A Proofs for "Tight Sample Complexity of Large-Margin Learning" (S. Sabato, N. Srebro and N. Tishby)

A.1 Proof of Lemma 3.3

Proof. The inequality $k_\gamma \leq k_{\alpha\gamma}$ is trivial from the definition of k_γ. For the other inequality, note first that we can always let $\mathbb{E}_{X\sim D_X}[XX']$ be diagonal by rotating the axes w.l.o.g. . Therefore $k_\gamma = \min\{k \mid \sum_{i=k+1}^{d} \lambda_i \leq \gamma^2 k\}$. Since $k_\gamma \leq k_{\alpha\gamma}$, we have $\gamma^2 k_\gamma \geq \sum_{i=k_\gamma+1}^{d} \lambda_i \geq \sum_{i=k_{\alpha\gamma+1}}^{d} \lambda_i$. In addition, by the minimality of $k_{\alpha\gamma}$, $\sum_{k_{\alpha\gamma}}^{d} \lambda_i > \alpha^2\gamma^2(k_{\alpha\gamma}-1)$. Thus $\sum_{i=k_{\alpha\gamma+1}}^{d} \lambda_i > \alpha^2\gamma^2(k_{\alpha\gamma}-1)-\lambda_{k_{\alpha\gamma}}$. Combining the inequalities we get $\gamma^2 k_\gamma > \alpha^2\gamma^2(k_{\alpha\gamma}-1)-\lambda_{k_{\alpha\gamma}}$. In addition, if $k_\gamma < k_{\alpha\gamma}$ then $\gamma^2 k_\gamma \geq \sum_{i=k_{\alpha\gamma}}^{d} \lambda_i \geq \lambda_{k_{\alpha\gamma}}$. Thus, either $k_\gamma = k_{\alpha\gamma}$ or $2\gamma^2 k_\gamma > \alpha^2\gamma^2(k_{\alpha\gamma}-1)$. □

A.2 Details omitted from the proof of Theorem 4.2

The proof of Theorem 4.2 is complete except for the construction of $\tilde{X}$ and $\tilde{P}$ in the first paragraph, which is disclosed here in full, using the following lemma:

Lemma A.1. *Let $S = (X_1, \ldots, X_m)$ be a sequence of elements in $\mathbb{R}^d$, and let X be a $m \times d$ matrix whose rows are the elements of S. If S is γ-shattered, then for every $\epsilon > 0$ there is a column vector $r \in \mathbb{R}^d$ such that for every $y \in \{\pm\gamma\}^m$ there is a $w_y \in \mathbf{B}_{1+\epsilon}^{d+1}$ such that $\widetilde{X} w_y = y$, where $\widetilde{X} = (X \quad r)$.*

Proof. if S is γ-shattered then there exists a vector $r \in \mathbb{R}^d$, such that for all $y \in \{\pm 1\}^m$ there exists $w_y \in \mathbf{B}_1^d$ such that for all $i \in [m], y_i(\langle X_i, w_y\rangle - r_i) \geq \gamma$. For $\epsilon > 0$ define $\widetilde{w}_y = (w_y, \sqrt{\epsilon}) \in \mathbf{B}_{1+\epsilon}$, and $\widetilde{r} = r/\sqrt{\epsilon}$, and let $\widetilde{X} = (X \quad \widetilde{r})$. For every $y \in \{\pm 1\}^m$ there is a vector $t_y \in \mathbb{R}^m$ such that $\forall i \in [m], \frac{1}{\gamma} t_y[i] y[i] \geq 1$, and $\frac{1}{\gamma}\widetilde{X}\widetilde{w}_y = \frac{1}{\gamma} t_y$. As in the proof of necessity in Theorem 5.2, it follows that there exists $\widehat{w}_y \in \mathbf{B}_{1+\epsilon}$ such that $\frac{1}{\gamma}\widetilde{X}\widehat{w}_y = y$. Scaling y by γ, we get the claim of the theorem. □

Now, Let X be a $m \times d$ matrix whose rows are a set of m points in $\mathbb{R}^d$ which is γ-shattered. By Lemma A.1, for any $\epsilon > 0$ there exists matrix $\tilde{X}$ of dimensions $m \times (d+1)$ such that the first d columns of $\tilde{X}$ are the respective columns of X, and for all $y \in \{pm\gamma\}^m$, there is a $w_y \in \mathbf{B}_{1+\epsilon}^{d+1}$ such that $\widetilde{X} w_y = y$. Since $\mathcal{X}$ is (B^2, k)-limited, there exists an orthogonal projection matrix P of size $d \times d$ and rank $d-k$ such that $\forall i \in [m], \|X_i' P\|^2 \leq B^2$. Let $\widetilde{P}$ be the embedding of P in a $(d+1) \times (d+1)$ zero matrix, so that $\widetilde{P}$ is of the same rank and projects onto the same subspace. The rest of the proof follows as in the body of the paper.

A.3 Proof of Theorem 4.4

Proof of Theorem 4.4. Let $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ be the covariance matrix of D_X, where $\forall i \in [d-1], \lambda_i \geq \lambda_{i-1}$. Define $\mathcal{X}_\alpha = \{x \in \mathbb{R}^d \mid \sum_{i=k_\gamma(D_X)+1}^{d} x[i]^2 \leq \alpha\}$.

Let $\{x_i\}_{i=1}^m$ be an i.i.d. sample of size m drawn from D_X. We will select α such that the probability that the whole sample is contained in $\mathcal{X}_\alpha$ is large. $\mathbb{P}[\forall i \in [m], x_i \in \mathcal{X}_\alpha] = (1 - \mathbb{P}[x_i \notin \mathcal{X}_\alpha])^m$. Let $X \sim D_X$. Then for all $t > 0$, $\mathbb{P}[X \notin \mathcal{X}_\alpha] = \mathbb{P}[\sum_{i=k_\gamma+1}^{d} X[i]^2 \geq \alpha] \leq \mathbb{E}[\exp(t\sum_{i=k_\gamma+1}^{d} X[i]^2)]\exp(-t\alpha)$.

Let $\lambda_{\max} = \lambda_{k_\gamma+1}$. Define $Y \in \mathbb{R}^d$ such that $Y[i] = X[i]\sqrt{\frac{\lambda_{\max}}{\lambda_i}}$. Then $\sum_{i=k_\gamma+1}^{d} X[i]^2 = \sum_{i=k_\gamma+1}^{d} \frac{\lambda_i}{\lambda_{\max}} Y[i]^2$, and by the definition of k_γ, $\sum_{i=k_\gamma+1}^{d} \frac{\lambda_i}{\lambda_{\max}} \leq \frac{k_\gamma}{\lambda_{\max}}$. Thus, by Lemma A.2

$$\mathbb{E}[\exp(t\sum_{i=k_\gamma+1}^{d} X[i]^2)] \leq \max_i(\mathbb{E}[\exp(3tY[i]^2)])^{\lceil k_\gamma/\lambda_{\max}\rceil}.$$

For every i, $Y[i]$ is a sub-Gaussian random variable with moment $B = \rho\sqrt{\lambda_{\max}}$. By [12], Lemma 1.1.6, $\mathbb{E}[\exp(3tY[i]^2)] \leq (1 - 6\rho^2\lambda_{\max}t)^{-\frac{1}{2}}$, for $t \in (0, (6\rho^2\lambda_{max})^{-1})$. Setting $t = \frac{1}{12\rho^2\lambda_{\max}}$,

$$\mathbb{P}[X \notin \mathcal{X}_\alpha] \leq 2^{k_\gamma/\lambda_{\max}} \exp(-\frac{\alpha}{12\rho^2\lambda_{\max}}).$$

Thus there is a constant C such that for $\alpha(\gamma) \triangleq C \cdot \rho^2(k_\gamma(D_X) + \lambda_{\max} \ln \frac{m}{\delta})$, $\mathbb{P}[X \notin \mathcal{X}_{\alpha(\gamma)}] \leq 1 - \frac{\delta}{2m}$. Clearly, $\lambda_{\max} \leq k_\gamma(D_X)$, and $k_\gamma(\mathcal{X}_{\alpha(\gamma)}) \leq \alpha(\gamma)$. Therefore, from Theorem 4.2, the γ-fat-shattering dimension of $\mathcal{W}(\mathcal{X}_{\alpha(\gamma)})$ is $O(\rho^2 k_\gamma(D_X) \ln \frac{m}{\delta})$. Define D_γ to be the distribution such that $\mathbb{P}_{D_\gamma}[(X, Y)] = \mathbb{P}_{D_X}[(X, Y) \mid X \in \mathcal{X}_{\alpha(\gamma)}]$. By standard sample complexity bounds [16], for any distribution D over $\mathbb{R}^d \times \{\pm 1\}$, with probability at least $1 - \frac{\delta}{2}$ over samples, $\ell_m(\mathcal{A}, D) \leq \tilde{O}(\sqrt{\frac{F(\gamma/8,D) \ln \frac{1}{\delta}}{m}})$, where $F(\gamma, D)$ is the γ-fat-shattering dimension of the class of linear functions with domain restricted to the support of D in $\mathbb{R}^d$. Consider $D_{\gamma/8}$. Since the support of $D_{\gamma/8}$ is $\mathcal{X}_{\alpha(\gamma/8)}$, $F(\gamma/8, D_{\gamma/8}) \leq O(\rho^2 k_{\gamma/8}(D_X) \ln \frac{m}{\delta})$. With probability $1 - \delta$ over samples from D_X, the sample is drawn from $D_{\gamma/8}$. In addition, the probability of the unlabeled example to be drawn from $\mathcal{X}_{\alpha(\gamma/8)}$ is larger than $1 - \frac{1}{m}$. Therefore $\ell_m(\mathcal{A}, D) \leq \tilde{O}(\sqrt{\frac{\rho^2 k_{\gamma/8}(D_X) \ln \frac{m}{\delta}}{m}})$. Setting $\delta = \epsilon/2$ and bounding the expected error, we get $m(\epsilon, \gamma, D) \leq \tilde{O}(\frac{\rho^2 k_{\gamma/8}(D_X)}{\epsilon^2})$. Lemma 3.3 allows replacing $k_{\gamma/8}$ with $O(k_\gamma)$. □

Lemma A.2. *Let $T_1, \ldots, T_d$ be independent random variables such that all the moments $\mathbb{E}[T_i^n]$ for all i are non-negative. Let $\lambda_1, \ldots, \lambda_d$ be real coefficients such that $\sum_{i=1}^d \lambda_i = L$, and $\lambda_i \in [0, 1]$ for all $i \in [d]$. Then for all $t \geq 0$*

$$\mathbb{E}[\exp(t \sum_{i=1}^d \lambda_i T_i)] \leq \max_{i \in [d]} (\mathbb{E}[\exp(3tT_i)])^{\lceil L \rceil}.$$

Proof. Let T_i be independent random variables. Then, by Jensen's inequality,

$$\mathbb{E}[\exp(t \sum_{i=1}^d \lambda_i T_i)] = \prod_{i=1}^d \mathbb{E}[\exp(t\lambda_i T_i)] \leq \prod_{i=1}^d \mathbb{E}[\exp(tT_i \sum_{j=1}^d \lambda_j)]^{\frac{\lambda_i}{\sum_{j=1}^d \lambda_j}} \leq \max_{i \in [d]} \mathbb{E}[\exp(tT_i \sum_{j=1}^d \lambda_j)].$$

Now, consider a partition $Z_1, \ldots, Z_k$ of $[d]$, and denote $L_j = \sum_{i \in Z_j} \lambda_i$. Then by the inequality above,

$$\mathbb{E}[\exp(t \sum_{i=1}^d \lambda_i T_i)] = \prod_{j=1}^k \mathbb{E}[\exp(t \sum_{i \in Z_j} \lambda_i T_i)] \leq \prod_{j=1}^k \max_{i \in Z_j} \mathbb{E}[\exp(tT_i L_j)].$$

Let the partition be such that for all $j \in [k]$, $L_j \leq 1$. There exists such a partition such that $L_j < \frac{1}{2}$ for no more than one j. Therefore, for this partition $L = \sum_{i=1}^d \lambda_i = \sum_{j \in [k]} L_j \geq \frac{1}{2}(k - 1)$. Thus $k \leq 2L + 1$.

Now, consider $\mathbb{E}[\exp(tT_i L_j)]$ for some i and j. For any random variable X

$$\mathbb{E}[\exp(tX)] = \sum_{n=0}^\infty \frac{t^n \mathbb{E}[X^n]}{n!}.$$

Therefore, $\mathbb{E}[\exp(tT_i L_j)] = \sum_{n=0}^\infty \frac{t^n L_j^n \mathbb{E}[T_i^n]}{(n)!}$. Since $\mathbb{E}[T_i^n] \geq 0$ for all n, and $L_j \leq 1$, it follows that $\mathbb{E}[\exp(tT_i L_j)] \leq \mathbb{E}[\exp(tT_i)]$. Thus

$$\mathbb{E}[\exp(t \sum_{i=1}^d \lambda_i T_i)] \leq \prod_{j=1}^k \max_{i \in Z_j} \mathbb{E}[\exp(tT_i)] \leq \max_{i \in [d]} \mathbb{E}[\exp(t \sum_{j=1}^k T_i[j])],$$

where $T_i[j]$ are independent copies of T_i.

It is easy to see that $\mathbb{E}[\exp[\frac{1}{a}\sum_{i=1}^{a} X_i]] \leq \mathbb{E}[\exp[\frac{1}{b}\sum_{i=1}^{b} X_i]]$, for $a \geq b$ and $X_1, \ldots, X_a$ i.i.d. random variables. Since $k \geq \lceil L \rceil$ it follows that

$$\mathbb{E}[\exp(t\sum_{i=1}^{d}\lambda_i T_i)] \leq \max_{i\in[d]}\mathbb{E}[\exp(t\sum_{j=1}^{k} T_i[j])] \leq \max_{i\in[d]}\mathbb{E}[\exp(t\frac{k}{\lceil L\rceil}\sum_{j=1}^{\lceil L\rceil} T_i[j])].$$

Since $k \leq 2L+1$ and all the moments of $T_i[j]$ are non-negative, it follows that

$$\mathbb{E}[\exp(t\sum_{i=1}^{d}\lambda_i T_i)] \leq \max_{i\in[d]}\mathbb{E}[\exp(t(2+\frac{1}{\lceil L\rceil})\sum_{j=1}^{\lceil L\rceil} T_i[j])].$$

□

A.4 Proof of Theorem 5.2

the following lemma, which allows converting the representation of the Gram-matrix to a different feature space while keeping the separation properties intact. For a matrix M, M^+ denotes its pseudo-inverse. If $(M'M)$ is invertible then $M^+ = (B'B)^{-1}B'$.

Lemma A.3. *Let X be an $m \times d$ matrix such that XX' is invertible, and Y such that $XX' = YY'$. Let $r \in \mathbb{R}^m$ be some real vector. If there exists a vector $\widetilde{w}$ such that $Y\widetilde{w} = r$, then there exists a vector w such that $Xw = r$ and $\|w\| = \|P\widetilde{w}\|$, where $P = Y'Y'^{+} = Y'(YY')^{-1}Y$ is the projection matrix onto the sub-space spanned by the rows of Y.*

Proof. Denote $K = XX' = YY'$. Set $T = Y'X'^{+} = Y'K^{-1}X$. Set $w = T'\widetilde{w}$. We have $Xw = XT'\widetilde{w} = XX'K^{-1}Y\widetilde{w} = Y\widetilde{w} = r$. In addition, $\|w\| = w'w = \widetilde{w}'TT'\widetilde{w}$. By definition of T, $TT' = Y'X'^{+}X^{+}Y = Y'K^{+}Y = Y'K^{-1}Y = Y'(YY')^{-1}Y = Y'Y'^{+} = P$. Since P is a projection matrix, we have $P^2 = P$. In addition, $P = P'$. Therefore $TT' = PP'$, and so $\|w\| = \widetilde{w}'PP'\widetilde{w} = \|P\widetilde{w}\|$. □

The next lemma will allow us to prove that if a set is shattered at the origin, it can be separated with the exact margin.

Lemma A.4. *Let $R = \{r_y \in \mathbb{R}^m \mid y \in \{\pm 1\}^m\}$ such that for all $y \in \{\pm 1\}^m$ and for all $i \in [m]$, $r_y[i]y[i] \geq 1$. Then $\forall y \in \{\pm 1\}^m, y \in$ conv(R).*

Proof. We will prove the claim by induction on the dimension m.

Induction base: For $m = 1$, we have $R = \{(a), (b)\}$ where $a \leq -1$ and $b \geq 1$. Clearly, conv$R = [a, b]$, and the two one-dimensional vectors $(+1)$ and (-1) are in $[a, b]$.

Induction step: For a vector $t = (t[1], \ldots, t[m]) \in \mathbb{R}^m$, denote by $\bar{t}$ its projection $(t[1], \ldots, t[m-1])$ on $\mathbb{R}^{m-1}$. Similarly, for a set of vectors $S \subseteq \mathbb{R}^m$, let $\bar{S} = \{\bar{s} \mid s \in S\} \subseteq \mathbb{R}^{m-1}$. Define

$$Y_+ = \{y \in \{\pm 1\}^m \mid y[m] = +1\}$$
$$Y_- = \{y \in \{\pm 1\}^m \mid y[m] = -1\}.$$

Let $R_+ = \{r_y \mid y \in Y_+\}$, and similarly for R_-. Then $\bar{R}_+$ and $\bar{R}_-$ satisfy the assumptions for R when $m-1$ is substituted for m.

Let $y^* \in \{\pm 1\}^m$. We wish to prove $y^* \in$ conv(R). From the induction hypothesis we have $\bar{y}^* \in$ conv$(\bar{R}_+)$ and $\bar{y}^* \in$ conv$(\bar{R}_-)$. Thus

$$\bar{y}^* = \sum_{y\in Y_+}\alpha_y \bar{r}_y = \sum_{y\in Y_-}\beta_y \bar{r}_y,$$

where $\alpha_y, \beta_y \geq 0$, $\sum_{y\in Y_+}\alpha_y = 1$, and $\sum_{y\in Y_-}\beta_y = 1$. Let $y^*_a = \sum_{y\in Y_+}\alpha_y r_y$ and $y^*_b = \sum_{y\in Y_-}\alpha_y r_y$. We have that $\forall y \in Y_+, r_y[m] \geq 1$, and $\forall y \in Y_-, r_y[m] \leq -1$. Therefore, $y^*_a[m] \geq 1$ and $y^*_b[m] \leq -1$. In addition, $\bar{y}^*_a = \bar{y}^*_b = \bar{y}$. Hence there is $\gamma \in [0, 1]$ such that $y^* = \gamma y^*_a + (1-\gamma)y^*_b$. Since $y^*_a \in$ conv(R_+) and $y^*_b \in$ conv(R_-), we have $y^* \in$ conv(R). □

Proof of Theorem 5.2. Let $XX' = U\Lambda U'$ be the SVD of XX', where U is an orthogonal matrix and Λ is a diagonal matrix. Let $Y = U\Lambda^{\frac{1}{2}}$. We have $XX' = YY'$. We show that the conditions are sufficient and necessary for the shattering of S.

Sufficient: Assume XX' is invertible. Then Λ is invertible, thus Y is invertible. For any $y \in \{\pm 1\}^m$, Let $\widetilde{w} = Y^{-1}y$. We have $Y\widetilde{w} = y$. In addition, $\|\widetilde{w}\|^2 = y'(YY')^{-1}y = y'(XX')^{-1}y \leq 1$. Therefore, by Lemma A.3, there exists a separator w such that $Xw = y$ and $\|w\| = \|P\widetilde{w}\| = \|\widetilde{w}\|$.

Necessary: If XX' is not invertible then the vectors in S are linearly dependent, thus by standard VC-theory [16] S cannot be shattered using linear separators. The first condition is therefore necessary. We assume S is 1-shattered at the origin and show that the second condition necessarily holds. Let $L = \{r \mid \exists w \in \mathbf{B}_1^d, Xw = r\}$. Since S is shattered, For any $y \in \{\pm 1\}^m$ there exists $r_y \in L$ such that $\forall i \in [m], r_y[i]y[i] \geq 1$. By Lemma A.4, $\forall y \in \{\pm 1\}^m, y \in \text{conv}(R)$ where $R = \{r_y \mid y \in \{\pm 1\}^m\}$. Since L is convex and $R \subseteq L$, $\text{conv}(R) \subseteq L$. Thus for all $y \in \{\pm 1\}^m$, $y \in L$, that is there exists $w_y \in \mathbb{R}^m$ such that $Xw_y = y$ and $\|w_y\| \leq 1$. From Lemma A.3 we thus have $\widetilde{w}_y$ such that $Y\widetilde{w}_y = y$ and $\|\widetilde{w}_y\| = \|Pw_y\| \leq \|w_y\| \leq 1$. Y is invertible, hence $\widetilde{w}_y = Y^{-1}y$. Thus $y'(XX')^{-1}y = y'(YY')^{-1}y = \|\widetilde{w}_y\| \leq 1$. □

A.5 Proof of Theorem 6.2

First, define:

- The unit sphere: $S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$.
- An ϵ-Net on the unit sphere: Denote by $\mathcal{N}_n(\epsilon)$ a minimal-size ϵ-Net for S^{n-1}, that is a set such that for all $y \in S^{n-1}, \exists x \in \mathcal{N}_n(\epsilon)$ such that $\|x - y\|_2 \leq \epsilon$.

For ϵ-Nets we have the following bound on their size:

Proposition A.5 ([19], Proposition 2.1)**.** *For any $\epsilon > 0$,*

$$|\mathcal{N}_n(\epsilon)| \leq 2n(1 + \frac{2}{\epsilon})^{n-1}.$$

The proof of the theorem follows. It relies on several lemmas which are disclosed subsequently.

Proof of Theorem 6.2. Let $\mathcal{N}_m(\epsilon)$ be an ϵ-Net for the unit sphere as defined above. Then for any matrix A of dimensions $m \times d$,

$$\lambda_m(AA') = \inf_{\|x\|_2=1} \|x'AA'x\|_2 = \inf_{\|x\|_2=1} \|A'x\|_2^2,$$

and $\inf_{\|x\|_2=1} \|A'x\|_2 \geq \min_{x \in \mathcal{N}} \|A'x\|_2 - \epsilon \|A'\|_{2,2}$.

We assume w.l.o.g. that Σ is not singular (otherwise the dimension of the space can be reduced appropriately), and let $Y = X_m\Sigma^{-1}$. Let $\beta \leq (c - K\epsilon)^2$ where c, K, ϵ are parameters to be fixed later, and let $m = \beta L$. Then

$$\begin{aligned}
&\mathbb{P}[\lambda_m(Y\Sigma Y') \leq m] \leq \mathbb{P}[\inf_{\|x\|_2=1} \|\sqrt{\Sigma}Y'x\|_2 \leq (c - K\epsilon)\sqrt{L}] \\
&\leq \mathbb{P}[\min_{x \in \mathcal{N}(\epsilon)} \|\sqrt{\Sigma}Y'x\|_2 - \epsilon\|\sqrt{\Sigma}Y'\| \leq (c - K\epsilon)\sqrt{L}] \\
&\leq \mathbb{P}[\min_{x \in \mathcal{N}(\epsilon)} \|\sqrt{\Sigma}Y'x\|_2 \leq c\sqrt{L}] + \mathbb{P}[\|\sqrt{\Sigma}Y'\| \geq K\sqrt{L}].
\end{aligned}$$

Since Y_{ij} is sub-Gaussian with moment B, $\mathbb{E}[Y_{ij}^4] \leq 5B^4$ [12, Lemma 1.4]. Thus, by Lemma A.11, there are α and η which depend only on B such that

$$\mathbb{P}_Y[\|\sqrt{\Sigma}Yx\|^2 \leq \alpha L] \leq \eta^L.$$

Therefore

$$\mathbb{P}[\min_{x \in \mathcal{N}(\epsilon)} \|\sqrt{\Sigma}Y'x\|_2 \leq \sqrt{\alpha L}] \leq \sum_{x \in \mathcal{N}_m(\epsilon)} \mathbb{P}_Y[\|\sqrt{\Sigma}Y'x\|_2 \leq \sqrt{\alpha L}] \leq |\mathcal{N}_m(\epsilon)|\eta^L.$$

We thus let $c = \min(\sqrt{\alpha}, 1)$.

By Lemma A.8, for any K and ρ,

$$\mathbb{P}[\|\sqrt{\Sigma}Y\| \geq K\sqrt{L}] \leq 2|\mathcal{N}_m(\rho)||\mathcal{N}_d(\rho)| \exp(-\frac{(1-\rho)^4 K^2 L}{2B^2}).$$

Combining the inequalities and setting $m = \beta L$,

$$\begin{aligned}\mathbb{P}[\lambda_m(Y\Sigma Y') \leq m] &\leq |\mathcal{N}_m(\epsilon)|\eta^L + 2|\mathcal{N}_m(\rho)||\mathcal{N}_d(\rho)| \exp(-\frac{(1-\rho)^4 K^2 L}{2B^2}) \\ &\leq 2\beta L(1+\frac{2}{\epsilon})^{\beta L - 1}\eta^L + 4\beta L^2(1+\frac{2}{\rho})^{(1+\beta)L-1} \exp(-\frac{(1-\rho)^4 K^2 L}{2B^2}),\end{aligned}$$

where the last inequality follows from Proposition A.5.

Fix $\rho = \frac{1}{2}$. Let K be a constant large enough such that for all $\beta < 1$, and for all $L \geq L_0$ (where $L_0 > 0$ is arbitrary)

$$4\beta L^2(1+\frac{2}{\rho})^{(1+\beta)L-1} \exp(-\frac{(1-\rho)^4 K^2 L}{2B^2}) \leq \frac{1}{2} - \delta/2.$$

Let $\epsilon = c/2K$, so that $c - K\epsilon > 0$, and let β such that for all $L > L_0$,

$$2\beta L(1+\frac{2}{\epsilon})^{\beta L - 1}\eta^L \leq \frac{1}{2} - \delta/2.$$

Then for the chosen β, and for any $L > L_0$,

$$\mathbb{P}[\lambda_m(Y\Sigma Y') \geq m] \geq 1 - \delta.$$

□

The following easy to prove facts are found in several places in the literature:

Proposition A.6 (See e.g. [22])**.** *For any linear operator $A : \mathbb{R}^m \to \mathbb{R}^d$,*

$$\|A\| \leq \frac{1}{1-\epsilon} \sup_{x \in \mathcal{N}_m(\epsilon)} \|Ax\|.$$

Proposition A.7 (See e.g. [19], proof of Proposition 2.2)**.** *For any linear operator $A : \mathbb{R}^m \to \mathbb{R}^d$, and any $x \in \mathbb{R}^m$,*

$$\|Ax\| \leq \frac{1}{1-\epsilon} \sup_{y \in \mathcal{N}_d(\epsilon)} \langle Ax, y\rangle.$$

In [19] this fact appears using the absolute value of $\langle Ax, y\rangle$, but the version above can be proved in the same manner.

The following is a variation on Prop 2.3 in [19].

Lemma A.8. *Let Y be a $d \times m$ matrix with $m \leq d$, such that Y_{ij} are independent sub-Gaussian variables with moment B. Let $\Sigma = diag(\lambda_1, \ldots, \lambda_d)$ be a diagonal $d \times d$ matrix with $\lambda_i \in [0,1]$ for all i. Then for all $t \geq 0$,*

$$\mathbb{P}[\|\sqrt{\Sigma}Y\| \geq t] \leq 2|\mathcal{N}_m(\epsilon)||\mathcal{N}_d(\epsilon)| \exp(-\frac{1}{2}\frac{(1-\epsilon)^4 t^2}{B^2}).$$

Proof. From Proposition A.6 it follows that

$$\|\sqrt{\Sigma}Y\| \leq \frac{1}{1-\epsilon} \sup_{x \in \mathcal{N}_m(\epsilon)} \|\sqrt{\Sigma}Yx\|.$$

Therefore

$$\mathbb{E}[\exp(t\|\sqrt{\Sigma}Y\|)] \leq \mathbb{E}[\exp(\frac{t}{1-\epsilon} \sup_{x \in \mathcal{N}_m(\epsilon)} \|\sqrt{\Sigma}Yx\|)] \leq \sum_{x \in \mathcal{N}_m(\epsilon)} \mathbb{E}[\exp(\frac{t}{1-\epsilon}\|\sqrt{\Sigma}Yx\|)].$$

Now, let $x \in \mathcal{N}_m(\epsilon)$. From Proposition A.7,

$$\mathbb{E}[\exp(t\|\sqrt{\Sigma}Yx\|)] \leq \mathbb{E}[\exp(\frac{t}{1-\epsilon}\sup_{y\in\mathcal{N}_d(\epsilon)}\langle\sqrt{\Sigma}Yx,y\rangle)]$$
$$\leq \sum_{y\in\mathcal{N}_d(\epsilon)}\mathbb{E}[\exp(\frac{t}{1-\epsilon}\langle\sqrt{\Sigma}Yx,y\rangle)].$$

For any $x \in S^{m-1}, y \in S^{d-1}$,

$$\mathbb{E}[\exp(t\langle\sqrt{\Sigma}Yx,y\rangle)] = \mathbb{E}[\exp(t\sum_{i=1}^{d}\sum_{j=1}^{m}\sqrt{\lambda_i}Y_{ij}x_jy_i)] = \prod_{i=1,j=1}^{d,m}\mathbb{E}[\exp(t\sqrt{\lambda_i}Y_{ij}x_jy_i]$$
$$\leq \prod_{i=1,j=1}^{d,m}\exp(\frac{1}{2}B^2t^2x_j^2\lambda_iy_i^2) = \exp(\frac{1}{2}B^2t^2\sum_{j=1}^{m}x_j^2\sum_{i=1}^{d}\lambda_iy_i^2) \leq \exp(\frac{1}{2}B^2t^2),$$

where the last inequality follows from the facts $\|x\|^2 = 1, \|y\|^2 = 1$, and $\forall i \in [d]\lambda_i \leq 1$. It follows that

$$\mathbb{E}[\exp(t\|\sqrt{\Sigma}Yx\|)] \leq |\mathcal{N}_d(\epsilon)|\exp(\frac{1}{2}\frac{B^2t^2}{(1-\epsilon)^2})$$

Thus, for all $t \in \mathbb{R}$,

$$\mathbb{E}[\exp(t\|\sqrt{\Sigma}Y\|)] \leq |\mathcal{N}_m(\epsilon)||\mathcal{N}_d(\epsilon)|\exp(\frac{1}{2}\frac{B^2t^2}{(1-\epsilon)^4}).$$

By Chernoff's method, for all $t \geq 0$

$$\mathbb{P}[\|\sqrt{\Sigma}Y\| \geq t] \leq 2|\mathcal{N}_m(\epsilon)||\mathcal{N}_d(\epsilon)|\exp(-\frac{1}{2}\frac{(1-\epsilon)^4t^2}{B^2}).$$

□

The following lemma is a variation of [20], Lemma 2.6.

Lemma A.9. *Let Y be a $d \times m$ matrix with $m \leq d$, such that the columns of Y are i.i.d. random vectors. Assume further that $\forall i \in [d], j \in [m]$ $E[Y_{ij}] = 0$, $\mathbb{E}[Y_{ij}^2] = 1$ and $\mathbb{E}[Y_{ij}^4] \leq B$ for some real number B. Let $\Sigma = diag(\lambda_1, \ldots, \lambda_d)$ be a diagonal $d \times d$ matrix such that $\forall i \in [d]$ $\lambda_i \geq 0$ and trace$(\Sigma) \leq L$. Then for every $x \in S^{m-1}$,*

$$\mathbb{P}[\|\sqrt{\Sigma}Yx\|_2^2 \leq \frac{L}{2}] \leq 1 - \frac{1}{196B}.$$

Proof. Let $T_i = (\sum_{j=1}^{m} Y_{ij}x_j)^2$, and let $T_\Sigma = \|\sqrt{\Sigma}Yx\|_2^2 = \sum_{i=1}^{d}\lambda_iT_i$.

First, since $\mathbb{E}[Y_{ij}] = 0$ and $\mathbb{E}[Y_{ij}] = 1$ for all i, j,

$$\mathbb{E}[T_i] = \sum_{i=1}^{m}x_j^2\mathbb{E}[Y_{ij}^2] = \|x\|_2^2 = 1.$$

Therefore $\mathbb{E}[T_\Sigma] = L$.

Second, since $Y_{i1}, \ldots, Y_{im}$ are independent, we can use a symmetrization argument as done in [20], proof of Lemma 2.6, to get

$$\mathbb{E}[T_i^2] = \mathbb{E}[(\sum_{j=1}^{m}Y_{ij}x_j)^4] \leq 16B_4B = 48B,$$

Where $B_4 = 3$ is the upper Khinchine constant for $p = 4$ [23]. Thus,

$$\mathbb{E}[T_\Sigma^2] = \mathbb{E}[(\sum_{i=1}^{d}\lambda_iT_i)^2] = \sum_{i,j=1}^{d}\lambda_i\lambda_j\mathbb{E}[T_iT_j]$$
$$\leq \sum_{i,j=1}^{d}\lambda_i\lambda_j\mathbb{E}[T_i^2]^{\frac{1}{2}}\mathbb{E}[T_j^2]^{\frac{1}{2}} \leq 48B(\sum_{i=1}^{d}\lambda_i)^2 = 48BL^2,$$

Where we have used the fact that $\mathbb{E}[XY]^2 \leq E[X^2]\mathbb{E}[Y^2]$ for any two random variables X and Y.

By the Paley-Zigmund inequality [24], if $0 < \theta < 1$,

$$\mathbb{P}[T_\Sigma \geq \theta\mathbb{E}[T_\Sigma]] \geq (1-\theta)^2 \frac{\mathbb{E}[T_\Sigma]^2}{\mathbb{E}[T_\Sigma^2]} \geq \frac{(1-\theta)^2}{48B}.$$

Therefore, setting $\theta = 1/2$,

$$\mathbb{P}[T_\Sigma \leq \frac{L}{2}] \leq 1 - \frac{1}{196B}.$$

□

We also use the following lemma:

Lemma A.10 (Lemma 2.2 item (2) in [20])**.** *Let $T_1, \ldots, T_n$ be independent non-negative random variables. Assume that there exists $\theta > 0$ and $\mu \in (0,1)$ such that for any i $\mathbb{P}[T_i \leq \theta] \leq \mu$. Then there exist $\alpha > 0$ and $\eta \in (0,1)$ that depend only on θ and μ such that*

$$\mathbb{P}(\sum_{i=1}^{n} T_i < \alpha n) \leq \eta^n.$$

The following lemma is used in the proof of the theorem above.

Lemma A.11. *Let Y be a $d \times m$ matrix with $m \leq d$, such that $Y_{i,j}$ are independent centered random variables with variance 1 and fourth moments no more than B. Let $\Sigma = diag(\lambda_1, \ldots, \lambda_d)$ be a diagonal $d \times d$ matrix with $\lambda_i \leq 1$ for all i, and let $L = \sum_{i=1}^{d} \lambda_i$. Let $x \in \mathbb{R}^m$ be a vector such that $\|x\|_2 = 1$. Then there exist $\alpha > 0$ and $\eta \in (0,1)$ that depend only on B such that*

$$\mathbb{P}_Y[\|\sqrt{\Sigma}Yx\|^2 \leq \alpha L] \leq \eta^{2L}.$$

Proof. Consider a partition $Z_1, \ldots, Z_k$ of $[d]$, and denote $L_j = \sum_{i \in Z_j} \lambda_i$. Let the partition be such that for all $j \in [k]$, $L_j \leq 1$. There exists such a partition such that $L_j > \frac{1}{2}$ for all but at most one j. Therefore, for this partition $L = \sum_{i=1}^{d} \lambda_i = \sum_{j \in [k]} L_j \geq \frac{1}{2}(k-1)$. Thus $k \leq 2L+1$.

Let $\Sigma[j]$ be the diagonal matrix whose diagonal elements are λ_i such that $i \in Z_j$, and let $Y[j]$ be the sub-matrix of Y which includes only the lines whose indexes are in Z_j.

Then

$$\|\sqrt{\Sigma}Yx\|^2 = \sum_{i=1}^{d} \lambda_i (\sum_{j=1}^{m} Y_{ij}x_j)^2 = \sum_{j \in [k]} \sum_{i \in Z_j} \lambda_i (\sum_{j=1}^{m} Y_{ij}x_j)^2 = \sum_{j \in [k]} \|\Sigma[j]Y[j]x\|^2.$$

By Lemma A.9, $\mathbb{P}[\|\Sigma[j]Y[j]x\|^2 \leq \frac{L_j}{2}] \leq 1 - \frac{1}{196B}$.

Let $J = \{j \in [k] \mid L_j > \frac{1}{2}\}$. For all $j \in J$,

$$\mathbb{P}[\|\Sigma[j]Y[j]x\|^2 \leq \frac{1}{4}] \leq 1 - \frac{1}{196B}.$$

In addition, $|J| \geq L$. Therefore, by Lemma A.10 there are $\alpha > 0$ and $\eta \in (0,1)$ that depend only on B such that

$$\begin{aligned}\mathbb{P}[\|\sqrt{\Sigma}Yx\|^2 < \alpha L] &\leq \mathbb{P}[\|\sqrt{\Sigma}Yx\|^2 < \alpha |J|] \\ &\leq \mathbb{P}[\sum_{j \in J} \|\Sigma[j]Y[j]x\|^2 < \alpha |J|] \leq \eta^{|J|} \leq \eta^L.\end{aligned}$$

□