Proceedings of Machine Learning Research vol 99:1–26, 2019 32nd Annual Conference on Learning Theory

# Space lower bounds for linear prediction in the streaming model

**Yuval Dagan** DAGAN@MIT.EDU
*Massachusetts Institute of Technology, CSAIL*

**Gil Kur** GILKUR@MIT.EDU
*Massachusetts Institute of Technology, CSAIL*

**Ohad Shamir** OHAD.SHAMIR@WEIZMANN.AC.IL
*Department of Computer Science and Applied Mathematics, Weizmann University.*

**Editors:** Alina Beygelzimer and Daniel Hsu

## Abstract

We show that fundamental learning tasks, such as finding an approximate linear separator or linear regression, require memory at least *quadratic* in the dimension, in a natural streaming setting. This implies that such problems cannot be solved (at least in this setting) by scalable memory-efficient streaming algorithms. Our results build on a memory lower bound for a simple linear-algebraic problem – finding approximate null vectors – and utilize the estimates on the packing of the Grassmannian, the manifold of all linear subspaces of fixed dimension.

**Keywords:** streaming, memory, communication, linear classification, linear regression

## 1. Introduction

The complexity of learning, as studied in classical learning theory, is mostly concerned about the number of data instances required to solve a given learning task (a.k.a. sample complexity). However, as data becomes increasingly abundant and plentiful, the bottleneck in many tasks has shifted to computational resources, such as running time and memory usage. In particular, our understanding of how memory constraints affect learning performance is still rather limited.

As of today, scalable supervised learning algorithms are characterized by being linear in the data dimension: In other words, the amount of required computer memory is not much larger than what is required to store a single data instance (represented as a vector in $\mathbb{R}^d$). Stochastic gradient-based methods, which are based on sequentially processing a single or a small mini-batch of examples, are a prominent member of this class. In contrast, algorithms whose memory usage is super-polynomial in $d$ are challenging to implement for high-dimensional data. It is thus an important theoretical problem to understand what are the inherent limitations of memory-constrained algorithms.

In this paper, we study several fundamental linear prediction problems in a natural streaming setting, and prove *quadratic* memory lower bounds using any, possibly randomized algorithms (in other words, for data in $d$ dimensions, one needs $\Omega(d^2)$ memory in order to solve them):

- **Linear Separators:** Given a stream of $\Omega(d)$ unit vectors $x_1, x_2, \ldots$, which are linearly separable (that is, $\min_i x_i^\top w > \gamma$ for some unit vector $w$ and margin $\gamma > 0$), find a linear separator. In fact, the lower bound is shown even if the margin $\gamma$ is as large as $\Theta(1/\sqrt{d})$, and even if the predictor is allowed to classify a small (constant) fraction of the points incorrectly.

- **Linear Regression:** Given a stream of $d$ labeled examples $\{(A_i, b_i)\}_{i=1}^{d}$ (which can be interpreted as rows of a $d \times d$ matrix $A$ and entries of a vector $b$), find a point $\hat{w}$ such that $\sum_i (A_i^\top \hat{w} - b_i)^2$ is smaller than some universal constant. It also applies for algorithms that are allowed to make a pass over the stream at a random order. The lower bound is shown even if there exists a solution $w^*$ ($\|w^*\| \leq 1$) such that $\forall i, A_i w^* = b_i$ and even if $\forall i, \|A_i\| \leq 1$ and $\|b\| \leq 1$.

Both problems are based on a reduction from the following simple linear-algebraic problem:

- **Approximate Null Vectors:** Given a stream of $d-1$ vectors $x_1, \ldots, x_{d-1}$ in $\mathbb{R}^d$, sampled i.i.d. from a standard Gaussian, find a unit vector approximately orthogonal to all of them. Specifically, we show that quadratic memory is required to find a vector $\hat{w}$ such that $\frac{1}{d}\sum_i (x_i^\top \hat{w})^2$ is less than some universal constant.

All of these lower bounds hold even for randomized algorithms which *succeed with probability exponentially small* in $d$. Furthermore, they are essentially tight in terms of parameter dependencies. First of all, in terms of memory, all of the problems are trivially solvable with $\widetilde{O}(d^2)$ memory (where $\widetilde{O}$ hides constants and logarithmic factors), simply by storing all the data and solving the problem offline (and in polynomial time) by phrasing them as a convex optimization problem. Moreover, our results are also tight in terms of the other problem parameters:

- For finding an approximate linear separator on $m$ samples, this problem can be solved in $\widetilde{O}(1/\gamma^4)$ memory, by drawing a random subsample of size $\widetilde{O}(1/\gamma^2)$, storing a random projection of this sample into $\widetilde{O}(1/\gamma^2)$ dimensions, finding a linear separator in that space, and translating it back to the original space (Blum, 2006). Thus, a memory of $\widetilde{\Theta}(1/\gamma^4)$ is sufficient, and necessary when $m = \Omega(1/\gamma^2)$, for some hard distribution over datasets.

- For the linear regression problem, we can trivially get $\sum_i (A_i^\top \hat{w} - b_i)^2 \leq 1$ (as opposed to some constant $\ll 1$) by picking $\hat{w} = \mathbf{0}$.

- For the approximate null vector problem, it is easy to get $\frac{1}{d}\sum_i (x_i^\top \hat{w})^2 \approx 1$ (rather than a constant $\ll 1$) by picking $\hat{w}$ uniformly at random from the unit sphere.

As mentioned earlier, our results are based on the lower bound we show for the approximate null vector problem. We rely on the existence of a collection of $\exp\left(\Omega(d^2)\right)$ linear subspaces (all $d/2$-dimensional in $\mathbb{R}^d$) which are pairwise far from each other, with respect to a standard distance (Dai et al., 2007). Using angles between vector spaces, symmetries, and the distribution over singular values of random matrices, we show that any successful algorithm for the above tasks should not confuse between two vector spaces from that collection. To allow storing each vector space at a different memory configuration, approximately $\log \exp(d^2)$ memory is required.

We emphasize that our results focus on a streaming setting, where only a single pass over the examples is allowed, and refer to performing some task on a given set of examples. (rather than over some underlying distribution, in a statistical learning setting). It would be interesting to study whether our results can be extended to such scenarios.

### Prior Work

As mentioned earlier, the memory complexity of learning problems has attracted increasing interest in recent years, and we survey some relevant results below. However, to the best of our knowledge, these results are different than our work, by either focusing on very small memory budgets

(e.g. insufficient to store even a single example), specialized data access models (which do not, for instance, allow for the natural setting of examples being streamed one-by-one), or apply to other, fundamentally different learning problems (except the recent independent work of Sharan et al. 2019, discussed below).

In a breakthrough result, Raz (2016) proved that learning parities – corresponding to linear regression *over finite fields* – in a statistical setting requires either quadratic memory or an exponential sample size. This was later improved and extended by several works, e.g. (Raz, 2017; Moshkovitz and Moshkovitz, 2017; Kol et al., 2017; Garg et al., 2017; Beame et al., 2017, 2018; Moshkovitz and Moshkovitz, 2018). But, all these are specific to finite fields, rather than regression over $\mathbb{R}$, where no exponential gap is known. Indeed, some of these hard problems can be solved over $\mathbb{R}$ in polynomial time and linear memory using gradient based optimization. In a recent related paper, Sharan et al. (2019) consider the problem of performing linear regression in the statistical learning setting where a stream of examples are drawn from a distribution, and show that any algorithm that uses sub-quadratic memory exhibits a slower rate of convergence to the true solution than can be achieved without memory constraints. Their result is stronger than ours on linear regression and it was studied independently using different techniques.

Another remarkable result (Clarkson and Woodruff, 2009) studies linear regression over $\mathbb{R}$, but in a different model than ours, where individual entries of the entire dataset matrix arrive at an arbitrary order (rather than row-by-row), and updates to the entries can be received (e.g. "add 1 to coordinate $(2, 3)$"). Chu and Schnitger (1991) studied a model of exact computations on matrices of integer entries, where no approximation error is allowed. Related to the problem of linear separation, but in a different setting than ours, Guha and McGregor (2008) show that a streaming algorithm for finding the intersection of $n$ halfspaces in 2 or 3 dimensions requires $\Omega(n)$ memory. In Dagan and Shamir (2018), an $\Omega(d^2)$ memory lower bound is proven for finding correlations in $d$-dimensional distributions with optimal sample complexity. This is an unsupervised statistical learning problem quite different than the ones we study here. Steinhardt and Duchi (2015) has studied the sample complexity for memory bounded sparse linear regression.

Memory lower bounds can be reduced from communication complexity lower bounds. We list two prior works on related settings, which are incomparable to ours and cannot derive quadratic memory lower bounds in the dimension. First, Kane et al. (2017) studied the communication complexity of classification problems, in a general setting which enables dealing with arbitrary classification problems. Secondly, Daniely and Feldman (2018) showed that in a distributed setting with limited communication, exponentially many samples are required to find a linear separator, if the margin is small.

We list some other linear algebraic works in streaming and communication settings over the real numbers: Balcan et al. (2019) and Zhang et al. (2015) studied the problem of finding approximate matrix ranks, Braverman et al. (2018) studied Schatter $p$-norms of matrices, Levin et al. (2018) studied the problem of finding a subspace which approximates the input data, Cohen et al. (2016) studied approximate matrix product, Braverman et al. (2016) studied sparse linear regression, and many other works exist. Relevant work studying related linear algebraic problems over finite fields includes Li et al. (2014), Chu and Schnitger (1995), Sun and Wang (2012) and many others.

**Paper organization.** Section 2 contains preliminaries, Section 3 contains the main results, Section 4 contains the proof summary, Appendix A contains auxiliary mathematical lemmas, and Appendix B contains the full proofs.

## 2. Preliminaries

**Notations.** We use $C, C', C_1, c, c'$ etc. to denote absolute positive constants which do not depend on the dimension nor on the other problem parameters. When uppercase $C$ appears, the statement is correct for any sufficiently large constant, and when lowercase $c$ appears it holds for any sufficiently small positive value.

Here are some linear algebraic definitions: The unit sphere is denoted by $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \colon \|x\|_2 = 1\}$. The *Grassmannain*, denoted by $\mathrm{Gr}(k,d)$, is the set of all subspaces of $\mathbb{R}^d$ of dimension $k$.

We use the following standard notations: The Euclidean norm $\|\cdot\|_2$ is denoted by $\|\cdot\|$. Given $V \in \mathrm{Gr}(k,d)$ and $u \in \mathbb{R}^d$, $\mathrm{Proj}_V(u)$ denotes the projection of $u$ into $V$.

For convenience, given linearly independent vectors $v_1, \ldots, v_{d-1} \in \mathbb{R}^d$, let $\ker(v_1 \cdots v_d)$ denote the unique unit vector orthogonal to $v_1 \cdots v_{d-1}$.

**One-pass low memory algorithms.** We assume a setting where samples $z_1, \ldots, z_m \in Z$ are obtained one after the other in a streaming fashion, and an algorithm has to compute some function of them, the output lying in a domain $O$. There is not enough memory to store all samples: only $b$ binary bits are available. The memory configuration $s_i \in \{0,1\}^b$ after receiving $z_i$ is some function $f_i$ of the previous memory configuration $s_{i-1}$ and the sample $z_i$. Here is a formal definition:

**Definition 1** *A* one-pass algorithm $\mathcal{A}$ *with memory usage* $b$ *is a collection of functions,* $(f_1, \ldots, f_m, o)$*, where* $f_i \colon Z \times \{0,1\}^b \to \{0,1\}^b$ *and* $o \colon \{0,1\}^b \to O$*. The output of* $\mathcal{A}$ *given the input* $(z_1, \ldots, z_m)$ *is* $o(s_m)$*, where* $s_m$ *is defined by the recursive formula:* $s_0 = 0$ *and* $s_i = f_i(z_i, s_{i-1})$*, for* $i \in \{1, \ldots, m\}$*.*

We also consider algorithms which use randomness: assume there exists a finite (but unbounded) collection of $N$ numbers drawn i.i.d uniformly from $[0,1]$ at the beginning of the execution. The algorithm is allowed to read these random numbers at any time, and they do not count towards the memory usage. Formally, these random numbers are now given to $f_i$ as additional inputs: $f_i \colon Z \times \{0,1\}^b \times [0,1]^N \to \{0,1\}^b$.

**Hard distributions and data arriving at a random order.** To prove lower bounds, we show that there is some hard distribution *over datasets* (over $Z^m$, rather than over $Z$), where any low memory algorithm fails. The samples $z_1, \ldots, z_m$ are either assumed to be shuffled beforehand, arriving at a *random order*, or at a *fixed order*. Formally, we say that they arrive at a random order if for any $(z_1, \ldots, z_m)$ and any permutation $\pi \colon \{1, \ldots, m\} \to \{1, \ldots, m\}$, the probability of $(z_1, \ldots, z_m)$ to arrive equals the probability of $\left(z_{\pi(1)}, \ldots, z_{\pi(m)}\right)$. While the main results on the approximate null vector problem and linear regression captures a random order of arrival, the impossibility results on linear separators requires them to arrive at a fixed order.

**One sided communication protocols.** This captures the setting where two parties receive inputs $z_1, z_2 \in Z$ (one input per party). The first party sends a short message based on its input. Then, the second party, upon receiving its input and looking on the message, decides on the output. We allow a finite unbounded collection of $N$ i.i.d random numbers, uniform in $[0,1]$.

**Definition 2** *A* communication protocol $\mathcal{A}$ *that communicates* $b$ *bits is a pair of functions,* $f \colon Z \times [0,1]^N \to \{0,1\}^b$ *and* $o \colon Z \times \{0,1\}^b \times [0,1]^N \to O$*. The output of* $\mathcal{A}$ *given the inputs* $z_1, z_2$ *and the randomness* $R \in [0,1]^N$ *equals* $o(z_2, f(z_1, R), R)$*.*

**Reducing between communication protocols and one-pass algorithms.** One can simulate a low memory algorithms using communication protocols: Fix a one-pass algorithm $\mathcal{A}$ with memory usage $b$, receiving samples $z_1, \ldots, z_m$. Assume the corresponding communication setting, where the first party receives $z_1, \ldots, z_{m/2}$ and the second party receives $z_{m/2+1}, \ldots, z_m$. There exists a communication protocol $\mathcal{A}'$ using $b$ bits of communication, which simulates $\mathcal{A}$, namely, given any input $(z_1, \ldots, z_m)$, $\mathcal{A}'$ outputs the same as $\mathcal{A}$. Indeed, this protocol $\mathcal{A}'$ proceeds as follows: the first party starts simulating $\mathcal{A}$, feeding the samples $z_1, \ldots, z_{m/2}$ into $\mathcal{A}$. Then, it sends the last memory configuration of $\mathcal{A}$, using $b$ bits. The second party continues simulating the algorithm on the points $z_{m/2+1}, \ldots, z_m$. Then, it outputs the same as $\mathcal{A}$. Hence, any lower bound on the communication of $\mathcal{A}'$ derives a lower bound on the memory usage of $\mathcal{A}$.

**Approximability and measurability.** To avoid dealing with the technicalities of bit representation, we assume that the inputs are real numbers, and the algorithms are allowed to compute any *measurable* function on them. However, both the upper and lower bounds apply also in the standard RAM model, where each number is rounded to logarithmically many bits. The lower bounds trivially apply, since the RAM model is weaker. The upper bounds apply as well: since we are dealing with approximate solutions and problems with large margin, rounding the numbers degrades the performance only by a negligible amount.

**Linear separators and margin.** Given a list of pairs $((x_i, y_i))_{i=1}^m$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, we say that $w \in \mathbb{S}^{d-1}$ is a *linear separator* if $w^\top x_i y_i > 0$ for all $i \in \{1, \ldots, m\}$. The *margin* of $w$ on this set equals $\min_{i=1}^m w^\top x_i y_i / \|x_i\|$. The *margin* of the dataset is the maximal margin over $w \in \mathbb{S}^{d-1}$. A *hyperplane* is any $w \in \mathbb{S}^{d-1}$ used for classification.

## 3. Main Results

First, we discuss the approximate null vector problem, then linear separators and lastly, linear regression.

### 3.1. The approximate null vector problem (ANV)

The following result shows that any one pass algorithm which receives vectors $x_1, \ldots, x_{d-1}$ and outputs a vector which is approximately orthogonal to all of them, has a memory requirement of $\Omega(d^2)$. We present two variants: one, where the vectors are drawn from a standard normal distribution, and a different variant which we is use in the reductions to linear separators and linear regression.

**Theorem 3** *Let $g_1, \ldots, g_{d-1}$ be i.i.d vectors drawn from $\mathcal{N}(0, I_d)$. Let $\mathcal{A}$ be a randomized one-pass algorithm which outputs a unit vector $\hat{w}$ such that:*

$$\frac{1}{d} \sum_{i=1}^{d-1} (\hat{w}^\top g_i)^2 \leq c', \tag{1}$$

*with probability at least $e^{-cd}$ (the randomness is over the algorithm and over $g_1 \cdots g_{d-1}$). Then, the memory usage of $\mathcal{A}$ is $\Omega(d^2)$.*

Thm. 3 is a direct corollary of the communication variant, Thm. 28, proved in Appendix B.1. A summary of the proof appears in Sec. 4.

Note that if $\hat{w}$ is drawn uniformly at random from $\mathbb{S}^{d-1}$, then $\sum_{i=1}^{d-1}(\hat{w}^\top g_i)^2 \approx d$. Hence, it is impossible to do significantly better than random, even with a tiny probability of $e^{-\Omega(d)}$.

Next, we state the second variant. Given linearly independent vectors $v_1, \ldots, v_{d-1}$. We show that it is hard to find an approximate null vector even if the first entry of $\ker(g_1 \cdots g_{d-1})$ (the unit vector orthogonal to $g_1 \cdots g_{d-1}$) is guaranteed to be least some constant.

**Theorem 4** *Let $P$ denote the distribution over $d-1$ i.i.d uniformly drawn vectors from $\mathbb{S}^{d-1}$, $\theta'_1 \cdots \theta'_{d-1}$. Let $E$ be the event that $e_1^\top \ker(\theta'_1 \cdots \theta'_{d-1}) \geq c_f$, where $c_f$ is some sufficiently small universal constant and $e_1 = (1, 0, \ldots, 0)$. Assume that the input $\theta_1 \cdots \theta_{d-1}$ is drawn from $(P \mid E)$ (from the distribution $P$ conditioned on $E$). Let $\mathcal{A}$ be a randomized one-pass algorithm which outputs a vector $\hat{w}$ that satisfies:*

$$\sum_{i=1}^{d-1} \left(\hat{w}^\top \theta_i\right)^2 \leq c_1, \tag{2}$$

*with probability at least $e^{-c_2 n}$. Then, the memory usage of $\mathcal{A}$ is $\Omega(d^2)$.*

Thm. 4 is a direct corollary of the communication variant, Lemma 30, proved in Subsection B.1. A summary of the proof appears in Sec. 4.

Both Thm. 3 and Thm. 4 follow from the following lemma, which regards the communication setting where two parties receive vector spaces from $\mathrm{Gr}(d/2, d)$ and $\mathrm{Gr}(d/2-1, d)$, respectively, and their goal is to find an approximately orthogonal $\hat{w}$.

**Lemma 5** *Assume the following communication setting: the first party receives a uniformly random vector space $V$ from $\mathrm{Gr}(d/2, d)$, and the second party receives a uniformly random vector space $U$ from $\mathrm{Gr}(d/2-1, d)$. Let $\mathcal{A}$ be randomized one-sided communication protocol which outputs $\hat{w} \in \mathbb{S}^{d-1}$ that satisfies:*

$$\max\left(\|\mathrm{Proj}_V(\hat{w})\|, \|\mathrm{Proj}_U(\hat{w})\|\right) \leq c,$$

*with probability at least $e^{-c'd}$. Then, the communication contains $\Omega(d^2)$ bits.*

The proof appears in Subsection B.1 and a summary appears in Sec. 4.

### 3.2. Linear separators (LSP)

Let $((x_i, y_i))_{i=1}^{2m}$ denote a dataset, where $x_i \in \mathbb{S}^{d-1}$, $y_i \in \{-1, 1\}$ and $m \geq Cd$, for some constant $C > 0$. Assume that the points are separable with a margin of $\gamma = \Theta(d^{-1/2})$. Given a specific dataset, the goal of the algorithm is to find a hyperplane which classifies a large fraction of the points correctly. For the lower bounds, we will fix some hard distribution over datasets (rather than on examples, which are assumed to arrive at a fixed order). We show that any algorithm which outputs a hyperplane which classifies more than $(1-c_2)2m$ points correctly ($c_2 > 0$ is a universal constant), with non-negligible probability, requires a memory of $\Omega(d^2)$.

**Theorem 6** *There exists a distribution over datasets $((x_i, y_i))_{i=1}^{2m}$ satisfying the above properties, such that the following holds: any randomized one-pass algorithm which outputs a hyperplane $\hat{w}$, that with probability at least $e^{-cd}$ classifies $(1-c_2)m$ points correctly, has a memory usage of $\Omega(d^2)$ (the randomness is over the algorithm and the distribution over datasets).*

This is a direct corollary of the following communication bound, for the setting where the first party receives $((x_i, y_i))_{i=1}^{m}$ and the second receives the remaining $m$ examples.

**Theorem 7** *There exists a distribution over datasets $((x_i, y_i))_{i=1}^{2m}$ satisfying the above properties, such that the following holds: any randomized one-sided communication protocol $\mathcal{A}$ which outputs a hyperplane $\hat{w}$ that, with probability at least $e^{-c_3 d}$ classifies $(1 - c_2)2m$ points correctly, has a memory usage of $\Omega(d^2)$.*

The proof of Theorem 7 appears in Appendix B.2, and its proof sketch appears in Sec. 4. To illustrate some proof ideas of Thm. 6, we prove a weaker version, on finding an exact separator:

**Theorem 8** *There exists a distribution over datasets $((x_i, y_i))_{i=1}^{2d-2}$ satisfying the above properties, such that any one-pass algorithm $\mathcal{A}$ which outputs with probability at least $e^{-cd}$ a linear separator (classifying all points correctly), has a memory usage of $\Omega(d^2)$.*

**Proof** We reduce Thm. 8 from Thm. 4, by showing that given an algorithm $\mathcal{A}$ for LSP which satisfies the requirements in Thm. 8, one can create an algorithm $\mathcal{A}'$ for ANV satisfying the requirements in Thm. 4, with the same memory usage. Thm. 4 states that the memory usage of $\mathcal{A}'$ is $\Omega(d^2)$, which implies that the memory usage of $\mathcal{A}$ is $\Omega(d^2)$ as well and concludes the proof.

Here is how $\mathcal{A}'$ is constructed, by simulating $\mathcal{A}$: Whenever $\mathcal{A}'$ receives a point $x_i$, it creates the points $x_{i+} = x_i + c_4 e_1/\sqrt{d}$ and $x_{i-} = x_i - c_4 e_1/\sqrt{d}$, where $e_1$ is the first vector in the standard basis and $c_4 = \sqrt{c_1}$ ($c_1$ is the constant defined in Eq. (2)). Then, $\mathcal{A}'$ feeds $\mathcal{A}$ with the two pairs $(x_{i+}, 1)$ and $(x_{i-}, -1)$. Once the last iteration terminates, $\mathcal{A}'$ outputs the output of $\mathcal{A}$ (assuming, without loss of generality, that $\mathcal{A}$ outputs a unit vector).

Note that the algorithm $\mathcal{A}$ is assumed to operate only if the margin is $\Omega(d^{-1/2})$: our theorem is only concerned with such datasets. Luckily, $\mathcal{A}$ is fed with a sufficiently separated dataset. Indeed, Theorem 4 states that $w^* := \ker(x_1 \cdots x_{d-1})$ satisfies $w^{*\top} e_1 \geq c_f$. The same $w^*$ is a linear separator with margin $c_f c_4/\sqrt{d}$:

$$w^{*\top} x_{i+} = w^{*\top} x_i + w^{*\top} e_1 c_4/\sqrt{d} \geq c_f c_4/\sqrt{d}\,; \qquad w^{*\top} x_{i-} = w^{*\top} x_i - w^{*\top} e_1 c_4/\sqrt{d} \leq -c_f c_4/\sqrt{d}.$$

We are left with showing that $\mathcal{A}'$ outputs a vector with a loss of at most $c_1$, satisfying Eq. (2). Indeed, since the output $\hat{w}$ of $\mathcal{A}$ is a linear separator:

$$0 < \hat{w}^\top x_{i+} = \hat{w}^\top x_i + \hat{w}^\top e_1 c_4/\sqrt{d}\,; \qquad 0 > \hat{w}^\top x_{i-} = \hat{w}^\top x_i - \hat{w}^\top e_1 c_4/\sqrt{d}$$

hence $|\hat{w}^\top x_i| < \hat{w}^\top e_1 c_4/\sqrt{d} \leq c_4/\sqrt{d}$. Therefore, $\sum_{i=1}^{d-1} (\hat{w}^\top x_i)^2 \leq c_4^2 = c_1$. ■

Thm. 7 shows that when the margin is $\gamma$ and $m, d = \Theta(\gamma^{-2})$, any algorithm classifying $(1 - \varepsilon)$ of the points correctly requires $\Omega(\gamma^{-4})$ memory (where $\varepsilon$ is a small constant). This bound is asymptotically tight up to logarithmic factors, and there exists a one-pass algorithm with memory $\widetilde{O}(\log^2 m/(\gamma^4 \varepsilon))$ (or, $\widetilde{O}(\log^2 m/\gamma^4)$ when $\varepsilon$ is a constant). This upper bound holds for any values of $m$ and $d$, where $m$ is the sample size. It is based on the following fact: if we randomly project all points to dimension $d' = O(\log m/\gamma^2)$, with high probability the dataset will still be separable with margin $\gamma/2$ (Blum, 2006). We sketch this algorithm below.

First, note that if $m \gg d' (= \widetilde{\Theta}(1/\gamma^2))$, it suffices to subsample $O(d'/\varepsilon)$ points, and with high probability, any linear separator on the subsample will classify $(1 - \varepsilon)$ of the points in the

original dataset correctly (this follows from the sample complexity of realizable learning over $\mathbb{R}^{d'}$, see Shalev-Shwartz and Ben-David (2014), Sec. 6.4).

Hence, it suffices to construct an algorithm with memory $O(m \log m/\gamma^2)$ which finds a hypothesis that classifies *all* points correctly. This algorithm is implemented as follows: first, a uniformly random projection $P$ from $\mathbb{R}^d$ to $\mathbb{R}^{d'}$ is drawn, where $d' = O(\log m/\gamma^2)$. The algorithm projects all points $x_i$ and stores the projection $Px_i$ up to a sufficient accuracy, together with the label $y_i$. Then, it finds a linear separator $w_p$ in the projected space. Lastly, it outputs a preimage of $w_p$, namely, a vector $\hat{w}$ which satisfies $P\hat{w} = w_p$. There are many preimages of $\hat{w}_p$, and we select the one which is orthogonal to the kernel of $P$. This ensures that $\hat{w}^\top x_i y_i = w_p^\top P x_i y_i > 0$, and $\hat{w}$ is a linear separator as required. Indeed, if $x_{i,k}$ is the projection of $x_i$ to the kernel of $P$ and $x_{i,p} = x_i - x_{i,k}$, the following holds:

$$\hat{w}^\top x_i = \hat{w}^\top x_{i,p} = (P\hat{w})^\top (P x_{i,p}) = w_p^\top P x_i,$$

where the second equality follows from the fact that $\hat{w}$ and $x_{i,p}$ are in the subspace orthogonal to the kernel of $P$, hence applying $P$ on them results in a rotation, and, in particular, the angle between $\hat{w}$ and $x_{i,p}$ is the same as the angle between $P\hat{w}$ and $Px_{i,p}$.

**Remark 9** *The lower bound shows that while the low-memory perceptron attains low* online mistake bound*, it does not guarantee low error on the training set.*

### 3.3. Linear regression (LR)

Let $A$ be a real matrix of dimension $d \times d$ where each row $A_i$ satisfies $\|A_i\| \leq 1$. Let $b \in \mathbb{R}^d$ where $\|b\| \leq 1$. Assume that there is a solution $w^* \in \mathbb{R}^d$ with $\|w^*\| \leq 1$ for the equation system $Aw = b$. We prove the following theorem, on algorithms which receive the linear equations one after the other in a random order:

**Theorem 10** *There exists a distribution $P$ over pairs $(A, b)$ satisfying the definition from above, where the equations arrive at a random order, such that the following holds: Any randomized one-pass algorithm $\mathcal{A}$ outputting $\hat{w}$ which satisfies $\|A\hat{w} - b\|^2 \leq c$ with probability at least $e^{-c'n}$, has a memory usage of $\Omega(d^2)$.*

**Proof** We reduce this theorem from Thm. 4, as in the proof of Thm. 8. Assume the existence of an algorithm $\mathcal{A}$ for LR which satisfies the conditions in Thm. 10 with $c = \min(c_1 c_f^2/4, c_f^2/4)$ and $c' = c_2$, where $c_1$, $c_f$ and $c_2$ are the constants from Theorem 4. We will show that there exists an algorithm $\mathcal{A}'$ for ANV with the same memory usage, obtained by simulating $\mathcal{A}$. Thm. 4 will imply that the memory usage of $\mathcal{A}'$ is $\Omega(d^2)$, hence the memory usage of $\mathcal{A}$ is $\Omega(d^2)$.

The algorithm $\mathcal{A}'$, given any input point $\theta_i$ for ANV ($i = 1, \ldots, d-1$), will feed $\mathcal{A}$ with the equation $\theta_i^\top w = 0$. Additionally, $\mathcal{A}'$ will feed $\mathcal{A}$ with the equation $e_1^\top w = c_f$, where $e_1 = (1, 0, \ldots, 0)$. This equation will be fed at a uniformly random location (right after feeding $\theta_i^\top w = 0$, where $i$ is drawn uniformly at random from $\{0, 1, \ldots, d-1\}$). After receiving the output $\hat{w}_{\mathrm{LR}}$ of $\mathcal{A}$, $\mathcal{A}'$ will normalize this vector, outputting $\hat{w} = \hat{w}_{\mathrm{LR}}/\|\hat{w}_{\mathrm{LR}}\|$.

Note that the dataset $A = (\theta_1 | \cdots | \theta_i | e_1 | \theta_{i+1} | \cdots | \theta_{d-1})^\top$ and $b = (0, \ldots, 0, c_f, 0, \ldots, 0)^\top$ satisfies the required assumptions: each row of $A$ is of norm at most 1 and $b$ as well. There exists a solution $w^*$ to $Aw = 0$, of $\|w^*\| \leq 1$ as required: $w^* = c_f \theta_d / (e_1^\top \theta_d)$, where $\theta_d = \ker(\theta_1 \cdots \theta_{d-1})$. It is guaranteed from the requirements in Subsection 3.1 that $e_1^\top \theta_d \geq c_f$, hence $\|w^*\| \leq 1$. Also, note that the samples arrive at a random order (see definition in Sec. 2).

Next, we will show that the outputted vector $\hat{w}$ is approximately orthogonal to all $\theta_i$, satisfying Eq. (2). From the guarantees of $\mathcal{A}$ as discussed above, it follows that with probability at least $e^{-c_2 d}$, $\|A\hat{w}_{\mathrm{LR}} - b\|^2 \leq c \leq \min(c_1 c_f^2/4, c_f^2/4)$. Assuming that this holds, then $\|e_1^\top \hat{w}_{\mathrm{LR}} - c_f\|^2 \leq c_f^2/4$, hence $\|\hat{w}_{\mathrm{LR}}\| \geq e_1^\top \hat{w}_{\mathrm{LR}} \geq c_f/2$. Therefore,

$$\sum_{i=1}^{d-1} (\hat{w}^\top \theta_i)^2 = \frac{1}{\|\hat{w}_{\mathrm{LR}}\|^2} \sum_{i=1}^{d-1} (\hat{w}_{\mathrm{LR}}^\top \theta_i)^2 \leq \frac{1}{\|\hat{w}_{\mathrm{LR}}\|^2} \|A\hat{w}_{\mathrm{LR}} - b\|^2 \leq \frac{c_1 c_f^2}{4\|\hat{w}_{\mathrm{LR}}\|^2} \leq c_1. \tag{3}$$

Eq. (2) is satisfied, as required, which concludes the reduction from LR to ANV, and the proof follows. ■

This problem can be stated as a convex optimization over the unit ball:

$$\arg\min_x \|Ax - b\|^2 ; \quad \text{s.t. } \|x\| \leq 1.$$

A solution $x^*$ with zero loss is guaranteed to exist, and the choice $\hat{x} = 0$ is guaranteed to have a loss of $\|b\|^2 \leq 1$. We show that in order to achieve a loss less than some constant with non-negligible probability, $\Omega(d^2)$ memory is required. For comparison, there are several gradient-based algorithms for this problem which require memory usage of only $\widetilde{O}(d)$, but at the cost of multiple passes over the data.

**Remark 11** *We suspect that when the condition number is small, there are efficient one-pass algorithms.*

## 4. Proof summary

We sketch some of our results. The full proofs can be found in Appendix B.

**Proof Sketch of Lemma 5.** We show that the message sent by the first party has to contain $\Omega(d^2)$ bits: There are $\exp(\Omega(d^2))$ linear subspaces in $\mathrm{Gr}(d/2, d)$ which are pairwise far from each other in a known metric over the Grassmannian (Dai et al., 2007). The first party has to send $\log_2 \exp(\Omega(d^2)) = \Omega(d^2)$ bits to specify the vector space $V$ up to a sufficient approximation factor, otherwise the second party would not be able to find an approximately null vector. Concretely, we show the following (Lemma 27):

*Let $V_1, V_2 \in \mathrm{Gr}(d/2, d)$ be fixed vector spaces which are far apart, and let $U$ be drawn uniformly from $\mathrm{Gr}(d/2 - 1, d)$. Then, with probability $1 - e^{-\Omega(d)}$, all vectors $w \in \mathbb{S}^{d-1}$ satisfy*

$$\|\mathrm{Proj}_{V_1}(w)\|^2 + \|\mathrm{Proj}_{V_2}(w)\|^2 + \|\mathrm{Proj}_U(w)\|^2 = \Omega(1).$$

Here is the proof outline for this statement: since $V_1$ is far from $V_2$, their orthogonal complementaries, $V_1^\perp$ and $V_2^\perp$, are far from each other. Hence, a uniformly random vector from $\mathbb{S}^{d-1} \cap V_1^\perp$ will be far from $V_2^\perp$, in expectation. Concentration of measure phenomena on the Euclidean sphere implies that we can improve from expectation, to high probability. Hence, a random vector from $\mathbb{S}^{d-1} \cap V_1^\perp$ will be far from $V_2^\perp$, with high probability.

For a typical $U$, the space of vectors $w \in \mathbb{S}^{d-1}$ satisfying $\|\mathrm{Proj}_{V_1}(w)\|^2 + \|\mathrm{Proj}_U(w)\|^2 = o(1)$ is approximately a low dimensional vector space. If $U$ is chosen uniformly at random, this vector

space can be approximated by a uniformly random subspace of $V_1^\perp$ of low dimension, denoted by $W$.

A standard technique to reduce a problem from a subspace $W$ to a finite set of points is by discretization, namely, to create a $\delta$-net of $W \cap \mathbb{S}^{d-1}$ of size exponential in the dimension of $W$. When the net is defined properly and the subspace $W$ is uniformly drawn from $V_1^\perp$, each element in the $\delta$-net is drawn uniformly from the sphere as well. We apply the union bound over the net, and derive that with high probability, each member of $W$ will be far from $V_2^\perp$, i.e. the subspaces are far from each other.

To summarize: all vectors $w \in \mathbb{S}^{d-1}$ which are approximately orthogonal to $V_1$ and $U$, lie close to the subspace $W$. The subspace $W$ is far from being orthogonal to $V_2$, namely, far from $V_2^\perp$. Hence, there exists no vector which is approximately orthogonal both to $V_1$, $V_2$ and $U$.

**Reducing Theorem 3 from Lemma 5.** We prove the communication variant of Theorem 3 (Thm. 28), where there are two parties, receiving $d/2$ and $d/2-1$ samples, respectively. We consider a scaled version, where the vectors $g_1 \cdots g_{d-1}$ are drawn $\mathcal{N}(0, I_d/d)$, and the goal is to show that a memory of $\Omega(d^2)$ is required in order to find $\hat{w}$ with $\sum_{i=1}^{d-1}(\hat{w}^\top g_i)^2 = o(1)$. We show the following (Lemma 29):

*Let $G$ be a matrix of dimension $d-1 \times d$ of entries $\mathcal{N}(0, I_d/d)$. Let $V$ and $U$ be the subspaces spanned by the first $d/2$ rows and the last $d/2-1$ rows of $G$, respectively. Then, with high probability, all vectors $w \in \mathbb{R}^n$ satisfy*

$$c\|Gw\|^2 \leq \|\mathrm{Proj}_V(w)\|^2 + \|\mathrm{Proj}_U(w)\|^2 \leq C\|Gw\|^2.$$

*Equivalently, if $V'$ and $U'$ are matrices with rows forming orthonormal bases for $V$ and $U$, respectively, then*

$$c\|Gw\|^2 \leq \left\| \begin{pmatrix} V' \\ U' \end{pmatrix} w \right\|^2 \leq C\|Gw\|^2.$$

The last statement implies that drawing orthonormal bases $V$ and $U$ is equivalent, up to absolute constants, to drawing random Gaussian vectors, and the reduction follows.

To sketch a proof of this statement, let $G_1$ and $G_2$ be the top and bottom halves of $G$, respectively. It is known that all singular values of each of these matrices are bounded by absolute constants, hence

$$\sigma_{\min}(G_i)^2\|\mathrm{Proj}_V(w)\|^2 \leq \|G_i w\|^2 \leq \sigma_{\max}(G_i)^2\|\mathrm{Proj}_V(w)\|^2,$$

where $\sigma_{\min}$ and $\sigma_{\max}$ denote the minimal and maximal singular values, respectively (for $i = 1, 2$).

**Reducing Theorem 4 from Theorem 3.** We consider here the streaming variants. As discussed in the previous paragraph, we consider a scaled variant of Theorem 3, where each vector is distributed $\mathcal{N}(0, I_d/d)$. First, we claim that each such Gaussian vector is approximately of unit norm, hence we can assume they are distributed uniformly in $\mathbb{S}^{d-1}$ instead, and denote them by $\theta_1 \cdots \theta_{d-1}$.

Next, Thm. 3 states that with insufficient memory, any algorithm may succeed in outputting a vector approximately orthogonal to $\theta_1 \cdots \theta_{d-1}$ only with a tiny probability of $e^{-cd}$. Since $w^* := \ker(\theta_1 \cdots \theta_{d-1})$ is distributed uniformly in $\mathbb{S}^{d-1}$, the distribution of $e_1^\top w^*$ is known to approximately equal $\mathcal{N}(0, 1/d)$ (Lemma 24). In particular, $e_1^\top w^* \geq c_f$ with probability greater than $e^{-cd/2}$ (Lemma 25). Since $e^{-cd/2} \gg e^{-cd}$, even conditioned on $e_1^\top w^* \geq c_f$ it is impossible to find an approximate separator.

**Reducing Thm. 7 from Lemma 5** We consider a variant of Lemma 5 where the vector $w^*$ orthogonal to $U$ and $V$ satisfies $e_1^\top w^* \geq c_f$ (Lemma 31). We show that if $\mathcal{A}$ is a protocol for finding a linear separator, there exists a protocol $\mathcal{A}'$ for finding an approximate null vector with the same amount of communication.

Here is how $\mathcal{A}'$ is created, based on $\mathcal{A}$. The first party, given $V \in \mathrm{Gr}(d/2, d)$, creates an auxiliary distribution $D_V$ over pairs $(x, y)$, with the following property: Any hyperplane $w \in \mathbb{S}^{d-1}$ with low classification error on $D_V$, satisfies $\|\mathrm{Proj}_V(w)\|^2 \approx 0$. Similarly, the second party will create an auxiliary distribution $D_U$, such that any approximate separator $w$ satisfies $\|\mathrm{Proj}_U(w)\|^2 \approx 0$. In particular, any hyperplane with low error on the uniform mixture of $D_V$ and $D_U$ satisfies: $\|\mathrm{Proj}_V(w)\|^2 + \|\mathrm{Proj}_U(w)\|^2 \approx 0$.

Each party draws $m = \Omega(d)$ samples from their corresponding distribution ($D_V$ or $D_U$). Then, they simulate $\mathcal{A}$ to find a hyperplane $\hat{w}$ with low classification error on the mixed sample. Since the class of linear separators over $\mathbb{R}^d$ is of VC dimension $d$, $\hat{w}$ has low classification error on the mixture of $D_V$ and $D_U$, hence it satisfies $\|\mathrm{Proj}_V(w)\|^2 + \|\mathrm{Proj}_U(w)\|^2 \approx 0$, as required. Lemma 5 states that the communication of $\mathcal{A}'$ is $\Omega(d^2)$, hence the communication of $\mathcal{A}$ is $\Omega(d^2)$ as well.

Here is how a random pair $(x, y)$ is drawn from $D_V$ ($D_U$ is analogously defined): First a random point $x'$ is drawn uniformly from $V \cap \mathbb{S}^{d-1}$. Then, set $(x, y) = (x_+, 1)$ with probability $1/2$ and $(x, y) = (x_-, -1)$ with probability $1/2$, where $x_+ = x' + \Theta(e_1/\sqrt{d})$ and $x_- = x' - \Theta(e_1/\sqrt{d})$. For any fixed $w \in \mathbb{S}^{d-1}$, if $x$ is drawn uniformly from $V \cap \mathbb{S}^{d-1}$ then $w^\top x \sim \mathcal{N}(0, \|\mathrm{Proj}_V(w)\|^2)$ (approximately, see Lemma 24). From the definition of $D_V$, any hyperplane $w$ with low classification error on $D_V$ satisfies $w^\top x \approx 0$ for most $x \in V \cap \mathbb{S}^{d-1}$, hence any such $w$ satisfies $\|\mathrm{Proj}_V(w)\|^2 \approx 0$, as required.

## Acknowledgments

This research is supported in part by European Research Council (ERC) grant 754705.

## References

Shiri Artstein-Avidan, Apostolos Giannopoulos, and Vitali D Milman. *Asymptotic geometric analysis, Part I*, volume 202. American Mathematical Soc., 2015.

Maria-Florina Balcan, Yi Li, David P. Woodruff, and Hongyang Zhang. Testing matrix rank, optimally. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019.*, pages 727–746, 2019. doi: 10.1137/1.9781611975482.46.

Paul Beame, Shayan Oveis Gharan, and Xin Yang. Time-space tradeoffs for learning from small test spaces: Learning low degree polynomial functions. *arXiv preprint arXiv:1708.02640*, 2017.

Paul Beame, Shayan Oveis Gharan, and Xin Yang. Time-space tradeoffs for learning finite functions from random evaluations, with applications to polynomials. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 843–856, 2018.

Avrim Blum. Random projection, margins, kernels, and feature-selection. In *Subspace, Latent Structure and Feature Selection*, pages 52–68. Springer, 2006.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L. Nguyen, and David P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 1011–1020, 2016. doi: 10.1145/2897518.2897582.

Vladimir Braverman, Stephen R. Chestnut, Robert Krauthgamer, Yi Li, David P. Woodruff, and Lin F. Yang. Matrix norms in data streams: Faster, multi-pass and row-order. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 648–657, 2018.

Jeff I Chu and Georg Schnitger. The communication complexity of several problems in matrix computation. *Journal of Complexity*, 7(4):395–407, 1991.

Jeff I Chu and Georg Schnitger. Communication complexity of matrix computation over finite fields. *Mathematical systems theory*, 28(3):215–228, 1995.

Kenneth L Clarkson and David P Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 205–214. ACM, 2009.

Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 11:1–11:14, 2016. doi: 10.4230/LIPIcs.ICALP.2016.11.

Yuval Dagan and Ohad Shamir. Detecting correlations with little memory and communication. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 1145–1198, 2018.

Wei Dai, Brian C Rider, and Youjian Liu. Volume growth and general rate quantization on grassmann manifolds. In *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE*, pages 1441–1445. IEEE, 2007.

Amit Daniely and Vitaly Feldman. Learning without interaction requires separation. *CoRR*, abs/1809.09165, 2018.

Sumegha Garg, Ran Raz, and Avishay Tal. Extractor-based time-space lower bounds for learning. *arXiv preprint arXiv:1708.02639*, 2017.

Sudipto Guha and Andrew McGregor. Tight lower bounds for multi-pass stream computation via pass elimination. In *International Colloquium on Automata, Languages, and Programming*, pages 760–772. Springer, 2008.

Daniel M Kane, Roi Livni, Shay Moran, and Amir Yehudayoff. On communication complexity of classification problems. *arXiv preprint arXiv:1711.05893*, 2017.

Gillat Kol, Ran Raz, and Avishay Tal. Time-space hardness of learning sparse parities. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1067–1080. ACM, 2017.

Roie Levin, Anish Prasad Sevekari, and David P. Woodruff. Robust subspace approximation in a stream. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 10706–10716, 2018.

Yi Li, Xiaoming Sun, Chengu Wang, and David P Woodruff. On the communication complexity of linear algebraic problems in the message passing model. In *International Symposium on Distributed Computing*, pages 499–513. Springer, 2014.

Dana Moshkovitz and Michal Moshkovitz. Mixing implies lower bounds for space bounded learning. In *Conference on Learning Theory*, pages 1516–1566, 2017.

Dana Moshkovitz and Michal Moshkovitz. Entropy samplers and strong generic lower bounds for space bounded learning. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 28:1–28:20, 2018. doi: 10.4230/LIPIcs.ITCS.2018.28.

Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 266–275. IEEE, 2016.

Ran Raz. A time-space lower bound for a large class of learning problems. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 732–742. IEEE, 2017.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Vatsal Sharan, Aaron Sidford, and Gregory Valiant. Memory-sample tradeoffs for linear regression with small error. *arXiv preprint arXiv:1904.08544*, 2019.

Jacob Steinhardt and John Duchi. Minimax rates for memory-bounded sparse linear regression. In *Conference on Learning Theory*, pages 1564–1587, 2015.

Xiaoming Sun and Chengu Wang. Randomized communication complexity for linear algebra problems over finite fields. In *STACS'12 (29th Symposium on Theoretical Aspects of Computer Science)*, volume 14, pages 477–488. LIPIcs, 2012.

Stanislaw J Szarek. Spaces with large distance to $\ell_\infty^n$ and random matrices. *American Journal of Mathematics*, 112(6):899–942, 1990.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Feng Wei. Upper bound for intermediate singular values of random matrices. *Journal of Mathematical Analysis and Applications*, 445(2):1530–1547, 2017.

Ke Ye and Lek-Heng Lim. Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197, 2016.

Yuchen Zhang, Martin Wainwright, and Michael Jordan. Distributed estimation of generalized matrix rank: Efficient algorithms and lower bounds. In *International Conference on Machine Learning*, pages 457–465, 2015.

## Appendix A. Auxiliary Mathematical results

**Notations.** Let $\mathrm{Proj}_V(v)$ denote the projection of a vector $v$ into a vector space $V$. For any subspace $V$ of $\mathbb{R}^d$ of dimension $k$, let $V^\perp$ denote the subspace of dimension $d-k$ orthogonal to $V$. For any two subspaces $U, V$ of $\mathbb{R}^d$, let $U \oplus V = \{u+v \colon u \in U, V \in V\}$ denote their direct sum.

### A.1. The Grassmannian

There exists a unique measure over $\mathbb{S}^{d-1}$ which is uniform under rotations, namely, that satisfies: $\Pr[UA] = \Pr[A]$ for any $A \subseteq \mathbb{S}^{d-1}$ and any orthogonal (unitary) transformation $U \in O(d)$. This measure is also called the *uniform* measure.

Next, we give some definitions:

**Definition 12** *For any positive integer $d$ and $0 \le k \le d$, the set of all linear subspaces of $\mathbb{R}^d$ of dimension $k$ is denoted $\mathrm{Gr}(k,d)$, and called the* Grassmannian.

**Definition 13** *The unique uniform probability measure (Haar measure) on the Grassmannian $\mathrm{Gr}(k,d)$ can defined as follows: Choose $k$ vectors independently and uniformly from $\mathbb{S}^{d-1}$ and take their linear span.*

Clearly, this measure is invariant under rotations, namely for any $A \subseteq \mathrm{Gr}(k,d)$ and any orthogonal transformation $U \in O(d)$, $\Pr(A) = \Pr(UA)$.

It is known that any two lines in $\mathbb{R}^3$ have an angle between then. A generalization of this statement holds for subspaces of $\mathbb{R}^d$: For any two linear subspaces $U, V \in \mathrm{Gr}(k,d)$ we define the $k$ principal angles between them, $0 \le \theta_k \le \cdots \le \theta_1 \le \pi/2$ as follows: First, we use a fact from linear algebra that there are two orthonormal (normalized orthogonal) bases of $U$ and $V$: $v_1, \ldots, v_k$ and $u_1, \ldots, u_k$ respectively, such that $\langle v_i, u_j \rangle = 0$ for all $i \neq j$. Assume without loss of generality that $|\langle u_1, v_1 \rangle| \le \cdots \le |\langle u_k, v_k \rangle|$. Then, the $i$'th principal angle is $\theta_i = \arccos |\langle u_i, v_i \rangle|$.

**Definition 14** *Let $U, V \in \mathrm{Gr}(k,d)$ be two linear subspaces and let $\theta_1, \ldots, \theta_k$ denote the $k$ principal angles between them. The* chordal distance *between $U$ and $V$ is defined as*

$$\mathrm{d}(U,V) =: \sqrt{\sum_{i=1}^{d} \sin^2 \theta_i}.$$

The Grassmannian can be regarded as a metric space with respect to the chordal distance. A result of Dai et al. (2007) shows that if $k$ is a constant fraction of $d$, then there is a collection of $e^{\Omega(d^2)}$ linear subspaces in $\mathrm{Gr}(k,d)$ such that all pairwise distances are $\Omega(\sqrt{d})$. The chordal distance has also the following nice property: (see, for example, Ye and Lim (2016))

**Lemma 15** *Let $U, V \in \mathrm{Gr}(d/2, d)$ be two linear subspaces, then*

$$\mathrm{d}(U,V) = \mathrm{d}(U^\perp, V^\perp).$$

**Theorem 16 (Dai et al. (2007))** *Let $0 < \alpha < 1$, then there exists a $c(\alpha)\sqrt{d}$-**separated set** $\mathcal{F} \subseteq \mathrm{Gr}(\lceil \alpha d \rceil, d)$ of size $2^{c'(\alpha)d^2}$. Namely, for any $V \neq U \in \mathcal{F}$ it holds that $\mathrm{d}(U,V) \ge c(\alpha)\sqrt{d}$.*

### A.2. Random matrix theory

Given a matrix $A$ of dimension $N \times d$, the *singular values* of $A$ are the square roots of the eigenvalues of $A^\top A$. We denote them by $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_d(A) \geq 0$.

**Claim 1** *For any matrix $A_{N\times d}$, there exists an orthonormal basis of $\mathbb{R}^d$ of singular vectors $v_1, \ldots, v_d$, such that for any $\lambda_1, \ldots, \lambda_d \in \mathbb{R}$,*

$$\left\| A \left( \sum_{i=1}^{d} \lambda_i v_i \right) \right\|^2 = \sum_{i=1}^{d} \lambda_i^2 \sigma_i(A)^2.$$

**Claim 2** *For any matrix $A$, the collection of non-zero singular values of $A$ equals the non-zero singular values of $A^\top$. Moreover, when we restrict the matrix $A$ to operate on its rows span, then the restricted operator has the same singular values as $A^\top$.*

Let $N \geq d$, let $A_{N\times d}$ be a random matrix. We say that $A$ is normal random matrix, when the all its entries are $N(0,1)$ independent random variables. Also let $\sigma_{\min}$ and let $\sigma_{\max}$ be the minimal and maximal singular values of $A$. The following are fundamental results in random matrix theory: (see for example the survey of Vershynin (2010))

**Theorem 17** *Let $A_{d\times d}$ be a normal random matrix. The following holds for its minimal and maximal singular values:*

$$\sqrt{N} - \sqrt{d} \leq \mathbb{E}[\sigma_{\min}] \leq \mathbb{E}[\sigma_{\max}] \leq \sqrt{N} + \sqrt{d}.$$

**Corollary 18** *Let $A$ be an $N \times d$ matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least of $1 - 2e^{-\frac{t^2}{2}}$ the following holds:*

$$\sqrt{N} - \sqrt{d} - t \leq \sigma_{\min} \leq \sigma_{\max} \leq \sqrt{N} + \sqrt{d} + t.$$

The final tool that we need gives results for the mid-singular values of a normal random matrix of size $A_{N\times d}$. The following result is from Szarek (1990) and was generalized by Wei (2017).

**Lemma 19** *Let $A_{N\times d}$ be a normal random matrix and let $0 \leq \tau \leq 1$. Then, the following holds*

$$c(1-\tau)\sqrt{d} \leq \sigma_{\tau d} \leq C(1-\tau)\sqrt{d},$$

*with probability of at least $1 - e^{-c\tau d}$.*

### A.3. Net on the Sphere and Concentration on the sphere

**Definition 20 (Nets, covering numbers)** *Let $(X, d)$ be a metric space and let $\delta > 0$. A subset $N_\delta$ of $X$ is called a $\delta$-net of $X$ if for every point $x \in X$ there exists a point $y \in N_\delta$, such that $d(x, y) \leq \delta$. The covering number of $X$ at scale $\delta$ is the size $N_\delta$ of the smallest $\delta$-net of $X$.*

The next lemma provides a bound on the size of a $\delta$-net of the Euclidean sphere, see for example Lemma 5.2 in Vershynin (2010).

**Lemma 21** *The unit Euclidean sphere $\mathbb{S}^{d-1}$ equipped with the Euclidean metric satisfies for every $\delta > 0$ that*

$$N_\delta \leq \left(1 + \frac{2}{\delta}\right)^d.$$

**Lemma 22 ( Lemma 5.3.5 in Artstein-Avidan et al. (2015))** *Let $\mathcal{N}$ be a $\delta$-net on $\mathbb{S}^{d-1}$, let $f$ be a 1- Lipshitz function. If for any $\epsilon \in (0,1)$, we know that*

$$\forall x \in \mathcal{N} \quad f(x) \leq 1 - \epsilon,$$

*then,*

$$\forall x \in \mathbb{S}^{d-1} \quad f(x) \leq \min\left\{\frac{1-\epsilon}{1-\delta}, (1-\epsilon) + \arcsin(\delta)\right\}.$$

The following two Lemmas are classical results from non-asymptotic geometry, see for example Artstein-Avidan et al. (2015). The first lemma states that any Lipschitz function on $\mathbb{S}^{d-1}$ is tightly concentrated around its mean:

**Lemma 23** *Let $y \sim U(\mathbb{S}^{d-1})$ and let $f : \mathbb{S}^{d-1} \to \mathbb{R}$ be a 1-Lipschitz function. The following holds:*

$$\Pr(|f(y) - \mathbb{E}[f(y)]| \geq \epsilon) \leq 2e^{-c_1 d\epsilon^2}$$

The next lemmas are on the distribution of a uniformly random unit vector:

**Lemma 24** *Let $\theta^{(d)}$ be a uniformly random vector from $\mathbb{S}^{d-1}$, and let $w \in \mathbb{S}^{d-1}$. Then, as $d \to \infty$, the distribution of $\sqrt{d}w^T\theta^{(d)}$ converges in distribution to $\mathcal{N}(0,1)$, namely, for any $\alpha \in \mathbb{R}$,*

$$\lim_{d\to\infty} \Pr\left[\sqrt{d}w^T\theta^{(d)} > \alpha\right] = \Pr_{g\sim\mathcal{N}(0,1)}[g > \alpha].$$

*Furthermore, the convergence rate does not depend on $w$.*

**Lemma 25** *Fix some constant $\alpha > 0$ and let $\theta$ be chosen uniformly from $\mathbb{S}^{d-1}$. Fix $w \in \mathbb{S}^{d-1}$. Then, there exists $c(\alpha) > 0$ which satisfies:*

$$\Pr(w^T\theta \geq c(\alpha)) \geq e^{-\alpha d}$$

*for any sufficiently large $n$.*

## Appendix B. Proofs

Proof of statements related to the approximate null vector problem appear in Subsection B.1; The proof of Thm. 7 on linear separation appears in Subsection B.2; and proofs of the mathematical statements appear in Subsection B.3.

### B.1. Approximate null vector problem

We prove results on the approximate null vector problem, providing reductions between different problem settings. Let $\mathcal{F}$ be a $\sqrt{d\delta/2}$-separated set on $\mathrm{Gr}(d/2, d)$ of size $e^{\Omega(d^2)}$, which exists from Theorem 16, where $\delta > 0$ is a universal constant.

**Lemma 26** *Assume the following communication setting: the first party receives a uniformly random vector space $V$ from $\mathcal{F}$, and the second party receives a uniformly random vector space $U$ from $\mathrm{Gr}(d/2-1, d)$. Let $\mathcal{A}$ be a randomized one-sided communication protocol which outputs a vector $\hat{w} \in \mathbb{S}^{d-1}$ which satisfies:*

$$\max\left(\|\mathrm{Proj}_V(\hat{w})\|, \|\mathrm{Proj}_U(\hat{w})\|\right) < c, \tag{4}$$

*with probability at least $e^{-c'd}$. Then, the communication contains $\Omega(d^2)$ bits.*

The proof of this theorem relies on the following lemma:

**Lemma 27** *Let $V_1, V_2$ be **fixed** linear subspaces in $\mathrm{Gr}(d, d/2)$ with distance $\mathrm{d}(V_1, V_2)^2 \geq \delta d/2$ ($\delta > 0$ is a universal constant). Let $U$ a uniformly random subspace that is drawn from $\mathrm{Gr}(d/2-1, d)$. Then with probability of at least $1 - e^{-cd}$, all vectors $v \in \mathbb{S}^{n-1}$ satisfy*

$$\max\left(\left\|\mathrm{Proj}_{V_1}(v)\right\|, \left\|\mathrm{Proj}_{V_2}(v)\right\|, \|\mathrm{Proj}_U(v)\|\right) \geq c. \tag{5}$$

*for a sufficiently small universal constant $c > 0$.*

The proof appears in Sec. B.3.

**Proof** [Proof of Lemma 26] First, we argue that it suffices to assume that $\mathcal{A}$ is randomized. Indeed, if there exists a randomized algorithm which outputs an approximately null vector with probability $e^{-c'd}$, then there exists a deterministic algorithm with the same guarantee: any randomized algorithm is a distribution over deterministic algorithms, hence there has to be a fixing of the randomness which outputs an approximate null vector with probability at least $e^{-c'd}$.

Recall that $|\mathcal{F}| \geq e^{\Omega(d^2)}$, and assume that the communication of $\mathcal{A}$ is at most $\log_2 |\mathcal{F}|/2$. We will show that with high probability, $\max(\mathrm{Proj}_V(\hat{w}), \mathrm{Proj}_U(\hat{w})) > c$, to conclude the proof. Denote $N = 2^b$ where $b$ is the communication $\mathcal{A}$, and note that $N \leq \sqrt{|\mathcal{F}|}$. For each $i \in \{1, \ldots, N\}$, let $\mathcal{X}_i$ denote the set of all vector spaces $V \in \mathcal{F}$ such that the first player sends the message $i$ after receiving $V$ as an input. Note that $\{\mathcal{X}_i\}_{i=1}^N$ is a partition of $\mathcal{F}$ to disjoint sets.

For any $V \in \mathcal{F}$ and $U \in \mathrm{Gr}(d/2-1, d)$, let $I_{V,U}$ be the indicator of whether the protocol $\mathcal{A}$ on inputs $V$ and $U$ outputs $\hat{w}$ which satisfies,

$$\max\left(\|\mathrm{Proj}_V(\hat{w})\|, \|\mathrm{Proj}_U(\hat{w})\|\right) < c,$$

where $c$ is the constant from Eq. (4) and Eq. (5) (we define the constant $c$ in Eq. (4) to equal the constant of Eq. (5)). For any $V_1 \neq V_2 \in \mathcal{F}$ and $U \in \mathrm{Gr}(d/2-1, d)$, let $J_{V_1,V_2,U}$ be the indicator of whether Eq. (5) is not satisfied, namely if there exists $v \in \mathbb{S}^{d-1}$ such that

$$\max\left(\left\|\mathrm{Proj}_{V_1}(v)\right\|, \left\|\mathrm{Proj}_{V_2}(v)\right\|, \|\mathrm{Proj}_U(v)\|\right) < c,$$

where $c$ is the value appearing in Eq. (5). From Lemma 27, for any $V_1 \neq V_2 \in \mathcal{F}$, it holds that $\mathbb{E}_{U \sim \mathrm{Gr}(d/2-1,d)} J_{V_1,V_2,U} \leq \xi$, where $\xi = e^{-\Omega(d)}$. Additionally, note that for all $\mathcal{X}_i$ and all

$V_1, V_2 \in \mathcal{X}_i$, the output of the protocol given the pair $(V_1, U)$ equals the output given $(V_2, U)$. Hence, if $J_{V_1,V_2,U} = 0$, then either $I_{V_1,U} = 0$ or $I_{V_2,U} = 0$. In other words, $I_{V_1,U} I_{V_2,U} \leq J_{V_1,V_2,U}$.

Note that the probability that Eq. (4) holds equals

$$\mathop{\mathbb{E}}_{V\sim\mathcal{F},\, U\sim \mathrm{Gr}(d/2-1,d)} [I_{V,U}] = \frac{1}{|\mathcal{F}|} \sum_{i=1}^{N} \mathop{\mathbb{E}}_{U} \left[ \sum_{V\in\mathcal{X}_i} I_{V,U} \right] = \frac{1}{|\mathcal{F}|} \sum_{i=1}^{N} \mathop{\mathbb{E}}_{U} [K_{i,U}] , \tag{6}$$

where $K_{i,U} = \sum_{V\in\mathcal{X}_i} I_{V,U}$. For any $i \in \{1, \ldots, N\}$, Jenssen's inequality implies:

$$\begin{aligned} \frac{(\mathbb{E}_U[K_{i,U}] - 1)^2}{2} &\leq \frac{\mathbb{E}_U[K_{i,U}](\mathbb{E}_U[K_{i,U}] - 1)}{2} \leq \mathop{\mathbb{E}}_{U} \left[ \frac{K_{i,U}(K_{i,U} - 1)}{2} \right] \\ &= \sum_{V_1\neq V_2\in\mathcal{X}_i} \mathop{\mathbb{E}}_{U} [I_{V_1,U} I_{V_2,U}] \leq \sum_{V_1\neq V_2\in\mathcal{X}_i} \mathop{\mathbb{E}}_{U} [J_{V_1,V_2,U}] \leq \binom{|\mathcal{X}_i|}{2} \xi \leq \frac{|\mathcal{X}_i|^2 \xi}{2}. \end{aligned}$$

Hence, $\mathbb{E}_U[K_{i,U}] \leq 1 + \sqrt{\xi}|\mathcal{X}_i|$. We conclude that the right hand side of Eq. (6) is bounded by

$$\frac{1}{|\mathcal{F}|} \sum_{i=1}^{N} \left(1 + \sqrt{\xi}|\mathcal{X}_i|\right) = \frac{N}{|\mathcal{F}|} + \sqrt{\xi} \leq 2^{-\Omega(n)},$$

using the fact that $N$ was defined to be significantly smaller than $|\mathcal{F}|$. ■

Instead of assuming that the input of the first party arrives uniformly from $\mathcal{F}$, we can assume that it arrives uniformly from $\mathrm{Gr}(d/2, d)$, as stated in Lemma 5, which we prove below:

**Proof** [Proof of Lemma 5] We reduce from Lemma 26. Fix a protocol $\mathcal{A}$ which solves the setting in Lemma 5 and we will show that there exists a protocol $\mathcal{A}'$ for the setting in Lemma 26 with the same amount of communication. The lower on the communication of $\mathcal{A}'$ implies a lower bound on the communication of $\mathcal{A}$.

Here is how $\mathcal{A}'$ is constructed: using the joint random bits[1], the parties will draw a uniformly random rotation $R$, namely, a unitary matrix of dimension $d \times d$. Then, they simulate $\mathcal{A}$ as if their inputs are $RV$ and $RU$ (where $RV$ and $RU$ are the results of applying $R$ on their vector spaces). Let $w$ be the output of the simulated protocol. The second party will output $\hat{w} = R^{-1}w$.

First, note that $RV$ and $RU$ are two i.i.d uniformly random vector spaces from $\mathrm{Gr}(d/2, d)$ and $\mathrm{Gr}(d/2 - 1, d)$, respectively, hence, the simulated protocol $\mathcal{A}$ receive inputs as stated in Lemma 5. In particular, it outputs an approximately null $w$ with a sufficiently large probability. Hence,

$$\begin{aligned} c \geq \max(\mathrm{Proj}_{RV}(w), \mathrm{Proj}_{RU}(w)) &= \max(\mathrm{Proj}_{RV}(R\hat{w}), \mathrm{Proj}_{RU}(R\hat{w})) \\ &= \max(\mathrm{Proj}_{V}(\hat{w}), \mathrm{Proj}_{U}(\hat{w})), \end{aligned}$$

with probability probability $e^{-c'd}$, as required. ■

Next, we prove the communication analogue of Theorem 3.

1. The parties are assumed to have shared random bits, as described in Section 2

**Theorem 28** *Let $g_1, \ldots, g_{d-1}$ be $d-1$ i.i.d vectors drawn from $\mathcal{N}(0, I_d)$. Assume the following communication setting: the first party receives $g_1, \ldots, g_{d/2}$ and the second party receives $g_{d/2+1}, \ldots, g_{d-1}$. Let $\mathcal{A}$ be a communication protocol outputting $\hat{w} \in \mathbb{S}^{d-1}$ which satisfies:*

$$\sum_{i=1}^{d-1} \left(\hat{w}^\top g_i\right)^2 \leq cd,$$

*with probability at least $e^{-c'd}$. Then, the communication of $\mathcal{A}$ is $\Omega(d^2)$.*

Theorem 28 follows from the following fact: $d/2$ random vectors are far from being linearly dependent, hence, a collection of such vectors behave as an approximate basis to a random vector space. Formally, we provide the following lemma:

**Lemma 29** *Let $g_1, g_2, \ldots, g_{d-1}$ be independent random normal vectors $\mathcal{N}(0, I_d/d)$. Let $G$ be the matrix of size $(d-1) \times d$ that its $i^{\text{th}}$ row is $g_i$. Also set $V = \text{span}\{g_1, \ldots, g_{d/2}\}$ and $U = \text{span}\{g_{n/2+1}, \ldots, g_{d-1}\}$.*

*Then, with probability $e^{-c_2 n}$, all $v \in \mathbb{S}^{n-1}$ satisfies*

$$c_1 \|Gv\|^2 \leq \|\text{Proj}_V(v)\|^2 + \|\text{Proj}_U(v)\|^2 \leq C_1 \|Gv\|^2 .$$

The proof of Lemma 29 appears in Subsection B.3.

**Proof** [Proof of Theorem 28] We will reduce to Lemma 5. Let $\mathcal{A}$ be a protocol for the setting in Lemma 5 and we will show how to create a protocol $\mathcal{A}'$ for the setting in Theorem 28 with the same amount of communication. The lower bound on the communication of $\mathcal{A}'$ implies a lower bound on the communication of $\mathcal{A}$.

Here is how $\mathcal{A}'$ is created. Let $P_{d/2} = \mathcal{N}(0, I_d)^{d/2}$ be the distribution over $d/2$ i.i.d copies of $\mathcal{N}(0, I_d)$, and for any $V \in \text{Gr}(d/2, d)$, let $E_V$ be the event that the span of these $d/2$ vectors equals $V$. Given an input $V \in \text{Gr}(d/2, d)$, the first party will draw $g_1, \ldots, g_{d/2}$ from the joint distribution $\left(P_{d/2} \mid E_V\right)$. Similarly, the second party, upon receiving $U$, will draw $g_{d/2+1}, \ldots, g_{d-1}$ from $\left(P_{d/2-1} \mid E_U\right)$, where $P_{d/2-1}$ and $E_U$ are similarly defined. The parties will simulate $\mathcal{A}$ as if the input is $g_1, \ldots, g_{d-1}$, and output the vector $\hat{w}$ outputted by $\mathcal{A}$.

For symmetrical reasons, since $U$ and $V$ are independent and uniform, the vectors $g_1, \ldots, g_{d-1}$ are distributed as $d-1$ i.i.d copies from $\mathcal{N}(0, I_d)$. We assumes that $\mathcal{A}$ satisfies the guarantees of Thm. 28, hence with probability at least $e^{-c'd}$,

$$\sum_{i=1}^{d-1} \left(\hat{w}^\top g_i\right)^2 \leq cd.$$

With probability at least $e^{-c'd} - e^{-c_2 n}$,

$$\|\text{Proj}_V(v)\|^2 + \|\text{Proj}_U(v)\|^2 \leq C_1 \sum_{i=1}^{d-1} \left(\hat{w}^\top \frac{g_i}{\sqrt{d}}\right)^2 \leq cC_1,$$

where the first inequality follows from Lemma 29 and holds with probability at least $1 - e^{-c_2 d}$ and the second with probability at least $e^{-c'd}$. If we select the constants $c$ and $c'$ in Theorem 28 to be

sufficiently small, we obtain that from Lemma 5, the memory requirement of $\mathcal{A}'$ is $\Omega(d^2)$, hence the memory requirement of $\mathcal{A}$ is $\Omega(d^2)$ as required. ■

Lastly, we provide the communication variant of Theorem 4. We remind the reader that given linearly independent vectors $v_1, \ldots, v_{d-1}$ we defined by $\ker(v_1 \cdots v_d)$ the unique unit vector orthogonal to $v_1 \cdots v_{d-1}$.

**Theorem 30** *Let $P$ denote the distribution over $d-1$ i.i.d uniformly drawn vectors from $\mathbb{S}^{d-1}$, $\theta'_1 \cdots \theta'_{d-1}$. Let $E$ be the event that $e_1^\top \ker(\theta'_1 \cdots \theta'_{d-1}) \geq c_f$, where $c_f$ is some sufficiently small universal constant. Let $\theta_1 \cdots \theta_{d-1}$ be random vectors drawn from $(P \mid E)$. Assume the following communication setting: the first party receives $\theta_1 \cdots \theta_{d/2}$ and the second receives $\theta_{d/2+1} \cdots \theta_{d-1}$. Let $\mathcal{A}$ be a communication protocol which outputs a vector $\hat{w}$ that satisfies:*

$$\sum_{i=1}^{d-1} \left(\hat{w}^\top \theta_i\right)^2 \leq c_1,$$

*with probability at least $e^{-c_2 d}$. Then, the communication of $\mathcal{A}$ is $\Omega(d^2)$.*

**Proof** We reduce from Thm. 28: Given a protocol $\mathcal{A}$ for satisfying the conditions in Lemma 30, we create a protocol $\mathcal{A}'$ with the same amount of communication that satisfies the conditions of Thm. 28. The protocol $\mathcal{A}'$ is defined as follows: given inputs $g_1 \cdots g_{d-1}$, the parties will normalize them to create $\theta_1 \cdots \theta_{d-1}$, where $\theta_i = g_i / \|g_i\|$. Then, they will simulate $\mathcal{A}$ as if their input is $\theta_1 \cdots \theta_{d-1}$. The second party will output the same output $\hat{w}$ outputted by $\mathcal{A}$.

Assume that $\mathcal{A}$ satisfies the conditions in Lemma 30 for sufficiently small constants $c_1$ and $c_2$. Let $c$ and $c'$ be the constants in Theorem 28. First, note from symmetry, that the inputs $\theta_1 \cdots \theta_{d-1}$ of $\mathcal{A}$ are distributed as $d-1$ i.i.d uniform copies from $\mathbb{S}^{d-1}$, hence $\ker(\theta_1 \cdots \theta_{d-1})$ is also uniformly distributed. From Lemma 25, with probability at least $e^{-c'd/2}$, $e_1^\top \ker(\theta_1 \cdots \theta_{d-1}) \geq c_f$ (assuming $c_f$ is sufficiently small). Recall that conditioned on this holding, $\mathcal{A}$ is guaranteed to output an approximate separator with probability at least $e^{-c'd}$. Hence,

$$\sum_{i=1}^{d-1} \left(\hat{w}^\top \theta_i\right)^2 \leq c_1,$$

with probability at least $e^{-c_2 d - c'd/2}$. Select $c_2$ to be sufficiently small such that $e^{-c_2 d - c'd/2} \geq 2e^{-c'd}$ (assuming that $d$ is sufficiently large). Since each $\|g_i\|^2$ is distributed as Chi-squared with $d$ degrees of freedom, there exists a constant $C > 0$, such that with probability at least $1 - e^{-c'd}/d$, $\|g_i\|^2 \leq C$. From union bound, with probability at least $1 - e^{-c'd}$, $\|g_i\|^2 \leq Cd$ for all $i \in \{1, \ldots, d-1\}$. Hence,

$$\sum_{i=1}^{d-1} \left(\hat{w}^\top g_i\right)^2 \leq Cd \sum_{i=1}^{d-1} \left(\hat{w}^\top \theta_i\right)^2 \leq Cdc_1,$$

where the first inequality holds with probability at least $1 - e^{-c'd}$ and the second inequality with probability at least $2e^{-c'd}$. Hence with probability at least $e^{-c'd}$, both inequalities hold, and if $c_1$ is sufficiently small, $\mathcal{A}'$ satisfies the requirements of Thm. 28. In particular, the memory usage of $\mathcal{A}'$ is $\Omega(d^2)$. ■

### B.2. Linear separators (Theorem 7)

We prove the Theorem 7. First, we present an auxiliary lemma, which is a variant of the approximate null vector problem. Given a vector space $W \in \mathrm{Gr}(d-1,d)$, denote by $\ker(W)$ the unique unit vector in $W^\perp$ and given subspaces $U$ and $V$ of $\mathbb{R}^n$, let $V \oplus U$ denote their direct sum.

**Lemma 31** *Let $P$ be a distribution over an independent pair of vector spaces: $V$ and $U$, drawn uniformly from $\mathrm{Gr}(d/2,d)$ and $\mathrm{Gr}(d/2-1,d)$, respectively. Let $E$ be the event that $e_1^\top \ker(V \oplus U) \geq c_f$, for some universal constant $c_f > 0$. Assume the communication setting where the inputs $U$ and $V$ are drawn from $(P \mid E)$. Let $\mathcal{A}$ be randomized one-sided communication protocol which outputs $\hat{w} \in \mathbb{S}^{d-1}$ that satisfies:*

$$\max\left(\|\mathrm{Proj}_V(\hat{w})\|, \|\mathrm{Proj}_U(\hat{w})\|\right) \leq c,$$

*with probability at least $e^{-c'd}$. Then, the communication is $\Omega(d^2)$.*

Note that Lemma 31 is the same as Lemma 5, expect that the inputs are drawn from $(P \mid E)$ rather than from $P$. One can reduce Lemma 31 from Lemma 5 the same way that Lemma 30 follows from Theorem 28.

We proceed with the following definition: Let $H$ be the set of linear separators over $\mathbb{R}^d$. Given a distribution $D$ over pairs $(x,y)$ where $x \in \mathbb{R}^d$ and $y \in \{-1,1\}$, an *$\varepsilon$-approximate net* for $H$ is a finite set $S$ of pairs $(x_i,y_i)$ such that each $w \in H$ satisfies:

$$\left| \Pr_{(x,y)\sim D}[w^\top xy > 0] - \Pr_{(x,y)\sim \mathrm{Uniform}(S)}[w^\top xy > 0] \right| \leq \varepsilon.$$

The following claim is equivalent to the standard uniform convergence theorems on the class of linear separators (Shalev-Shwartz and Ben-David, 2014):

**Claim 3** *For any $m \geq Cd/\varepsilon^2$, there exists an $\varepsilon$-approximate net $S$ of size $m$ for the hypothesis class $H$ of linear separators over $\mathbb{R}^d$ ($C > 0$ is a universal constant).*

**Proof** [Proof of Theorem 7] We will reduce from Lemma 31. Given an algorithm $\mathcal{A}$ for finding a linear separator, we will create an algorithm $\mathcal{A}'$ for the approximate null vector problem, as follows: the first party, upon receiving $V \in \mathrm{Gr}(d/2,d)$, creates a distribution $D_V$ (as defined below), and selects a $c_2$-approximate net $S_V$ for $D_V$ of size $m \geq Cd$ (arbitrarily). Similarly, the second party, upon receiving $U \in \mathrm{Gr}(d/2-1,d)$, selects a $c_2$-approximate net $S_U$ for the corresponding distribution $D_U$. Then, they simulate the protocol $\mathcal{A}$ on the combined dataset $S_V \cup S_U$, and output the output $\hat{w}$ outputted by $\mathcal{A}$.

Next, we define $D_V$. Here is how a random point $(x',y')$ is drawn from $D_V$: first, a point $x$ is drawn uniformly from $V \cap \mathbb{S}^{d-1}$. Then, with probability $1/2$, $(x',y') = (x_+,1)$ and with probability $1/2$, $(x',y') = (x_-,-1)$, where $x_+ = x + c/(4\sqrt{d})$ and $x_- = x - c/(4\sqrt{d})$ (where $c$ is the constant from Lemma 31). The distribution $D_U$ is defined similarly with respect to $U$.

First, note that the created dataset is guaranteed to have a margin of $\Omega(\sqrt{d})$. Indeed, $\ker(V \oplus U)$ is a linear separator achieving this margin (see the proof of Theorem 8 for a similar argument). We will show that if $\mathcal{A}$ finds a classifier which classifies $(1-c_2)2m$ points correctly, then $\mathcal{A}'$ satisfies the conditions of Lemma 31, and derive the communication lower bound.

We will show that any $w \in \mathbb{S}^{d-1}$ which classifies correctly a random point from $D_V$ with probability at least $1 - 3c_2$, satisfies $\|\mathrm{Proj}_V(w)\| < c/2$ (where $c$ is the constant from Lemma 31).

We will prove the contrapositive: that if $\|\mathrm{Proj}_V(w)\| \geq c/2$, then $w$ classifies a constant fraction of the points in $D_V$ incorrectly. Indeed, fix such $w$ and let $\alpha = \|\mathrm{Proj}_V(w)\|$. Note that if $x$ is drawn uniformly from $V$, Lemma 24 implies that $\sqrt{d}w^\top x$ is distributed approximately as a random variable $\mathcal{N}(0, \alpha^2)$. In particular, with constant probability, $w^\top x \geq \alpha/\sqrt{d} \geq c/(2\sqrt{d})$. For these values of $x$, $w^\top x_- > 0$, hence, $w$ classifies $(x_-, -1)$ incorrectly. This implies that $w$ classifies incorrectly a constant fraction of the points, namely, it classifies incorrectly a random point from $D_V$ with probability $3c_2$ of the points, if $c_2$ is sufficiently small. We conclude that any $w$ which classifies a random point from $D_V$ with probability at least $1 - 3c_2$, satisfies $\|\mathrm{Proj}_V(w) < c/2\|$.

Since $S_V$ is a $c_2$ approximate net for $D_V$, any $w$ which classifies a $(1-2c_2)$ fraction of the points in $S_V$ correctly, satisfies $\|\mathrm{Proj}_V(w)\| < c/2$. We derive that any $w$ which classifies $(1 - c_2)2m$ points correctly for the combined dataset $S_V \cup S_U$, satisfies $\|\mathrm{Proj}_V(w)\| < c/2$. For analogous reasoning, any such classifies satisfies $\|\mathrm{Proj}_U(w)\| < c/2$. Assuming that $\mathcal{A}$ outputs a hypothesis which classifies $(1 - c_2)2m$ points correctly, this implies that $\mathcal{A}'$ outputs $\hat{w}$ which satisfies $\|\mathrm{Proj}_V(w)\| + \|\mathrm{Proj}_U(w)\| \leq c$. From Lemma 31, it follows that the communication of $\mathcal{A}'$ is $\Omega(d^2)$. ■

### B.3. Proofs of the mathematical statements (Lemma 29 and Lemma 27)

#### B.3.1. PROOF OF LEMMA 29

We prove a result that is more general than Lemma 29.

**Lemma 32** *Let $g_1, g_2, \ldots, g_{(k_1+k_2)d}$ be independent random normal vectors $\mathcal{N}(0, I_d/d)$, where $c_0 d \leq k_1 d, k_2 d \leq d/2$ are integers. And Let $G$ be the matrix of size $(k_1 + k_2)d \times d$ that its $i^{th}$ row is $g_i$. Also set $U_1$ and $U_2$ be the bases of $\mathrm{span}\{g_1, \ldots, g_{k_1}\}$ and $\mathrm{span}\{g_{k_1+1}, \ldots, g_{(k_1+k_2)d}\}$ respectively.*

*Then, for all $t > 0$, with probability $1 - 2e^{-0.5\min\{k_1,k_2\}dt^2}$, all $v \in \mathbb{S}^{d-1}$ satisfy*

$$\left(1 + (1+t)\sqrt{\max\{k_1,k_2\}}\right)^{-2} \|Gv\|^2 \leq \left\|\mathrm{Proj}_{\mathrm{span}\{U_1\}}(v)\right\|^2 + \left\|\mathrm{Proj}_{\mathrm{span}\{U_2\}}(v)\right\|^2 \leq \left(1 - (1+t)\sqrt{\max\{k_1,k_2\}}\right)^{-2} \|Gv\|^2 .$$

*Or equivalently, in a matrix formulation*

$$\left(1 + (1+t)\sqrt{\max\{k_1,k_2\}}\right)^{-2} \|Gv\|^2 \leq \left\|\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} v\right\|^2 \leq \left(1 - (1+t)\sqrt{\max\{k_1,k_2\}}\right)^{-2} \|Gv\|^2 .$$

Observe that Lemma 29 follows when $k_1 = 1/2$ and $k_2 = 1/2 - 1/d$, and for $t$ that is small enough.

**Proof** Let $v \in \mathbb{S}^{d-1}$. Denote by $W_1 = \mathrm{span}\{g_1, \ldots, g_{k_1 d}\}$ and $W_2 = \mathrm{span}\{g_{k_1 d}, \ldots, g_{(k_1+k_2)d}\}$. Decompose $v$ in two different ways: $v = w_1 + w_1^\perp$ and to $v = w_2 + w_2^\perp$, where $w_i \in W_i$ and $w_i^\perp \in W_i^\perp$. Clearly,

$$\begin{aligned} \left\|\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} v\right\|^2 &= \|U_1 v\|^2 + \|U_2 v\|^2 = \left\|U_1(w_1 + w_1^\perp)\right\|^2 + \left\|U_2(w_2 + w_2^\perp)\right\|^2 \\ &= \|U_1 w_1\|^2 + \|U_2 w_2\|^2 = \left\|\mathrm{Proj}_{W_1}(w_1)\right\|^2 + \left\|\mathrm{Proj}_{W_2}(w_2)\right\|^2 , \end{aligned} \tag{7}$$

where we used the fact that $U_1, U_2$ are orthonormal bases. Similarly, split the rows of $G$ into two blocks with the same sizes as the number of rows of $U_1$ and $U_2$: $G = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}$. Similarly to Eq. (7),

$$\left\| \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} v \right\|^2 = \|G_1 v\|^2 + \|G_2 v\|^2 = \left\|G_1 \mathrm{Proj}_{W_1}(w_1)\right\|^2 + \left\|G_2 \mathrm{Proj}_{W_2}(w_2)\right\|^2, \tag{8}$$

where we use the fact that the span of the rows of $G_i$ equals to $U_i$. Now, in order to prove the lemma, we need to connect the last two equations. Observe that $G_1, G_2$ are singular matrices, however when we restrict them to operate on the span of their rows, the restricted linear operators have the singular values of $G_1^\top$ and $G_2^\top$ respectively (Claim 2). Thus, by Claim 1 it is enough to to bound the minimal and singular values of $G_1^\top$ and $G_2^\top$. By Corollary 18 applied to $G_1^\top$ and $G_2^\top$, the following holds for $t > 0$ and $i \in \{1, 2\}$:

$$1 - (1+t)\sqrt{k_i} \le \sigma_{\min}(G_i^\top) \le \sigma_{\max}(G_i^\top) \le 1 + (1+t)\sqrt{k_i},$$

with probability of at least $1 - 2e^{-0.5\min\{k_1,k_2\}dt^2}$. Thus by Eqs. (7) and (8) we derive that for all $v \in \mathbb{S}^{d-1}$,

$$\left(1 + (1+t)\sqrt{\max\{k_1,k_2\}}\right)^{-2} \le \frac{\left\| \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} v \right\|^2}{\|Gv\|^2} \le \left(1 - (1+t)\sqrt{\max\{k_1,k_2\}}\right)^{-2}.$$

and the the claim follows. ■

### B.3.2. PROOF OF LEMMA 27

**Note 1** *We will sometimes abuse notation as follows: given some subspace $U$, the same notation will be used to denote both the subspace and an arbitrary matrix whose rows form an orthonormal basis for the same subspace.*

We begin with a direct corollary of Lemma 29 and Lemma 19.

**Corollary 33** *Fix some $U_1 \in \mathrm{Gr}(d/2, d)$ and let $U_2$ be drawn uniformly from $\mathrm{Gr}(d/2, d)$. Fix some constant $0 < \eta < 1$, then with probability of at least $1 - e^{c(\eta)d}$ the top $(1-\eta)d$ singular values of $\begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$ are at least $c_1(\eta)$.*

Recall that by Lemma 15 for any $V_1, V_2 \in \mathrm{Gr}(d/2, d)$ we know that the chordal distance satisfies that $d(V_1, V_2) = d(V_1^\perp, V_2^\perp)$. We continue with another auxiliary lemma:

**Lemma 34** *Let $V_1^\perp$ and $V_2^\perp$ be any two vector spaces from $\mathrm{Gr}(d/2, d)$ that their distance is at least $\sqrt{\delta d/2}$. Let $W^\perp$ be a subspace drawn uniformly from the subspaces of $V_1^\perp$ of dimension $\eta d$, where $\eta(\delta) > 0$ is a sufficiently small constant. Then, with probability at least $1 - e^{-c_3\delta^2 d}$, any $w^\perp \in W^\perp \cap \mathbb{S}^{d-1}$ satisfies that $\|\mathrm{Proj}_{V_2}(w^\perp)\|_2^2 \ge \delta/16$.*

**Proof** Let $U_2$ be a matrix whose rows form an orthonormal base of $V_2^\perp$. By Pythagorean law it is enough to show that

$$\|U_2 w^\perp\|_2^2 \le 1 - \delta/16.$$

Let $v_i^{\perp,1}, \ldots, v_i^{\perp,d/2}$ be the orthonormal basis of $V_i^\perp$ for $i = 1, 2$ with respect to the decomposition according to the principal angles, namely, $v_i^{\perp,j}$ corresponds to $\theta_j$ for $i = 1, 2$, see Definition 14. Let $y$ be a random vector chosen uniformly from $V_1^\perp \cap \mathbb{S}^{d-1}$. Then

$$\begin{aligned}\|U_2 y\|^2 &= \sum_{i=1}^{0.5d} \langle v_2^{\perp,i}, y\rangle^2 = \sum_{i=1}^{0.5d} (\langle v_2^{\perp,i}, \sum_{j=1}^{0.d} \langle v_1^{\perp,j}, y\rangle v_1^{\perp,j}\rangle)^2 \\ &= \sum_{i=1}^{0.5d} \langle y, v_1^{\perp,i}\rangle^2 \langle v_2^{\perp,i}, v_1^{\perp,i}\rangle^2,\end{aligned}$$

where we used the fact that $\langle v_1^{\perp,i} v_2^{\perp,i}\rangle = 0$ for $i \neq j$.

Recall that $y \sim \mathrm{Unif}(V_1^\perp \cap \mathbb{S}^{d-1}) \sim \mathrm{Unif}(\mathbb{S}^{d/2})$, which implies that $\mathbb{E}_y[\langle y, v\rangle^2]$ is identical for all $v \in V_1^\perp \cap \mathbb{S}^{d-1}$. Moreover,

$$\mathbb{E}[\langle y, v_1^{\perp,i}\rangle^2] = \int_{\mathbb{S}^{d/2}} \langle y, e_1\rangle^2 d\sigma(y) = \frac{1}{d/2} \int_{\mathbb{S}^{d/2}} \sum_{i=1}^{d/2} \langle y, e_i\rangle^2 d\sigma(y) = \frac{1}{d/2} \int_{\mathbb{S}^{d/2}} \|y\|^2 d\sigma(y) = \frac{1}{d/2}.$$

Hence,

$$\begin{aligned}\mathbb{E}[\|U_2 y\|^2] &= \mathbb{E}[\langle y, v_1^{\perp,1}\rangle^2] \sum_{i=1}^{d/2} \langle v_2^{\perp,i}, v_1^{\perp,i}\rangle^2 \\ &= \frac{1}{d/2} \sum_{i=1}^{d/2} \cos(\theta_i)^2 = \frac{1}{d/2} \sum_{i=1}^{d/2} \left(1 - \sin^2(\theta_i)\right) = 1 - (2/d) d(V_1^\perp, V_2^\perp)^2 = 1 - \delta,\end{aligned}$$

where we used the fact that the distance between the subspaces is $\sqrt{\delta d/2}$. Using the fact that $\sqrt{1-\delta} \le 1 - \frac{\delta}{2}$, we derive

$$\mathbb{E}[\|U_2 y\|] \le \sqrt{\mathbb{E}[\|U_2 y\|^2]} \le \sqrt{1-\delta} \le 1 - \delta/2. \tag{9}$$

Now, since the rows of $U_2$ are orthonormal, its largest singular value is at most 1, therefore, $\|U_2 x\|$ is a 1- Lipschitz function. Hence, by Lemma 23, for any $\epsilon > 0$

$$\Pr_{y \in \mathbb{S}^{d-1} \cap V_1^\perp} (\|U_2 y\| \le \mathbb{E}[\|U_2 y\|] + \epsilon) \ge 1 - e^{-c_1 d \epsilon^2}.$$

Let

$$A_\delta := \{y \in \mathbb{S}^{d-1} \cap V_1^\perp : \|U_2 y\| \le 1 - \delta/16\}.$$

Using Eq. (9) and taking $\epsilon = \min\{c_1 \delta, \delta/4\}$, we derive that

$$\Pr_{y \in \mathbb{S}^{d-1} \cap V_1^\perp} (\|U_2 y\| \le 1 - \delta/4) \ge 1 - e^{-c_2 d \delta^2} \tag{10}$$

i.e. the measure of $A_\delta$ is at least $1 - e^{-c_2 d\delta^2}$. Informally speaking, if we show that "most" of the subspaces of $V_1^\perp \cap \mathbb{S}^{d-1}$ of dimension $\eta d$ lie in the set $A_\delta$, then we are done. For this purpose, we choose a $\delta/16$-net $\mathcal{N}$ of $W_0 \cap \mathbb{S}^{d-1}$, where $W_0$ is a fixed subspace in $V_1^\perp$ of dimension $\eta d$ (see Definition 20). By Lemma 21 we can assume that its size is bounded by

$$|\mathcal{N}| \leq \left(\frac{48}{\delta}\right)^{\eta d} \leq e^{-\ln(\delta)\eta d + \ln(32)d} \leq e^{C\ln(\delta^{-1})\eta d}.$$

Now, let $W^\perp$ be defined as in this lemma (a uniform random subspace of $V_1^\perp$). It can be written as $UW_0$ for a random uniform rotation $U$ on $V_1^\perp$. Note that $\mathcal{N}_{W^\perp} := U\mathcal{N}$ is a $\delta/16$-net of $W^\perp \cap \mathbb{S}^{d-1}$. Notice that this net is a random set of points, and moreover, each point is distributed uniformly on $\mathbb{S}^{d-1} \cap V_1^\perp$.

Now we set $\eta = c_2\delta^2 \ln(\delta^{-1})^{-1}$ for some small enough $c_2$. Now, in order to prove this Lemma we first estimate the probability that all the points in $\mathcal{N}_{W^\perp}$ lie in $A_\delta$. Using the union bound and Eq. (10), we derive that

$$\Pr_{W^\perp}[\forall x \in \mathcal{N}_{W^\perp} : x \in A_\delta] \geq 1 - |\mathcal{N}_{W^\perp}| \Pr_{y \in \mathbb{S}^{d-1} \cap V_1^\perp}[y \notin A_\delta] \geq 1 - |\mathcal{N}_{W^\perp}| e^{-c_2\delta^2 d} \geq 1 - e^{-c_3\delta^2 d}.$$

where we used the fact that each point in the net distributed uniformly on $\mathbb{S}^{d-1} \cap V_1^\perp$. Finally, by Lemma 22 with $\epsilon = \delta/16$ we derive that

$$\Pr_{W^\perp}\left[\forall y \ \in W^\perp \colon \|U_2 y\| \leq 1 - \delta/16\right] \geq 1 - e^{-c_2 n\delta^2}.$$

Thus the claim follows for $c(\delta) = \delta/16$ and for $\eta = c_2\delta^2 \ln(\delta^{-1})^{-1}$. ■

Finally, we conclude the proof:

**Proof** [Proof of Lemma 27] Let $u \in \mathbb{S}^{d-1}$ and let $\eta(\delta) > 0$ be a fixed constant such that Lemma 34 is valid. Denote by $W$ the subspaces of the top $(1 - \eta(\delta))d$ singular vectors of $\begin{pmatrix} V_1 \\ U \end{pmatrix}$. Also from Corollary 33, note that the top $(1 - \eta(\delta))d$ singular values of $\begin{pmatrix} V_1 \\ U \end{pmatrix}$ are greater than some constant $c_5$. Write $u$ as $u = w + w^\perp$, where $w \in W$ and $w^\perp \in W^\perp$. If $\|w\|^2 > c$, for some constant that will be defined later, then

$$\left\| \begin{pmatrix} V_1 \\ U \end{pmatrix} u \right\| \geq c_5 \cdot \sqrt{c}, \tag{11}$$

and we are done. Otherwise, $\|w^\perp\|^2 \geq 1 - c$. Then, if $\|\mathrm{Proj}_{V_1}(w^\perp)\|^2 \geq (1 - c) \cdot c$, the following holds:

$$\begin{aligned} \left\| \begin{pmatrix} V_1 \\ U \end{pmatrix} u \right\|^2 &= \left\| \begin{pmatrix} V_1 \\ U \end{pmatrix} w \right\|^2 + \left\| \begin{pmatrix} V_1 \\ U \end{pmatrix} w^\perp \right\|^2 \geq \left\| \begin{pmatrix} V_1 \\ U \end{pmatrix} w^\perp \right\|^2 \geq \left\| \mathrm{Proj}_{V_1}(w^\perp) \right\|^2 \\ &\geq (1 - c) \cdot c, \end{aligned} \tag{12}$$

and we are done. The last option is that $\|w^\perp\|^2 \geq 1 - c$ and $\|\mathrm{Proj}_{V_1^\perp}(w^\perp)\|^2 \geq (1 - c)^2$ (or equivalently $\|\mathrm{Proj}_{V_1}(w^\perp)\|^2 \leq (1 - c) \cdot c$). Now, we project the subspace $W^\perp$ on $V_1^\perp$, and denote the new subspace as $E$. Since the subspace $U$ was chosen uniformly, then clearly $E$ is a uniform

subspace of $V_1^{\perp}$ ($V_1$ is a fixed subspace). From Lemma 34, with probability $1 - e^{-c(\delta)d}$, any $e \in E \cap \mathbb{S}^{d-1}$ satisfies: $\|\mathrm{Proj}_{V_2}(e)\| \geq c(\delta)$, and assume for the rest of the proof that this holds. Since $\|\mathrm{Proj}_{V_1^{\perp}}(w^{\perp})\|^2 \geq (1-c)^2$, we also know that $\|\mathrm{Proj}_{V_2}(w^{\perp})\|^2 \geq (1-c)^2 c(\delta)$. Finally,

$$\begin{aligned} \left\| \begin{pmatrix} V_2 \\ U \end{pmatrix} u \right\| = \left\| \begin{pmatrix} V_2 \\ U \end{pmatrix} (w + w^{\perp}) \right\| &\geq \left\| \begin{pmatrix} V_2 \\ U \end{pmatrix} w^{\perp} \right\| - \sqrt{c} \\ &\geq \left\| \mathrm{Proj}_{V_2}(w^{\perp}) \right\| - \sqrt{c} = (1-c)^2 c(\delta) - \sqrt{c}, \end{aligned} \tag{13}$$

By Eqs. (11), (12), (13) choose $c = \min\{0.01, c(\delta)^4\}$ and the claim follows. ■