Supplementary Materials

A Proof of Theorem 1: Upper Bound for Sign-Consistent JL Matrices

A.1 Proof of Lemma 4.1

Racall that we want to upper bound

$$s^t \cdot \mathbb{E}[Z^t] = \sum_{\substack{i_1,\ldots,i_t,j_1,\ldots,j_t\in[d]\\ i_1\neq j_1,\ldots,i_t\neq j_t}} \left(\prod_{u=1}^{t} x_{i_u}x_{j_u}\right)\left(\mathop{\mathbb{E}}_{\sigma}\prod_{u=1}^{t}\sigma_{i_u}\sigma_{j_u}\right)\left(\mathop{\mathbb{E}}_{\eta}\prod_{u=1}^{t}\sum_{r=1}^{m}\eta_{r,i_u}\eta_{r,j_u}\right), \tag{A.1}$$

We first show that

Lemma 4.1.

$$s^t \cdot \mathbb{E}[Z^t] \le e^t \sum_{v=2}^{t}\sum_{G\in\mathcal{G}''_{v,t}} \left(\frac{1}{t^t}\prod_{p=1}^{v}\sqrt{d_p}^{d_p}\right)\cdot \sum_{r_1,\ldots,r_t\in[m]}\prod_{i=1}^{w}\left(\frac{s}{m}\right)^{v_i} .$$

Here,

- $\mathcal{G}''_{v,t}$ *is a set of directed multigraphs with* v *labeled vertices (*1 *to* v*) and* t *labeled edges (*1 *to* t*).*
- d_p *is the total degree of vertex* $p \in [v]$ *in a graph* $G \in \mathcal{G}''_{v,t}$*.*[11]
- w *and* $v_1, \ldots, v_w$ *are defined by* G *and* $r_1, \ldots, r_t$ *as follows. Let an edge* $u \in [t]$ *be colored with* $r_u \in [m]$*, then we define* w *to be the number of distinct colors used in* $r_1, \ldots, r_t$*, and* v_i *to be the number of vertices incident to an edge with color* $i \in [w]$*.*

Proof. We prove the desired inequality from (A.1) in three steps. The first step removes the random variables of σ in (A.1). The second step removes x from (A.1) using the assumption of $\|x\|_2 = 1$. The third step removes the random variables η in (A.1) by carefully exploiting the independence or negative correlation among different η terms.

In the first step, we use a standard trick to map each summand

$$\left(\prod_{u=1}^{t} x_{i_u}x_{j_u}\right)\left(\mathop{\mathbb{E}}_{\sigma}\prod_{u=1}^{t}\sigma_{i_u}\sigma_{j_u}\right)\left(\mathop{\mathbb{E}}_{\eta}\prod_{u=1}^{t}\sum_{r=1}^{m}\eta_{r,i_u}\eta_{r,j_u}\right)$$

in expression (A.1) to a directed multigraph. That is, for each pair of (i_u, j_u) where $u \in [t]$, we associate it with a directed edge $i_u \to j_u$. It is easy to see that it suffices for us to consider only graphs with all the vertices having *even* total degree, since otherwise the expectation $\mathbb{E}_\sigma \prod_{u=1}^{t}\sigma_{i_u}\sigma_{j_u}$ becomes zero (e.g., $\mathbb{E}_\sigma[\sigma_1^3\sigma_2^2\sigma_4] = 0$).

To make this precise, let us define $\mathcal{G}_t$ to be the set of directed multigraphs G with the following properties:

- G has between 2 and t (identical) vertices.
- G has exactly t *distinct* edges, labels by $1, 2, \ldots, t$.
- There are no self-loops.

[11] The total degree of a vertex is defined as the number of incident edges regardless of direction.

- Each vertex has a non-zero and even total degree (sum of in- and out-degrees).

Note that we intentionally made the vertices *identical (i.e., unlabeled)* in the above definition, and we will separately enumerate over the vertex labeling.

Let f be a map from $(i_u, j_u)_{u\in[t]}$ to its underlying graph G by adding a directed edge $i_u \to j_u$ as the u-th edge of a graph. Our argument above shows that in order to enumerate $(i_u, j_u)_{u\in[t]}$ in (A.1), it suffices to enumerate $G \in \mathcal{G}_t$ and the vertex labeling as follows

$$s^t \cdot \mathbb{E}[Z^t] = \sum_{G\in\mathcal{G}_t} \sum_{\substack{i_1\neq j_1,\dots,i_t\neq j_t\in[d] \\ f((i_u,j_u)_{u=1}^t)=G}} \left(\prod_{u=1}^t x_{i_u} x_{j_u}\right) \left(\mathbb{E}_{\eta} \prod_{u=1}^t \sum_{r=1}^m \eta_{r,i_u}\eta_{r,j_u}\right)$$

In the above expression, the $\mathbb{E}_\sigma \prod_{u=1}^t \sigma_{i_u}\sigma_{j_u}$ factors have disappeared because they equal to one if G has even total degrees for all of its vertices. Also, the second summation —the one over all choices of $(i_u, j_u)_u$ such that $f((i_u, j_u)) = G$— is in fact an enumeration over the missing vertex labeling of the graph G.

In the second step, we observe that $\eta_{\star,i}$ and $\eta_{\star,j}$ for $i \neq j$ are independent because they are for different columns, and generated by the same random process. Thus, for a given graph $G \in \mathcal{G}_t$, the $\mathbb{E}_\eta \prod_{u=1}^t \sum_{r=1}^m \eta_{r,i_u}\eta_{r,j_u}$ factor has the same value for all mappings with $f((i_u, j_u)_{u=1}^t) = G$ (i.e., for all the vertex labeling).[12] Let us call this function $\hat{\eta}(G)$ and write:

$$s^t \cdot \mathbb{E}[Z^t] = \sum_{G\in\mathcal{G}_t} \sum_{\substack{i_1\neq j_1,\dots,i_t\neq j_t\in[d] \\ f((i_u,j_u)_{u=1}^t)=G}} \left(\prod_{u=1}^t x_{i_u} x_{j_u}\right) \hat{\eta}(G) = \sum_{G\in\mathcal{G}_t} \hat{\eta}(G) \cdot \sum_{\substack{i_1\neq j_1,\dots,i_t\neq j_t\in[d] \\ f((i_u,j_u)_{u=1}^t)=G}} \left(\prod_{u=1}^t x_{i_u} x_{j_u}\right) . \tag{A.2}$$

Next, for a fixed graph $G \in \mathcal{G}_t$, let v be the number of vertices in G and d_p the total degree of vertex $p \in [v]$. We observe a simple fact that

$$\binom{t}{d_1/2, \dots, d_v/2} \cdot \sum_{\substack{i_1\neq j_1,\dots,i_t\neq j_t\in[d] \\ f((i_u,j_u)_{u=1}^t)=G}} \left(\prod_{u=1}^t x_{i_u} x_{j_u}\right) \le \left(\sum_{l=1}^d x_l^2\right)^t \cdot v! = v! . \tag{A.3}$$

The above inequality holds as each (distinct) monomial in $\sum_{i_1\neq j_1,\dots,i_t\neq j_t\in[d], f((i_u,j_u)_{u=1}^t)=G} \left(\prod_{u=1}^t x_{i_u} x_{j_u}\right)$, for instance appears at most $v!$ times in this summation due to vertex re-labeling, and thus $\binom{t}{d_1/2,\dots,d_v/2} \cdot v!$ times in total on the left hand side; each of these monomials also appear on the right hand side exactly $\binom{t}{d_1/2,\dots,d_v/2} \cdot v!$ times; and finally, each monomial is non-negative and $\|x\|_2 = 1$.

Now we are ready to plug (A.3) to (A.2) and get

$$\begin{aligned} s^t \cdot \mathbb{E}[Z^t] = \sum_{G\in\mathcal{G}_t} \sum_{\substack{i_1\neq j_1,\dots,i_t\neq j_t\in[d] \\ f((i_u,j_u)_{u=1}^t)=G}} \left(\prod_{u=1}^t x_{i_u} x_{j_u}\right) \hat{\eta}(G) &\le \sum_{G\in\mathcal{G}_t} \frac{v!}{\binom{t}{d_1/2,\dots,d_v/2}} \hat{\eta}(G) \\ &= \sum_{G\in\mathcal{G}'_t} \frac{1}{\binom{t}{d_1/2,\dots,d_v/2}} \hat{\eta}(G) \qquad \text{(A.4)} \\ &\le e^t \sum_{G\in\mathcal{G}'_t} \frac{1}{t^t} \prod_{p=1}^v \sqrt{d_p}^{d_p} \hat{\eta}(G) \qquad \text{(A.5)} \end{aligned}$$

[12]In fact, the value does not depend on the edge labeling of G as well, but we are not going to use this fact.

Here in (A.4), we have defined $\mathcal{G}_t'$ to be the same as $\mathcal{G}_t$ except that we require the v vertices to have distinct labels in $[v]$, and (A.4) follows because each there are $v!$ distinct ways to label each $G \in \mathcal{G}_t$.[13] For (A.5), we use that $t! \geq t^t/e^t$ and $\prod_{p=1}^{v}(d_p/2)! \leq \prod_{p=1}^{v}\sqrt{d_p}^{d_p}$. We have been ambiguous when writing $\hat{\eta}(G)$ because G may either be vertex-labelled or not vertex-labelled; its value is independent of such a labeling.

In the third step, we give an upper bound on $\hat{\eta}(G)$ by carefully exploiting the independence or negative correlation among the random variables in it. We first rewrite

$$\hat{\eta}(G) = \mathop{\mathbb{E}}_{\eta} \prod_{u=1}^{t} \sum_{r=1}^{m} \eta_{r,i_u}\eta_{r,j_u} = \sum_{r_1,\ldots,r_t \in [m]} \mathop{\mathbb{E}}_{\eta} \prod_{u=1}^{t} \eta_{r_u,i_u}\eta_{r_u,j_u}$$

From this point, whenever we fix a graph G and a sequence $r = (r_1, \ldots, r_t) \in [m]^t$, we would like to view them together as a *directed and edge-colored multigraph* (G, r) —i.e., graph G appended with edge colors such that its u-th edge $i_u \to j_u$ is given the color $r_u \in [m]$.

The big advantage of such edge coloring is to allow us to exploit the negative correlation between graphs of different colors. Indeed, for any fixed $G \in \mathcal{G}_t$ and $r \in [m]^t$, let us define

$$\tilde{\eta}_c(G, r) \stackrel{\text{def}}{=} \prod_{u \in [t], r_u = c} \eta_{r_u,i_u}\eta_{r_u,j_u}$$

to be the factors of η associated with color $c \in [m]$. Then we have

$$\hat{\eta}(G) = \sum_{r_1,\ldots,r_t \in [m]} \mathop{\mathbb{E}}_{\eta} \prod_{c=1}^{m} \tilde{\eta}_c(G, r) \leq \sum_{r_1,\ldots,r_t \in [m]} \prod_{c=1}^{m} \mathop{\mathbb{E}}_{\eta}[\tilde{\eta}_c(G, r)]$$

Here the inequality is owing to the fact that different rows of η are negatively correlated.[14]

Next, let us denote by $w \in [t]$ the number of distinct colors in (G, r). For notational simplicity, we can assume that the used colors in G are $1, 2, \ldots, w$ (so $w + 1, \ldots, m$ are unused). Let G_i be the subgraph of G containing all the edges of color $i \in [w]$, and suppose that G_i has $v_i \geq 2$ vertices and $c_i \geq 1$ edges.

It is straightforward to see that for a fixed color $i \in [w]$, there are precisely v_i distinct η factors in the definition of $\tilde{\eta}_i(G, r)$ (by the definition that G_i has v_i "vertices"). Since these η factors are across different columns, they are independent and each has a probability of $\frac{s}{m}$ to be 1 (due to our probabilistic construction of $\mathcal{A}$). We therefore can simply write $\mathbb{E}_\eta[\tilde{\eta}_i(G, r)] = \left(\frac{s}{m}\right)^{v_i}$ and conclude that

$$\hat{\eta}(G) \leq \sum_{r_1,\ldots,r_t \in [m]} \prod_{i=1}^{w} \left(\frac{s}{m}\right)^{v_i} \tag{A.6}$$

[13] All these labelings are distinct in $\mathcal{G}_t'$ because there is a canonical way to label the vertices of each $G \in \mathcal{G}_t$: since G does not have isolated vertices, to get a canonical labeling, we can order the directed edges in G in increasing order and label the vertices in this order as well.

[14] As a simple example, we have $\mathbb{E}[\eta_{2,4} \cdot \eta_{2,5} \cdot \eta_{3,4} \cdot \eta_{3,7}] \leq \mathbb{E}[\eta_{2,4} \cdot \eta_{2,5}] \cdot \mathbb{E}[\eta_{3,4} \cdot \eta_{3,7}]$ because: (a) $\eta_{2,4}$ is negatively correlated with $\eta_{3,4}$, and independent with $\eta_{3,7}$, and (b) $\eta_{2,5}$ is independent with both $\eta_{3,4}$ and $\eta_{3,7}$. In general, if an indicator variable is set to 1, the probability of other indicator variables being set to 1 in the same column and different row, decreases. Therefore, the product of expectations is always no less than the expectation of product of corresponding negatively correlated terms.

At last, we incorporate (A.6) in (A.5) and get

$$
\begin{aligned}
s^t \cdot \mathbb{E}[Z^t] &\le e^t \sum_{G \in \mathcal{G}'_t} \left(\frac{1}{t^t} \prod_{p=1}^{v} \sqrt{d_p}^{d_p} \right) \cdot \sum_{r_1,\dots,r_t \in [m]} \prod_{i=1}^{w} \left(\frac{s}{m} \right)^{v_i} \\
&\le e^t \sum_{v=2}^{t} \sum_{G \in \mathcal{G}''_{v,t}} \left(\frac{1}{t^t} \prod_{p=1}^{v} \sqrt{d_p}^{d_p} \right) \cdot \sum_{r_1,\dots,r_t \in [m]} \prod_{i=1}^{w} \left(\frac{s}{m} \right)^{v_i} \quad . \qquad \text{(A.7)}
\end{aligned}
$$

Here $\mathcal{G}''_{v,t}$ contains graphs with v labeled vertices and t labeled edges, without the restriction (like we did in $\mathcal{G}_t$ and $\mathcal{G}'_t$) that a vertex has a positive or even degree. We can have $v \le t$ because in $\mathcal{G}'_t$ each vertex must degree no less than 2, while the total degree over all vertices equal to $2t$. Therefore, going from $\mathcal{G}'$ to $\mathcal{G}''$ we only add non-negative terms and the inequality goes through. This concludes the proof of Lemma 4.1. □

A.2 Proof of Lemma 4.2

Recall that in Section 4 we proceed from Lemma 4.1 as follows. Instead of enumerating $G \in \mathcal{G}''_{v,t}$ as a whole, we now enumerate subgraphs of different colors separately, and then combine the results. Below is one way (and perhaps the only way the authors believe without incurring a $\log(1/\delta)$ factor loss in m) to enumerate G that can lead to tight upper bounds

$$
s^t \cdot \mathbb{E}[Z^t] \le e^t \underbrace{\sum_{v=2}^{t}}_{\text{i}} \underbrace{\sum_{w=1}^{t} \binom{m}{w}}_{\text{ii}} \underbrace{\sum_{\substack{c_1,\dots,c_w \\ c_1+\dots+c_w=t \\ c_i \ge 1}} \binom{t}{c_1,\dots,c_w}}_{\text{iii}} \underbrace{\sum_{\substack{v_1,\dots,v_w \\ 2 \le v_i \le 2c_i}}}_{\text{iv}} \left(\frac{s}{m} \right)^{v_1+\dots+v_w} \underbrace{\sum_{f_1,\dots,f_w}}_{\text{v}} \underbrace{\sum_{\forall i, G_i \in \mathcal{G}''_{v_i,c_i}}}_{\text{vi}} \frac{1}{t^t} \prod_{p=1}^{v} \sqrt{d_p}^{d_p} \qquad (4.1)
$$

This gigantic expression enumerates all graphs $G \in \mathcal{G}''_{v,t}$ and its coloring $r_1, \dots, r_t \in [m]$ in six steps:

(i). Number of graph vertices, $v \in \{2, \dots, t\}$; the vetices are labelled by $1, 2, \dots, v$.

(ii). Number of used edge colors, $w \in \{1, \dots, t\}$, and all $\binom{m}{w}$ possibilities of choosing w colors.

(iii). Edge colorings of the graph using selected w colors: how many (denoted by $c_i \ge 1$) edges are colored in color i and which of the t edges are colored in color i.

(iv). Number of vertices $v_i \in \{2, \dots, 2c_i\}$ in each G_i, the subgraph containing edges of color i.

(v). All possible increasing functions $f_i : [v_i] \to [v]$, such that $f_i(j)$ maps vertex j in G_i to the $f_i(j)$-th global vertex. (And we ensure $f_i(j) < f_i(k)$ for $j < k$ to reduce double counting.)

(vi). All graphs $G_i \in \mathcal{G}''_{v_i,c_i}$ with v_i labeled vertices (1 to v_i) and c_i labeled edges (1 to c_i).

(Using all the information above, d_p, the degree of vertex $p \in [v]$ is well defined.)

We emphasize here that any pair of graph $G \in \mathcal{G}''_{v,t}$ and coloring $r_1, \dots, r_t \in [m]$ will be generated *at least once* in the above procedure.[15] Thus, (4.1) follows from Lemma 4.1, since the summation terms also have the same value $\left(\frac{s}{m}\right)^{v_1+\dots+v_w} \frac{1}{t^t} \prod_{p=1}^{v} \sqrt{d_p}^{d_p}$.

It is now possible to consider G_i's separately in (4.1) and prove the following lemma:

[15]This follows from the fact that G and $r_1, \dots, r_t$ together determine (a) w, the number of used colors, (b) G_i for each $i \in [w]$ (with v_i vertices and c_i edges), the subgraph of G of the i-th used color, and (c) f_i, the vertex mapping from G_i back to G. Any such triple will be generated at least once in (4.1). Note also, we may have double counts but it will not affect our asymptotic upper bound.

Lemma 4.2. *From (4.1) we can get*

$$s^t \cdot \mathbb{E}[Z^t] \le 2^{O(t)} \sum_{v=2}^{t} \sum_{w=1}^{t} \binom{m}{w} \sum_{\substack{c_1,\ldots,c_w \\ c_1+\ldots+c_w=t \\ c_i\ge 1}} \binom{t}{c_1,\ldots,c_w} \sum_{\substack{v_1,\ldots,v_w \\ 2\le v_i \le 2c_i}} \prod_{j=1}^{w} \left(\frac{s}{m}\right)^{v_j} v_j^{c_j} \binom{v-1}{v_j-1}$$

Proof. From (4.1) it suffices to show that

$$\sum_{f_1,\ldots,f_w} \sum_{\forall i, G_i \in \mathcal{G}''_{v_i,c_i}} \frac{1}{t^t} \prod_{p=1}^{v} \sqrt{d_p}^{d_p} \le 2^{O(t)} \cdot \prod_{j=1}^{w} v_j^{c_j} \binom{v-1}{v_j-1} \tag{A.8}$$

Recall that here d_p remains to be the total degree of vertex $p \in [v]$ in the combined graph G, which is essentially $G_1 \cup \cdots \cup G_w$ but glued together using the vertex mappings $f_1, \ldots, f_w$.

To show (A.8), let us define:

$$\text{for any } \vec{\gamma} \in \mathbb{Z}^w_{\ge 0} \text{ and } \vec{a} \in \mathbb{R}^v_{>0}: \qquad S(\vec{\gamma}, \vec{a}) \stackrel{\text{def}}{=} \sum_{\forall i, G_i \in \mathcal{G}''_{v_i,\gamma_i}} \prod_{p=1}^{v} \sqrt{a_p}^{d_p} \ ,$$

where as before d_p is the degree of the p-th vertex in the combined graph $G = G_1 \cup \cdots \cup G_w$, but a_p is a constant. Ideally, we want an upper bound on $S(\vec{\gamma}, \vec{a})$ for the choice of $\vec{\gamma} = \vec{c}$ and $\vec{a} = \vec{d}$, so that $S(\vec{\gamma}, \vec{a})$ becomes identical to the left hand side of (A.8).[16] Thus, let us now shoot for an upper bound of $S(\vec{\gamma}, \vec{a})$ using induction on $\vec{\gamma}$.

When $\vec{\gamma} = \vec{0}$, observe that $S(\vec{0}, \vec{a}) = 1$ since each G_i has no edge in it and $d_p = 0$ for all $p \in [v]$.

Now, consider adding an edge to G of some color l. for any $\vec{\gamma}$, define $\vec{\gamma}'$ so that $\gamma'_l = \gamma_l + 1$ and $\forall j \ne l : \gamma'_j = \gamma_j$. Then,

$$\frac{S(\vec{\gamma}', \vec{a})}{S(\vec{\gamma}, \vec{a})} \le \sum_{\alpha \ne \beta \in [v_l]} \sqrt{a_{f_l(\alpha)}} \sqrt{a_{f_l(\beta)}} \le \sum_{\alpha=1}^{v_l} \left(\sqrt{a_{f_l(\alpha)}}\right)^2 \le \sum_{\alpha=1}^{v_l} \left(a_{f_l(\alpha)}\right) \cdot v_l$$

where the first inequality is because this new edge may be added anywhere between two vertices $f_l(\alpha)$ and $f_l(\beta)$ for $\alpha, \beta \in [v_l]$, the second inequality is by the simple expansion of square of sum, the last inequality is by Cauchy-Schwartz. Therefore, by induction we conclude that

$$\sum_{\forall i, G_i \in \mathcal{G}''_{v_i,c_i}} \prod_{p=1}^{v} \sqrt{a_p}^{d_p} = S(\vec{c}, \vec{a}) \le \prod_{j=1}^{w} \Big(\sum_{\alpha \in [v_j]} a_{f_j(\alpha)} \Big)^{c_j} \cdot v_j^{c_j} \ . \tag{A.9}$$

It is worth noting that (A.9) would be sufficient for us to show (A.8), if one could replace $\vec{a}$ by $\vec{d}$. However, since the degree vector $\vec{d}$ is determined *after* the choices of G_j for $j \in [w]$, this simple substitution is impossible and we need a different approach.

Indeed, we fix this by enumerating $G_i \in \mathcal{G}''_{v_i,c_i}$ in two steps: first enumerating the degrees $d'_1, \ldots, d'_v$ and then enumerating the possible G_i's satisfying such degree spectrum (i.e., $d_p = d'_p$ for all $p \in [v]$)

$$\sum_{\forall i, G_i \in \mathcal{G}''_{v_i,c_i}} \prod_{p=1}^{v} \sqrt{d_p}^{d_p} = \sum_{\substack{d'_1,\ldots,d'_v \ge 0 \\ d'_1+\cdots+d'_v = 2t}} \left(\sum_{\substack{\forall i, G_i \in \mathcal{G}''_{v_i,c_i} \\ \text{s.t.} \forall p, d_p = d'_p}} \prod_{p=1}^{v} \sqrt{d'_p}^{d_p} \right)$$

[16] However, this mission is non-trivial because the values of $\vec{d}$ are decided *after* $G_i \in \mathcal{G}''_{v_i,c_i}$ are chosen. Let us anyways ignore this issue for a moment and resolve it later.

This seemingly redundant separation in fact enables us to prove (A.8). Indeed, we proceed the above equation as follows

$$\begin{aligned}\sum_{\forall i, G_i \in \mathcal{G}''_{v_i,c_i}} \prod_{p=1}^{v} \sqrt{d_p}^{d_p} &\le \sum_{\substack{d'_1,\dots,d'_v \ge 0 \\ d'_1+\cdots+d'_v=2t}} \left(\sum_{\forall i, G_i \in \mathcal{G}''_{v_i,c_i}} \prod_{p=1}^{v} \sqrt{d'_p}^{d_p} \right) \\ &= \sum_{\substack{d'_1,\dots,d'_v \ge 0 \\ d'_1+\cdots+d'_v=2t}} S(\vec{c}, \vec{d'}) \le \sum_{\substack{d'_1,\dots,d'_v \ge 0 \\ d'_1+\cdots+d'_v=2t}} \prod_{j=1}^{w} \Big(\sum_{\alpha \in [v_j]} d'_{f_j(\alpha)} \Big)^{c_j} \cdot v_j^{c_j} \ . \end{aligned} \tag{A.10}$$

Here the first inequality gets rid of the $d_p = d'_p$ constraint, and the second one is from (A.9).

To proceed from here, we make use of the summation over $f_1, \dots, f_w$ that we intentionally ignored when defining $S(\vec{\gamma}, \vec{a})$, and get

$$\begin{aligned}\sum_{f_1,\dots,f_w} \prod_{j=1}^{w} \Big(\sum_{\alpha \in [v_j]} d'_{f_j(\alpha)} \Big)^{c_j} \cdot v_j^{c_j} &= \prod_{j=1}^{w} v_j^{c_j} \sum_{f_j} \Big(\sum_{\alpha \in [v_j]} d'_{f_j(\alpha)} \Big)^{c_j} \\ &\le \prod_{j=1}^{w} v_j^{c_j} \binom{v-1}{v_j-1} \cdot (2t)^{c_j} = (2t)^t \prod_{j=1}^{w} v_j^{c_j} \binom{v-1}{v_j-1} \end{aligned} \tag{A.11}$$

Above, the first equality is a simple swap between adacant $\sum$ and $\prod$. The inequality in (A.11) needs some justifications:

> Recall that the mapping f_j chooses v_j vertex labels out of $[v]$. If we represent $2t$ as the summation $d'_1 + \cdots + d'_v$, we have that $\sum_{\alpha\in[v_j]} d'_{f_j(\alpha)}$ is the partial sum over only the selected v_j vertices under f_j. Hence, for a fixed f_j, each monomial in the expansion of $\big(\sum_{\alpha\in[v_j]} d'_{f_j(\alpha)}\big)^{c_j}$ also appears in $(2t)^{c_j} = (d'_1+\cdots+d'_v)^{c_j}$ with the same coefficient. However, any such monomial can appear in at most $\binom{v-1}{v_j-1}$ different f_j mappings: each such monimal contains at least one vertex (so may look like $(d'_p)^{c_j}$ for some $p \in [v]$), and f_j could have the freedom to pick at most the $v_j - 1$ more vertices out of $v - 1$ to complete as an increasing mapping $[v_j] \to [v]$.

Finally, we plug (A.11) into (A.10) and get

$$\sum_{f_1,\dots,f_w} \sum_{\forall i, G_i \in \mathcal{G}''_{v_i,c_i}} \frac{1}{t^t} \prod_{p=1}^{v} \sqrt{d_p}^{d_p} \le \sum_{\substack{d'_1,\dots,d'_v \ge 0 \\ d'_1+\cdots+d'_v=2t}} \frac{(2t)^t}{t^t} \prod_{j=1}^{w} v_j^{c_j} \binom{v-1}{v_j-1} \le 2^{O(t)} \cdot \prod_{j=1}^{w} v_j^{c_j} \binom{v-1}{v_j-1}$$

where the last inequality is because the number of ways to partition $2t$ into $d'_1 + \cdots + d'_v$ less than $2^{O(t+v)} = 2^{O(t)}$. This concludes (A.8) and thus the proof of Lemma 4.2. □

A.3 Proof of Lemma 4.3

The last lemma of our proof is essentially to handle algebra manipulations in a careful way.

Lemma 4.3. *We can rearrange the inequality in (4.2) and get*

$$s^t \cdot \mathbb{E}[Z^t] \le 2^{O(t)} \cdot t^t \left(\frac{s^2}{m}\right)^t \ .$$

Proof. Simplifying the result of Lemma 4.2, we get:

$$s^t \cdot \mathbb{E}[Z^t] \le 2^{O(t)} \sum_{v=2}^{t} \sum_{w=1}^{t} \binom{m}{w} \sum_{\substack{c_1,\dots,c_w \\ c_1+\dots+c_w=t \\ c_i \ge 1}} \binom{t}{c_1,\dots,c_w} \sum_{\substack{v_1,\dots,v_w \\ 2\le v_i \le 2c_i}} \prod_{j=1}^{w} \left(\frac{s}{m}\right)^{v_j} v_j^{c_j} v^{v_j-1} \tag{A.12}$$

$$= 2^{O(t)} \sum_{v=2}^{t} \sum_{w=1}^{t} \binom{m}{w} \sum_{\substack{c_1,\dots,c_w \\ c_1+\dots+c_w=t \\ c_i \ge 1}} \binom{t}{c_1,\dots,c_w} \prod_{j=1}^{w} \frac{1}{v} \sum_{v_j=2}^{2c_j} \left(\frac{sv}{m}\right)^{v_j} v_j^{c_j}$$

$$\le 2^{O(t)} \sum_{v=2}^{t} \sum_{w=1}^{t} \binom{m}{w} \sum_{\substack{c_1,\dots,c_w \\ c_1+\dots+c_w=t \\ c_i \ge 1}} \binom{t}{c_1,\dots,c_w} \prod_{j=1}^{w} \frac{1}{v} \left(\frac{sv}{m}\right)^{2} (2c_j)^{c_j+1} \tag{A.13}$$

$$\le 2^{O(t)} \sum_{v=2}^{t} \sum_{w=1}^{t} \binom{m}{w} \sum_{\substack{c_1,\dots,c_w \\ c_1+\dots+c_w=t \\ c_i \ge 1}} \frac{t^t}{c_1^{c_1} \cdot \dots \cdot c_w^{c_w}} \prod_{j=1}^{w} \left(\frac{s^2 v}{m^2}\right) c_j^{c_j+1} \tag{A.14}$$

$$= 2^{O(t)} \sum_{v=2}^{t} \sum_{w=1}^{t} \binom{m}{w} \sum_{\substack{c_1,\dots,c_w \\ c_1+\dots+c_w=t \\ c_i \ge 1}} t^t \left(\frac{s^2 v}{m^2}\right)^{w} \prod_{j=1}^{w} c_j \tag{A.15}$$

Here, (A.12) uses the upper bound on binomial coefficients. To get (A.13), we require $st < m$.[17] Then, since $v \le t$, it satisfies that $\frac{sv}{m} < 1$ and we can replace the power on $\left(\frac{sv}{m}\right)^{v_j}$ by 2, to get an upper bound $\left(\frac{sv}{m}\right)^2$. To obtain (A.14), we use Stirling's formula to bound the factorials in $\binom{t}{c_1,\dots,c_w}$, and $2^{c_1+\dots+c_w+w} = 2^{O(t)}$.

The multiplicant $\prod_{j=1}^{w} c_j$ in (A.15) can be upper bounded by $\left(\frac{t}{w}\right)^w$, since $c_1 + \dots + c_w = t$. Also, the number of choices of positive integers $c_1, \dots, c_w$ summing up to t is $\binom{t-1}{w-1}$, upper bounded by $2^{O(w)} \left(\frac{t}{w}\right)^w \le 2^{O(t)} \left(\frac{t}{w}\right)^w$. Incorporating these in (A.15) gives:

$$s^t \cdot \mathbb{E}[Z^t] \le 2^{O(t)} \sum_{v=2}^{t} \sum_{w=1}^{t} \binom{m}{w} t^t \left(\frac{s^2 v}{m^2}\right)^{w} \left(\frac{t}{w}\right)^{2w}$$

$$\le 2^{O(t)} \sum_{v=2}^{t} \sum_{w=1}^{t} t^t \left(\frac{s^2}{m}\right)^{w} \left(\frac{t}{w}\right)^{3w} \tag{A.16}$$

$$\le 2^{O(t)} \cdot t^t \left(\frac{s^2}{m}\right)^{t} \tag{A.17}$$

Here to get (A.16), we again use the upper bound on binomial coefficients for $\binom{m}{w}$. For (A.17), note that $\left(\frac{t}{w}\right)^{3w}$ is maximized when $w = t/e$ (which can be seen by taking the derivative), so is upper bounded by $e^{3t/e} = 2^{O(t)}$. Therefore, we can replace $\left(\frac{s^2}{m}\right)^w$ by $\left(\frac{s^2}{m}\right)^t$ since this is at this moment the only term that depends on w. This concludes the proof of Lemma 4.3. □

[17] For our setting of parameters to be chosen later, this will correspond to $\varepsilon^{-1} \cdot 2C > 1$ for a large constant $C > 1$.

B Proof of Theorem 2

B.1 Strengthening the Sparsity Lower Bound of [32]

We begin with a simple fact connecting JL matrices to ε-*incoherence* matrices. Given a matrix $A \in \mathbb{R}^{m\times n}$, let us denote its columns by $v_1, \dots, v_n \in \mathbb{R}^m$. A is said to be ε-incoherent if for all $i \neq j$, $|\langle v_i, v_j\rangle| \leq \varepsilon$, and for all i, $\|v_i\|_2 = 1$. Then,

Fact B.1 ([26])**.** *Let $\{e_1, \dots, e_n\}$ be the first n unit basis vectors of $\mathbb{R}^d$ where $n \in [d]$. Given any matrix $A \in \mathbb{R}^{m\times d}$ satisfying for any $x, y \in \{0, e_1, \dots, e_n\}$, $\|A(x-y)\|_2 = (1\pm\varepsilon)\|x-y\|_2$, we have that the first n columns of A (after normalization) form an $O(\varepsilon)$-incoherent submatrix.*

Owing to this fact above, lower bounds on ε-incoherent matrices directly translate to that of JL matrices, after choosing appropriate values of n (and we will eventually choose $n = \min\{d, \frac{1}{\delta^{1/4}}\}$). In particular, Nelson and Nguyễn [32] show that in an ε-incoherent matrix, there exists at least some column whose ℓ_0-sparsity —i.e., number of non-zero entries— is $\Omega(\varepsilon^{-1}\log n/\log(m/\log n))$.

We prove a strengthened version of this ℓ_0-sparsity lower bound. Namely, we show a lower bound on the ℓ_1 norm (which implies the same lower bound on the ℓ_0-sparsity), on at least half of the columns of A rather than a single column. More precisely, we show that (whose proof is deferred to Appendix C):

Theorem 3. *There is some fixed $0 < \varepsilon_0 < 1/2$ so that the following holds. For any $1/\sqrt{n} < \varepsilon < \varepsilon_0$ and $m < O(n/\log(1/\varepsilon))$, let $A \in \mathbb{R}^{m\times n}$ be an ε-incoherent matrix. Then, at least half of the columns A must have ℓ_1 norm being $\Omega(\sqrt{\varepsilon^{-1}\log n/\log(m/\log n)})$.*

It is worth noting that our strengthened lower bound implies: (1) the average ℓ_1 norm of the columns of A is $\Omega(\sqrt{\varepsilon^{-1}\log n/\log(m/\log n)})$, (2) at least half of the columns of A must have ℓ_0-sparsity $\Omega(\varepsilon^{-1}\log n/\log(m/\log n))$.

B.2 Dimension Lower Bound for Sign-Consistent JL Matrices

The lower bound in Appendix B.1 works as follows. There is a fixed hard instance of vectors, i.e., $\{0, e_1, \dots, e_n\}$, so that *even if* the adversary knows this hard instance, he cannot produce a good ε-incoherent matrix (and thus a JL matrix), unless the sparsity reaches the desired lower bound.

In this section, we lower bound m in a conceptually different way. We will choose the hard instance *after* the JL construction $\mathcal{A}$ (i.e., the distribution of the matrices) is determined, and then show that $\mathcal{A}$ must perform bad on this hard instance, unless m is large. This is a major difference between our proof and the related lower bounds for JL matrices, see instance [26, 32].

High Level Proof Sketch. Let us assume for simplicity that $\delta = 1/\mathsf{poly}(d)$ and $n = d$; the general case needs to be done more carefully. Take an arbitrary distribution $\mathcal{A}$ of $m \times n$ matrices satisfying the JL property with ε and δ. We divide our proof into three steps.

- In the first step, we use our Theorem 3 to conclude that almost all A drawn from $\mathcal{A}$ (being ε-incoherent) must have an average ℓ_1-sparsity (over the columns) $\geq \sqrt{s}$, where $s = \tilde{\Omega}(\varepsilon^{-1}\log n)$. For simplicity, assume that all matrices $A \sim \mathcal{A}$ have such property.
- In the second step, we use this ℓ_1-sparsity lower bound on $A \sim \mathcal{A}$ to deduce that A must have a large pairwise *column correlation.* Namely, $\sum_{i,j}|\langle v_i, v_j\rangle| \geq sn^2/m$ where v_i represents the i-th column of A. By an averaging argument, we can pick some subset $S \subset [d]$ of the columns where $|S| = N \stackrel{\text{def}}{=} 1/\log n$, such that the correlations between columns in S are also large: namely, $\sum_{i,j\in S}|\langle v_i, v_j\rangle| \geq \Omega(sN^2/m)$. This is formally proved as Lemma B.2 below.

- In the third step, we begin with a wishful thinking. By the property of JL, A must satisfy $\|\sum_i v_i\|_2^2 = (1\pm\varepsilon)^2 N$ because A must preserve the ℓ_2-norm on vector $x = \sum_{i\in S} e_i$. If all columns v_i for $i \in S$ *had* positive signs, then $\|\sum_i v_i\|_2^2 = N + \sum_{i,j\in S}\langle v_i, v_j\rangle \le N + \varepsilon N$. This formula, when combined with the previous step of $\sum_{i,j\in S} |\langle v_i, v_j\rangle| \ge \Omega(sN^2/m)$, would give $sN^2/m \le \varepsilon N$ so we have $m \ge \varepsilon^{-1}\log n \cdot s = \tilde{\Omega}(\varepsilon^{-2}\log^2 n)$ and we *would* be done.

  To fix this, we need to construct the hard instance more carefully. Instead of letting a single vector x be the hard instance, we want all the 2^N possible combinations $X = \big\{ \sum_{i\in S} s_i e_i : s_i \in \{-1, 1\}\big\}$ to present in the hard instance. We can afford this since $2^N = n$. Therefore, although the the signs of the columns in S in a matrix A may vary as $A \sim \mathcal{A}$, there is always some $x \in X$ that makes all the correlation to go positive, and the above sign issue goes away.

In sum, our hard instance so constructed depends on S, a set chosen *after* we see the distribution $\mathcal{A}$; and it contains $\mathsf{poly}(n)$ vectors. We now begin with our averaging lemma for the second step.

Lemma B.2. *For any distribution of $m \times n$ matrices $\mathcal{A}$ such that (1) $m < n/2$, (2) each column of $A \in \mathcal{A}$ is normalized and (3) $\mathbb{E}_{A\sim\mathcal{A}}\left[\frac{1}{n}\sum_{i,j}|A_{i,j}|\right] = \sqrt{s}$, there exist a subset $S \subseteq [n]$ of columns with cardinality $|S| = N$ (for any $N \in [n]$) such that*

$$\mathop{\mathbb{E}}_{A\sim\mathcal{A}}\left[\textstyle\sum_{i,j\in S, i\neq j} |\langle v_i, v_j\rangle|\right] \ge \Omega(sN^2/m) \enspace .$$

Here, as usual, we denote by v_i the i-th column of A.

Proof. We compute this quantity via an averaging argument. On one hand, for a matrix A:

$$\sum_{i,j\in[n], i\neq j} |\langle v_i, v_j\rangle| = \sum_{r=1}^m \left(\Big(\sum_{i\in[n]} |A_{r,i}|\Big)^2 - \sum_{i\in[n]} A_{r,i}^2\right) = \sum_{r=1}^m \Big(\sum_{i\in[n]} |A_{r,i}|\Big)^2 - n \ge \frac{1}{m}\Big(\sum_{r,i} |A_{r,i}|\Big)^2 - n$$

and therefore when taking over the distribution of $A \sim \mathcal{A}$ we have

$$\mathop{\mathbb{E}}_{A\sim\mathcal{A}}\Bigg[\sum_{i,j\in[n], i\neq j} |\langle v_i, v_j\rangle|\Bigg] \ge \mathop{\mathbb{E}}_{A\sim\mathcal{A}}\left(\frac{1}{m}\Big(\sum_{r,i}|A_{r,i}|\Big)^2 - n\right) \ge \frac{1}{m}\Big(\mathop{\mathbb{E}}_{A\sim\mathcal{A}}\sum_{r,i}|A_{r,i}|\Big)^2 - n = sn^2/m - n = \Omega(sn^2/m) \enspace .$$

On the other hand, we note that

$$\sum_{S\subset[n], |S|=N} \; \sum_{i,j\in S, i\neq j} |\langle v_i, v_j\rangle| = \binom{n-2}{N-2} \cdot \sum_{i,j\in[n], i\neq j} |\langle v_i, v_j\rangle|$$

and there are a total number of $\binom{n}{N}$ subsets S of cardinality N. By an averaging argument, there exist some subset $S^* \subset [n]$ satisfying

$$\mathop{\mathbb{E}}_{A\sim\mathcal{A}}\Bigg[\sum_{i,j\in S^*, i\neq j} |\langle v_i, v_j\rangle|\Bigg] \ge \frac{1}{\binom{n}{N}}\binom{n-2}{N-2} \cdot \mathop{\mathbb{E}}_{A\sim\mathcal{A}}\Bigg[\sum_{i,j\in[n], i\neq j} |\langle v_i, v_j\rangle|\Bigg] \ge \Omega(sN^2/m) \enspace . \qquad \square$$

Proof of Theorem 2. We are now ready to implement the aforementioned high level proof sketch. Given any such distribution $\mathcal{A}$, we let $n = \min\{d, \frac{1}{\delta^{1/4}}\}$. Using union bound, with probability at least $1 - O(\delta n^2) \ge 1 - O(\frac{1}{n^2})$, a matrix A drawn from $\mathcal{A}$ will preserve ℓ_2 norms with ε distortion for all vector $x = v_1 - v_2$ where $v_1, v_2 \in \{0, e_1, \ldots, e_n\}$.

In other words, owing to Fact B.1, with probability at least $1-O(\delta n^2)\geq 1-O(\frac{1}{n^2})$, a matrix A drawn from $\mathcal{A}$ satisfies that its first n columns form an $O(\varepsilon)$-incoherent $m\times n$ submatrix (after column normalizations). Let $\mathcal{A}'$ be this subdistribution of $m\times n$ $O(\varepsilon)$-incoherent matrices.

Thanks to our strengthened Theorem 3 on the ℓ_1-sparsity,[18] letting $\sqrt{s}\stackrel{\text{def}}{=}\mathbb{E}_{A'\sim\mathcal{A}'}\left[\frac{1}{n}\sum_{i,j}|A'_{i,j}|\right]$, we must have $s=\Omega(\varepsilon^{-1}\log n/\log(m/\log n))$. We plug this distribution $\mathcal{A}'$ into Lemma B.2 along with the choice of $N\stackrel{\text{def}}{=}\log(1/\delta^{1/2})$, and deduce that

$$\mathop{\mathbb{E}}_{A'\sim\mathcal{A}'}\left[\sum_{i,j\in S,i\neq j}|\langle v_i,v_j\rangle|\right]\geq\Omega(sN^2/m)\ . \tag{B.1}$$

Now comes the important construction. Let us study define the following set of 2^N vectors,

$$X=\Big\{\sum_{i\in S}s_ie_i\ :\ \forall i\ s_i\in\{-1,1\}\Big\}\subset\mathbb{R}^n\subset\mathbb{R}^d\ .$$

Because $1-O(\delta 2^N)\geq 1-O(\frac{1}{n^2})$, with probability at least $1-O(\frac{1}{n^2})$, all vectors in $x\in X$ must have their ℓ_2 norm preserved within ε-distortion over the choice of $A\sim\mathcal{A}$. This is also true with probability at least $1-O(\frac{1}{n^2})$ over the choice of $A'\sim\mathcal{A}'$ by union bound. Let us denote by $\mathcal{A}''$ this subdistribution of matrices $A'\sim\mathcal{A}'$ such that

$$\forall x\in X,\quad \|A'x\|_2=(1\pm O(\varepsilon))\|x\|_2\ .$$

By the above argument, $\mathcal{A}''$ contributes to at least $1-O(\frac{1}{n^2})$ probability mass in both $\mathcal{A}'$ and $\mathcal{A}$.

Next, for each matrix $A''\in\mathcal{A}''$, we claim that $\sum_{i,j\in S,i\neq j}|\langle v_i,v_j\rangle|=O(\varepsilon N)$. This is because, letting $x=\sum_{i\in S}s_ie_i\in X$ be a vector where s_i coincides with the column sign of v_i, then

$$O(\varepsilon N)\geq\|A''x\|_2^2-\|x\|_2^2=\|\sum_{i\in S}s_iv_i\|_2^2-\|\sum_{i\in S}e_i\|_2^2=\sum_{i,j\in S,i\neq j}|\langle v_i,v_j\rangle|\ .$$

Therefore, we must have

$$\mathop{\mathbb{E}}_{A'\sim\mathcal{A}'}\left[\sum_{i,j\in S,i\neq j}|\langle v_i,v_j\rangle|\right]\leq\mathop{\mathbb{E}}_{A''\sim\mathcal{A}''}\left[\sum_{i,j\in S,i\neq j}|\langle v_i,v_j\rangle|\right]+O\Big(\frac{1}{n^2}\Big)N^2\leq O\Big(\varepsilon N+\frac{N^2}{n^2}\Big)$$

when comparing the above lower bound and (B.1), we get $m=\Omega(\varepsilon^{-1}N\cdot s)=\Omega(\varepsilon^{-1}\log(1/\delta)s)$. Substituting the ℓ_1-sparsity lower bound for s we have

$$\begin{aligned}&m\geq\Omega(\varepsilon^{-2}\log(1/\delta)\log n/\log(m/\log n))\implies\\&m\geq\Omega(\varepsilon^{-2}\log(1/\delta)\log n/\log(\varepsilon^{-2}\log(1/\delta)))\\&\quad=\Omega\Big(\frac{\varepsilon^{-2}\log(1/\delta)}{\log(\varepsilon^{-2}\log(1/\delta))}\min\big\{\log d,\log(1/\delta)\big\}\Big)\end{aligned}$$

■

[18] To be precise, we need to verify that $1/\sqrt{n}<\varepsilon$. This is easy given our assumption of $1/\sqrt{d}<\varepsilon$ and $\delta<\varepsilon^{12}$. In addition, we need to verify that $m<O(n/\log(1/\varepsilon))=O\big(\min\{d,1/\delta^{1/4}\}/\log(1/\varepsilon)\big)$. The first term in min is true by assumption: $m\leq O(d/\log(1/\varepsilon))$. For the second term, suppose it is false, then we get $m\geq\Omega(\frac{1}{\delta^{1/4}}/\log(1/\varepsilon))\geq\Omega(\frac{\varepsilon^{-2.5}}{\log(1/\varepsilon)}\cdot\frac{1}{\delta^{1/24}})\geq\Omega(\varepsilon^{-2}\log^2(1/\delta))$, using $\delta\leq\varepsilon^{12}$ and δ and ε smaller than some sufficiently small constant.

C Proof of Theorem 3: ℓ_1-Sparsity Lower Bound for JL Matrices

Our proof to Theorem 3 follows from the same proof framework as [32]; however, since the ℓ_1-sparsity guarantee is a stronger one, this strengthening needs a lot of careful care.

Given a matrix $A \in \mathbb{R}^{m\times n}$, let us denote its columns by $v_1, \dots, v_n \in \mathbb{R}^m$. Throughout this section, we assume that A is *ε-incoherent*: for all $i \neq j$, $|\langle v_i, v_j\rangle| \le \varepsilon$, and for all i, $\|v_i\|_2 = 1$.

Like in the sparsity case from [32], we first need a weaker lower bound on the ℓ_1 norm:

Lemma C.1. *Suppose $m < n/(40\log(1/2\varepsilon))$, and $A \in \mathbb{R}^{m\times n}$ is ε-incoherent. If A has $n/2$ columns with ℓ_1 norm at most $\sqrt{s/2}$ each, then $s \ge 1/(4\varepsilon)$.*

Proof. For the sake of contradiction, assume $s < 1/(4\varepsilon)$.

Let $W \subset [n]$ be a set of the $n/2$ columns each with ℓ_1 norm at most $\sqrt{s/2}$. Define $L(W) = \{(i,j) : A_{i,j}^2 \ge 2\varepsilon, j \in W\}$ to be the set of large entries, and $S(W) = \{(i,j) : A_{i,j}^2 < 2\varepsilon, j \in W\}$ to be that of small entries. Clearly, $S(W)$ and $L(W)$ are disjoint and span all the entries in W. Let us bound the sum of squares of matrix entries in $S(W)$ and $L(W)$ respectively.

- Small entries in W are less than $\sqrt{2\varepsilon}$ in absolute magnitude, so their squares sum up to at most $\sqrt{2\varepsilon}$ times the ℓ_1 norm of such entries:

$$\begin{aligned}\sum_{(i,j)\in S(W)} A_{i,j}^2 &\le \sqrt{2\varepsilon}\cdot\Big(\sum_{(i,j)\in S(W)} |A_{i,j}|\Big) = \sqrt{2\varepsilon}\cdot\Big(\sum_{i\in[m], j\in W} |A_{i,j}| - \sum_{(i,j)\in L(W)} |A_{i,j}|\Big) \\ &\le \sqrt{2\varepsilon}\cdot\Big(\sum_{j\in W}\|v_j\|_1 - \sqrt{2\varepsilon}|L(W)|\Big) \le \frac{n}{2}\sqrt{\varepsilon s} - 2\varepsilon|L(W)| < \frac{n}{4} - 2\varepsilon|L(W)| \quad \text{(C.1)}\end{aligned}$$

- For large entries, we reuse the same analysis as [32]. Let X denote the square of a random entry from $L(W)$. Then,

$$\begin{aligned}\sum_{(i,j)\in L(W)} A_{i,j}^2 &= |L(W)|\cdot\mathbb{E}[X] = |L(W)|\cdot\int_{x=0}^{1}\Pr[X > x]dx \le 2\varepsilon|L(W)| + m\cdot\int_{x=2\varepsilon}^{1}\frac{10}{x}dx \\ &= 2\varepsilon|L(W)| + 10m\log(1/2\varepsilon) \le 2\varepsilon|L(W)| + \frac{n}{4}\ . \quad \text{(C.2)}\end{aligned}$$

  Here the first inequality is due to a simple fact about ε-incoherent matrices: there cannot be more than $\frac{10}{x}$ entries in each row of absolute value more than $\sqrt{x}$, for any $x \ge 2\varepsilon$. See for instance [32, Lemma 3]. The second inequality is owing to the choice of $m < n/(40\log(1/2\varepsilon))$.

Now we can combine equations (C.1) and (C.2) to get $\sum_{(i,j)\in S(W)\cup L(W)} A_{i,j}^2 < \frac{n}{2}$. On the other hand, $\sum_{(i,j)\in S(W)\cup L(W)} A_{i,j}^2 = \sum_{j\in W}\|v_j\|_2^2 = \frac{n}{2}$ and we get a contradiction. □

The above lower bound is weak since it is obtained merely from a counting argument. Let us now strengthen it into a stronger form, by using the pigeon-hole principle to find N columns that pairwisely and positively correlate to each other. This cannot happen if the original matrix is ε-incoherent. Let us explain:

Lemma C.2. *Let $0 < \varepsilon < 1/2$, $A \in \mathbb{R}^{m\times n}$ be an ε-incoherent matrix, and s be any value such that half of A's columns have ℓ_1 norm at most $\sqrt{s}/2$. Define $C = 2/(1-1/\sqrt{2})$. Then, for any $t \in [s/2]$ with $t/s > C\varepsilon$, we must have $s \ge t(N-1)/C$ with $N = \left\lceil \frac{n}{2^{t+1}\binom{m}{t}\binom{2(s+t)}{t}} \right\rceil$.*

Proof. The proof structure is similar to that of [32, Lemma 9]: for any vector v_i, consider its t largest coordinates in absolute magnitude, and define its *t-type* to be a triple containing:

- the locations of the top t coordinates (there are at most $\binom{m}{t}$ choices);
- the signs of the top t coordinates (there are at most 2^t different choices); and
- the rounding of the top t values so that their squares round to the nearest integer multiple of $1/(2s)$. Values halfway between two multiples can be rounded arbitrarily. (There are at most $\binom{2s+2t}{t}$ number of different roundings.[19])

All in all, there are $2^t\binom{m}{t}\binom{2s+2t}{t}$ possible *t-types*. By the pigeon-hole principle, out of $n/2$ column vectors that have ℓ_1 norm at most $\sqrt{s}/2$, we can select N vectors $\tilde{v}_1, \ldots, \tilde{v}_N$, such that they all have the same *t-type*.

Let $S \subset [n]$ be the set of the largest coordinates for these vectors, and we have $|S| = t$. Now define $u_i = (\tilde{v}_i)_{[m]-S} \in \mathbb{R}^{m-t}$, with the coordinates in S zeroed out. Then, for $j \neq k \in [N]$, since $\tilde{v}_j$ and $\tilde{v}_k$ have the same type, we must have

$$\begin{aligned} \langle u_j, u_k \rangle &= \langle \tilde{v}_j, \tilde{v}_k \rangle - \sum_{r\in S} (\tilde{v}_j)_r (\tilde{v}_k)_r \leq \varepsilon - \sum_{r \in S} (\tilde{v}_j)_r ((\tilde{v}_j)_r \pm 1/\sqrt{2s}) \\ &\leq \varepsilon - \sum_{r\in S} \Big((\tilde{v}_j)_r^2 - |(\tilde{v}_j)_r| / \sqrt{2s} \Big) = \varepsilon - \|(\tilde{v}_j)_S\|_2^2 + \|(\tilde{v}_j)_S\|_1/\sqrt{2s} \\ &\leq \varepsilon - (1 - \sqrt{t/2s}) \|(\tilde{v}_j)_S\|_2^2 \ . \end{aligned} \tag{C.3}$$

Here, the last inequality follows from the Cauchy-Schwarz inequality $\|(\tilde{v}_j)_S\|_1 \leq \sqrt{t} \cdot \|(\tilde{v}_j)_S\|_2$.

Now we use a simple proposition on the relationship between ℓ_1 and ℓ_2 norms: given $t \leq s/2$, ℓ_1 norm $\|\tilde{v}_j\|_1 \leq \sqrt{s}/2$, and ℓ_2 norm $\|\tilde{v}_j\|_2 = 1$, we must have $\|(\tilde{v}_j)_S\|_2 \geq \sqrt{t/s}$, i.e., much of the ℓ_2 mass must lie on it t largest coordinates (see Proposition C.3 below for its proof).

Combining this with (C.3) and $t/s > C\varepsilon$ gives:

$$\langle u_j, u_k \rangle \leq \varepsilon - \left(1 - \frac{1}{\sqrt{2}}\right) t/s < t/s \cdot (1/C - 2/C) = -\frac{t}{sC}$$

Now we can write

$$0 \leq \Big\| \sum_{j=1}^N u_j \Big\|_2^2 = \sum_{j=1}^N \|u_j\|_2^2 + \sum_{j\neq k} \langle u_j, u_k \rangle \leq N - \frac{tN(N-1)}{sC} \ ,$$

which gives $s \geq \frac{t(N-1)}{C}$ and completes the proof. ∎

Proposition C.3. *Let $x \in \mathbb{R}^m$ with $\|x\|_1 \leq \sqrt{s}/2$ and $\|x\|_2 = 1$. Also, assume that $|x_1| \geq \ldots \geq |x_m|$. Then, for any $t \in [s/2]$, $\sum_{i=1}^t x_i^2 \geq \frac{t}{s}$ holds.*

Proof. Assume contrary: $\exists t \leq s/2 : \sum_{i=1}^t x_i^2 < \frac{t}{s}$. Then, since the absolute values of components are sorted, $x_t^2 < \frac{1}{s} \Rightarrow \forall j \geq t : x_j^2 < \frac{1}{s}$ and we have

$$\sqrt{\frac{1}{s}} \|x\|_1 > x_{t+1}^2 + \ldots + x_m^2 = 1 - \sum_{i=1}^t x_i^2 > 1 - \frac{t}{s} \geq \frac{1}{2}$$

However, this implies $\|x\|_1 > \frac{\sqrt{s}}{2}$, leading to a contradiction. ∎

[19] The amount of ℓ_2 mass contained in the top t coordinates is at most $1 + t/(2s)$, so the sum of integer multiples of $1/(2s)$ that correspond to the rounded t coordinates can be at most $2s + t$. Consider the representations of $2s + t$ as the sum of $t + 1$ non-negative integers. Then, each possible rounding has an unique representation, where the first t summands correspond to the integer multiples of $1/(2s)$ and the last summand is the residual.

At last, we put in the right parameter of t and conclude the proof. The following theorem resembles [32, Theorem 10].

Theorem 3. *There is some fixed $0 < \varepsilon_0 < 1/2$ so that the following holds. For any $1/\sqrt{n} < \varepsilon < \varepsilon_0$ and $m < O(n/\log(1/\varepsilon))$, let $A \in \mathbb{R}^{m\times n}$ be an ε-incoherent matrix. Then, at least half of the columns A must have ℓ_1 norm being $\Omega(\sqrt{\varepsilon^{-1}\log n/\log(m/\log n)})$.*

Proof. Let s be a value such that half of A's columns have ℓ_1 norm at most $\sqrt{s}/2$, then we want to show that $s \geq \Omega(\varepsilon^{-1}\log n/\log(m/\log n))$.

By Lemma C.1, we have a weak lower bound $4\varepsilon s \geq 1$, allowing us to chose $t = 7\varepsilon s \geq 1$. We are now ready to prove that:

$$s \geq \frac{\log(7\varepsilon n/(4C))}{7\varepsilon\log\left(\frac{8e^2m}{49\varepsilon^2 s}\right)} \ , \tag{C.4}$$

where C is as in Lemma C.2. Assume contrary, then we get:

$$\left(\frac{8e^2m}{49\varepsilon^2 s}\right)^{7\varepsilon s} < \frac{7\varepsilon n}{4C} \ .$$

Furthermore, for small enough ε,

$$2^{t+1}\binom{m}{t}\binom{2(s+t)}{t} \leq 2^{t+1}\frac{(em)^t}{t^t}\frac{(2e)^t(s+t)^t}{t^t} \leq 2\cdot\left(\frac{8e^2m}{49\varepsilon^2 s}\right)^{7\varepsilon s} < \frac{7\varepsilon n}{2C} \leq \frac{n}{2} \ ,$$

so we can now apply Lemma C.2 and get:

$$\frac{sC}{t} \geq N - 1 \geq \frac{n}{2\cdot 2^{t+1}\binom{m}{t}\binom{2(s+t)}{t}}.$$

By rearranging terms, it directly follows that

$$7\varepsilon n = \frac{tn}{s} \leq 2C\cdot 2^{t+1}\binom{m}{t}\binom{2(s+t)}{t} \leq 4C\cdot\left(\frac{8e^2m}{49\varepsilon^2 s}\right)^{7\varepsilon s} < 7\varepsilon n \ ,$$

giving a contradiction. This completes the proof of (C.4).

Let us now define $r = \log(7\varepsilon n/(4C))/(7\varepsilon)$ and $q = 8e^2m/(49\varepsilon^2)$. Then we have $s\log(q/s) \geq r$ and for $\varepsilon < 1/2$, $q/e \geq m \geq s$. By [26], $m = \Omega(\log n)$ and hence for small enough ε, $q/r > 2$ also holds. Using Proposition C.4 below, we get $s \geq \Omega(r/\log(q/r)) = \Omega(\varepsilon^{-1}\log n/\log(\varepsilon^{-1}m/\log n))$, since $\log(\varepsilon n) = \Theta(\log n)$ as $\varepsilon > 1/\sqrt{n}$. This is be equivalent to our theorem statement, since $m = \Omega(\frac{1}{\varepsilon})$ (using for instance the general lower bound on m from [26], or our weak sparsity lower bound Lemma C.1 as $m \geq \Omega(s)$). ☐

Proposition C.4. *Let s, q, r be positive reals with $q \geq \max(2r, es)$. Then, if $s\log(q/s) \geq r$ it must be the case that $s = \Omega(r/\log(q/r))$.*

Proof. The function $f(s) = s\log(q/s)$ is non-decreasing for $s \leq q/e$ since $f'(s) = \log(q/(es)) \geq 0$. Since we are proving a lower bound on s, we can without the loss of generality consider $s\log(q/s) = r$. From here with $q/s \geq e$ immediately follows that $s \leq r$. , $r/s = \log(q/s) = \log(q/r) + \log(r/s)$.

Finally, we can write:

$$\frac{s}{r/\log(q/r)} = \frac{s((r/s) - \log(r/s))}{r} = 1 - \frac{s}{r}\log(\frac{r}{s}) \geq 1 - \frac{1}{e}$$ ☐

D Simple Lower Bound for Non-Negative JL Matrices

In this section we show a simple fact: at least in the interesting parameter regime of $\delta = 1/\mathsf{poly}(d)$, we must have $m \geq \Omega(d)$ in order to construct a non-negative JL matrix. Since we cannot find the proof of this simple fact anywhere else, we provide it below.

Fact D.1. *Let $\mathcal{A}$ be a distribution over $m \times d$* non-negative *matrices such that, for any $x \in \mathbb{R}^d$, with probability at least $1-\delta$, the ℓ_2 embedding $\|Ax\|_2 = (1 \pm \varepsilon)\|x\|_2$ has ε-distortion. Then,*

$$m \geq (1 - 4\varepsilon) \min\left\{d, \frac{1}{\delta} - 2\right\} .$$

Proof. Given any such distribution $\mathcal{A}$, we choose $n = \min\{d, \frac{1}{\delta} - 2\}$. Using union bound, with probability at least $1 - \delta(n+1) > 0$, a matrix A drawn from $\mathcal{A}$ will preserve ℓ_2 norms with ε distortion for all vector $x \in \{e_1, \dots, e_n\} \cup \{e_1 + e_2 + \dots + e_n\}$.

This implies that, the ℓ_2-norm of each of the first n columns of A is at least $1-\varepsilon$: this is because for every $j \in [n]$, $\sqrt{\sum_{i \in [m]} A_{i,j}^2} = \|Ae_j\|_2 \geq (1-\varepsilon)\|e_j\|_2 = 1 - \varepsilon$.

Next, we check the norm preservation on $x = e_1 + e_2 + \dots + e_n \in \mathbb{R}^d$. Its ℓ_2 norm is $\|x\|_2 = \sqrt{n}$, so we must have $\|Ax\|_2^2 \leq n(1+\varepsilon)^2$. On the other hand,

$$\|Ax\|_2^2 = \sum_{i=1}^m \left(\sum_{j=1}^n A_{i,j}\right)^2 \geq \frac{1}{m}\left(\sum_{i=1}^m \sum_{j=1}^n A_{i,j}\right)^2 \geq \frac{1}{m}\left(\sum_{j=1}^n \|Ae_j\|_1\right)^2 \geq \frac{1}{m}\left(\sum_{j=1}^n \|Ae_j\|_2\right)^2$$
$$\geq \frac{1}{m}\left((1-\varepsilon)n\right)^2 .$$

Together, they imply $m \geq (1 - 4\varepsilon)n$. □

References

[1] Peter Sterling and Jonathan B. Demb. Retina. In Gordon M. Shepherd, editor, *The synaptic organization of the brain*, chapter 6, pages 217–270. Oxford University Press Oxford, 5th edition, 2004.

[2] Ghez C. Wolpert D. M., Pearson K. The organization and planning of movement. In Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell, editors, *Principles of neural science*, chapter 33, pages 653–673. McGraw-Hill New York, 5th edition, 2012.

[3] Horace B Barlow. Possible principles underlying the transformation of sensory messages. In Walter A. Rosenblith, editor, *Sensory communication*, chapter 13, pages 217–234. MIT press, 1961.

[4] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[5] Surya Ganguli and Haim Sompolinsky. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annual review of neuroscience*, 35:485–508, 2012.

[6] Roozbeh Kiani, Hossein Esteky, Koorosh Mirpour, and Keiji Tanaka. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of neurophysiology*, 97(6):4296–4309, 2007.

[7] Timothy T Rogers and James L McClelland. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004.

[8] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[9] Kanaka Rajan, L. F. Abbott, and Haim Sompolinsky. Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical Review E*, 82(1):011903, July 2010.

[10] Kanaka Rajan and L. F. Abbott. Eigenvalue spectra of random matrices for neural networks. *Physical review letters*, 97(18):188104, 2006.

[11] Richard T Gray and Peter a Robinson. Stability and structural constraints of random brain networks with excitatory and inhibitory neural populations. *Journal of computational neuroscience*, 27(1):81–101, August 2009.

[12] Eric R Kandel, James H Schwartz, and Thomas M Jessell. *Principles of neural science*. McGraw-Hill New York, 2000.

[13] Gyorgy Buzsaki. *Rhythms of the Brain*. Oxford University Press, 2006.

[14] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, June 2003.

[15] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlos. A Sparse Johnson-Lindenstrauss Transform. In *Proceedings of the 42nd ACM symposium on Theory of computing - STOC '10*, pages 341–350, New York, New York, USA, 2010. ACM Press.

[16] Daniel M. Kane and Jelani Nelson. A Derandomized Sparse Johnson-Lindenstrauss Transform. Technical report, June 2010.

[17] Vladimir Braverman, Rafail Ostrovsky, and Yuval Rabani. Rademacher Chaos, Random Eulerian Graphs and The Sparse Johnson-Lindenstrauss Transform. Technical report, November 2010.

[18] Daniel M. Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms - SODA '12*, pages 1195–1206, New York, New York, USA, December 2012. ACM Press.

[19] Felix Krahmer and Rachel Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[20] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958.

[21] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[22] Leslie G Valiant. *Circuits of the Mind*. Oxford University Press, Inc., New York, NY, USA, 1994.

[23] Leslie G. Valiant. Memorization and association on a realistic neural model. *Neural computation*, 17(3):527–555, 2005.

[24] Vitaly Feldman and Leslie G. Valiant. Experience-induced neural circuits that achieve high capacity. *Neural computation*, 21(10):2715–2754, 2009.

[25] Leslie G. Valiant. The hippocampus as a stable memory allocator for cortex. *Neural Computation*, 24(11):2873–2899, 2012.

[26] Noga Alon. Perturbed identity matrices have high rank: Proof and applications. *Combinatorics, Probability & Computing*, 18(1-2):3–15, 2009.

[27] Nir Ailon and Edo Liberty. Fast Dimension Reduction Using Rademacher Series on Dual BCH Codes. *Discrete & Computational Geometry*, 42(4):615–630, September 2008.

[28] Nir Ailon and Bernard Chazelle. The Fast JohnsonCLindenstrauss Transform and Approximate Nearest Neighbors. *SIAM Journal on Computing*, 39(1):302–322, January 2009.

[29] Aicke Hinrichs and Jan Vybral. Johnson-lindenstrauss lemma for circulant matrices. *Random Structures & Algorithms*, 39(3):391–398, 2011.

[30] Jan Vybíral. A variant of the JohnsonCLindenstrauss lemma for circulant matrices. *Journal of Functional Analysis*, 260(4):1096–1105, February 2011.

[31] Nir Ailon and Edo Liberty. An Almost Optimal Unrestricted Fast Johnson-Lindenstrauss Transform. *ACM Transactions on Algorithms*, 9(3):1–12, June 2013.

[32] Jelani Nelson and Huy L. Nguyễn. Sparsity lower bounds for dimensionality reducing maps. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing - STOC '13*, pages 101–110, New York, New York, USA, 2013. ACM Press.

[33] Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing - STOC '13*, pages 81–90, New York, New York, USA, July 2013. ACM Press.

[34] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, pages 10–33, 2001.

[35] Santosh S. Vempala. *The random projection method*, volume 65. American Mathematical Society, 2004.

[36] EJ Candes and T Tao. Decoding by Linear Programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.

[37] EJ Candes, JK Romberg, and T Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[38] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

[39] Avrim Blum. Random projection, margins, kernels, and feature-selection. In *Proceedings of the 2005 international conference on Subspace, Latent Structure and Feature Selection*, pages 52–68. Springer, 2005.

[40] Shuheng Zhou, John Lafferty, and Larry Wasserman. Compressed and privacy-sensitive sparse regression. *Information Theory, IEEE Transactions on*, 55(2):846–866, 2009.