Hadamard Wirtinger Flow for Sparse Phase Retrieval: Supplementary Materials

1 Understanding the Dynamics of Hadamard Wirtinger Flow

As discussed in Section 3, the Hadamard parametrization has previously been applied to problems such as sparse recovery (Hoff, 2017; Vaškevičius et al., 2019; Zhao et al., 2019) and matrix factorization (Gunasekar et al., 2017; Li et al., 2018; Arora et al., 2019), where it turns the additive updates of gradient descent into multiplicative updates. The combination of multiplicative updates and a small initialization was shown to lead to sparsity in the aforementioned problems, under the assumption of the restricted isometry property (RIP).

The problem of sparse phase retrieval that we consider is known to satisfy the RIP property, cf. (Voroninski and Xu, 2016) for instance, and a similar explanation on why on-support variables and off-support variables can be made to grow at different speeds also holds in our setting. We now provide the main intuition behind the convergence properties of HWF by considering the evolution of the algorithm at the population level, i.e. in the case when $m = \infty$. While a rigorous convergence investigation of HWF is outside the scope of the current submission, the analysis that we now provide is instrumental to construct a good initialization for Algorithm 1.

Consider the simplified setting where $\mathbf{x}^*$ is non-negative, i.e. $x_i^* \geq 0$ for all i. We can set $\mathbf{v} = \mathbf{0}$ in the parametrization, so that $\mathbf{x} = \mathbf{u}^2$. Further, assume that we have access to the population risk $f(\mathbf{x}) := \mathbb{E}[\ell(\mathbf{x}, \mathbf{Z})]$ (in other words, $m = \infty$), where $\mathbf{Z} = (Y, \mathbf{A})$ is defined by $Y = (\mathbf{A}^T\mathbf{x}^*)^2$. Its gradient can be computed as

$$\nabla f(\mathbf{x}) = \big(3\|\mathbf{x}\|_2^2 - 1\big)\mathbf{x} - 2\big(\mathbf{x}^T\mathbf{x}^*\big)\mathbf{x}^*. \tag{1}$$

Under these two assumptions, first consider the initialization $\mathbf{x}^0 = \alpha^2\mathbf{1}_n$ for some small constant $\alpha > 0$. We can directly track the evolution of the estimates $\mathbf{x}^t$ (note that we use lowercase letters, since with $m = \infty$ the sequence is not random anymore) generated by Algorithm 1 via

$$x_i^{t+1} = x_i^t\big(1 - 2\eta\big[\big(3\|\mathbf{x}^t\|_2^2 - 1\big)x_i^t - 2\big((\mathbf{x}^t)^T\mathbf{x}^*\big)x_i^*\big]\big)^2.$$

This suggests that the evolution of $\mathbf{x}^t$ can be divided into two phases: if $\|\mathbf{x}^t\|_2^2 < \frac{1}{3}$, all coordinates grow ($x_i^{t+1} > x_i^t$), while coordinates $i \in \mathcal{S}$ on the support do so at a faster rate. If $\|\mathbf{x}^t\|_2^2 > \frac{1}{3}$, coordinates $i \notin \mathcal{S}$ decrease ($x_i^{t+1} < x_i^t$), while coordinates on the support increase if the product of the signal component x_i^* and the inner product $(\mathbf{x}^t)^T\mathbf{x}^*$ is larger than the term $(3\|\mathbf{x}^t\|_2^2 - 1)x_i^t$.

If we choose $\alpha > 0$ small enough, we expect x_j^t to still be small (e.g. $< 1/n$) for $j \notin \mathcal{S}$ when $\|\mathbf{x}^t\|_2^2 \geq \frac{1}{3}$ first occurs, as x_i^t grows at a faster rate than x_j^t for $i \in \mathcal{S}$. Since x_j^t decreases for $j \notin \mathcal{S}$ when $\|\mathbf{x}^t\|_2^2 \geq \frac{1}{3}$, we expect x_j^t to stay small throughout the algorithm for $j \notin \mathcal{S}$.

The smaller the step size η is, the more iterations are needed for the algorithm to converge. On the other hand, η cannot be too large; to illustrate this, consider the simplest case $n = 1$. The (scalar) gradient update becomes $x^{t+1} = x^t(1 - 6\eta[(x^t)^3 - x^t)])$, and x^t diverges if η is too large. We found a constant step size $\eta = 0.1$ to work well in our simulations.

This recursion has three types of fixed points: $\mathbf{x}^{(1)} = \mathbf{0}$, any $\mathbf{x}^{(2)}$ satisfying $\|\mathbf{x}^{(2)}\|_2^2 = \frac{1}{3}$ and $(\mathbf{x}^{(2)})^T\mathbf{x}^* = 0$, and $\mathbf{x}^{(3)} = \pm\mathbf{x}^*$. The first fixed point $\mathbf{x}^{(1)}$ is repelling, as all coordinates grow if $\|\mathbf{x}^t\|_2^2 < \frac{1}{3}$. Similarly, the second fixed point $\mathbf{x}^{(2)}$ is repelling as x_i^t grows at a faster rate than x_j^t for $i \in \mathcal{S}, j \notin \mathcal{S}$. This leaves only $\mathbf{x}^{(3)}$, which is an attracting fixed point of the recursion. Thus, we expect Algorithm 1 to converge to $\mathbf{x}^*$ if m is sufficiently large.

Guided by this intuition, we aim to construct an initialization $\mathbf{X}^0$ with $(\mathbf{X}^0)^T\mathbf{x}^*$ large (more precisely, we will have $|(\mathbf{X}^0)^T\mathbf{x}^*| \geq \frac{1}{4}x_{max}^*$), while at the same time $\|\mathbf{X}^0\|_2^2$ should not be too large (e.g. fixed to $\|\mathbf{X}^0\|_2^2 = \frac{1}{3}\|\mathbf{x}^*\|_2$; note that any other constant would also work, and that we use the estimate $\hat{\theta} = (\frac{1}{m}\sum_{j=1}^m Y_j)^{1/2}$ of the signal size

$\|\mathbf{x}^*\|_2$, see e.g. (Candès et al., 2015; Wang et al., 2017)). In order to obtain such an initialization, it suffices to find a coordinate $i \in [n]$ with $|x_i^*| \geq \frac{1}{2}x_{max}^*$. Then, we can set $X_i^0 = \hat{\theta}/\sqrt{3}$ and $X_j^0 = 0$ for all $j \neq i$. Note that such an initialization is not necessary, and even with a random initialization (e.g. U_i^0, V_i^0 set to small random noise for all $i = 1, \ldots, n$), the above intuition that coordinates $i \in \mathcal{S}$ on the support grow at a faster rate than coordinates $i \notin \mathcal{S}$ not on the support, continues to hold. However, the initial inner product $(\mathbf{X}^0)^T\mathbf{x}^*$ is closer to zero with random initialization compared to our proposed initialization, which leads to the population gradient $\nabla f(\mathbf{X}^t)_i$ initially being close to zero for all $i = 1, \ldots, n$, and therefore slow convergence.

Define the random variables $R_i = \frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2$ for $i = 1, \ldots, n$. These quantities were also used in (Wang et al., 2018) for support recovery, as one can compute $\mathbb{E}[R_i] = \|\mathbf{x}^*\|_2^2 + 2x_i^2$ using the assumption $\mathbf{A}_j \sim \mathcal{N}(0, \mathbf{I}_n)$ i.i.d.. Hence, if the number of measurements m is large, the random variables $\{R_i\}_{i=1}^n$ will concentrate around their means, separating them for $i \in \mathcal{S}$ and $i \notin \mathcal{S}$. This intuition suggests the initialization proposed in Section 3.

2 Proof of Lemma 1

In the following, we assume, without loss of generality, that $\|\mathbf{x}^*\|_2 = 1$; this assumption is made purely for notational simplicity, since we then have $x_{max}^* = \max_i \frac{|x_i^*|}{\|\mathbf{x}^*\|_2} = \max_i |x_i^*|$ and $x_{min}^* = \min_{i:x_i^* \neq 0} \frac{|x_i^*|}{\|\mathbf{x}^*\|_2} = \min_{i:x_i^* \neq 0} |x_i^*|$. If $\|\mathbf{x}^*\|_2 \neq 1$ is unknown, then we only need to replace x_{max}^* and x_{min}^* with $\max_i |x_i^*| = x_{max}^*\|\mathbf{x}^*\|_2$ and $\min_{i:x_i^* \neq 0} |x_i^*| = x_{min}^*\|\mathbf{x}^*\|_2$ respectively in the following proof. Further, note that knowledge of $\|\mathbf{x}^*\|_2$ is not required for HWF.

The proof of Lemma 1 relies on the following result, which is a combination of Theorems 3.6 and 3.7 of (Chung and Lu, 2006).

Theorem 3. *(Chung and Lu, 2006) Let X_i be independent random variables satisfying $|X_i| \leq M$ for all $i \in [n]$. Let $X = \sum_{i=1}^n X_i$ and $\|X\| = \sqrt{\sum_{i=1}^n \mathbb{E}[X_i^2]}$. Then, we have*

$$\mathbb{P}[|X - \mathbb{E}[X]| > \lambda] \leq 2\exp\left(-\frac{\lambda^2}{2(\|X\|^2 + M\lambda/3)}\right).$$

Proof of the first claim.

We first show that by choosing the largest instance in $\{\frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2\}_{i=1}^n$, we obtain an index i with $|x_i^*| \geq \frac{x_{max}^*}{2}$ with high probability. Recall that we write $R_i = \frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2$. We can compute

$$\begin{aligned}
\mathbb{E}[R_i] &= \mathbb{E}[(\mathbf{A}_1^T\mathbf{x}^*)^2 A_{1i}^2] \\
&= \mathbb{E}[A_{1i}^4 (x_i^*)^2 + (\mathbf{A}_{1,-i}^T \mathbf{x}_{-i}^*)^2 A_{1i}^2] \\
&= 3(x_i^*)^2 + \|\mathbf{x}_{-i}^*\|_2^2 \\
&= \|\mathbf{x}^*\|_2^2 + 2(x_i^*)^2,
\end{aligned}$$

where we denote by $\mathbf{x}_{-i} \in \mathbb{R}^{n-1}$ the vector obtained by deleting the i-th entry from $\mathbf{x} \in \mathbb{R}^n$ and use the fact that $A_{ji} \sim \mathcal{N}(0,1)$ i.i.d. and hence $\mathbf{A}_{j,-i}^T \mathbf{x}_{-i}^* \sim \mathcal{N}(0, \|\mathbf{x}_{-i}^*\|_2^2)$, as $\mathbf{x}^* \in \mathbb{R}^n$ is a fixed vector independent of the measurement vectors $\{\mathbf{A}_j\}_{j=1}^m$.

Let $I_{max} = \operatorname{argmax}_i R_i$. By definition, $R_{I_{max}} \geq R_i$ holds for all $i \in [n]$. If we can show $|R_i - \mathbb{E}[R_i]| \leq \frac{3}{4}(x_{max}^*)^2$ for all $i \in [n]$, then this would imply

$$\begin{aligned}
\|\mathbf{x}^*\|_2^2 + 2(x_{I_{max}}^*)^2 &= \mathbb{E}[R_{I_{max}}] \\
&= \mathbb{E}[R_i] + (R_i - \mathbb{E}[R_i]) + (\mathbb{E}[R_{I_{max}}] - R_{I_{max}}) + (R_{I_{max}} - R_i) \\
&\geq \mathbb{E}[R_i] - 2\max_j |R_j - \mathbb{E}[R_j]| \\
&\geq \|\mathbf{x}^*\|_2^2 + 2(x_i^*)^2 - \frac{3}{2}(x_{max}^*)^2,
\end{aligned}$$

for any $i \in [n]$. In particular, if we choose $i = \operatorname{argmax}_j |x_j^*|$, this implies $|x_{I_{max}}^*| \geq \frac{1}{2}x_{max}^*$, which concludes the proof of the first claim.

In order to show $|R_i - \mathbb{E}[R_i]| \le \frac{3}{4}(x^*_{max})^2$, we use the following truncation argument: for any $i \in [n]$, we write

$$R_i = \frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2 = \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_j^T\mathbf{x}^*)^2 A_{ji}^2 = \frac{1}{m}\sum_{j=1}^m (Z_{1,j} + Z_{2,j}),$$

where $Z_{1,j} = (\mathbf{A}_j^T\mathbf{x}^*)^2 A_{ji}^2 \cdot \mathbf{1}(\max\{|\mathbf{A}_j^T\mathbf{x}^*|, |A_{ji}|\} < \sqrt{44\log n})$ and $Z_{2,j} = (\mathbf{A}_j^T\mathbf{x}^*)^2 A_{ji}^2 - Z_{1,j}$. Since $Z_{1,j}$ is bounded, we can apply Theorem 3. To this end, compute the second moment

$$\sum_{j=1}^m \frac{1}{m^2}\mathbb{E}[Z_{1,j}^2] \le \sum_{j=1}^m \frac{1}{m^2}\mathbb{E}\big[(\mathbf{A}_j^T\mathbf{x}^*)^4 A_{ji}^4\big] \le \sum_{j=1}^m \frac{1}{m^2}\sqrt{\mathbb{E}\big[(\mathbf{A}_j^T\mathbf{x}^*)^8\big]\mathbb{E}\big[A_{ji}^8\big]} \le \frac{105}{m},$$

where we used the Cauchy-Schwarz inequality and the fact that $\mathbf{A}_j^T\mathbf{x}^* \sim \mathcal{N}(0,1)$. With this, we have

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{j=1}^m Z_{1,j} - \mathbb{E}[Z_{1,j}]\right| > \frac{3}{8}(x^*_{max})^2\right] \le 2\exp\left(-\frac{\frac{9}{64}(x^*_{max})^4}{2(\frac{105}{m} + \frac{44^2\log^2 n}{m}\cdot\frac{3}{8}(x^*_{max})^2/3)}\right) \le \mathcal{O}(n^{-11})$$

since $m \ge \mathcal{O}(\max\{k\log n, \log^3 n\}(x^*_{max})^{-2})$.

For the second term $Z_{2,j}$, we can use the Chebyshev inequality: we have

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{m}\sum_{j=1}^m Z_{2,j}\right) &\le \frac{1}{m}\mathbb{E}\Big[(\mathbf{A}_1^T\mathbf{x}^*)^4 A_{1i}^4 \cdot \mathbf{1}\Big(\max\{|\mathbf{A}_1^T\mathbf{x}^*|, |A_{1i}|\} > \sqrt{44\log n}\Big)\Big] \\
&\le \frac{1}{m}\sqrt{\mathbb{E}\big[(\mathbf{A}_1^T\mathbf{x}^*)^8 A_{1i}^8\big]\cdot\mathbb{P}\Big[\max\{|\mathbf{A}_1^T\mathbf{x}^*|, |A_{1i}|\} > \sqrt{44\log n}\Big]} \\
&\le \frac{45\sqrt{1001}}{m}\cdot 2n^{-11},
\end{aligned}$$

and hence, by the Chebyshev inequality,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{j=1}^m Z_{2,j} - \mathbb{E}[Z_{2,j}]\right| > \frac{3}{8}(x^*_{max})^2\right] \le \frac{\frac{45\sqrt{1001}}{m}\cdot 2n^{-11}}{\frac{9}{64}(x^*_{max})^4} \le \mathcal{O}(n^{-11}).$$

Put together, this implies that

$$\mathbb{P}\left[|R_i - \mathbb{E}[R_i]| > \frac{3}{4}(x^*_{max})^2\right] \le \mathcal{O}(n^{-11}).$$

Taking the union bound over all $i \in [n]$ implies that $|R_i - E[R_i]| \le \frac{3}{4}(x^*_{max})^2$ holds for all $i \in [n]$ with probability at least $1 - O(n^{-10})$. This concludes the proof of the first claim of Lemma 1.

Proof of the second claim.

First, note that the gradient of the empirical risk $F(\mathbf{x})$ is given by

$$\nabla F(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^m \big((\mathbf{A}_j^T\mathbf{x})^2 - (\mathbf{A}_j^T\mathbf{x}^*)^2\big)(\mathbf{A}_j^T\mathbf{x})\mathbf{A}_j.$$

By the dominated convergence theorem, the gradient of the population risk $f(\mathbf{x})$ can then be computed as

$$\begin{aligned}
\nabla f(\mathbf{x}) = \mathbb{E}[\nabla F(\mathbf{x})] &= \mathbb{E}\big[\big((\mathbf{A}_1^T\mathbf{x})^2 - (\mathbf{A}_1^T\mathbf{x}^*)^2\big)(\mathbf{A}_1^T\mathbf{x})\mathbf{A}_1\big] \\
&= \big(3\|\mathbf{x}\|_2^2 - 1\big)\mathbf{x} - 2\big(\mathbf{x}^T\mathbf{x}^*\big)\mathbf{x}^*
\end{aligned}$$

for any fixed vector $\mathbf{x} \in \mathbb{R}^n$. Further, we have the initialization

$$\begin{aligned}
U_i^0 &= \begin{cases} \left(\frac{\hat{\theta}}{\sqrt{3}} + \alpha^2\right)^{\frac{1}{2}} & i = I_{max} \\ \alpha & i \ne I_{max} \end{cases} \\
V_i^0 &= \alpha
\end{aligned}$$

which leads to

$$X_i^0 = \begin{cases} \frac{\hat{\theta}}{\sqrt{3}} & i = I_{max} \\ 0 & i \neq I_{max} \end{cases}$$

Hence, we have

$$\nabla f(\mathbf{X}^0)_i = \big(\hat{\theta}^2 - 1\big) X_i^0 - \frac{2\hat{\theta}}{\sqrt{3}} x^*_{I_{max}} x^*_i.$$

In particular, we have $\nabla f(\mathbf{X}^0)_j = 0$ for $j \notin \mathcal{S}$. Using standard concetration bounds for sub-exponential random variables (see e.g. Prop. 5.16 of (Vershynin, 2012)), we can bound with probability $1 - \mathcal{O}(n^{-10})$ (recall that we have assumed $\|\mathbf{x}^*\|_2 = 1$ for notational simplicity),

$$|\hat{\theta}^2 - 1| = \left| \frac{1}{m} \sum_{j=1}^{m} (\mathbf{A}_j^T \mathbf{x}^*)^2 - 1 \right| \leq 9\sqrt{\frac{\log n}{m}}.$$

This bound implies $2\hat{\theta} \geq \sqrt{3}$, where we used that $m \geq \mathcal{O}(\max\{k \log n, \log^3 n\}(x^*_{max})^{-2})$.

For the second claim we need to show that $|X_i^1| > |X_j^1|$ holds whenever $i \in \mathcal{S}$ and $j \notin \mathcal{S}$. We can assume without loss of generality that $x^*_{I_{max}} > 0$. First, consider the case $i \neq I_{max}$. Since $|X_i^1| = |(U_i^1)^2 - (V_i^1)^2|$, it suffices to show, assuming $x_i^* > 0$, that

$$U_i^1 > \max\{U_j^1, V_j^1\} \tag{2}$$

$$V_i^1 < \min\{U_j^1, V_j^1\} \tag{3}$$

holds simultaneously. The case $x_i^* < 0$ can be dealt with the same way, exchanging the roles of U_i^1 and V_i^1. We can bound

$$\begin{aligned} U_i^1 &= \alpha\big(1 - 2\eta \nabla F(\mathbf{X}^0)_i\big) \\ &\geq \alpha\big(1 - 2\eta \nabla f(\mathbf{X}^0)_i - 2\eta |\nabla F(\mathbf{X}^0)_i - \nabla f(\mathbf{X}^0)_i|\big), \end{aligned}$$

and

$$U_j^1 \leq \alpha\big(1 + 2\eta |\nabla F(\mathbf{X}^0)_j - \nabla f(\mathbf{X}^0)_j|\big).$$

We have shown above that (recall that $X_i^0 = 0$ for $i \neq I_{max}$)

$$-\nabla f(\mathbf{X}^0)_i = \frac{2\hat{\theta}}{\sqrt{3}} x^*_{I_{max}} x^*_i \geq \frac{1}{2} x^*_{max} x^*_{min},$$

since from the first part we know that $x^*_{I_{max}} \geq \frac{1}{2} x^*_{max}$ and we assumed $x_i^* > 0$. Hence, if we can show

$$\max_i |\nabla F(\mathbf{X}^0)_i - \nabla f(\mathbf{X}^0)_i| \leq \frac{1}{4} x^*_{max} x^*_{min}, \tag{4}$$

then $U_i^1 \geq U_j^1$ follows. We also have

$$V_j^1 \leq \alpha\big(1 + 2\eta |\nabla F(\mathbf{X}^0)_j - \nabla f(\mathbf{X}^0)_j|\big),$$

which then implies $U_i^1 \geq V_j^1$, completing the proof of (2); (3) can be shown the same way.

The case $i = I_{max}$ also follows from the bound (4). Since $m \geq \mathcal{O}(k(x^*_{max})^{-2} \log n)$, we can bound

$$\begin{aligned} |\nabla F(\mathbf{X}^0)_i| &\leq |\nabla f(\mathbf{X}^0)_i| + |\nabla F(\mathbf{X}^0)_i - \nabla f(\mathbf{X}^0)_i| \\ &\leq 9\sqrt{\frac{\log n}{m}} \frac{\hat{\theta}}{\sqrt{3}} + \frac{2\hat{\theta}}{\sqrt{3}} (x^*_{max})^2 + \frac{1}{2\sqrt{3}} x^*_{max} x^*_{min} \\ &\leq 2, \end{aligned}$$

where we used that $x^*_{max} \leq 1$. Since we assume $\eta \leq 0.1$, we can bound

$$|2\eta \nabla F(\mathbf{X}^0)_i| \leq 0.4,$$

which, since also $\alpha \leq 0.1$, implies

$$U_i^1 \geq \left(\frac{\hat{\theta}}{\sqrt{3}} + \alpha^2\right)^{\frac{1}{2}}(1 - 0.4) \geq 2\max\{U_j^1, V_j^1, V_i^1\},$$

and hence

$$|X_i^1| = (U_i^1)^2 - (V_i^1)^2 \geq \max\left\{(U_j^1)^2, (V_j^1)^2\right\} \geq |X_j^1|.$$

What is left to show is (4). Since $\mathbf{X}^0$ is not independent from $\{\mathbf{A}_j\}_{j=1}^m$, we cannot immediately apply the truncation argument from the proof of the first claim. Therefore, define the (deterministic) vectors $\mathbf{x}^{(l)} \in \mathbb{R}^n$ for $l = 1, ..., n$ by

$$x_i^{(l)} = \begin{cases} \frac{1}{\sqrt{3}} & i = l \\ 0 & i \neq l \end{cases}$$

Now, we need to show that the empirical gradient

$$\nabla F(\mathbf{x}^{(l)})_i = \frac{1}{m}\sum_{j=1}^m ((\mathbf{A}_j^T\mathbf{x}^{(l)})^2 - (\mathbf{A}_j^T\mathbf{x}^*)^2)(\mathbf{A}_j^T\mathbf{x}^{(l)})A_{ji}$$

is close to its expectation $\nabla f(\mathbf{x}^{(l)})_i$. Using the same truncation argument as in the proof of the first claim, we can show

$$\mathbb{P}\left[\left|\nabla f(\mathbf{x}^{(l)})_i - \nabla F(\mathbf{x}^{(l)})_i\right| \geq \frac{1}{8}x_{max}^* x_{min}^*\right] \leq \mathcal{O}(n^{-12}),$$

Taking the union bound over all i and l implies that

$$\max_l \max_i |\nabla F(\mathbf{x}^{(l)})_i - \nabla f(\mathbf{x}^{(l)})_i| \leq \frac{1}{8}x_{max}^* x_{min}^*$$

holds with probability $1 - \mathcal{O}\left(n^{-10}\right)$. The bound (4) now follows since $\mathbf{X}^0$ is close to $\mathbf{x}^{(I_{max})}$. We can write

$$\left|\nabla F(\mathbf{X}^0) - \nabla F(\mathbf{x}^{(I_{max})})\right| \leq \left|\frac{1}{m}\sum_{j=1}^m A_{ji}\left((\mathbf{A}_j^T\mathbf{X}^0)^3\right) - \left(\mathbf{A}_j^T\mathbf{x}^{(I_{max})}\right)^3\right| + \left|\frac{1}{m}\sum_{j=1}^m A_{ji}\left(\mathbf{A}_j^T\mathbf{x}^*\right)^2\left(\mathbf{A}_j^T\left(\mathbf{X}^0 - \mathbf{x}^{(I_{max})}\right)\right)\right|.$$

As both terms can be bounded the same way, we only demonstrate the following computations for the first term. Using the definitions and Hölder's inequality, we can bound

$$\begin{aligned}
\left|\frac{1}{m}\sum_{j=1}^m A_{ji}\left((\mathbf{A}_j^T\mathbf{X}^0)^3\right) - \left(\mathbf{A}_j^T\mathbf{x}^{(I_{max})}\right)^3\right| &= \left|\frac{1}{m}\sum_{j=1}^m A_{ji}A_{jI_{max}}^3\frac{\hat{\theta}^3 - 1}{3\sqrt{3}}\right| \\
&\leq \frac{1}{m}\sum_{j=1}^m \left|A_{ji}A_{jI_{max}}^3\right| \cdot \left|\frac{\hat{\theta}^3 - 1}{3\sqrt{3}}\right| \\
&\leq \left(\frac{1}{m}\sum_{j=1}^m A_{ji}^4\right)^{1/4}\left(\frac{1}{m}\sum_{j=1}^m A_{jI_{max}}^4\right)^{3/4}\left|\frac{\hat{\theta}^3 - 1}{3\sqrt{3}}\right|.
\end{aligned}$$

It follows from standard Gaussian concentration that the first two sums are bounded by $\mathcal{O}(1)$ with high probability. As shown above, we can bound

$$\left|\frac{\hat{\theta}^3 - 1}{3\sqrt{3}}\right| \leq \mathcal{O}\left(\sqrt{\frac{\log n}{m}}\right) \leq \mathcal{O}(x_{max}^* x_{min}^*),$$

where we used the assumption $x_{min}^* \geq \Omega(1/\sqrt{k})$. Repeating the same computation for the second term, we can show that

$$\left|\nabla F(\mathbf{X}^0) - \nabla F(\mathbf{x}^{(I_{max})})\right| \leq \frac{1}{16}x_{max}^* x_{min}^*.$$

Recalling the definition of the population gradient ∇f, we can also bound

$$\left|\nabla f(\mathbf{X}^0) - \nabla f(\mathbf{x}^{(I_{max})})\right| \leq \frac{1}{16}x_{max}^* x_{min}^*,$$

which completes the proof of (4) and therefore also completes the proof of Lemma 1. □

References

Arora, S., Cohen, N., Hu, W., and Luo, Y. (2019). Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7411–7422.

Candès, E. J., Li, X., and Soltanolkotabi, M. (2015). Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Chung, F. and Lu, L. (2006). Concentration inequalities and martingale inequalities: a survey. *Internet Math.*, 3(1):79–127.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017). Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159.

Hoff, P. D. (2017). Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198.

Li, Y., Ma, T., and Zhang, H. (2018). Algorithmic regularization in over-parametrized matrix sensing and neural networks with quadratic activation. In *Conference on Learning Theory*, pages 2–47.

Vaškevičius, T., Kanade, V., and Rebeschini, P. (2019). Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, pages 2968–2979.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. and Kutyniok, G., editors, *Compressed Sensing, Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, Cambridge.

Voroninski, V. and Xu, Z. (2016). A strong restricted isometry property, with an application to phaseless compressed sensing. *Applied and Computational Harmonic Analysis*, 40(2):386–395.

Wang, G., Giannakis, G. B., and Eldar, Y. C. (2017). Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794.

Wang, G., Zhang, L., Giannakis, G. B., Akçakaya, M., and Chen, J. (2018). Sparse phase retrieval via truncated amplitude flow. *IEEE Transactions on Signal Processing*, 66(2):479–491.

Zhao, P., Yang, Y., and He, Q.-C. (2019). Implicit regularization via Hadamard product over-parametrization in high-dimensional linear regression. *arXiv preprint arXiv:1903.09367*.