshape}\Big(\widehat{\boldsymbol{\mathcal{X}}}_{(1)}, \prod_{k=1}^{j} n_{i_k}, \prod_{k=j+1}^{K} n_{i_k}\Big). \tag{C.1}$$

Lemma 4 and Theorem 6 can also be easily extended. As shown by Theorem 6 (after modification), to maximize the effect of our square model, we would like to choose $\mathcal{I}$ to minimize the quantity,

$$\text{rank}(\boldsymbol{\mathcal{X}}_{\mathcal{I}}) \cdot \max\Big\{\prod_{k=1}^{j} n_{i_k}, \prod_{k=j+1}^{K} n_{i_k}\Big\}. \tag{C.2}$$

In practice, normally we do not know the exact rank of each mode, and hence (C.2) cannot be computed directly. However, prior knowledege of the physical properties of the underlying tensor can provide some guidance. For example, in multi-spectral video data, the video tensor tends to be low rank in both the wavelength and the temporal modes. Thus grouping these two modes would lead to a low-rank matrix. Hence, practically, we should set $\mathcal{I}$ by taking both the size and the physical characteristics of the true tensor into consideration.

D. Experimental and Algorithmic Details for Section 5

D.1. Video Data Description for Section 5.2

The Ocean video (source: `http://pages.cs.wisc.edu/~ji-liu/`) is of size $112 \times 160 \times 3 \times 32$. It records the movements of the ocean and has been used in (Liu et al., 2009) to demonstrate the efficacy of the SNN model.

The Campus video (Li et al., 2004) (source: `http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html`) is of size $128 \times 160 \times 3 \times 199$. It records a campus scene in the day time.

The Face video (source: `http://www.youtube.com/watch?v=Ew1i2zY9IEA`) is of size $96 \times 65 \times 3 \times 994$. It is a YOUTUBE video that records the face of a lady aging from young to old.

D.2. On the Equivalence between Problem (5.3) and Problem (5.5)

In this part, we argue that the unconstrained problem (5.3):

$$\min \ \tfrac{1}{2} \left\| \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] - \boldsymbol{\mathcal{D}} \right\|_F^2 + \sum_{i=1}^4 \lambda_i \left\| \boldsymbol{\mathcal{X}}_{(i)} \right\|_*$$

and the norm constrained problem (5.5):

$$\min \ \tfrac{1}{2} \left\| \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] - \boldsymbol{\mathcal{D}} \right\|_F^2 \quad \text{s.t.} \quad \left\| \boldsymbol{\mathcal{X}}_{(i)} \right\|_* \le \beta_i, \ i = 1 \dots 4,$$

are equivalent, in the following sense:

> *If $\boldsymbol{\mathcal{X}}^\star$ is an optimal solution to (5.3) for some choice of $\lambda_1, \dots, \lambda_4 \ge 0$, then there exists a choice of $\beta_1, \dots, \beta_4 \ge 0$ such that $\boldsymbol{\mathcal{X}}^\star$ is also an optimal solution to problem (5.5). Conversely, if $\boldsymbol{\mathcal{X}}^\star$ is an optimal solution to problem (5.5) for some $\beta_1, \dots, \beta_4$, then there exist $\lambda_1, \dots, \lambda_4 \ge 0$ such that $\boldsymbol{\mathcal{X}}^\star$ is also an optimal solution to (5.3).*

To prove the equivalence, we first characterize the optimality conditions for the two problems [1].

For problem (5.3), $\boldsymbol{\mathcal{X}}^\star$ is an optimal solution if and only if

$$\mathbf{0} \in \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}^\star] - \boldsymbol{\mathcal{D}} + \sum_{i=1}^4 \lambda_i \partial \left\| \boldsymbol{\mathcal{X}}^\star_{(i)} \right\|_* . \tag{D.1}$$

[1] Here we always assume the optimal solution $\boldsymbol{\mathcal{X}}^\star$ to either (5.3) or (5.5) is not trivially $\mathbf{0}$, though the equivalence can also be established under this not interesting scenario. Therefore, we would assume $\beta_i > 0$ for any $i \in [4]$ in problem (5.5).

For problem (5.5), since $\mathbf{0}$ is in the interior of the feasible region ($\boldsymbol{\beta} > \mathbf{0}$), Slater's condition is satisfied. Therefore, $\boldsymbol{\mathcal{X}}^\star$ is an optimal solution to problem (5.5) if and only if the KKT conditions are satisfied, i.e.

$$\begin{cases} \left\|\boldsymbol{\mathcal{X}}^\star_{(i)}\right\|_* \le \beta_i, \quad \text{for any } i \in [4] \\ \mathbf{0} \in \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}^\star] - \boldsymbol{\mathcal{D}} + \sum_{i=1}^4 \mu_i^\star \partial \left\|\boldsymbol{\mathcal{X}}^\star_{(i)}\right\|_* \\ \mu_i^\star \left(\left\|\boldsymbol{\mathcal{X}}^\star_{(i)}\right\|_* - \beta_i\right) = 0, \quad \text{for all } i \in [4] \\ \mu_i^\star \ge 0, \quad \text{for any } i \in [4]. \end{cases} \tag{D.2}$$

Suppose $\boldsymbol{\mathcal{X}}^\star$ is an optimal solution to problem (5.5). Then due to the optimality condition, there exists $\{\mu_i^\star\}$ together with $\boldsymbol{\mathcal{X}}^\star$ satisfying (D.2), which implies that $\boldsymbol{\mathcal{X}}^\star$ is also an optimal solution to problem (5.3) with $\lambda_i = \mu_i^\star$ for all $i \in [4]$.

On the other hand, suppose $\boldsymbol{\mathcal{X}}^\star$ is an optimal solution to problem (5.3). Then set $\beta_i = \left\|\boldsymbol{\mathcal{X}}^\star_{(i)}\right\|_*$ for each $i \in [4]$ and set $\boldsymbol{\mu}^\star = \boldsymbol{\lambda}$. It can be easily verified that $(\boldsymbol{\mathcal{X}}^\star,\ \boldsymbol{\mu}^\star)$ satisfies (D.2), which implies that $\boldsymbol{\mathcal{X}}^\star$ is an optimal solution to problem (5.5) with $\beta_i = \left\|\boldsymbol{\mathcal{X}}^\star_{(i)}\right\|_*$.

Therefore, we have proved the equivalence between problem (5.3) and problem (5.5), i.e. $\boldsymbol{\mathcal{X}}^\star$ is an optimal solution to (5.3) with some $\boldsymbol{\lambda} \ge \mathbf{0}$ if and only if $\boldsymbol{\mathcal{X}}^\star$ is an optimal solution to (5.5) with some $\boldsymbol{\beta} \ge \mathbf{0}$.

In a similar vein, the equivalence between problem (5.4) and (5.6) can also be established.

D.3. Accelerated linearized Bregman algorithm for problem (5.1)

Recall the SNN model (5.1):

$$\text{minimize}_{\boldsymbol{\mathcal{X}}} \quad \sum_{i=1}^K \|\boldsymbol{\mathcal{X}}_{(i)}\|_* \quad \text{subject to} \quad \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0]. \tag{D.3}$$

By introducing auxiliary variable $\boldsymbol{\mathcal{W}}$ and splitting $\boldsymbol{\mathcal{X}}$ into $\boldsymbol{\mathcal{X}}_1,\ \boldsymbol{\mathcal{X}}_2,\ \cdots,\ \boldsymbol{\mathcal{X}}_K$, it can be easily verified that problem (D.3) is equivalent to

$$\begin{aligned} \min_{(\{\boldsymbol{\mathcal{X}}_i\},\boldsymbol{\mathcal{W}})} \quad & \sum_{i=1}^K \|(\boldsymbol{\mathcal{X}}_i)_{(i)}\|_* \\ \text{s.t.} \quad & \boldsymbol{\mathcal{X}}_i = \boldsymbol{\mathcal{W}}, \quad i = 1, 2, \cdots, K, \\ & \mathcal{P}_\Omega[\boldsymbol{\mathcal{W}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0], \end{aligned} \tag{D.4}$$

whose objective function is now separable.

The accelerated linearized Bregman (ALB) algorithm, proposed in (Huang et al., 2013), is an efficient first-order method designed for solving convex optimization problems with nonsmooth objective functions and linear constraints. It has been successfully applied to solve ℓ_1 and nuclear norm minimization problems (Huang et al., 2013). The ALB algorithm solves nonsmooth problem

Algorithm 1 Accelerated linearized Bregman algorithm for the SNN model (5.1)

Initialization: $\boldsymbol{\mathcal{Y}}_i^0 = \tilde{\boldsymbol{\mathcal{Y}}}_i^0 = \mathbf{0}$ for each $i \in [K]$, $\boldsymbol{\mathcal{Z}}^0 = \tilde{\boldsymbol{\mathcal{Z}}}^0 = \mathbf{0}$, $\mu > 0$, $\tau > 0$, $t_0 = 1$;
for $k = 0,\ 1,\ 2,\ \cdots$ **do**
  for $i = 1,\ 2,\ \cdots,\ K$ **do**
    $\boldsymbol{\mathcal{X}}_i^{k+1} = \mu \cdot \text{Shrinkage}(\boldsymbol{\mathcal{Y}}_i^k, 1)$;
  end for
  $\boldsymbol{\mathcal{W}}^{k+1} = \mu \cdot \left(\mathcal{P}_\Omega\left[\boldsymbol{\mathcal{Z}}^k\right] - \sum_i \boldsymbol{\mathcal{Y}}_i^k\right)$;
  for $i = 1,\ 2,\ \cdots,\ K$ **do**
    $\tilde{\boldsymbol{\mathcal{Y}}}_i^k = \boldsymbol{\mathcal{Y}}_i^k - \tau \cdot \left(\boldsymbol{\mathcal{X}}_i^{k+1} - \boldsymbol{\mathcal{W}}^{k+1}\right)$;
  end for
  $\tilde{\boldsymbol{\mathcal{Z}}}^k = \boldsymbol{\mathcal{Z}}^k - \tau \cdot \mathcal{P}_\Omega\left[\boldsymbol{\mathcal{W}}^{k+1} - \boldsymbol{\mathcal{X}}_0\right]$;
  $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  for $i = 1,\ 2,\ \cdots,\ K$ **do**
    $\boldsymbol{\mathcal{Y}}_i^{k+1} = \tilde{\boldsymbol{\mathcal{Y}}}_i^k + \frac{t_k - 1}{t_{k+1}}\left(\tilde{\boldsymbol{\mathcal{Y}}}_i^k - \tilde{\boldsymbol{\mathcal{Y}}}_i^{k-1}\right)$;
  end for
  $\boldsymbol{\mathcal{Z}}^{k+1} = \tilde{\boldsymbol{\mathcal{Z}}}^k + \frac{t_k - 1}{t_{k+1}}\left(\tilde{\boldsymbol{\mathcal{Z}}}^k - \tilde{\boldsymbol{\mathcal{Z}}}^{k-1}\right)$;
end for

by first smoothing the objective function (e.g. adding a small l_2 perturbation), and then exploiting Nesterov's accelerated scheme (Nesterov, 1983) to the dual problem, which can be verified to be unconstrained and Lipschitz differentiable. In Algorithm 1, we describe our ALB algorithm adapted to problem (D.4). Algorithm 1 solves exactly the smoothed version of problem (D.4):

$$\begin{aligned} \min_{(\{\boldsymbol{\mathcal{X}}_i\},\boldsymbol{\mathcal{W}})} \quad & \sum_{i=1}^{K}\left(\|(\boldsymbol{\mathcal{X}}_i)_{(i)}\|_* + \frac{1}{2\mu}\|(\boldsymbol{\mathcal{X}}_i)_{(i)}\|_F^2\right) + \frac{1}{2\mu}\|\boldsymbol{\mathcal{W}}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{\mathcal{X}}_i = \boldsymbol{\mathcal{W}}, \quad i = 1, 2, \cdots, K, \\ & \mathcal{P}_\Omega[\boldsymbol{\mathcal{W}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0], \end{aligned} \tag{D.5}$$

where we denote $\boldsymbol{\mathcal{Y}}_i$ as the dual variable for the constraint $\boldsymbol{\mathcal{X}}_i = \boldsymbol{\mathcal{W}}$ and denote $\boldsymbol{\mathcal{Z}}$ as the dual variable for the last constraint $\mathcal{P}_\Omega[\boldsymbol{\mathcal{W}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0]$. Since the objective function in (D.5) is separable, each setup of the ALB algorithm is easy to solve as we can see from Algorithm 1 [2].

For our numerical experiment ($K = 4$), we choose smoothing parameter $\mu = 50\|\boldsymbol{\mathcal{X}}_0\|_F$ and step size $\tau = \frac{1}{5\mu}$. Empirically, we observe that larger values of μ do not result in a better recovery performance. This is consistent with the theoretical results established in (Lai & Yin, 2013; Zhang et al., 2012).

[2] The Shrinkage operator in line 4 of Algorithm 1 performs the regular shrinkage on the singular values of the ith unfolding matrix of $\boldsymbol{\mathcal{Y}}_i^k$, i.e. $(\boldsymbol{\mathcal{Y}}_i^k)_{(i)}$, and then folds the resulting matrix back into a tensor.

Algorithm 2 Frank-Wolfe method for problem (D.6)

Initialization: $\boldsymbol{x}^0 \in \mathcal{D}$;
for $k = 0,\ 1,\ 2,\ \cdots$ **do**
  Compute $\mathbf{s} = \operatorname{argmin}_{\mathbf{s}\in\mathcal{D}} \left\langle \mathbf{s}, \nabla f(\boldsymbol{x}^k)\right\rangle$;
  $\gamma = \frac{2}{k+2}$;
  Update $\boldsymbol{x}^{k+1} = (1-\gamma)\boldsymbol{x}^k + \gamma\mathbf{s}$;
end for

D.4. Frank-Wolfe algrorithm for problem (5.6)

Frank-Wolfe algorithm deals with the following general optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x}), \tag{D.6}$$

where the objective function f is convex and differentiable with Lipschitz continuous gradient, and the feasible set $\mathcal{D}$ is a compact and convex set in $\mathbb{R}^n$. The Frank-Wolfe method, described in Algorithm 2, is a simple iterative scheme that can be dated back to 1956 (Frank & Wolfe, 1956). Regarding its convergence, it is well known that the iterates of Algorithm 2 satisfy $f(\boldsymbol{x}^k) - f^\star \leq O(\frac{1}{k})$, where $f^\star$ is the optimal value to problem (D.6). Recently, due to its good scalability, use of it has resurged in machine learning (Jaggi, 2013; Harchaoui et al., 2013; Mu et al., 2014).

Recall the nuclear norm constrained problem (5.6):

$$\text{minimize}_{\boldsymbol{\mathcal{X}}} \quad \frac{1}{2}\left\|\mathcal{P}_\Omega[\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{D}}]\right\|_F^2 \quad \text{subject to} \quad \left\|\mathcal{A}[\boldsymbol{\mathcal{X}}]\right\|_* \leq \beta. \tag{D.7}$$

Here we denote $\mathcal{A}[\cdot]$ as the matricization operator, $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}_{\mathcal{I}}$, with $\mathcal{A}^*[\cdot]$ as its adjoint operator. Let the matrix $\mathbf{D} = \mathcal{A}[\boldsymbol{\mathcal{D}}]$ and $\widehat{\Omega} = \mathcal{A}[\Omega]$. It can be easily verified that

$$\mathbf{X}^\star \in \arg\min_{\mathbf{X}} \quad \frac{1}{2}\left\|\mathcal{P}_{\widehat{\Omega}}[\mathbf{X} - \mathbf{D}]\right\|_F^2 \quad \text{subject to} \quad \left\|\mathbf{X}\right\|_* \leq \beta, \tag{D.8}$$

if and only if $\mathcal{A}^*[\mathbf{X}^\star]$ is an optimal solution to problem (D.7). Thus it is sufficient to solve problem (D.8). A direct implementation of Frank-Wolfe method for problem (D.8) is presented in Algorithm 3.

Algorithm 3 Frank-Wolfe method for problem (5.6)

Initialization: $\mathbf{X}^0 = \mathbf{0}$;
for $k = 0,\ 1,\ 2,\ \cdots$ **do**
  Compute $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively as the left- and right- singular vectors corresponding to the largest singular value of the matrix $\mathcal{P}_{\widehat{\Omega}}[\mathbf{X}^k - \mathbf{D}]$;
  $\gamma = \frac{2}{k+2}$;
  Update $\mathbf{X}^{k+1} = (1-\gamma)\mathbf{X}^k - \gamma\beta\boldsymbol{u}\boldsymbol{v}^T$;
end for

D.5. Frank-Wolfe algrorithm for problem (5.5)

Recall our nuclear norm constrained problem (5.5),

$$\text{minimize}_{\boldsymbol{\mathcal{X}}} \quad \frac{1}{2}\left\|\mathcal{P}_{\Omega}[\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{D}}]\right\|_F^2 \qquad \text{subject to} \qquad \left\|\mathcal{A}_i[\boldsymbol{\mathcal{X}}]\right\|_* \leq \beta_i \text{ for all } i \in [K]. \quad \text{(D.9)}$$

Here we consider $\mathcal{A}_i[\cdot]$ as the mode-i unfolding operator, $\mathcal{A}_i : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}_{(i)}$, with $\mathcal{A}_i^*[\cdot]$ as its adjoint operator. By considering $\mathbf{M}_i = \mathcal{A}_i[\boldsymbol{\mathcal{X}}]$, it can be easily verified that problem (D.9) is equivalent to

$$\begin{aligned} \text{minimize}_{\{\mathbf{M}_i\}_{i=1}^K} \quad & \frac{1}{2}\left\|\mathcal{P}_{\mathcal{A}_1[\Omega]}\left[\mathbf{M}_1 - \mathcal{A}_1[\boldsymbol{\mathcal{D}}]\right]\right\|_F^2 \\ \text{subject to} \quad & \|\mathbf{M}_i\|_* \leq \beta_i, \quad i = 1, 2, \cdots, K \\ & \mathcal{A}_1^*[\mathbf{M}_1] = \mathcal{A}_j^*[\mathbf{M}_j], \quad j = 2, 3, \cdots, K. \end{aligned} \quad \text{(D.10)}$$

Consider the following penalized version of problem (D.10),

$$\begin{aligned} \text{minimize}_{\{\mathbf{M}_i\}_{i=1}^K} \quad & \frac{1}{2}\left\|\mathcal{P}_{\mathcal{A}_1[\Omega]}\left[\mathbf{M}_1 - \mathcal{A}_1[\boldsymbol{\mathcal{D}}]\right]\right\|_F^2 + \frac{\rho}{2}\sum_{j=2}^{K}\left\|\mathcal{A}_1^*[\mathbf{M}_1] - \mathcal{A}_j^*[\mathbf{M}_j]\right\|_F^2 \\ \text{subject to} \quad & \|\mathbf{M}_i\|_* \leq \beta_i, \quad i = 1, 2, \cdots, K, \end{aligned} \quad \text{(D.11)}$$

where $\rho > 0$ is the penalty parameter. Then a direct implementation of Frank-Wolfe method for problem (D.11) would lead to Algorithm 4.

Algorithm 4 Frank-Wolfe method for problem (D.11)

Initialization: $\mathbf{M}_1^0 = \mathbf{0}, \mathbf{M}_2^0 = \mathbf{0}, \cdots, \mathbf{M}_K^0 = \mathbf{0}$;
for $k = 0,\ 1,\ 2,\ \cdots$ **do**
  $\gamma = \frac{2}{k+2}$;
  Compute $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ as the left- and right- singular vectors corresponding to the largest singular value of the matrix $\mathcal{P}_{\mathcal{A}_1[\Omega]}\left[\mathbf{M}_1^k - \mathcal{A}_1[\boldsymbol{\mathcal{D}}]\right] + \rho\sum_{j=2}^{K}\left(\mathbf{M}_1^k - \mathcal{A}_1\mathcal{A}_j^*[\mathbf{M}_j^k]\right)$;
  Update $\mathbf{M}_1^{k+1} = (1-\gamma)\mathbf{M}_1^k - \gamma\beta_1\boldsymbol{u}_1\boldsymbol{v}_1^T$;
  for $j = 2,\ 3, \cdots, K$ **do**
    Compute $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ as the left- and right- singular vectors corresponding to the largest singular value of the matrix $\rho(\mathbf{M}_j^k - \mathcal{A}_j\mathcal{A}_1^*[\mathbf{M}_1^k])$;
    Update $\mathbf{M}_j^{k+1} = (1-\gamma)\mathbf{M}_j^k - \gamma\beta_j\boldsymbol{u}_j\boldsymbol{v}_j^T$;
  end for
end for

In our video completion experiments, we set the number of iterations as 10^4 for both Algorithm 3 and Algorithm 4.

References

Amelunxen, Dennis. *Geometric analysis of the condition of the convex feasibility problem*. PhD thesis, PhD Thesis, Univ. Paderborn, 2011.

Amelunxen, Dennis, Lotz, Martin, McCoy, Michael B, and Tropp, Joel A. Living on the edge: A geometric theory of phase transitions in convex optimization. *Inform. Inference, 2014. To appear. Available at arXiv.org/abs/1303.6672*, 2013.

Eldar, YC, Needell, D, and Plan, Y. Uniqueness conditions for low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 33(2):309–314, 2012.

Frank, Marguerite and Wolfe, Philip. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Harchaoui, Zaid, Juditsky, Anatoli, and Nemirovski, Arkadi. Conditional gradient algorithms for norm-regularized smooth convex optimization. *arXiv preprint arXiv:1302.2325*, 2013.

Huang, Bo, Ma, Shiqian, and Goldfarb, Donald. Accelerated linearized bregman method. *Journal of Scientific Computing*, 54(2-3):428–453, 2013.

Jaggi, Martin. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 427–435, 2013.

Lai, Ming-Jun and Yin, Wotao. Augmented \ell_1 and nuclear-norm models with a globally linearly convergent algorithm. *SIAM Journal on Imaging Sciences*, 6(2):1059–1091, 2013.

Li, Liyuan, Huang, Weimin, Gu, IY-H, and Tian, Qi. Statistical modeling of complex backgrounds for foreground object detection. *Image Processing, IEEE Transactions on*, 13(11):1459–1472, 2004.

Liu, Ji, Musialski, Przemyslaw, Wonka, Peter, and Ye, Jieping. Tensor completion for estimating missing values in visual data. In *ICCV*, pp. 2114–2121, 2009.

Mu, Cun, Zhang, Yuqian, Wright, John, and Goldfarb, Donald. Scalable robust matrix recovery: Frank-wolfe meets proximal methods. *arXiv preprint arXiv:1403.7588*, 2014.

Nesterov, Yurii. A method for unconstrained convex minimization problem with the rate of convergence o (1/k2). In *Doklady AN SSSR*, volume 269, pp. 543–547, 1983.

Vershynin, Roman. Math 280 lectue notes. http://www-personal.umich.edu/~romanv/teaching/2006-07/280/lec6.pdf, 2007.

Vershynin, Roman. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Zhang, Hui, Cai, Jian-Feng, Cheng, Lizhi, and Zhu, Jubo. Strongly convex programming for exact matrix completion and robust principal component analysis. *Inverse Problems & Imaging*, 6(2), 2012.