A PROOFS OF MAIN THEOREMS

A.1 Proof of Theorem 1

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

1. For $i \in [n]$, $B_i$ is the event that $K_{\setminus i}$ is non-singular and
   \[ y_i y_i^T K_{\setminus i}^{-1} Z_i^T \Lambda z_i \geq 1. \]
2. For $i \in [n]$, $S_i$ is the event that $K_{\setminus i}$ is singular.
3. $S$ is the event that $K$ is singular.
4. $B := S \cup \bigcup_{i=1}^n (B_i \cup S_i)$.

Additionally, we define the event $E_i(t)$, for every $i \in [n]$ and a given $t > 0$, that $K_{\setminus i}$ is non-singular and
\[ \|\Lambda Z_i^T K_{\setminus i}^{-1} y_i\|_2^2 \geq \frac{1}{t}. \]

Note that if the event $B$ does not occur, then $Z \Lambda Z^T$ is non-singular, each $K_{\setminus i}$ is non-singular, and
\[ y_i y_i^T y_i^T \Lambda z_i < 1, \quad \text{for all } i = 1, \ldots, n. \]

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

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

\[
Pr(B) \leq Pr(S) + \sum_{i=1}^n Pr(B_i \cup S_i)
\]
\[
= Pr(S) + \sum_{i=1}^n \left( Pr((B_i \cap S_i^c \cap E_i(t)^c) \cup (S_i \cap E_i(t)^c)) + Pr((B_i \cup S_i) \cap E_i(t)) \right)
\]
\[
\leq Pr(S) + \sum_{i=1}^n \left( Pr(B_i \mid S_i^c \cap E_i(t)^c) Pr(S_i^c \cap E_i(t)^c) + Pr(S_i \cap E_i(t)^c) + Pr((B_i \cup S_i) \cap E_i(t)) \right)
\]
\[
\leq Pr(S) + \sum_{i=1}^n \left( Pr(B_i \mid S_i^c \cap E_i(t)^c) + Pr(S_i) + Pr(E_i(t)) \right). \quad (7)
\]

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
\[
\max\{Pr(S), Pr(S_1), \ldots, Pr(S_n)\} \leq 2 \cdot 9^n \cdot \exp\left( -c \cdot \min \left\{ \frac{d_2}{v^2}, \frac{d_\infty}{v} \right\} \right)
\]
where $c > 0$ is the universal constant in the statement of Lemma 8.

**Proof.** It suffices to bound $Pr(S)$, since each $K_{\setminus i}$ is a principal submatrix of $K$, and hence $\lambda_{\min}(K_{\setminus i}) \geq \lambda_{\min}(K)$ for all $i \in [n]$. Observe that
\[
Z \Lambda Z^T = \sum_{j=1}^d \lambda_j v_j v_j^T
\]
where \( v_j \) is the \( j \)th column of \( Z \). Recall that the columns of \( Z \) are independent, and so these vectors satisfy the conditions of Lemma 8. Moreover, since \( ZZ^T \) is positive semi-definite, its singularity would require

\[
\|ZZ^T − \|\lambda\|_1 I\|_2 ≥ \|\lambda\|_1.
\]

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

The next lemma upper bounds the probability of the event \( B_i \) conditioned on the non-singularity event \( S_i \) and the complement of the event \( E_i(t) \).

**Lemma 3.** For any \( t > 0 \),

\[
\Pr(B_i \mid S_i \cap E_i(t)^c) ≤ 2 \exp \left( \frac{-t}{2v} \right).
\]

**Proof.** Let \( B_i^\prime \) be the event that \( K_{\setminus i} \) is non-singular and

\[
|y_i^\dagger K_{\setminus i}^{-1} Z_{\setminus i} \lambda z_i| = \max \{ −y_i^\dagger K_{\setminus i}^{-1} Z_{\setminus i} \Lambda z_i, y_i^\dagger K_{\setminus i}^{-1} Z_{\setminus i} \Lambda z_i \} ≥ 1.
\]

Since \( |y_i| = 1 \), it follows that \( B_i \subseteq B_i^\prime \), so

\[
\Pr(B_i \mid S_i \cap E_i(t)^c) ≤ \Pr(B_i^\prime \mid S_i \cap E_i(t)^c).
\]

Conditional on the event \( S_i^\prime \cap E_i(t)^c \), we have that \( K_{\setminus i} \) is non-singular and \( \|\Lambda Z_{\setminus i}^\dagger K_{\setminus i}^{-1} y_{\setminus i}\|_2^2 ≤ 1/t \). Since \( z_i \) is independent of \( \{z_j, y_j \} : j \neq i \), it follows that

\[
y_i^\dagger K_{\setminus i}^{-1} Z_{\setminus i} \Lambda z_i = (\Lambda Z_{\setminus i}^\dagger K_{\setminus i}^{-1} y_{\setminus i})^\dagger z_i
\]

is (conditionally) sub-Gaussian with parameter at most \( v \cdot \|\Lambda Z_{\setminus i}^\dagger K_{\setminus i}^{-1} y_{\setminus i}\|_2^2 ≤ v/t \). Then, the standard sub-Gaussian tail bound gives us

\[
\Pr(B_i \mid S_i \cap E_i(t)^c) ≤ \Pr(B_i^\prime \mid S_i \cap E_i(t)^c) ≤ 2 \exp \left( \frac{-t}{2v} \right).
\]

This completes the proof of the lemma.

Finally, the following lemma upper bounds the probability of the event \( E_i(t) \) for \( t := d_∞/2n \).

**Lemma 4.**

\[
\Pr(E_i(d_∞/(2n))) ≤ 2 \cdot 9^{n−1} \cdot \exp \left( −c \cdot \min \left\{ \frac{d_2}{4v^2}, \frac{d_∞}{v} \right\} \right)
\]

where \( c > 0 \) is the universal constant from Lemma 8.

**Proof.** Let \( E_i^\prime(t) \) be the event that

\[
\lambda_{\min}(K_{\setminus i}) ≤ n\|\lambda\|_∞ t.
\]

Under \( S_i^\prime \), the matrix \( K_{\setminus i} \) is non-singular. We get

\[
\|\Lambda Z_{\setminus i}^\dagger K_{\setminus i}^{-1} y_{\setminus i}\|_2^2 ≤ \|\Lambda^{1/2}\|_\infty \|\Lambda^{1/2} Z_{\setminus i}^\dagger K_{\setminus i}^{-1} y_{\setminus i}\|_2^2
\]

\[
= \|\lambda\|_\infty y_i^\dagger K_{\setminus i}^{-1} Z_{\setminus i} \Lambda Z_{\setminus i}^\dagger K_{\setminus i}^{-1} y_{\setminus i}
\]

\[
≤ n\|\lambda\|_\infty \sup_{u \in \mathbb{R}^{n−1}, \|u\|_2 = 1} u^\dagger K_{\setminus i}^{-1} u
\]

\[
= \frac{n\|\lambda\|_\infty}{\lambda_{\min}(K_{\setminus i})}.
\]
It follows that $E_i(t) \subseteq E_i'(t)$. Observe that for $t := d_\infty/(2n)$, the event $E_i'(t)$ is that where

$$\lambda_{\min}(K_{-i}) \leq \frac{1}{2} \|\lambda\|_1.$$ 

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

$$\Pr(E_i'(d_\infty/(2n))) = \Pr\left(\lambda_{\min}(K_{-i}) \leq \frac{1}{2} \|\lambda\|_1\right) \leq 2 \cdot 9^{n-1} \cdot \exp\left(-c \cdot \min\left\{\frac{d_2}{4n^2}, \frac{d_\infty}{v}\right\}\right).$$ 

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 $\frac{1}{\sqrt{d}} z_i | (Z_{-i}, y_{-i})$ is a uniformly random unit vector in $S^{d-1}$ restricted to the subspace orthogonal to the row space of $Z_{-i}$. That is, it has the same (conditional) distribution as $B_i u_i$, where:

1. $B_i$ is a $d \times (n - d + 1)$ matrix whose columns form an orthonormal basis for the orthogonal complement of $Z_{-i}$’s row space;
2. $u_i$ is a uniformly random unit vector in $S^{d-n}$.

As before, for every $i \in [n]$, we define the event $B_i$ that $K_{-i}$ is non-singular and

$$y_i y_i^T K_{-i}^{-1} Z_{-i} \Lambda z_i \geq 1.$$ 

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

$$\|B_i^T \Lambda Z_{-i} K_{-i}^{-1} y_{-i}\|_2^2 \geq \frac{d-n+1}{d} \cdot \frac{1}{t}.$$ 

By the union bound, we get

$$\Pr(B) \leq \sum_{i=1}^n \Pr(B_i) \leq \sum_{i=1}^n \Pr(B_i \mid E_i(t)^c) + \Pr(E_i(t)),$$

and so we need to upper bound the probabilities $\Pr(B_i \mid E_i(t)^c)$ and $\Pr(E_i(t))$ for every $i \in [n]$.

The following lemma upper bounds $\Pr(B_i \mid E_i(t)^c)$, and is analogous to Lemma 3 in the proof of Theorem 1.

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

$$\Pr(B_i \mid E_i(t)^c) \leq 2 \exp(-t).$$
On the proliferation of support vectors in high dimensions

Proof. First, as discussed above, we have
\[
\Pr \left( y_i y_i^T K_{\setminus i}^{-1} Z_{\setminus i} \Lambda z_i \geq 1 \right) = \Pr \left( \sqrt{d} \cdot y_i y_i^T K_{\setminus i}^{-1} Z_{\setminus i} \Lambda B_i u_i \geq 1 \right)
\]
\[
\leq \Pr \left( \sqrt{d} \cdot \| B_i^T \Lambda Z_{\setminus i} K_{\setminus i}^{-1} y_{\setminus i} \|^2 \cdot u_i \| \geq 1 \right).
\]
Moreover, \( u_i \) is independent of \( Z_{\setminus i} \), and as established in Lemma 9, the random vector \( u_i \) is sub-Gaussian with parameter at most \( O(1/(d - n + 1)) \). Therefore, \( \sqrt{d} \cdot (B_i^T \Lambda Z_{\setminus i} K_{\setminus i}^{-1} y_{\setminus i}) u_i \) is conditionally sub-Gaussian with parameter at most \( \frac{d}{d - n + 1} \cdot \| B_i^T \Lambda Z_{\setminus i} K_{\setminus i}^{-1} y_{\setminus i} \|^2 \leq \frac{1}{2} \). Here, the last inequality follows because we have conditioned on \( \mathcal{E}_i(t)^c \). Therefore, the standard sub-Gaussian tail bound gives us
\[
\Pr \left( B_i \mid \mathcal{E}_i(t)^c \right) \leq 2 \exp \left( -t \right).
\]

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
\[
\| B_i^T \Lambda Z_{\setminus i} K_{\setminus i}^{-1} y_{\setminus i} \|^2 \leq \| B_i^T \|_2^2 \cdot \| \Lambda Z_{\setminus i} K_{\setminus i}^{-1} y_{\setminus i} \|^2
\]
\[
= \| \Lambda Z_{\setminus i} K_{\setminus i}^{-1} y_{\setminus i} \|^2
\]
\[
\leq \frac{n \| \Lambda \|_{\infty}}{\lambda_{\min}(K_{\setminus i})},
\]
where we used the fact that \( B_i \) has orthonormal columns, and the last inequality follows by an identical argument to the proof of Lemma 4. We will show in particular that
\[
\Pr \left( \lambda_{\min}(K_{\setminus i}) \geq \frac{1}{2} \| \Lambda \|_1 \right) \geq 1 - \exp \left( -c_1 \cdot d \right) - 2 \cdot 9^n \cdot \exp \left( -c_2 \cdot \min \{d_2, d_{\infty}\} \right).
\]
Given Eq. (8), we can complete the proof of Lemma 6. This is because we get
\[
\| B_i^T \Lambda Z_{\setminus i} K_{\setminus i}^{-1} y_{\setminus i} \|^2 \leq \frac{2n \| \Lambda \|_{\infty}}{\| \Lambda \|_1} = \frac{2n}{d_{\infty}} = \frac{d - n + 1}{d} \cdot \frac{1}{t}
\]
for
\[
t := \frac{d - n + 1}{d} \cdot \frac{d_{\infty}}{2n}.
\]
We complete the proof by proving Eq. (8). Let \( S \in \mathbb{R}^{m \times d} \) be a random matrix with iid standard Gaussian entries with \( m := n - 1 \), and let the singular value decomposition of \( S \) be \( S = V \Lambda_S U^T \) where \( V \in \mathbb{R}^{m \times m} \) and \( U \in \mathbb{R}^{d \times m} \) are orthonormal matrices. Then, it is well-known that \( \sqrt{d} \cdot U^T \) follows the same distribution as \( Z_{\setminus i} \), and hence \( \lambda_{\min}(K_{\setminus i}) \) has the same distribution as \( d \cdot \lambda_{\min}(U^T \Lambda U) \). Moreover,
\[
d \cdot \lambda_{\min}(U^T \Lambda U) = \min_{v \in \mathbb{R}^m, \| v \|_2 = 1} v^T \Lambda_S^{-1} V^T V A_S U^T A_S V^T V \Lambda_S^{-1} v
\]
\[
\geq \frac{d}{\| \Lambda_S \|_{\text{op}}} \cdot \min_{v \in \mathbb{R}^m, \| v \|_2 = 1} v^T S A S^T v
\]
\[
= \frac{d}{\| \Lambda_S \|_{\text{op}}} \cdot \lambda_{\min}(S A S^T).
\]
By classical operator norm tail bounds on Gaussian random matrices (e.g., Vershynin, 2010, Corollary 5.35), we note that \( \|A_S\|_2^2 \leq \frac{3}{2} d \) with probability at least \( 1 - \exp(-c_1 \cdot d) \). Now, we note that the matrix \( SAS^T := \sum_{j=1}^d \lambda_j s_j s_j^T \) where the \( s_j \)'s are iid standard Gaussian random vectors in \( \mathbb{R}^n \). So, we directly substitute Lemma 8 with \( \tau := \frac{1}{4} \|\lambda\|_1 \), and get \( \lambda_{\min}(SAS^T) \geq \frac{3}{4} \|\lambda\|_1 \) 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

\[
y_i \left( \frac{Z_i}{Z_i^\top Z_i} \right)^{-1} Z_i z_i \geq 1.
\]

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:

\[
\Pr \left( \exists i \in [n] \text{ s.t. } y_i^\top \left( \frac{Z_i}{Z_i^\top Z_i} \right)^{-1} Z_i z_i \geq 1 \right) \geq \Pr \left( y_1 \left( \frac{Z_1}{Z_1^\top Z_1} \right)^{-1} Z_1 z_1 \geq 1 \right).
\]

Because \( z_1 \) is a standard Gaussian random vector independent of \( Z \setminus 1 \), the conditional distribution of

\[
y_1 \left( \frac{Z_1}{Z_1^\top Z_1} \right)^{-1} Z_1 z_1 \mid \frac{Z_1}{Z_1^\top Z_1}
\]

is Gaussian with mean zero and variance \( \sigma^2 := \|\frac{Z_1}{Z_1^\top Z_1} y_1\|_2^2 \). Therefore, for any \( t > 0 \), we have

\[
\Pr \left( y_1 \left( \frac{Z_1}{Z_1^\top Z_1} \right)^{-1} Z_1 z_1 \geq 1 \right) = \mathbb{E} \left[ \Pr \left( \sigma g \geq 1 \mid \sigma \right) \right] = \mathbb{E} \left[ \Phi \left( -1/\sigma \right) \right] \geq \mathbb{E} \left[ \Phi \left( -1/\sigma \right) \mid \sigma^2 \geq 1/t \right] \Pr \left( \sigma^2 \geq 1/t \right) \geq \Phi(-\sqrt{t}) \cdot \Pr(\mathcal{E}_1(t)),
\]

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

\[
\sigma^2 = y_1^\top \left( \frac{Z_1}{Z_1^\top Z_1} \right)^{-1} Z_1 y_1 = y_1^\top \frac{Z_1}{Z_1^\top Z_1} y_1 \geq \frac{1}{t}
\]

(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 \( K_{\setminus 1} = \frac{Z_1}{Z_1^\top Z_1} \) 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 \( y_1^\top K_{\setminus 1}^{-1} y_1 \) has the same distribution as that of \( (\sqrt{n - 1} e_1)^\top \frac{Z_1}{Z_1^\top Z_1} (\sqrt{n - 1} e_1) = (n - 1)e_1^\top K_{\setminus 1}^{-1} e_1 \). It is known that \( 1/e_1^\top K_{\setminus 1}^{-1} e_1 \) 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

\[
\Pr \left( y_1 \left( \frac{Z_1}{Z_1^\top Z_1} \right)^{-1} Z_1 z_1 \geq 1 \right) \geq \sup_{t \geq 0} \Phi(-\sqrt{t}) \cdot F_{d-n+2}(t(n-1)).
\]

For \( t := \frac{d-n+2 + 2\sqrt{d-n+2}}{n-1} \), 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

\[
\Pr \left( y_1 \left( \frac{Z_1}{Z_1^\top Z_1} \right)^{-1} Z_1 z_1 \geq 1 \right) \geq \Phi \left( -\sqrt{\frac{d-n+2 + 2\sqrt{d-n+2}}{n-1}} \right) \cdot \left( 1 - \frac{1}{e} \right)
\]

as claimed.
On the proliferation of support vectors in high dimensions

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 \( \lambda \).

**Theorem 4.** There are absolute constants \( c > 0 \) and \( c' > 0 \) such that the following hold. Let the training data \((x_1, y_1), \ldots, (x_n, y_n)\) follow the model from Section 2.2, with \( z_1, \ldots, z_n \) 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

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

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 \( \lambda \) 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:

\[
\Pr \left( \exists i \in [n] \text{ s.t. } y_i y_i^T K_{\setminus i}^{-1} Z_{\setminus i} \Lambda z_i \geq 1 \right) \geq \Pr \left( y_i y_i^T K_{\setminus i}^{-1} Z_{\setminus i} \Lambda z_i \geq 1 \right)
\]

\[
\geq \Phi(-\sqrt{t}) \cdot \Pr(\mathcal{E}_1(t))
\]

(10)

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

\[
\|\Lambda Z_{\setminus 1}^T K_{\setminus 1}^{-1} y_{\setminus 1}\|_2^2 \geq \frac{1}{t}.
\]

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 \( Z_{\setminus 1} \) is the same as that of \( QZ_{\setminus 1} \), where \( Q \) is a uniformly random \((n-1) \times (n-1)\) orthogonal matrix independent of \( Z_{\setminus 1} \). Therefore, \( \Lambda Z_{\setminus 1}^T K_{\setminus 1}^{-1} y_{\setminus 1} \) has the same distribution as

\[
\Lambda (QZ_{\setminus 1})^T (QZ_{\setminus 1} \Lambda Z_{\setminus 1}^T Q^T)^{-1} y_{\setminus 1} = \Lambda Z_{\setminus 1}^T Q^T Q (Z_{\setminus 1} \Lambda Z_{\setminus 1})^{-1} Q^T y_{\setminus 1}
\]

\[
= \sqrt{n-1} \Lambda Z_{\setminus 1}^T K_{\setminus 1}^{-1} u
\]

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

\[
\Pr(\mathcal{E}_1(t)) = \Pr \left( \|\sqrt{n-1} M u\|_2^2 > 1/t \right)
\]

\[
\geq \Pr \left( \|\sqrt{n-1} M u\|_2^2 > 1/t \mid \text{tr} \left( M^T M \right) \geq 2/t \right) \cdot \Pr \left( \text{tr} \left( M^T M \right) \geq 2/t \right).
\]

(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

\[
\Pr \left( \|\sqrt{n-1} M u\|_2^2 > \frac{1}{2} E \left[ \|\sqrt{n-1} M u\|_2^2 \right] \mid Z_{\setminus 1} \right) \geq \frac{1}{4} \cdot \frac{E \left[ \|\sqrt{n-1} M u\|_2^2 \right]^2}{E \left[ \|\sqrt{n-1} M u\|_2^4 \right]}. \]

(12)
Since $\sqrt{n-1}u$ is isotropic, we have
\[
E \left[ \|\sqrt{n-1}M u\|_2^2 \mid Z_1 \right] = (n-1) \text{tr} (M^T M \mathbb{E}[u u^T]) = \text{tr} (M^T M).
\]
Furthermore, by Lemma 9,
\[
E \left[ \|\sqrt{n-1}M u\|_2^4 \mid Z_1 \right] \leq C \text{tr} (M^T M)^2
\]
for some universal constant $C > 0$. Therefore, plugging back into Eq. (12), we obtain
\[
\Pr \left( \|\sqrt{n-1}M u\|_2^2 > \frac{1}{2} \text{tr} (M^T M) \mid Z_1 \right) \geq \frac{1}{4} \cdot \frac{\text{tr} (M^T M)^2}{C \text{tr} (M^T M)^2} = \frac{1}{4C}.
\]
Thus we also have the following for arbitrary $t > 0$:
\[
\Pr \left( \|\sqrt{n-1}M u\|_2^2 > \frac{1}{t} \mid \text{tr} (M^T M) \geq 2/t \right) \geq \frac{1}{4C}.
\] (13)

We next consider the second probability in Eq. (11), namely $\Pr (\text{tr} (M^T M) \geq 2/t)$.

Recall that we assume there exists $k < (n-1)/c$ and $b > 1$ such that
\[
\sum_{j=k+1}^d \lambda_j \leq b(n-1).
\] (14)
We claim that for $t := \frac{2ck^2(n-1)}{k+1}$,
\[
\Pr \left( \text{tr} (M^T M) \geq \frac{2}{t} \right) \geq 1 - 10e^{-(n-1)/c}.
\] (15)

Indeed, this claim follows from Lemma 16 of (Bartlett et al., 2020), where their matrix $C$ is our matrix $M^T M$, except our matrix is $(n-1) \times (n-1)$ instead of $n \times n$, and their matrix $\Sigma$ is our matrix $\Lambda$; 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 $r_k(\Sigma) < b(n-1)$ (with the same $k$ and $b$). Therefore, the conclusion of their lemma implies, in our notation, that with probability at least $1 - 10e^{-(n-1)/c}$,
\[
\text{tr} (M^T M) \geq \frac{k+1}{c b^2 (n-1)} = \frac{2}{t}.
\]
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
\[
\Phi \left( -\sqrt{\frac{2ck^2(n-1)}{k+1}} \right) \cdot \frac{1}{4C} \cdot \left( 1 - 10e^{-(n-1)/c} \right)
\]
as claimed.

\section{C TIGHTNESS OF ARGUMENT IN THEOREM 3}

We show below that our bound on $\Pr (y_1 y_1^T K_{\{1\}}^{-1} Z_{\{1\}} z_1 \geq 1)$ 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:

\[
\Pr \left( \mathbf{y}_1 \mathbf{y}_1^\top \mathbf{K}_{1:1}^{-1} \mathbf{Z}_{1:1} \mathbf{z}_1 \geq 1 \right) = \mathbb{E} \left( \Phi \left( -\frac{1}{\sigma} \right) \right)
\]

\[
\leq \inf_{t > 0} \left\{ \Phi(-\sqrt{t}) + \Pr (\mathcal{E}_1(t)) \right\}.
\]

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

\[
\sigma^2 = \mathbf{y}_1^\top \mathbf{K}_{1:1}^{-1} \mathbf{y}_1 \geq \frac{1}{t}.
\]

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

\[
\Pr \left( \lambda_{\min}(\mathbf{K}_{1:1}) \leq (\sqrt{d} - \sqrt{n} - \delta)^2 \right) \leq e^{-\delta^2/2}.
\]

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

\[
\Pr \left( \sigma^2 \geq \frac{n-1}{(\sqrt{d} - \sqrt{n} - \delta)^2} \right) \leq \Pr \left( \lambda_{\min}(\mathbf{K}_{1:1}) \leq (\sqrt{d} - \sqrt{n} - \delta)^2 \right) \leq e^{-\delta^2/2}.
\]

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

\[
\Pr \left( \mathbf{y}_1 \mathbf{y}_1^\top \mathbf{K}_{1:1}^{-1} \mathbf{Z}_{1:1} \mathbf{z}_1 \geq 1 \right) \leq \Phi(-\sqrt{t}) + \Pr (\mathcal{E}_1(t))
\]

\[
\leq \Phi \left( -\frac{\sqrt{d} - 2\sqrt{n}}{\sqrt{n-1}} \right) + e^{-n/2},
\]

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

\[\tag{D} \text{PROBABILISTIC INEQUALITIES} \]

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

\[
\|\mathbf{M}\|_2 \leq \frac{1}{1 - 2\epsilon} \max_{\mathbf{u} \in \mathcal{N}} |\mathbf{u}^\top \mathbf{M} \mathbf{u}|.
\]

\[\text{Proof.} \text{ See (Vershynin, 2010, Lemma 5.4).} \]

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

\[
\mathbb{E}(\mathbf{v}_j) = \mathbf{0}, \quad \mathbb{E}(\mathbf{v}_j \mathbf{v}_j^\top) = \mathbf{I}_n, \quad \mathbb{E}(\exp(\mathbf{u}^\top \mathbf{v}_j)) \leq \exp(v\|\mathbf{u}\|_2^2/2) \quad \text{for all } \mathbf{u} \in \mathbb{R}^n
\]

for all \(j = 1, \ldots, d\). For any \(\tau > 0\),

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

where \(\|\mathbf{\lambda}\|_1 := \sum_{j=1}^d \lambda_j, \|\mathbf{\lambda}\|_2^2 := \sum_{j=1}^d \lambda_j^2 , \text{ and } \|\mathbf{\lambda}\|_\infty := \max_{j \in [d]} \lambda_j. \)
Proof. Let $N$ 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 $N$ with $|N| \leq 9^n$. By Lemma 7, we have for any $t > 0$,
\[
\Pr \left( \left\| \sum_{j=1}^{d} \lambda_j v_j v_j^\top - \|\lambda\|_1 I_n \right\|_2 \geq \tau \right) \leq \Pr \left( \max_{u \in N} \left\| \sum_{j=1}^{d} \lambda_j (u^\top v_j)^2 - \|\lambda\|_1 \right\|_2 \geq \tau/2 \right).
\]

Next, observe that for any $u \in S^{n-1}$, the random variables $u^\top v_1, \ldots, u^\top v_d$ 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 $u \in S^{n-1}$ and any $\tau > 0$,
\[
\Pr \left( \max_{u \in N} \left\| \sum_{j=1}^{d} \lambda_j (u^\top v_j)^2 - \|\lambda\|_1 \right\|_2 \geq \tau/2 \right) \leq 2 \cdot 9^n \cdot \exp \left( -c \cdot \min \left\{ \frac{\tau^2}{v^2\|\lambda\|_2^2}, \frac{\tau}{v\|\lambda\|_\infty} \right\} \right).
\]

The claim follows. \qed

**Lemma 9.** Let $\theta$ be a uniformly random unit vector in $S^{m-1}$. For any unit vector $u \in S^{m-1}$, the random variable $u^\top \theta$ is sub-Gaussian with parameter $v = O(1/m)$. Moreover, for any matrix $M \in \mathbb{R}^{m \times m}$, we have
\[
\mathbb{E} \left[ \|M\theta\|_2^4 \right] \leq \frac{C}{m^2} \text{tr} (M^\top M)^2
\]
where $C > 0$ is a universal constant.

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

It follows that $u^\top \theta$ is sub-Gaussian with parameter $v = 1/\mu^2 = O(1/m)$. Similarly, again by Jensen’s inequality,
\[
\mu^4 \cdot \mathbb{E} \left[ \|M\theta\|_2^4 \right] = \mathbb{E} \left[ \mathbb{E}[L \mid \theta]^4 \|M\theta\|_2^4 \right]
\]
\[
\leq \mathbb{E} \left[ L^4 \|M\theta\|_2^4 \right]
\]
\[
= \mathbb{E} \left[ \|Mz\|_2^4 \right].
\]

Furthermore, a direct computation shows that
\[
\mathbb{E} \left[ \|Mz\|_2^4 \right] = 2 \text{tr} ((M^\top M)^2) + \text{tr} (M^\top M)^2
\]
\[
\leq 3 \text{tr} (M^\top M)^2.
\]

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