# A PROOFS OF MAIN THEOREMS

## A.1 Proof of Theorem 1

We fix $t = t(n, \boldsymbol{\lambda}) > 0$ to a positive value depending on $\boldsymbol{\lambda}$ and $n$ that will be determined later. We define the following events:

1. For $i \in [n]$, $\mathcal{B}_i$ is the event that $\boldsymbol{K}_{\setminus i}$ is non-singular and
$$y_i \boldsymbol{y}_{\setminus i}^{\mathsf{T}} \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda} \boldsymbol{z}_i \geq 1.$$
2. For $i \in [n]$, $\mathcal{S}_i$ is the event that $\boldsymbol{K}_{\setminus i}$ is singular.
3. $\mathcal{S}$ is the event that $\boldsymbol{K}$ is singular.
4. $\mathcal{B} := \mathcal{S} \cup \bigcup_{i=1}^{n} (\mathcal{B}_i \cup \mathcal{S}_i)$.

Additionally, we define the event $\mathcal{E}_i(t)$, for every $i \in [n]$ and a given $t > 0$, that $\boldsymbol{K}_{\setminus i}$ is non-singular and

$$\left\| \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus i}^{\mathsf{T}} \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i} \right\|_2^2 \geq \frac{1}{t}.$$

Note that if the event $\mathcal{B}$ does not occur, then $\boldsymbol{Z} \boldsymbol{\Lambda} \boldsymbol{Z}^{\mathsf{T}}$ is non-singular, each $\boldsymbol{K}_{\setminus i}$ is non-singular, and

$$y_i \boldsymbol{y}_{\setminus i}^{\mathsf{T}} \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda} \boldsymbol{z}_i < 1, \quad \text{for all } i = 1, \ldots, n.$$

Hence, by Lemma 1, if $\mathcal{B}$ does not occur, then every training example is a support vector.

So, it suffices to upper-bound the probability of the event $\mathcal{B}$. We bound $\Pr(\mathcal{B})$ as follows:

$$\begin{aligned}
\Pr(\mathcal{B}) &\leq \Pr(\mathcal{S}) + \sum_{i=1}^{n} \Pr(\mathcal{B}_i \cup \mathcal{S}_i) \\
&= \Pr(\mathcal{S}) + \sum_{i=1}^{n} \left( \Pr((\mathcal{B}_i \cap \mathcal{S}_i^{\mathrm{c}} \cap \mathcal{E}_i(t)^{\mathrm{c}}) \cup (S_i \cap \mathcal{E}_i(t)^{\mathrm{c}})) + \Pr((\mathcal{B}_i \cup \mathcal{S}_i) \cap \mathcal{E}_i(t)) \right) \\
&\leq \Pr(\mathcal{S}) + \sum_{i=1}^{n} \left( \Pr(\mathcal{B}_i \mid \mathcal{S}_i^{\mathrm{c}} \cap \mathcal{E}_i(t)^{\mathrm{c}}) \Pr(\mathcal{S}_i^{\mathrm{c}} \cap \mathcal{E}_i(t)^{\mathrm{c}}) + \Pr(S_i \cap \mathcal{E}_i(t)^{\mathrm{c}}) + \Pr((\mathcal{B}_i \cup \mathcal{S}_i) \cap \mathcal{E}_i(t)) \right) \\
&\leq \Pr(\mathcal{S}) + \sum_{i=1}^{n} \left( \Pr(\mathcal{B}_i \mid \mathcal{S}_i^{\mathrm{c}} \cap \mathcal{E}_i(t)^{\mathrm{c}}) + \Pr(S_i) + \Pr(\mathcal{E}_i(t)) \right). \qquad (7)
\end{aligned}$$

Above, the first two inequalities follow from the union bound, and the rest uses the law of total probability.

We first upper bound the probability of the singularity events in the following lemma.

**Lemma 2.** *We have*

$$\max\{\Pr(\mathcal{S}), \Pr(\mathcal{S}_1), \ldots, \Pr(\mathcal{S}_n)\} \leq 2 \cdot 9^n \cdot \exp\left( -c \cdot \min\left\{ \frac{d_2}{v^2}, \frac{d_\infty}{v} \right\} \right)$$

*where $c > 0$ is the universal constant in the statement of Lemma 8.*

*Proof.* It suffices to bound $\Pr(\mathcal{S})$, since each $\boldsymbol{K}_{\setminus i}$ is a principal submatrix of $\boldsymbol{K}$, and hence $\lambda_{\min}(\boldsymbol{K}_{\setminus i}) \geq \lambda_{\min}(\boldsymbol{K})$ for all $i \in [n]$. Observe that

$$\boldsymbol{Z} \boldsymbol{\Lambda} \boldsymbol{Z}^{\mathsf{T}} = \sum_{j=1}^{d} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^{\mathsf{T}}$$

where $\boldsymbol{v}_j$ is the $j^{\text{th}}$ column of $\boldsymbol{Z}$. Recall that the columns of $\boldsymbol{Z}$ are independent, and so these vectors satisfy the conditions of Lemma 8. Moreover, since $\boldsymbol{Z\Lambda Z}^\intercal$ is positive semi-definite, its singularity would require

$$\left\|\boldsymbol{Z\Lambda Z}^\intercal - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2 \geq \|\boldsymbol{\lambda}\|_1.$$

The probability of this latter event can be bounded by Lemma 8 with $\tau = \|\boldsymbol{\lambda}\|_1$, thereby giving the claimed bound on $\Pr(\mathcal{S})$. This completes the proof of the lemma. ☐

The next lemma upper bounds the probability of the event $\mathcal{B}_i$ conditioned on the non-singularity event $\mathcal{S}_i$ and the complement of the event $\mathcal{E}_i(t)$.

**Lemma 3.** *For any $t > 0$,*

$$\Pr(\mathcal{B}_i \mid \mathcal{S}_i^{\mathrm{c}} \cap \mathcal{E}_i(t)^{\mathrm{c}}) \leq 2\exp\left(-\frac{t}{2v}\right).$$

*Proof.* Let $\mathcal{B}_i'$ be the event that $\boldsymbol{K}_{\setminus i}$ is non-singular and

$$\left|\boldsymbol{y}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda} \boldsymbol{z}_i\right| = \max\left\{-\boldsymbol{y}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda} \boldsymbol{z}_i,\, \boldsymbol{y}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda} \boldsymbol{z}_i\right\} \geq 1.$$

Since $|y_i| = 1$, it follows that $\mathcal{B}_i \subseteq \mathcal{B}_i'$, so

$$\Pr(\mathcal{B}_i \mid \mathcal{S}_i^{\mathrm{c}} \cap \mathcal{E}_i(t)^{\mathrm{c}}) \leq \Pr(\mathcal{B}_i' \mid \mathcal{S}_i^{\mathrm{c}} \cap \mathcal{E}_i(t)^{\mathrm{c}}).$$

Conditional on the event $\mathcal{S}_i^{\mathrm{c}} \cap \mathcal{E}_i(t)^{\mathrm{c}}$, we have that $\boldsymbol{K}_{\setminus i}$ is non-singular and $\|\boldsymbol{\Lambda Z}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i}\|_2^2 \leq 1/t$. Since $\boldsymbol{z}_i$ is independent of $\{(\boldsymbol{z}_j, y_j) : j \neq i\}$, it follows that

$$\boldsymbol{y}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda} \boldsymbol{z}_i = (\boldsymbol{\Lambda Z}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i})^\intercal \boldsymbol{z}_i$$

is (conditionally) sub-Gaussian with parameter at most $v \cdot \|\boldsymbol{\Lambda Z}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i}\|_2^2 \leq v/t$. Then, the standard sub-Gaussian tail bound gives us

$$\Pr\left(\mathcal{B}_i \mid \mathcal{S}_i^{\mathrm{c}} \cap \mathcal{E}_i(t)^{\mathrm{c}}\right) \leq \Pr\left(\mathcal{B}_i' \mid \mathcal{S}_i^{\mathrm{c}} \cap \mathcal{E}_i(t)^{\mathrm{c}}\right) \leq 2\exp\left(-\frac{t}{2v}\right).$$

This completes the proof of the lemma. ☐

Finally, the following lemma upper bounds the probability of the event $\mathcal{E}_i(t)$ for $t := d_\infty/2n$.

**Lemma 4.**

$$\Pr(\mathcal{E}_i(d_\infty/(2n))) \leq 2 \cdot 9^{n-1} \cdot \exp\left(-c \cdot \min\left\{\frac{d_2}{4v^2}, \frac{d_\infty}{v}\right\}\right)$$

*where $c > 0$ is the universal constant from Lemma 8.*

*Proof.* Let $\mathcal{E}_i'(t)$ be the event that

$$\lambda_{\min}(\boldsymbol{K}_{\setminus i}) \leq n\|\boldsymbol{\lambda}\|_\infty t.$$

Under $\mathcal{S}_i^{\mathrm{c}}$, the matrix $\boldsymbol{K}_{\setminus i}$ is non-singular. We get

$$\begin{aligned}
\|\boldsymbol{\Lambda Z}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i}\|_2^2 &\leq \|\boldsymbol{\Lambda}^{1/2}\|_{\mathrm{op}}^2 \|\boldsymbol{\Lambda}^{1/2} \boldsymbol{Z}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i}\|_2^2 \\
&= \|\boldsymbol{\lambda}\|_\infty \boldsymbol{y}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda Z}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i} \\
&\leq n\|\boldsymbol{\lambda}\|_\infty \sup_{\boldsymbol{u} \in \mathbb{R}^{n-1} : \|\boldsymbol{v}\|_2 = 1} \boldsymbol{u}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{u} \\
&= \frac{n\|\boldsymbol{\lambda}\|_\infty}{\lambda_{\min}(\boldsymbol{K}_{\setminus i})}.
\end{aligned}$$

It follows that $\mathcal{E}_i(t) \subseteq \mathcal{E}'_i(t)$. Observe that for $t := d_\infty/(2n)$, the event $\mathcal{E}'_i(t)$ is that where

$$\lambda_{\min}(\boldsymbol{K}_{\setminus i}) \leq \frac{1}{2}\|\boldsymbol{\lambda}\|_1.$$

Therefore (as in the proof of Lemma 2), Lemma 8 with $\tau = \|\boldsymbol{\lambda}\|_1/2$ implies that

$$\begin{aligned}\Pr(\mathcal{E}'_i(d_\infty/(2n))) &= \Pr\left(\lambda_{\min}(\boldsymbol{K}_{\setminus i}) \leq \frac{1}{2}\|\boldsymbol{\lambda}\|_1\right) \\ &\leq 2 \cdot 9^{n-1} \cdot \exp\left(-c \cdot \min\left\{\frac{d_2}{4v^2}, \frac{d_\infty}{v}\right\}\right).\end{aligned}$$

This completes the proof of the lemma. □

Plugging the probability bounds from Lemma 2, Lemma 3 and Lemma 4 (with $t = d_\infty/(2n)$) into Eq. (7) completes the proof of Theorem 1. □

### A.2 Proof of Theorem 2

The proof follows a similar sequence of steps to that of Theorem 1 with slight differences in the events that we condition on. We first observe that $\frac{1}{\sqrt{d}}\boldsymbol{z}_i \mid (\boldsymbol{Z}_{\setminus i}, \boldsymbol{y}_{\setminus i})$ is a uniformly random unit vector in $S^{d-1}$ restricted to the subspace orthogonal to the row space of $\boldsymbol{Z}_{\setminus i}$. That is, it has the same (conditional) distribution as $\boldsymbol{B}_i\boldsymbol{u}_i$, where:

1. $\boldsymbol{B}_i$ is a $d \times (n-d+1)$ matrix whose columns form an orthonormal basis for the orthogonal complement of $\boldsymbol{Z}_{\setminus i}$'s row space;
2. $\boldsymbol{u}_i$ is a uniformly random unit vector in $S^{d-n}$.

As before, for every $i \in [n]$, we define the event $\mathcal{B}_i$ that $\boldsymbol{K}_{\setminus i}$ is non-singular and

$$y_i \boldsymbol{y}_{\setminus i}^{\mathsf{T}} \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda} \boldsymbol{z}_i \geq 1.$$

The Haar measure ensures that the matrices $\boldsymbol{Z}$ and $\boldsymbol{Z}_{\setminus i}$ always have full row rank. Therefore, because $\boldsymbol{\Lambda} \succ \boldsymbol{0}$, the matrices $\boldsymbol{K}$ and $\boldsymbol{K}_{\setminus i}$ are always non-singular. So we do not need to worry about singularity (c.f. the events $\mathcal{S}$ and $\mathcal{S}_i$). We accordingly consider the event $\mathcal{B} := \bigcup_{i=1}^n \mathcal{B}_i$. As before, we also define the event $\mathcal{E}_i(t)$ for every $i \in [n]$ and a given $t > 0$, that

$$\|\boldsymbol{B}_i^{\mathsf{T}} \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus i}^{\mathsf{T}} \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i}\|_2^2 \geq \frac{d-n+1}{d} \cdot \frac{1}{t}.$$

By the union bound, we get

$$\begin{aligned}\Pr(\mathcal{B}) &\leq \sum_{i=1}^n \Pr(\mathcal{B}_i) \\ &\leq \sum_{i=1}^n \Pr(\mathcal{B}_i \mid \mathcal{E}_i(t)^{\mathrm{c}}) + \Pr(\mathcal{E}_i(t)),\end{aligned}$$

and so we need to upper bound the probabilities $\Pr(\mathcal{B}_i \mid \mathcal{E}_i(t)^{\mathrm{c}})$ and $\Pr(\mathcal{E}_i(t))$ for every $i \in [n]$.

The following lemma upper bounds $\Pr(\mathcal{B}_i \mid \mathcal{E}_i(t)^{\mathrm{c}})$, and is analogous to Lemma 3 in the proof of Theorem 1.

**Lemma 5.** *For any $t > 0$, we have*

$$\Pr(\mathcal{B}_i | \mathcal{E}_i(t)^{\mathrm{c}}) \leq 2\exp(-t).$$

*Proof.* First, as discussed above, we have

$$\begin{aligned}\Pr\left(y_i \boldsymbol{y}_{\setminus i}^\top \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda} \boldsymbol{z}_i \geq 1\right) &= \Pr\left(\sqrt{d} \cdot y_i \boldsymbol{y}_{\setminus i}^\top \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda} \boldsymbol{B}_i \boldsymbol{u}_i \geq 1\right) \\ &\leq \Pr\left(\sqrt{d} \left|(\boldsymbol{B}_i^\top \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus i}^\top \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i})^\top \boldsymbol{u}_i\right| \geq 1\right).\end{aligned}$$

Moreover, $\boldsymbol{u}_i$ is independent of $\boldsymbol{Z}_{\setminus i}$, and as established in Lemma 9, the random vector $\boldsymbol{u}_i$ is sub-Gaussian with parameter at most $\mathcal{O}(1/(d-n+1))$. Therefore, $\sqrt{d} \cdot (\boldsymbol{B}_i^\top \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus i}^\top \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i})^\top \boldsymbol{u}_i$ is conditionally sub-Gaussian with parameter at most $\frac{d}{d-n+1} \cdot \|\boldsymbol{B}_i^\top \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus i}^\top \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i}\|_2^2 \leq \frac{1}{t}$. Here, the last inequality follows because we have conditioned on $\mathcal{E}_i(t)^{\mathrm{c}}$. Therefore, the standard sub-Gaussian tail bound gives us

$$\Pr\left(\mathcal{B}_i \mid \mathcal{E}_i(t)^{\mathrm{c}}\right) \leq 2 \exp\left(-t\right). \qquad \square$$

The next lemma upper bounds $\Pr\left(\mathcal{E}_i(t)\right)$ for $t := \frac{d-n+1}{d} \cdot \frac{d_\infty}{2n}$, and is analogous to Lemma 4 in the proof of Theorem 1.

**Lemma 6.** *We have*

$$\Pr\left(\mathcal{E}_i\left(\frac{d-n+1}{d} \cdot \frac{d_\infty}{2n}\right)\right) \leq \exp\left(-c_1 \cdot d\right) + 2 \cdot 9^n \cdot \exp\left(-c_2 \cdot \min\{d_2, d_\infty\}\right)$$

*where $c_1 > 0$ and $c_2 > 0$ are universal constants.*

*Proof.* We get

$$\begin{aligned}\|\boldsymbol{B}_i^\top \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus i}^\top \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i}\|_2^2 &\leq \|\boldsymbol{B}_i^\top\|_2^2 \cdot \|\boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus i}^\top \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i}\|_2^2 \\ &= \|\boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus i}^\top \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i}\|_2^2 \\ &\leq \frac{n\|\boldsymbol{\lambda}\|_\infty}{\lambda_{\min}(\boldsymbol{K}_{\setminus i})},\end{aligned}$$

where we used the fact that $\boldsymbol{B}_i$ has orthonormal columns, and the last inequality follows by an identical argument to the proof of Lemma 4. We will show in particular that

$$\Pr\left(\lambda_{\min}(\boldsymbol{K}_{\setminus i}) \geq \frac{1}{2}\|\boldsymbol{\lambda}\|_1\right) \geq 1 - \exp(-c_1 \cdot d) - 2 \cdot 9^{n-1} \cdot \exp(-c_2 \cdot \min\{d_2, d_\infty\}). \tag{8}$$

Given Eq. (8), we can complete the proof of Lemma 6. This is because we get

$$\|\boldsymbol{B}_i^\top \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus i}^\top \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{y}_{\setminus i}\|_2^2 \leq \frac{2n\|\boldsymbol{\lambda}\|_\infty}{\|\boldsymbol{\lambda}\|_1} = \frac{2n}{d_\infty} = \frac{d-n+1}{d} \cdot \frac{1}{t}$$

for

$$t := \frac{d-n+1}{d} \cdot \frac{d_\infty}{2n}.$$

We complete the proof by proving Eq. (8). Let $\boldsymbol{S} \in \mathbb{R}^{m \times d}$ be a random matrix with iid standard Gaussian entries with $m := n-1$, and let the singular value decomposition of $\boldsymbol{S}$ be $\boldsymbol{S} = \boldsymbol{V} \boldsymbol{\Lambda}_S \boldsymbol{U}^\top$ where $\boldsymbol{V} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{U} \in \mathbb{R}^{d \times m}$ are orthonormal matrices. Then, it is well-known that $\sqrt{d} \cdot \boldsymbol{U}^\top$ follows the same distribution as $\boldsymbol{Z}_{\setminus i}$, and hence $\lambda_{\min}(\boldsymbol{K}_{\setminus i})$ has the same distribution as $d \cdot \lambda_{\min}(\boldsymbol{U}^T \boldsymbol{\Lambda} \boldsymbol{U})$. Moreover,

$$\begin{aligned}d \cdot \lambda_{\min}(\boldsymbol{U}^\top \boldsymbol{\Lambda} \boldsymbol{U}) &= \min_{\boldsymbol{v} \in \mathbb{R}^n, \|\boldsymbol{v}\|_2 = 1} \boldsymbol{v}^\top \boldsymbol{\Lambda}_S^{-1} \boldsymbol{V}^\top \boldsymbol{V} \boldsymbol{\Lambda}_S \boldsymbol{U}^\top \boldsymbol{\Lambda} \boldsymbol{U} \boldsymbol{\Lambda}_S \boldsymbol{V}^\top \boldsymbol{V} \boldsymbol{\Lambda}_S^{-1} \boldsymbol{v} \\ &\geq \frac{d}{\|\boldsymbol{\Lambda}_S\|_{\mathrm{op}}^2} \cdot \min_{\boldsymbol{v} \in \mathbb{R}^n, \|\boldsymbol{v}\|_2 = 1} \boldsymbol{v}^\top \boldsymbol{S} \boldsymbol{\Lambda} \boldsymbol{S}^\top \boldsymbol{v} \\ &= \frac{d}{\|\boldsymbol{\Lambda}_S\|_{\mathrm{op}}^2} \cdot \lambda_{\min}(\boldsymbol{S} \boldsymbol{\Lambda} \boldsymbol{S}^\top).\end{aligned}$$

By classical operator norm tail bounds on Gaussian random matrices (e.g., Vershynin, 2010, Corollary 5.35), we note that $\|\boldsymbol{\Lambda}_S\|_2^2 \leq \frac{3}{2}d$ with probability at least $1 - \exp(-c_1 \cdot d)$. Now, we note that the matrix $\boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^\intercal := \sum_{j=1}^{d} \lambda_j \boldsymbol{s}_j \boldsymbol{s}_j^\intercal$ where the $\boldsymbol{s}_j$'s are iid standard Gaussian random vectors in $\mathbb{R}^n$. So, we directly substitute Lemma 8 with $\tau := \frac{1}{4}\|\boldsymbol{\lambda}\|_1$, and get $\lambda_{\min}(\boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^\intercal) \geq \frac{3}{4}\|\boldsymbol{\lambda}\|_1$ with probability at least $1 - 2 \cdot 9^m \cdot \exp(-c_2 \cdot \min\{d_2, d_\infty\})$. Putting both of these inequalities together directly gives us Eq. (8) with the desired probability bound, and completes the proof. □

Finally, putting the high probability statements of Lemma 5 and Lemma 6 together completes the proof of Theorem 2.

### A.3 Proof of Theorem 3

By Lemma 1, our task is equivalent to lower-bounding the probability that there exists $i \in [n]$ such that $y_i \boldsymbol{y}_{\setminus i}^\intercal \left(\boldsymbol{Z}_{\setminus i}\boldsymbol{Z}_{\setminus i}^\intercal\right)^{-1} \boldsymbol{Z}_{\setminus i}\boldsymbol{z}_i \geq 1$. This event is the union of $n$ (possibly overlapping) events, and hence its probability is at least the probability of one of the events, say, the first one:

$$\Pr\left(\exists i \in [n] \text{ s.t. } y_i \boldsymbol{y}_{\setminus i}^\intercal \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i}\boldsymbol{z}_i \geq 1\right) \geq \Pr\left(y_1 \boldsymbol{y}_{\setminus 1}^\intercal \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{Z}_{\setminus 1}\boldsymbol{z}_1 \geq 1\right).$$

Because $\boldsymbol{z}_1$ is a standard Gaussian random vector independent of $\boldsymbol{Z}_{\setminus 1}$, the conditional distribution of $y_1 \boldsymbol{y}_{\setminus 1}^\intercal \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{Z}_{\setminus 1}\boldsymbol{z}_1 \mid \boldsymbol{Z}_{\setminus 1}$ is Gaussian with mean zero and variance $\sigma^2 := \|\boldsymbol{Z}_{\setminus 1}^\intercal \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{y}_{\setminus 1}\|_2^2$. Therefore, for any $t > 0$, we have

$$\begin{aligned}
\Pr\left(y_1 \boldsymbol{y}_{\setminus 1}^\intercal \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{Z}_{\setminus 1}\boldsymbol{z}_1 \geq 1\right) &= \mathbb{E}\left[\Pr\left(\sigma g \geq 1 \mid \sigma\right)\right] \quad \text{(where } g \sim \mathrm{N}(0,1),\ g \perp\!\!\!\perp \sigma\text{)} \\
&= \mathbb{E}\left[\Phi\left(-1/\sigma\right)\right] \\
&\geq \mathbb{E}\left[\Phi\left(-1/\sigma\right) \mid \sigma^2 \geq 1/t\right] \Pr\left(\sigma^2 \geq 1/t\right) \\
&\geq \Phi(-\sqrt{t}) \cdot \Pr(\mathcal{E}_1(t)),
\end{aligned}$$

where $\Phi$ is the standard Gaussian cumulative distribution function, and $\mathcal{E}_1(t)$ is the event that

$$\sigma^2 = \boldsymbol{y}_{\setminus 1} \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{Z}_{\setminus 1} \boldsymbol{Z}_{\setminus 1}^\intercal \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{y}_{\setminus 1} = \boldsymbol{y}_{\setminus 1} \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{y}_{\setminus 1} \geq \frac{1}{t}$$

(as in the proofs of Theorem 1 and Theorem 2). We now lower-bound the probability of $\mathcal{E}_1(t)$. Observe that the $(n-1) \times (n-1)$ random matrix $\boldsymbol{K}_{\setminus 1} = \boldsymbol{Z}_{\setminus 1}\boldsymbol{Z}_{\setminus 1}^\intercal$ follows a Wishart distribution with identity scale matrix and $d$ degrees-of-freedom. Moreover, by the rotational symmetry of the standard Gaussian distribution, the random variable $\boldsymbol{y}_{\setminus 1}^\intercal \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{y}_{\setminus 1}$ has the same distribution as that of $(\sqrt{n-1}\boldsymbol{e}_1)^\intercal \boldsymbol{K}_{\setminus 1}^{-1} (\sqrt{n-1}\boldsymbol{e}_1) = (n-1)\boldsymbol{e}_1^\intercal \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{e}_1$. It is known that $1/\boldsymbol{e}_1^\intercal \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{e}_1$ follows a $\chi^2$ distribution with $d - (n-2)$ degrees-of-freedom; we denote its cumulative distribution function by $F_{d-n+2}$. Therefore,

$$\Pr(\mathcal{E}_1(t)) = F_{d-n+2}(t(n-1)).$$

So, we have shown that

$$\Pr\left(y_1 \boldsymbol{y}_{\setminus 1}^\intercal \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{Z}_{\setminus 1}\boldsymbol{z}_1 \geq 1\right) \geq \sup_{t \geq 0} \Phi(-\sqrt{t}) \cdot F_{d-n+2}(t(n-1)).$$

For $t := \frac{d-n+4+2\sqrt{d-n+2}}{n-1}$, we obtain $F_{d-n+2}(t) \geq 1 - 1/e$ by a standard $\chi^2$ tail bound (Laurent and Massart, 2000, Lemma 1). In this case, we obtain

$$\Pr\left(y_1 \boldsymbol{y}_{\setminus 1}^\intercal \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{Z}_{\setminus 1}\boldsymbol{z}_1 \geq 1\right) \geq \Phi\left(-\sqrt{\frac{d-n+4+2\sqrt{d-n+2}}{n-1}}\right) \cdot \left(1 - \frac{1}{e}\right) \tag{9}$$

as claimed. □

# B ANISOTROPIC VERSION OF THEOREM 3

Below, we give a version of Theorem 3 that applies to certain anisotropic settings, depending on some conditions on $\boldsymbol{\lambda}$.

**Theorem 4.** *There are absolute constants $c > 0$ and $c' > 0$ such that the following hold. Let the training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ follow the model from Section 2.2, with $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ being iid standard Gaussian random vectors in $\mathbb{R}^d$, and $y_1, \ldots, y_n \in \{\pm 1\}$ being arbitrary but fixed (i.e., non-random) values. Assume $d > n$ and that there exists $k \in \mathbb{N}$ and $b > 1$ such that $k < (n-1)/c$ and*

$$\frac{\sum_{j=k+1}^{d} \lambda_j}{\lambda_{k+1}} \leq b(n-1)$$

*where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Then the probability that at least one training example is not a support vector is at least*

$$c' \cdot \Phi\left(-\sqrt{\frac{2cb^2(n-1)}{k+1}}\right) \cdot \left(1 - 10e^{-(n-1)/c}\right),$$

*where $\Phi$ is the standard Gaussian cumulative distribution function.*

Note that the probability bound in Theorem 4 is at least a positive constant for sufficiently large $n$ provided that the $(k, b)$ obtained as a function of $\boldsymbol{\lambda}$ satisfy $k + 1 \geq c'' b^2 (n-1)$ for some absolute constant $c'' > 0$.

*Proof.* The proof begins in the same way as in that of Theorem 3. Using the same arguments, we obtain the following lower bound:

$$\begin{aligned}
\Pr\left(\exists i \in [n] \text{ s.t. } y_i \boldsymbol{y}_{\setminus i}^{\mathsf{T}} \boldsymbol{K}_{\setminus i}^{-1} \boldsymbol{Z}_{\setminus i} \boldsymbol{\Lambda} \boldsymbol{z}_i \geq 1\right) &\geq \Pr\left(y_1 \boldsymbol{y}_{\setminus 1}^{\mathsf{T}} \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{Z}_{\setminus 1} \boldsymbol{\Lambda} \boldsymbol{z}_1 \geq 1\right) \\
&\geq \Phi(-\sqrt{t}) \cdot \Pr(\mathcal{E}_1(t))
\end{aligned} \tag{10}$$

where $\mathcal{E}_1(t)$ is the event that

$$\left\| \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus 1}^{\mathsf{T}} \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{y}_{\setminus 1} \right\|_2^2 \geq \frac{1}{t}.$$

We next focus on lower-bounding the probability of $\mathcal{E}_1(t)$. (This part is more involved than in the proof of Theorem 3.) Observe that the (rotationally invariant) distribution of $\boldsymbol{Z}_{\setminus 1}$ is the same as that of $\boldsymbol{Q}\boldsymbol{Z}_{\setminus 1}$, where $\boldsymbol{Q}$ is a uniformly random $(n-1) \times (n-1)$ orthogonal matrix independent of $\boldsymbol{Z}_{\setminus 1}$. Therefore, $\boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus 1}^{\mathsf{T}} \boldsymbol{K}^{-1} \boldsymbol{y}_{\setminus 1}$ has the same distribution as

$$\begin{aligned}
\boldsymbol{\Lambda}(\boldsymbol{Q}\boldsymbol{Z}_{\setminus 1})^{\mathsf{T}} (\boldsymbol{Q}\boldsymbol{Z}_{\setminus 1} \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus 1}^{\mathsf{T}} \boldsymbol{Q}^{\mathsf{T}})^{-1} \boldsymbol{y}_{\setminus 1} &= \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus 1}^{\mathsf{T}} \boldsymbol{Q}^{\mathsf{T}} \boldsymbol{Q} (\boldsymbol{Z}_{\setminus 1} \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus 1}^{\mathsf{T}})^{-1} \boldsymbol{Q}^{\mathsf{T}} \boldsymbol{y}_{\setminus 1} \\
&= \sqrt{n-1} \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus 1}^{\mathsf{T}} \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{u}
\end{aligned}$$

where $\boldsymbol{u} := \boldsymbol{Q}^{\mathsf{T}} \boldsymbol{y}_{\setminus 1} / \sqrt{n-1}$ is a uniformly random unit vector, independent of $\boldsymbol{Z}_{\setminus 1}$. Letting $\boldsymbol{M} := \boldsymbol{\Lambda} \boldsymbol{Z}_{\setminus 1}^{\mathsf{T}} \boldsymbol{K}_{\setminus 1}^{-1}$, we can thus lower-bound the probability of $\mathcal{E}_1(t)$ using

$$\begin{aligned}
\Pr(\mathcal{E}_1(t)) &= \Pr\left(\|\sqrt{n-1}\boldsymbol{M}\boldsymbol{u}\|_2^2 > 1/t\right) \\
&\geq \Pr\left(\|\sqrt{n-1}\boldsymbol{M}\boldsymbol{u}\|_2^2 > 1/t \mid \operatorname{tr}(\boldsymbol{M}^{\mathsf{T}}\boldsymbol{M}) \geq 2/t\right) \cdot \Pr\left(\operatorname{tr}(\boldsymbol{M}^{\mathsf{T}}\boldsymbol{M}) \geq 2/t\right).
\end{aligned} \tag{11}$$

We lower-bound each of the probabilities on the right-hand side of Eq. (11).

We begin with the first probability in Eq. (11), which we handle for arbitrary $t > 0$. By the Paley-Zygmund inequality, we have

$$\Pr\left(\|\sqrt{n-1}\boldsymbol{M}\boldsymbol{u}\|_2^2 > \frac{1}{2}\mathbb{E}\left[\|\sqrt{n-1}\boldsymbol{M}\boldsymbol{u}\|_2^2\right] \mid \boldsymbol{Z}_{\setminus 1}\right) \geq \frac{1}{4} \cdot \frac{\mathbb{E}\left[\|\sqrt{n-1}\boldsymbol{M}\boldsymbol{u}\|_2^2\right]^2}{\mathbb{E}\left[\|\sqrt{n-1}\boldsymbol{M}\boldsymbol{u}\|_2^4\right]}. \tag{12}$$

Since $\sqrt{n-1}\boldsymbol{u}$ is isotropic, we have

$$\mathbb{E}\left[\|\sqrt{n-1}\boldsymbol{M}\boldsymbol{u}\|_2^2 \mid \boldsymbol{Z}_{\setminus 1}\right] = (n-1)\operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^\top\right]\right) = \operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right).$$

Furthermore, by Lemma 9,

$$\mathbb{E}\left[\|\sqrt{n-1}\boldsymbol{M}\boldsymbol{u}\|_2^4 \mid \boldsymbol{Z}_{\setminus 1}\right] \leq C\operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right)^2$$

for some universal constant $C > 0$. Therefore, plugging back into Eq. (12), we obtain

$$\Pr\left(\|\sqrt{n-1}\boldsymbol{M}\boldsymbol{u}\|_2^2 > \frac{1}{2}\operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right) \mid \boldsymbol{Z}_{\setminus 1}\right) \geq \frac{1}{4}\cdot\frac{\operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right)^2}{C\operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right)^2} = \frac{1}{4C}.$$

Thus we also have the following for arbitrary $t > 0$:

$$\Pr\left(\|\sqrt{n-1}\boldsymbol{M}\boldsymbol{u}\|_2^2 > 1/t \mid \operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right) \geq 2/t\right) \geq \frac{1}{4C}. \tag{13}$$

We next consider the second probability in Eq. (11), namely $\Pr\left(\operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right) \geq 2/t\right)$. Recall that we assume there exists $k < (n-1)/c$ and $b > 1$ such that

$$\frac{\sum_{j=k+1}^d \lambda_j}{\lambda_{k+1}} \leq b(n-1). \tag{14}$$

We claim that for $t := \frac{2cb^2(n-1)}{k+1}$,

$$\Pr\left(\operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right) \geq \frac{2}{t}\right) \geq 1 - 10e^{-(n-1)/c}. \tag{15}$$

Indeed, this claim follows from Lemma 16 of (Bartlett et al., 2020), where their matrix $\boldsymbol{C}$ is our matrix $\boldsymbol{M}^\top\boldsymbol{M}$, except our matrix is $(n-1)\times(n-1)$ instead of $n \times n$, and their matrix $\boldsymbol{\Sigma}$ is our matrix $\boldsymbol{\Lambda}$; see the definitions in their Lemma 8. The universal constant $c > 0$ in their lemma is the same as ours, and Eq. (14) is precisely their condition $r_k(\boldsymbol{\Sigma}) < b(n-1)$ (with the same $k$ and $b$). Therefore, the conclusion of their lemma implies, in our notation, that with probability at least $1 - 10e^{-(n-1)/c}$,

$$\operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right) \geq \frac{k+1}{cb^2(n-1)} = \frac{2}{t}.$$

This proves the claimed probability bound.

We conclude from Eq. (10), Eq. (11), Eq. (13), and Eq. (15), that the probability that at least one training example is not a support vector is bounded below by

$$\Phi\left(-\sqrt{\tfrac{2cb^2(n-1)}{k+1}}\right)\cdot\frac{1}{4C}\cdot\left(1 - 10e^{-(n-1)/c}\right)$$

as claimed. $\square$

# C TIGHTNESS OF ARGUMENT IN THEOREM 3

We show below that our bound on $\Pr\left(y_1\boldsymbol{y}_{\setminus 1}^\top\boldsymbol{K}_{\setminus 1}^{-1}\boldsymbol{Z}_{\setminus 1}\boldsymbol{z}_1 \geq 1\right)$ from the proof of Theorem 3 is essentially tight. This means that in order to improve our converse result, we cannot only improve our bound on the aforementioned probability. It seems important to be able to handle simultaneously the conditions corresponding to multiple training examples, which our present arguments do not do. In particular, resolving this gap would require reasoning about whether the indicator random variables, that the conditions are violated, are highly correlated or not. If they are, we should expect the phase transition to happen at $d \sim n$ (as predicted by the converse); if they are not, we should expect the phase transition to happen at $d \sim n\log n$ (as predicted by the upper bound).

Carrying over the notation from the proof above, we have the following upper-bound:

$$\begin{aligned}\Pr\left(y_1 \boldsymbol{y}_{\setminus 1}^\top \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{Z}_{\setminus 1} \boldsymbol{z}_1 \geq 1\right) &= \mathbb{E}\left(\Phi\left(-1/\sigma\right)\right) \\ &\leq \inf_{t>0}\left\{\Phi(-\sqrt{t}) + \Pr\left(\mathcal{E}_1(t)\right)\right\}.\end{aligned}$$

The last step follows by the law of total probability, and noting that $\Phi(-x)$ is a decreasing function in $x$ as well as being bounded above by one. We will bound the second term for a suitable choice of $t$. Recall that $\mathcal{E}_1(t)$ is the event that

$$\sigma^2 = \boldsymbol{y}_{\setminus 1}^\top \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{y}_{\setminus 1} \geq \frac{1}{t}.$$

Observe that $\sigma^2 \leq \frac{n-1}{\lambda_{\min}(\boldsymbol{K}_{\setminus 1})}$, where the $(n-1)\times(n-1)$ random matrix $\boldsymbol{K}_{\setminus 1} = \boldsymbol{Z}_{\setminus 1}\boldsymbol{Z}_{\setminus 1}^\intercal$ follows a Wishart distribution with identity scale matrix and $d$ degrees of freedom. Directly quoting (Vershynin, 2018, Theorem 5.32), we get

$$\Pr\left(\lambda_{\min}(\boldsymbol{K}_{\setminus 1}) \leq (\sqrt{d} - \sqrt{n} - \delta)^2\right) \leq e^{-\delta^2/2}.$$

for any value of $\delta$ such that $0 < \delta < \sqrt{d} - \sqrt{n}$. Therefore,

$$\Pr\left(\sigma^2 \geq \frac{n-1}{(\sqrt{d}-\sqrt{n}-\delta)^2}\right) \leq \Pr\left(\lambda_{\min}(\boldsymbol{K}_{\setminus 1}) \leq (\sqrt{d} - \sqrt{n} - \delta)^2\right) \leq e^{-\delta^2/2}.$$

Assuming $d > 4n$, we set $\delta := \sqrt{n}$ and $t := \frac{(\sqrt{d}-2\sqrt{n})^2}{n-1}$, and obtain

$$\begin{aligned}\Pr\left(y_1 \boldsymbol{y}_{\setminus 1}^\top \boldsymbol{K}_{\setminus 1}^{-1} \boldsymbol{Z}_{\setminus 1} \boldsymbol{z}_1 \geq 1\right) &\leq \Phi(-\sqrt{t}) + \Pr(\mathcal{E}_1(t)) \\ &\leq \Phi\left(-\frac{\sqrt{d}-2\sqrt{n}}{\sqrt{n-1}}\right) + e^{-n/2},\end{aligned}$$

which can be directly compared to Eq. (9).

# D PROBABILISTIC INEQUALITIES

**Lemma 7.** *Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a symmetric matrix, and let $\mathcal{N}$ be an $\epsilon$-net of $S^{n-1}$ with respect to the Euclidean metric for some $\epsilon < 1/2$, Then*

$$\|\boldsymbol{M}\|_2 \leq \frac{1}{1-2\epsilon} \max_{\boldsymbol{u}\in\mathcal{N}} |\boldsymbol{u}^\intercal \boldsymbol{M} \boldsymbol{u}|.$$

*Proof.* See (Vershynin, 2010, Lemma 5.4). □

**Lemma 8.** *There is a universal constant $c > 0$ such that the following holds. Let $\lambda_1, \dots, \lambda_d > 0$ be given. Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_d$ be independent random vectors taking values in $\mathbb{R}^n$ such that, for some $v > 0$,*

$$\mathbb{E}(\boldsymbol{v}_j) = \boldsymbol{0}, \qquad \mathbb{E}(\boldsymbol{v}_j\boldsymbol{v}_j^\intercal) = \boldsymbol{I}_n, \qquad \mathbb{E}(\exp(\boldsymbol{u}^\intercal \boldsymbol{v}_j)) \leq \exp(v\|\boldsymbol{u}\|_2^2/2) \quad \text{for all } \boldsymbol{u} \in \mathbb{R}^n$$

*for all $j = 1, \dots, d$. For any $\tau > 0$,*

$$\Pr\left(\left\|\sum_{j=1}^d \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\intercal - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}_n\right\|_2 \geq \tau\right) \leq 2\cdot 9^n \cdot \exp\left(-c\cdot\min\left\{\frac{\tau^2}{v^2\|\boldsymbol{\lambda}\|_2^2}, \frac{\tau}{v\|\boldsymbol{\lambda}\|_\infty}\right\}\right).$$

*where $\|\boldsymbol{\lambda}\|_1 := \sum_{j=1}^d \lambda_j$, $\|\boldsymbol{\lambda}\|_2^2 := \sum_{j=1}^d \lambda_j^2$, and $\|\boldsymbol{\lambda}\|_\infty := \max_{j\in[d]} \lambda_j$.*

*Proof.* Let $\mathcal{N}$ be an (1/4)-net of $S^{n-1}$ with respect to the Euclidean metric. A standard volume argument of Pisier (1999) allows a choice of $\mathcal{N}$ with $|\mathcal{N}| \leq 9^n$. By Lemma 7, we have for any $t > 0$,

$$\Pr\left(\left\|\sum_{j=1}^{d} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^{\top} - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}_n\right\|_2 \geq \tau\right) \leq \Pr\left(\max_{\boldsymbol{u}\in\mathcal{N}}\left|\sum_{j=1}^{d} \lambda_j (\boldsymbol{u}^{\top}\boldsymbol{v}_j)^2 - \|\boldsymbol{\lambda}\|_1\right| \geq \tau/2\right).$$

Next, observe that for any $\boldsymbol{u} \in S^{n-1}$, the random variables $\boldsymbol{u}^{\top}\boldsymbol{v}_1, \dots, \boldsymbol{u}^{\top}\boldsymbol{v}_d$ are independent random variables, each with mean-zero, unit variance, and sub-Gaussian with parameter $v$. By the Hanson-Wright inequality of Rudelson and Vershynin (2013) and a union bound, there exists a universal constant $c > 0$ such that, for any unit vector $\boldsymbol{u} \in S^{n-1}$ and any $\tau > 0$,

$$\Pr\left(\max_{\boldsymbol{u}\in\mathcal{N}}\left|\sum_{j=1}^{d} \lambda_j (\boldsymbol{u}^{\top}\boldsymbol{v}_j)^2 - \|\boldsymbol{\lambda}\|_1\right|_2 \geq \tau/2\right) \leq 2 \cdot 9^n \cdot \exp\left(-c \cdot \min\left\{\frac{\tau^2}{v^2\|\boldsymbol{\lambda}\|_2^2}, \frac{\tau}{v\|\boldsymbol{\lambda}\|_\infty}\right\}\right).$$

The claim follows. ∎

**Lemma 9.** *Let $\boldsymbol{\theta}$ be a uniformly random unit vector in $S^{m-1}$. For any unit vector $\boldsymbol{u} \in S^{m-1}$, the random variable $\boldsymbol{u}^{\top}\boldsymbol{\theta}$ is sub-Gaussian with parameter $v = O(1/m)$. Moreover, for any matrix $\boldsymbol{M} \in \mathbb{R}^{m\times m}$, we have*

$$\mathbb{E}\left[\|\boldsymbol{M}\boldsymbol{\theta}\|_2^4\right] \leq \frac{C}{m^2} \operatorname{tr}\left(\boldsymbol{M}^{\top}\boldsymbol{M}\right)^2$$

*where $C > 0$ is a universal constant.*

*Proof.* Let $L$ be a $\chi$ random variable with $m$ degrees-of-freedom, independent of $\boldsymbol{\theta}$, so the distribution of $\boldsymbol{z} := L\boldsymbol{\theta}$ is the standard Gaussian in $\mathbb{R}^m$. Let $\mu := \mathbb{E}[L] = \mathbb{E}[L \mid \boldsymbol{\theta}] = \sqrt{2}\frac{\Gamma((m+1)/2)}{\Gamma(m/2)} = \Omega(\sqrt{m})$. By Jensen's inequality, for any $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[\exp(t\boldsymbol{u}^{\top}\boldsymbol{\theta})\right] &= \mathbb{E}\left[\exp\left(\left(\frac{t}{\mu}\boldsymbol{u}\right)^{\top}\left(\mathbb{E}[L \mid \boldsymbol{\theta}]\boldsymbol{\theta}\right)\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\left(\frac{t}{\mu}\boldsymbol{u}\right)^{\top}(L\boldsymbol{\theta})\right)\right] \\
&= \mathbb{E}\left[\exp\left(\left(\frac{t}{\mu}\boldsymbol{u}\right)^{\top}\boldsymbol{z}\right)\right] \\
&= \exp\left(\frac{t^2}{2\mu^2}\right).
\end{aligned}$$

It follows that $\boldsymbol{u}^{\top}\boldsymbol{\theta}$ is sub-Gaussian with parameter $v = 1/\mu^2 = O(1/m)$.

Similarly, again by Jensen's inequality,

$$\begin{aligned}
\mu^4 \cdot \mathbb{E}\left[\|\boldsymbol{M}\boldsymbol{\theta}\|_2^4\right] &= \mathbb{E}\left[\mathbb{E}[L \mid \boldsymbol{\theta}]^4\|\boldsymbol{M}\boldsymbol{\theta}\|_2^4\right] \\
&\leq \mathbb{E}\left[L^4\|\boldsymbol{M}\boldsymbol{\theta}\|_2^4\right] \\
&= \mathbb{E}\left[\|\boldsymbol{M}\boldsymbol{z}\|_2^4\right].
\end{aligned}$$

Furthermore, a direct computation shows that

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{M}\boldsymbol{z}\|_2^4\right] &= 2\operatorname{tr}\left((\boldsymbol{M}^{\top}\boldsymbol{M})^2\right) + \operatorname{tr}\left(\boldsymbol{M}^{\top}\boldsymbol{M}\right)^2 \\
&\leq 3\operatorname{tr}\left(\boldsymbol{M}^{\top}\boldsymbol{M}\right)^2.
\end{aligned}$$

The conclusion follows since $\mu^4 = \Omega(m^2)$. ∎