Electronic Colloquium on Computational Complexity, Report No. 276 (2026)

Learning Positive Linear Combinations of ReLUs in Fully Polynomial Time

Ilias Diakonikolas[†]
UW Madison
ilias@cs.wisc.edu

Jelena Diakonikolas[‡]
UW Madison
jelena@cs.wisc.edu

Puqian Wang[§]
UW Madison
pwang333@wisc.edu

Nikos Zarifis[¶]
CSAIL, MIT
zarifis@mit.edu

Abstract

We study the fundamental problem of PAC learning positive linear combinations of k arbitrary ReLU activations on $\mathbb{R}^d$ with respect to the Gaussian distribution, in L_2 loss. We give the first fully-polynomial time algorithm for this learning task. Specifically, our algorithm draws $\mathrm{poly}(d, k, 1/\epsilon)$ samples, runs in $\mathrm{poly}(d, k, 1/\epsilon)$ time, and outputs a hypothesis that is ϵ-close to the target concept with high probability. All prior algorithms for this problem require super-polynomial time in k or $1/\epsilon$, or leverage strong well-conditioned assumptions on the weight matrix.

On a technical level, our approach develops a novel boosting procedure of a "weak" learner that recovers the hidden weight vectors to constant error, in a list-decodable manner. To develop this "weak" learning procedure, we leverage implicit tensor computation techniques to construct a low-degree filtering procedure based on Hermite tensors of the target function. This procedure amplifies separation among (clusters of) weight vectors to learn each of them to constant error. We then argue that the same procedure can be used for improving the estimation of hidden weight vectors, using gradient descent-like updates on the Euclidean sphere. A crucial ingredient of our algorithm is a pruning approach that ensures only relevant vectors that approximate the hidden weight vectors at target accuracy are maintained by the algorithm.

[*]Authors are listed in alphabetical order.

[†]Supported by NSF Medium Award CCF-2107079, ONR Award number N00014-25-1-2268, an H.I. Romnes Faculty Fellowship and a Guggenheim Fellowship.

[‡]Supported in part by the AFOSR YIP award FA9550-24-1-0076, by the NSF MFAI Award DMS-2502282, and by the NSF CAREER Award CCF-2440563. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Defense.

[§]Supported in part by the AFOSR YIP award FA9550-24-1-0076.

[¶]Supported by MIT METEOR fellowship. Some of this research was done while the author was at UW Madison, supported by NSF Medium Award CCF-2107079.

ISSN 1433-8092

Contents

1 Introduction **1**
- 1.1 Problem Formulation and Main Result . . . 1
- 1.2 Technical Overview . . . 2
  - 1.2.1 Coarse Direction Recovery . . . 3
  - 1.2.2 Boosting Coarse Direction Recovery . . . 5
- 1.3 Related Work . . . 7
- 1.4 Concurrent Work and AI Use . . . 9
- 1.5 Paper Organization . . . 9

2 Notation and Preliminaries **9**

3 Optimization via Coarse Direction Recovery, Refinement, and Pruning **11**
- 3.1 Learning λ-Heavy Hidden Weight Vectors . . . 12
  - 3.1.1 Boosting Hidden Vector Estimates . . . 12
  - 3.1.2 Closeness Tester . . . 16
  - 3.1.3 Covering and Pruning . . . 19
- 3.2 Proof of Theorem 3.1 . . . 23

4 Efficient Coarse Direction Recovery **24**
- 4.1 Low-Degree Filtering: Revealing Hidden Directions . . . 24
- 4.2 Warm-up: Quasi-Polynomial Time CDR Algorithm . . . 26
- 4.3 Polynomial Time CDR via Implicit Tensor Computation . . . 32
  - 4.3.1 Determining the Probability of the Probing Event . . . 33
  - 4.3.2 Estimating the Target Matrix Using Implicit Tensor Computation . . . 34
- 4.4 Proof of the Main Result of the Section: Polynomial Complexity CDR . . . 47

5 Putting Everything Together: Proof of Theorem 1.3 **47**

1 Introduction

Understanding the computational complexity of learning neural networks is a central question in the foundations of machine learning. Over the past decade, a substantial body of work has developed provable learning algorithms under various structural and distributional assumptions; see, e.g., [JSA15, ZLJ16, ZSJ+17, GLM18, GK19, CKM22, WZDD23, ZWDD25, DIKR25, DLB25]. Despite this progress, a basic question has remained open even for networks with a single hidden layer: Can positive linear combinations of ReLUs be learned in polynomial time under Gaussian inputs, without strong structural assumptions on the hidden weights? This longstanding question is explicitly highlighted, for example, by Chen, Klivans, and Meka [CKM22] (see also Chen's thesis [Che21]):

> "Ideally, we would like an algorithm with sample complexity and running time that is polynomial in all the relevant parameters. Even for learning arbitrary sums of ReLUs, [...], it remains **a major open question** to obtain a polynomial-time algorithm [...]."

Positive linear combinations of ReLUs are arguably the simplest extension of learning a single neuron—the setting of generalized linear models. A fully polynomial-time algorithm for this basic class is an important step towards understanding the learnability of more general networks. Moreover, the techniques developed for this problem may also be useful for learning other latent variable models.

In this paper, we study the task of PAC learning positive linear combinations of at most k ReLUs on $\mathbb{R}^d$ under the standard Gaussian distribution. Specifically, the learner has access to independent labeled examples from an unknown function in the concept class, and its goal is to output a hypothesis that approximates the target in L_2 norm. The relevant parameters are the dimension d, the number of hidden units k, and the inverse error $1/\epsilon$. It is well-known that polynomially many samples information-theoretically suffice for this learning task, ignoring computational complexity. The challenge is to develop a computationally efficient learning algorithm: despite a sequence of algorithmic advances, the best previous running-time bounds were super-polynomial in at least one of these parameters.

The work of [DKKZ20] gave the first learner for this problem with running time polynomial in d and $1/\epsilon$ for every fixed k. Their algorithm uses $\widetilde{O}(d)\mathrm{poly}(k/\epsilon)$ samples and runs in time $\mathrm{poly}(d/\epsilon)+(k/\epsilon)^{O(k^2)}$. The work of [DK20] improved the running time using a structural theorem on small covers for near-zero sets of polynomials. Its sample and computational complexities are $\mathrm{poly}(dk/\epsilon) + (k/\epsilon)^{O(\log^2 k)}$, i.e., quasi-polynomial in k even for constant error. More recently, [DK25] developed an algorithm based on implicit high-order moment estimation with sample and computational complexity $\mathrm{poly}(d,k)\, 2^{\mathrm{poly}(1/\epsilon)}$. That is, while this achieves polynomial dependence on d and k, the dependence on $1/\epsilon$ is exponential. In summary, neither of these results gives a fully-polynomial time algorithm.

The above discussion motivates the following question:

Can positive linear combinations of k ReLUs on $\mathbb{R}^d$ be learned under Gaussian inputs with sample and computational complexity $\mathrm{poly}(d,k,1/\epsilon)$*?*

As our main contribution, we answer this question in the affirmative. Specifically, we give a learning algorithm with sample and computational complexity $\mathrm{poly}(d,k,1/\epsilon)$ that works for arbitrary positive linear combinations of ReLUs. Our algorithm outputs a simple interpretable hypothesis. Moreover, the technical tools we develop in the process may be useful for learning other latent variable models.

1.1 Problem Formulation and Main Result

Before stating our main result, we define the underlying concept class and the PAC learning task.

Definition 1.1 (Positive linear combinations of ReLUs)**.** *Let $\sigma(t) \coloneqq \max\{0,t\}$ denote the ReLU activation. For $d,k \in \mathbb{Z}_+$, let $\mathcal{C}_{d,k}$ denote the class of functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form $f(\mathbf{x}) = \sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$, where $\beta_i \in [0,1]$ and $\|\mathbf{w}_i\|_2 = 1$ for all $i \in [k]$. We call the $\mathbf{w}_i$'s the* hidden weight vectors *and the β_i's the* coefficients.

By positive homogeneity of the ReLU, the norm of a nonzero hidden weight vector can be absorbed into its coefficient. The constraint $\beta_i \le 1$ fixes the scale (without loss of generality); see Remark 1.5. For a function $g : \mathbb{R}^d \to \mathbb{R}$, we write $\|g\|_{L_2} \coloneqq (\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[g(\mathbf{x})^2])^{1/2}$.

Definition 1.2 (PAC learning $\mathcal{C}_{d,k}$ under Gaussian marginals)**.** *The learner is given accuracy and confidence parameters $\epsilon, \delta \in (0,1)$ and access to independent examples $(\mathbf{x}, y) \sim \mathcal{D}$, where $\mathcal{D}$ is a distribution supported on $\mathbb{R}^d \times \mathbb{R}$ such that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $y = f(\mathbf{x})$ for an unknown $f \in \mathcal{C}_{d,k}$. The goal of the learner is to output a hypothesis $h : \mathbb{R}^d \to \mathbb{R}$ that, with probability at least $1 - \delta$, satisfies*

$$\|h - f\|_{L_2} \leq \epsilon \|f\|_{L_2}.$$

The hypothesis need not belong to $\mathcal{C}_{d,k}$. The learner is called proper *if it outputs a hypothesis in $\mathcal{C}_{d,k}$.*

Our main result is summarized in the following theorem (see Theorem 5.1 for a detailed full version).

Theorem 1.3 (Main result, informal)**.** *There is an algorithm that, given $d, k \in \mathbb{Z}_+$, $\epsilon \in (0,1)$, and access to independent examples $(\mathbf{x}, f(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $f \in \mathcal{C}_{d,k}$, uses $\mathrm{poly}(d, k, 1/\epsilon)$ samples and time and outputs a hypothesis h such that, with probability at least $2/3$,*

$$\|h - f\|_{L_2} \leq \epsilon \|f\|_{L_2}.$$

A few remarks are in order. While our learner is not proper, its hypothesis has a simple form: a linear function plus a linear combination of at most k absolute-value functions, i.e., $h(\mathbf{x}) = \mathbf{g} \cdot \mathbf{x} + \sum_{i=1}^{k} \alpha_i |\mathbf{v}_i \cdot \mathbf{x}|$. Equivalently, using $|t| = \sigma(t) + \sigma(-t)$, it is a linear function plus a linear combination of at most $2k$ ReLUs. Thus, the hypothesis is itself a small network with $O(k)$ nonlinear units. For comparison, the improper learner from [DK20] outputs a positive linear combination of quasi-polynomially many ReLUs, while the learner from [DK25] outputs a polynomial represented through compressed Hermite coefficients. For these representations, the bounds on the time needed to even *evaluate* the hypothesis on a single new example are also super-polynomial in k or $1/\epsilon$.

Remark 1.4 (Additive label noise)**.** More generally, one may consider examples $(\mathbf{x}, y)$ with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $y = f(\mathbf{x}) + \xi$, where ξ is independent of $\mathbf{x}$, has mean zero, and satisfies $\mathbf{E}[\xi^2] \leq \sigma_\xi^2$. Following the convention of [DKKZ20], the objective in this setting is to output a hypothesis h satisfying

$$\|h - f\|_{L_2}^2 \leq \epsilon^2 \big(\|f\|_{L_2}^2 + \sigma_\xi^2 \big)$$

with high probability. The realizable setting of Definition 1.2 corresponds to $\sigma_\xi = 0$. Our algorithm straightforwardly extends to this setting with a $\mathrm{poly}(\sigma_\xi)$ overhead in its complexity.

Remark 1.5 (Normalization)**.** Our result applies to arbitrary positive linear combinations of at most k ReLUs, without bounds on the coefficients or the norms of the hidden weight vectors. By positive homogeneity, we can absorb the norm of each hidden weight vector into its coefficient. Writing b for the resulting coefficient sum, we have $\mathbf{E}[f(\mathbf{x})] = \frac{b}{\sqrt{2\pi}}$ and $\mathbf{E}[f(\mathbf{x})^2] \leq \frac{b^2}{2}$. By Chebyshev's inequality, the empirical mean $\overline{y}$ of a sufficiently large constant number of labels gives an estimate $B = 2\sqrt{2\pi}\,\overline{y}$ satisfying $b \leq B \leq 3b$ with high constant probability. For $B > 0$, we can apply our learner to fresh examples with labels divided by B and multiply its output by B. The normalized target has unit hidden weight vectors and coefficients in $[0,1]$, and rescaling preserves the relative-error guarantee. If $B = 0$, we return the zero hypothesis. Thus, our normalization is without loss of generality and requires no prior knowledge of the target's scale.

Terminology. Throughout this paper, we will say that a distribution $(\mathbf{x}, y) \sim \mathcal{D}$ supported on $\mathbb{R}^d \times \mathbb{R}$ is a *Gaussian sum of k ReLUs* if the $\mathbf{x}$-marginals are distributed according to the standard normal distribution, i.e., $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and if $y(\mathbf{x}) = \sum_{i=1}^{k} \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$, where σ is the ReLU function, $\beta_i \in [0,1]$, and $\|\mathbf{w}_i\|_2 = 1$, for all $i \in [k]$.

1.2 Technical Overview

To explain the main ideas behind our results, we begin with the following observations. First, low-coefficient ReLU components in the sum have little influence on the predicted label: in particular, if the coefficient β_i associated with component i is smaller than $\sqrt{\epsilon}/k$, then zeroing out all such components can contribute at most an additive ϵ to the L_2^2 loss we aim to minimize. Thus, setting a threshold $\epsilon_0 \approx \sqrt{\epsilon}/k$ and focusing on approximating the ReLU components with coefficients above this threshold, comes at no loss of generality. In the remainder of the technical overview, we only consider such nontrivial components with coefficients $\beta_i \geq \epsilon_0$, and refer to them as "ϵ_0-heavy" in the rest of the paper.

Second, we claim that it suffices to recover the target vectors up to their orientation (i.e., up to their signs). The reason is the decomposition $\sigma(z) - \frac{z}{2} = \frac{|z|}{2}$. Specifically, if we calculate the first-moment parameter $\mathbf{g} = \mathbf{E}[y\mathbf{x}] = \frac{1}{2}\sum_{i=1}^{k} \beta_i \mathbf{w}_i$, we get

$$y - \mathbf{g} \cdot \mathbf{x} = \frac{1}{2} \sum_{i=1}^{k} \beta_i |\mathbf{w}_i \cdot \mathbf{x}| \ . \tag{1}$$

Thus, after subtracting the linear part, the remaining function is a sum of absolute values. Its terms are unchanged when $\mathbf{w}_i$ is replaced by $-\mathbf{w}_i$, so the directional part of the algorithm only needs to recover one of these two orientations. We measure the error of the angle using $\theta_{\pm}(\mathbf{v}, \mathbf{w}_i) = \min\{\theta(\mathbf{v}, \mathbf{w}_i), \pi - \theta(\mathbf{v}, \mathbf{w}_i)\}$, rather than by the ordinary angle for this reason.

The birds-eye view of our algorithm is the following. The algorithm operates in rounds. Each round $t = \{1, 2, \ldots, T\}$ of computation maintains the following invariant:

Algorithm Invariant. There is a set of vectors $\mathcal{V}_t$, $|\mathcal{V}_t| \leq k$, maintained by the algorithm such that for every ϵ_0-heavy hidden vector $\mathbf{w}_i$, there exists a vector $\mathbf{v}_i \in \mathcal{V}_t$ that is close to it; in particular, $\theta_{\pm}(\mathbf{v}_i, \mathbf{w}_i) \leq R_t$, $R_t > 0$.

Note that this mapping of vectors need not be one-to-one: one vector $\mathbf{v}$ can be close to multiple $\mathbf{w}_i$'s. Initially, $R_0 \in (0, 1)$ is a constant, and this is achieved using the Coarse Direction Recovery (CDR) subroutine described in Section 1.2.1 below. We argue that we can maintain the stated invariant with $R_{t+1} \leq \rho R_t$ for a constant $\rho \in (0, 1)$ and all $t \geq 0$, so that the algorithm approximates all ϵ_0-heavy hidden weight vectors $\mathbf{w}_i$ with error $\theta_{\pm}(\mathbf{v}_i, \mathbf{w}_i) \leq \sqrt{\epsilon}/k$, after $T = O(\log_{\rho}(k/\epsilon))$ rounds. The last step is then to estimate the coefficients for the absolute-sum representation (1). It is not hard to see that this part boils down to solving a linear regression problem of the form

$$\min_{\boldsymbol{\alpha} \in [0,1]^m} \mathop{\mathbf{E}}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\left(y - \mathbf{g} \cdot \mathbf{x} - \frac{1}{2} \sum_{j=1}^{m} \alpha_j |\mathbf{v}_j \cdot \mathbf{x}| \right)^2 \right] ,$$

where $m = |\mathcal{V}_T| \leq k$. This problem is convex and it is not hard to argue that it is efficiently solvable in terms of both sample and computational complexity.

The technical crux of the approach thus lies in implementing the rounds of the above high-level algorithm, which relies upon multiple core technical components described below.

1.2.1 Coarse Direction Recovery

A key technical tool in our algorithmic approach is a procedure that, given a failure probability $\delta_{\mathsf{CDR}} \in (0, 1)$ and access to i.i.d. examples $(\mathbf{x}, y(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $y(\mathbf{x}) = f(\mathbf{x})$ for a fixed but unknown $f \in \mathcal{C}_{d,k}$, outputs a polynomially-sized list in $k, 1/\delta_{\mathsf{CDR}}$ with the following property: with probability at least $1 - \delta_{\mathsf{CDR}}$, for each ϵ_0-heavy hidden weight vector $\mathbf{w}_i$ defining the labeling function, there exists at least one list element $\mathbf{r}_i$ such that $|\mathbf{r}_i \cdot \mathbf{w}_i| \geq 0.99$. Crucially, this procedure, to which we refer as the Coarse Direction Recovery (CDR) can be implemented using polynomially many labeled examples (in $d, k, 1/\epsilon_0, 1/\delta_{\mathsf{CDR}}$), and in time polynomial in d and the number of samples. CDR can be understood as a particularly "weak"[1] learner for our problem, where each ϵ_0-heavy hidden weight vector is learned to *constant* error, in a *list-decodable* manner. The latter means that the output provided by CDR is a polynomially-sized list containing useful, constant-error outputs, but without any information on how to separate useful outputs from the rest of the list. It is not hard to argue that a stronger learner, which would only output vectors $\mathbf{r}_i$ with sufficiently strong correlation $|\mathbf{r}_i \cdot \mathbf{w}_i| \geq 1 - \Theta(\sqrt{\epsilon}/k)$, would solve our learning problem stated in Definition 1.2. Yet, with the techniques described below, such a stronger requirement would lead only to a quasi-polynomial algorithm, at best.

The first insight for developing the CDR subroutine comes from the properties of Hermite tensors of positive linear combinations of ReLUs (see Section 2 for relevant definitions); namely, that Hermite moments of a linear combination of activations can be expressed as a sum of rank-one tensors, an

[1] Here, "weak" is in quotes to distinguish it from the standard notion in PAC learning theory, where weak learning is understood as slightly better than random, whereas we require constant correlation of vectors.

observation used extensively in prior work [DKKZ20, CDG+23, DK25]. More specifically, for a positive linear combination of ReLUs, we have

$$\mathbf{Q}_{2p+2} := \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{He}_{2p+2}(\mathbf{x})] \propto \sum_{i=1}^{k} \beta_i \mathbf{w}_i^{\otimes 2p+2} \ , \tag{2}$$

where $\mathbf{He}_{2p+2}(\mathbf{x})$ is the Hermite tensor of order $2p+2$, for $p \geq 0$. The moment formulation here is what enables estimation from samples; the fact that it is proportional to the sum of rank-one tensors $\sum_{i=1}^{k} \beta_i \mathbf{w}_i^{\otimes 2p+2}$ is what we use to obtain the guarantee of CDR.

Directly recovering the individual rank-one components of a high-order tensor is, in general, a computationally challenging problem. Instead, a useful approach is to contract the tensor along all but two indices, obtaining a matrix whose leading eigenvector can reveal one of the hidden directions. To be precise: taking the inner product of Equation (2) with a tensor $\mathbf{u}^{\otimes 2p}$ and reshaping the result as a matrix, we get that

$$\mathbf{Q}_2(\mathbf{u}) = \left\langle \mathbf{Q}_{2p+2}, \mathbf{u}^{\otimes 2p} \right\rangle \propto \sum_{i=1}^{k} \beta_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p} \mathbf{w}_i \mathbf{w}_i^\top \ . \tag{3}$$

The key observation is that the coefficients of the rank-one matrices in Equation (3) depend on *powers of the correlations*, $|\mathbf{w}_i \cdot \mathbf{u}|^{2p}$. Consequently, even a small multiplicative advantage in the correlation with one hidden weight vector can be amplified by choosing a sufficiently large degree p. To illustrate this idea, suppose that $\mathbf{w}_1$ is well-separated from the remaining hidden vectors $\mathbf{w}_2, \ldots, \mathbf{w}_k$ (e.g., $\max_{i \neq 1} |\mathbf{w}_1 \cdot \mathbf{w}_i| \leq 1/2$). Then, we can write $\mathbf{Q}_2(\mathbf{u})$ as a sum of two matrices $\mathbf{Q}_2(\mathbf{u}) \propto \beta_1 |\mathbf{w}_1 \cdot \mathbf{u}|^{2p} \mathbf{w}_1 \mathbf{w}_1^\top + \mathbf{Q}_\epsilon(\mathbf{u})$. If a vector $\mathbf{u}$ is such that the spectral norm of $\beta_1 |\mathbf{w}_1 \cdot \mathbf{u}|^{2p} \mathbf{w}_1 \mathbf{w}_1^\top$ is much larger than the spectral norm of $\mathbf{Q}_\epsilon(\mathbf{u})$, then the top eigenvector of $\mathbf{Q}_2(\mathbf{u})$ must have constant correlation with $\mathbf{w}_1$. A sufficient condition for this to hold is to have $\beta_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p} \gg \sum_{j \neq i}^{k} \beta_j |\mathbf{w}_j \cdot \mathbf{u}|^{2p}$. Given that we can choose p to be a (sufficiently) large power, it suffices to find a $\mathbf{u}$ so that for all $i \neq 1$, we have $|\mathbf{u} \cdot \mathbf{w}_1| \geq (1+\xi)|\mathbf{u} \cdot \mathbf{w}_i|$ for some $\xi > 0$ (for simplicity of this exposition, assume $\xi = 1$). Then, by choosing $p \gg \log(k \max_i \beta_i / \beta_1)$, we have that the spectral norm of $\beta_1 |\mathbf{w}_1 \cdot \mathbf{u}|^{2p} \mathbf{w}_1 \mathbf{w}_1^\top$ is at least two times larger than the spectral norm of $\mathbf{Q}_\epsilon(\mathbf{u})$. Hence, the top eigenvector of $\mathbf{Q}_2(\mathbf{u})$ is almost parallel to $\mathbf{w}_1$. Importantly, this intuition extends beyond the well-separated case. Indeed, directions $\mathbf{w}_i$ that are close to $\pm\mathbf{w}_1$, i.e., $|\mathbf{w}_i \cdot \mathbf{w}_1| \gg 0$, can be grouped together with $\mathbf{w}_1$ while retaining the same properties, since their contributions align with the desired rank-one matrix. Therefore, we only need to suppress directions that are far from $\pm\mathbf{w}_1$, rather than isolate a single component. The hope here is that by sampling $\mathbf{u}$ from an appropriately selected probability distribution, the above event would happen for all $\mathbf{w}_i$, with some nontrivial probability. This is the main intuition for why the stated guarantee of CDR should be possible.

There are at least two main technical challenges with the above argument. First, since $\mathbf{u}$ is chosen from a probability distribution, to ensure that we can produce a polynomially-sized list containing at least one vector that aligns well with some $\mathbf{w}_i$, the event ensuring separation in alignment among hidden weight vectors with $\mathbf{u}$ described above must happen with probability no smaller than an inverse polynomial in the problem parameters. (Ensuring that such an event is satisfied for all $\mathbf{w}_i$, $i \in [k]$, for at least one of the drawn probing vectors $\mathbf{u}$ follows by the union bound.) We argue that this is indeed the case for the simple distribution where $\mathbf{u}$ is chosen uniformly at random from the k-dimensional unit Euclidean sphere, defined over the space containing the space spanned by the targets $\mathbf{w}_i$, $i \in [k]$. The same can also be argued for other related distributions, such as standard Gaussians.

Second, as hinted before and more importantly, we need to argue that the matrix in (3) can be estimated efficiently from labeled examples with bounded error in the operator norm, so the top eigenvector is preserved (up to constant error). For example, when $\mathbf{u}$ is sampled uniformly from a k-dimensional sphere, its typical correlation with a fixed unit vector is of the order $1/\sqrt{k}$. Raising this correlation to degree $2p$ produces a factor of order k^{-p}, which is quasipolynomially small when $p = \Theta(\log(k/\epsilon))$. Estimating the resulting matrix may therefore require quasipolynomially small accuracy, and this, as a result, leads to quasipolynomial sample complexity. We provide full technical details for this approach in Section 4.2.

To obtain a fully polynomial CDR implementation, the main challenge is to ensure that the estimation task can be implemented efficiently, in terms of both samples and time. A major obstacle when sampling $\mathbf{u}$ uniformly from the sphere or from the standard Gaussian distribution is that the variance of the empirical estimate of $\mathbf{Q}_2 = \left\langle \mathbf{Q}_{2p+2}, \mathbf{u}^{\otimes 2p} \right\rangle$ is large because the variance of the tensor $\mathbf{u}^{\otimes 2p}$ is large

for large degree p, which leads to error amplification. To illustrate this point, consider $p = 1$ and decompose the estimation tensor into $\mathbf{Q}_4 \in (\mathbb{R}^d)^{\otimes 4}$ and an error tensor $\zeta\mathbf{H}$, $\widehat{\mathbf{Q}}_4 = \mathbf{Q}_4 + \zeta\mathbf{H}$. For the sake of discussion, suppose $\mathbf{H} = \mathbf{I}_d \otimes \mathbf{I}_d$. Then the error term after tensor contraction with the filter $\mathbf{u}^{\otimes 2}$ becomes $\langle \mathbf{H}, \mathbf{u}^{\otimes 2} \rangle = (\mathbf{u}^\top \mathbf{u})\mathbf{I}_d$. If $\mathbf{u}$ is selected from a real Gaussian distribution, then the second moment of the spectral norm of this error matrix is $\mathbf{E}[\|\langle \mathbf{H}, \mathbf{u}^{\otimes 2} \rangle\|_2^2] = \mathbf{E}[\|\mathbf{u}\|_2^4] = O(d^2)$. On the other hand, if $\mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$ is a complex Gaussian vector where $\mathbf{p}, \mathbf{q}$ are i.i.d. standard real Gaussians, then the second order moment becomes $\mathbf{E}[\|\langle \mathbf{H}, \mathbf{u}^{\otimes 2} \rangle\|_2^2] = O(d)$. This is the main intuition for why complex vectors are more helpful for keeping the estimation error small. The advantage of complex vectors leading to lower growth of the error also persists for higher moments; this is formally established for real and complex Gaussian vectors in Lemma 4.26 and Lemma 4.27, in Section 4. The choice of Gaussian vectors in our analysis is for technical convenience. Full technical details for this polynomial sample and time implementation based on complex Gaussians are included in Section 4.3.

1.2.2 Boosting Coarse Direction Recovery

The CDR procedure described in Section 1.2.1 gives a concrete approach for constructing initial, constant-correlation candidate vectors. However, several questions remain on how to make this procedure useful for constructing a hypothesis with arbitrarily small target error $\epsilon > 0$. Most importantly, we need a way to reduce the error of candidate vectors and isolate useful candidate vectors from the rest of the produced list. We begin with providing the intuition for how to improve the error.

Ideal (and nearly-ideal) update directions. Fix one candidate vector $\mathbf{v} \in \mathbb{S}^{d-1}$ and one target $\mathbf{w}_i \in \mathbb{S}^{d-1}$. If we knew $\mathbf{w}_i^{\perp\mathbf{v}}$, the component of $\mathbf{w}_i$ orthogonal to $\mathbf{v}$, then we would immediately know $\mathbf{w}_i$, since $\mathbf{w}_i = \sqrt{1 - \left\|\mathbf{w}_i^{\perp\mathbf{v}}\right\|_2^2}\mathbf{v} + \mathbf{w}_i^{\perp\mathbf{v}}$, so one update of $\mathbf{v}$ would suffice. Notice that $\mathbf{w}_i^{\perp\mathbf{v}}$ corresponds to the "shortest path" direction for rotating $\mathbf{v}$ towards $\mathbf{w}_i$: rotating $\mathbf{v}$ in the direction of $\mathbf{w}_i^{\perp\mathbf{v}}$ leads it to $\mathbf{w}_i$, following the geodesic that connects $\mathbf{v}$ and $\mathbf{w}_i$. Of course, such an ideal update direction is not available to the algorithm. However, it is also unnecessary for contracting the distance between $\mathbf{v}$ and $\mathbf{w}_i$. If, instead, we had access to a unit vector $\mathbf{r}$ that correlates well with $\mathbf{w}_i^{\perp\mathbf{v}}$, in the sense that $\mathbf{r} \cdot \frac{\mathbf{w}_i^{\perp\mathbf{v}}}{\left\|\mathbf{w}_i^{\perp\mathbf{v}}\right\|_2} \geq 0.9$, then $\mathbf{r}$ would be a "nearly-ideal" update direction: although not moving $\mathbf{v}$ on the geodesic connecting it to $\mathbf{w}_i$, it would allow for a large rotation that contracts the angle between $\mathbf{v}$ and $\mathbf{w}_i$ by a constant factor; this is the primary idea underlying efficient optimization on the Euclidean sphere [WZD26]. Of course, the question remains of how to construct such update directions.

Observe that the "constant correlation" requirement $\mathbf{r} \cdot \frac{\mathbf{w}_i^{\perp\mathbf{v}}}{\left\|\mathbf{w}_i^{\perp\mathbf{v}}\right\|_2} \geq 0.9$ is similar to what we had stated for CDR. In particular, if we had access to samples from a problem instance with $\widehat{\mathbf{w}_i^{\perp\mathbf{v}}} := \frac{\mathbf{w}_i^{\perp\mathbf{v}}}{\left\|\mathbf{w}_i^{\perp\mathbf{v}}\right\|_2}$ as the hidden weight vectors, CDR would apply directly. We argue that this is indeed possible; the small loss in correlation (0.99 vs 0.9) is a small error slack needed for the technical argument. Methodologically, the use of moments beyond the first to construct update directions that enable contracting the distance to target vectors is related to [WZDD25]; in the present work, the use of higher moments in (2) is crucial to updating a set of $k \geq 1$ vectors *simultaneously* (in [WZDD25], $k = 1$, and $p = 0$ in (2) suffices).

To argue that CDR could likewise apply to estimating $\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}$, in Section 3.1.1, we show that a simple rejection sampling-based procedure applied to labeled examples drawn from the input distribution leads to a new sampling distribution that itself is a Gaussian sum of ReLUs with hidden weight vectors $\widetilde{\mathbf{w}}_i = \overline{\mathbf{w}}_i / \|\overline{\mathbf{w}}_i\|_2$, where $\overline{\mathbf{w}}_i = \mathbf{w}_i^{\perp\mathbf{v}} + \gamma(\mathbf{v} \cdot \mathbf{w}_i)\mathbf{v}$ and $\gamma \in (0, 1)$ is a parameter of the rejection sampling procedure that we can choose. Then, choosing γ sufficiently small (relative to the relevant accuracy/heaviness threshold), the contribution of $\gamma(\mathbf{v} \cdot \mathbf{w}_i)\mathbf{v}$ in $\widetilde{\mathbf{w}}_i$ becomes negligible, and we can ensure that constant-alignment estimates for $\widetilde{\mathbf{w}}_i$ are also constant-alignment estimates for $\mathbf{w}_i^{\perp\mathbf{v}}$, with just a slightly smaller alignment constant (this is where the loss from 0.99 to 0.9 occurs). Thus, to obtain vectors with constant-factor alignment with $\mathbf{w}_i^{\perp\mathbf{v}}$s, it suffices to apply CDR to the samples from $\mathcal{D}$ retained after the rejection sampling.

Now, having access to a procedure that can produce near-ideal update directions, we discuss how the algorithm can maintain the invariant stated at the beginning of technical overview. The process of generating candidate updates and pruning the generated list is illustrated in Figure 1.

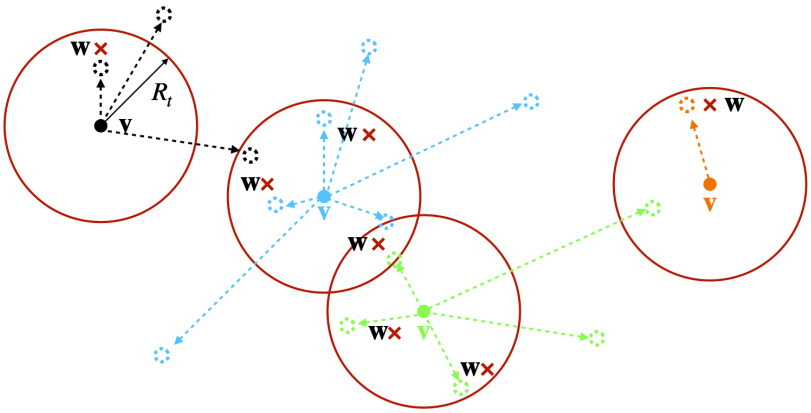
Constructing a large list of candidate updates. The process of constructing new candidate vectors is now as follows. In each round of computation t, we maintain a list $\mathcal{V}_t$, which contains vectors $\mathbf{v}_i$ that approximate targets $\mathbf{w}_i$. At the beginning of each round t, for each $\mathbf{v} \in \mathcal{V}_t$, we call the Refine procedure, which performs the following operations. We generate a polynomially-sized list $\mathcal{L}(\mathbf{v})$ that contains near-ideal update directions for all $\mathbf{w}_i$, $i \in [k]$ (and possibly many other vectors with no specific guarantee). For each vector in $\mathbf{r} \in \mathcal{L}(\mathbf{v})$, we perform candidate projected gradient descent-style updates $\mathrm{proj}_{\mathbb{S}^{d-1}}(\mathbf{u} + \eta_1 \mathbf{r})$, $\mathrm{proj}_{\mathbb{S}^{d-1}}(\mathbf{u} - \eta_1 \mathbf{r})$, and add both obtained vectors to $\mathcal{L}(\mathbf{v})$, where η_1 is an appropriately chosen step size. For technical reasons (namely, to ensure sufficient contraction), we repeat this procedure three times. As a result, we obtain a large but still polynomially sized list $\mathcal{V}_t^{\mathrm{new}}$ of vectors, with the following property: for each hidden weight vector $\mathbf{w}_i$, $i \in [k]$, the list $\mathcal{V}_t^{\mathrm{new}}$ contains at least one vector $\mathbf{v}_i$ such that $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \leq cR_{t+1} = c\rho R_t$, $\rho \in (0,1)$, and $c \in (0,1)$ is a constant that provides slack for introducing small errors when pruning the list. That is, there are new approximate vectors $\mathbf{v}_i$ within this large list $\mathcal{V}_t^{\mathrm{new}}$ that ensure constant contraction $c\rho$ of distance to the target vectors $\mathbf{w}_i$. However, as should be clear, at this point these "useful" vectors $\mathbf{v}_i$ are like a few needles in a large haystack. The next step thus involves identifying the useful vectors and pruning the rest.

Testing and pruning. As a first step in removing irrelevant vectors from $\mathcal{V}_t^{\mathrm{new}}$, in Section 3.1.2, we develop a testing procedure that can be used to remove all irrelevant vectors in $\mathcal{V}_t^{\mathrm{new}}$; namely, all vectors with alignment to all $\mathbf{w}_i$ larger than a small constant $1+\tau$, $\tau > 0$, times cR_{t+1}. This testing procedure is based on using properties of Hermite coefficients of ReLUs, with similar intuition as in construction of CDR discussed in Section 1.2.1. In particular, it is possible to argue that for even integer p, we have the following identity:

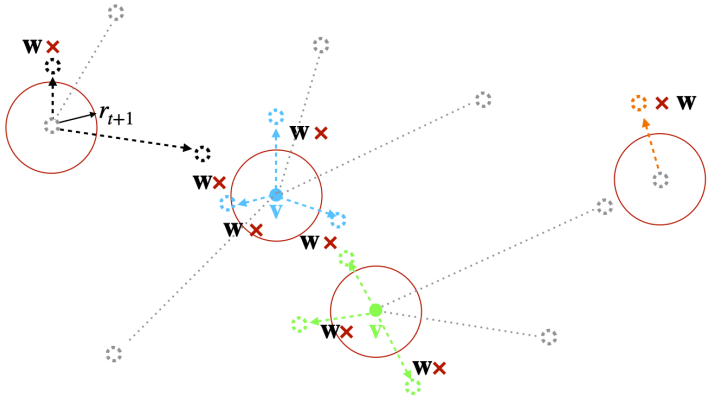
$$a_p(\mathbf{v}) := \frac{1}{\widehat{\sigma}_p} \mathbf{E}[y \mathrm{he}_p(\mathbf{v} \cdot \mathbf{x})] = \sum_i \beta_i \cos^p \theta_\pm(\mathbf{v}, \mathbf{w}_i), \tag{4}$$

where $\mathrm{he}_p(\mathbf{v} \cdot \mathbf{x})$ is the normalized probabilist's Hermite polynomial of order p (see Section 2). The first term $(1/\widehat{\sigma}_p)\,\mathbf{E}[y\mathrm{he}_p(\mathbf{v} \cdot \mathbf{x})]$ tells us that a_p can be estimated from samples; the second term $\sum_i \beta_i \cos^p \theta_\pm(\mathbf{v}, \mathbf{w}_i)$ that for sufficiently large degree p ($p \approx \log(k/\epsilon)/(cR_{t+1})^2$), we can determine if $\mathbf{v}$ correlates nontrivially with any target direction $\mathbf{w}_i$ (as measured by $\theta_\pm(\mathbf{v}, \mathbf{w}_i)$) or if it can be eliminated. More specifically, (with high probability), our tester returns FALSE for all $\mathbf{v} \in \mathcal{L}_t$ such that $\theta_\pm(\mathbf{v}, \mathbf{w}_i) > (1+\tau)cR_{t+1}$ and TRUE for all $\mathbf{v} \in \mathcal{L}_t$ such that $\theta_\pm(\mathbf{v}, \mathbf{w}_i) \leq cR_{t+1}$. As a result, the tester can be used as a rough pruning method that removes all $\mathbf{v} \in \mathcal{L}_t$ that do not approximate any of the target directions within angle $(1+\tau)cR_{t+1}$.

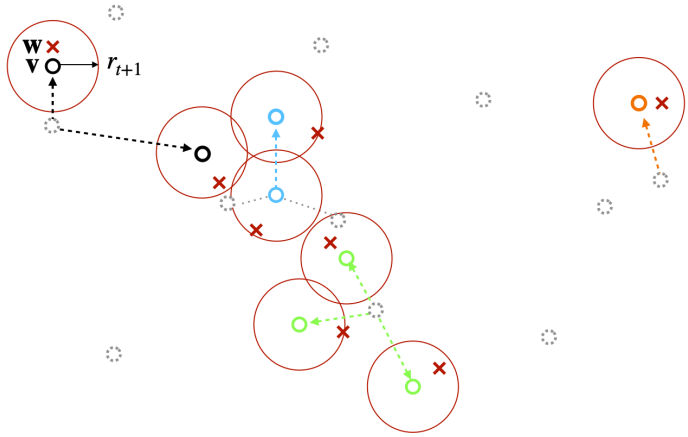
Ideally, the tester alone would suffice to move on to the next round of computation. However, it is possible that one $\mathbf{w}_i$ is approximated by many vectors $\mathbf{v}_i$ in the list, and if all were retained in $\mathcal{V}_{t+1}$, the size of $\mathcal{V}_T$ would not be guaranteed to be polynomial, as the algorithm takes $T = \Theta(\log(k/\epsilon))$ rounds and each call to Refine (for each vector in the list) produces a polynomially-sized list. Thus, the final step, in Section 3.1.3, is to prune the list $\mathcal{V}_t^{\mathrm{new}}$ down to at most k vectors (and assign it to $\mathcal{V}_{t+1}$). The reason we can do this is geometric: since the previous steps ensured sufficient contraction with leeway for additional error, we make the following observation. If multiple vectors $\mathbf{v}$ are close to the same target $\mathbf{w}_i$, then they must be close to each other. The final pruning step is thus the following: pick any candidate vector $\mathbf{v}$ in the pruned list and add it to $\mathcal{V}_{t+1}$. The testing procedure ensures that this vector is within the target distance of at least one hidden weight vector. Now, iterate through the remaining vectors $\mathbf{v}$ in the pruned list $\mathcal{V}_t^{\mathrm{new}}$ and add $\mathbf{v}$ to $\mathcal{V}_{t+1}$ if and only if it is at least $2(1+2\tau)cR_{t+1}$-far from all the vectors already in the list. We claim that for each $\mathbf{w}_i$, there is exactly one $\mathbf{v}_i \in \mathcal{V}_{t+1}$ such that $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \leq R_{t+1}$ (but the same $\mathbf{v}_i$ can satisfy $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \leq R_{t+1}$ for multiple $\mathbf{w}_i$), thus $|\mathcal{V}_{t+1}| \leq k$. Fix one $\mathbf{w}_i$ and consider all $\mathbf{v}$ such that $\theta_\pm(\mathbf{v}, \mathbf{w}_i) \leq (1+\tau)cR_{t+1}$; these are the candidates $\mathbf{v}$ that survive the testing prune, and among them, there is at least one vector $\mathbf{v}_i^*$ such that $\theta_\pm(\mathbf{v}_i^*, \mathbf{w}_i) \leq cR_{t+1}$, as guaranteed by the refinement procedure. Because, by the triangle inequality, all such $\mathbf{v}$ are within $2(1+\tau)cR_{t+1}$ of each other, at most one can get added to $\mathcal{V}_{t+1}$. If exactly $\mathbf{v}_i^*$ is added, we are done, so suppose $\mathbf{v}_i^*$ is not added to $\mathcal{V}_{t+1}$. This can only happen if there exists $\mathbf{v}' \in \mathcal{V}_{t+1}$ such that $\theta_\pm(\mathbf{v}_i^*, \mathbf{v}') \leq 2(1+2\tau)cR_{t+1}$. Thus, again by the triangle inequality, $\theta_\pm(\mathbf{w}_i, \mathbf{v}') \leq (2(1+2\tau)+1)cR_{t+1}$. Parameters c, τ are chosen so that $(2(1+2\tau)+1)c \leq 1$, and thus we have that each $\mathbf{w}_i$ can be mapped to one $\mathbf{v}_i \in \mathcal{V}_{t+1}$ such that $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \leq R_{t+1}$.



(a) After the Refine procedure, each candidate produces a list of vectors (marked in a dotted circle), of which some are closer to the hidden weight vectors. But there are also potentially many other redundant vectors.



(b) After the Test procedure, those vectors that do not "cover" hidden weight vectors at a target contracted distance are discarded (marked in gray).



(c) There might be many candidate vectors that are close to each other, covering the same hidden weight vector. After the Prune procedure, the candidate vectors are separated (the pruned vector is also marked in gray). Yet, we guarantee that each hidden weight vector is still covered by the candidate vectors after pruning.

Figure 1: The update procedure of the CoverAndPrune($\epsilon_0, \delta; \mathcal{D}$) algorithm. $\times$ denotes the hidden weight, $\bullet$ denotes the candidate vector at step t, $\circ$ denotes the vectors generated after step t. Only the hidden weight vectors with not-too-small coefficients are displayed.

1.3 Related Work

Our work belongs to the literature on learning structured multi-index models, and we review related work within this context.

Learning positive linear combinations of ReLUs. The most directly related works are [DKKZ20, DK20, DK25], whose complexity guarantees were discussed earlier in the introduction. The algorithm of [DKKZ20] uses the degree-2 Chow parameters of the target to find a low-dimensional subspace sufficient for prediction, and then searches over candidate networks in that subspace. The work of [DK20] uses higher-degree moments to find a vector space of polynomials that nearly vanish on the hidden weight vectors. Its structural result gives a small cover for the near-zero set of these polynomials, which is then used to construct a hypothesis. The work of [DK25] estimates high-order moment tensors implicitly, without writing down their entries, and uses these estimates to approximate the target's Hermite expansion. This generalizes the implicit moment-estimation technique of Li and Liu [LL22], developed for clustering separated mixtures of spherical Gaussians.

Learning general linear combinations of ReLUs. PAC learning general linear combinations of ReLUs, with both positive and negative coefficients, has also been studied. Without structural assumptions on the hidden weight vectors, [CDG+23] gave the first algorithm with running time polynomial in d and $1/\epsilon$ for every fixed k. Subsequent work improved the running time to $(dk/\epsilon)^{O(k^2)}$ [CN24] and $(dk/\epsilon)^{O(k)}$ [DK24]. These algorithms handle a broader class of functions than the one studied here, but their complexity scales exponentially in the number of hidden units.[2]

Learning multi-index models and deeper networks. The work of [CKM22] gives fixed-parameter learning algorithms for homogeneous ReLU networks of arbitrary depth under Gaussian inputs, with polynomial dependence on the ambient dimension but exponential dependence on network size and inverse accuracy. Recent work on multi-index models gives learning guarantees for more general classes of functions that depend only on an unknown low-dimensional subspace [DIKZ25, DIKR25, DLB25]. Similarly, the latter algorithms have polynomial dependence on the ambient dimension but exponential dependence on the other relevant parameters.

Parameter learning under nondegeneracy assumptions. A number of earlier works focus on parameter learning, namely recovering the parameters of the data-generating network. Examples include tensor-based methods [JSA15] and algorithms combining initialization with local optimization [ZSJ+17]. Related tensor methods were developed for mixtures of generalized linear models in [SJA16]. For positive linear combinations of ReLUs under Gaussian inputs, [GLM18] gives a parameter-learning algorithm assuming a full-rank hidden-weight matrix, with complexity depending polynomially on its condition number. [BJW19, GKLW19] give parameter-recovery algorithms for vector-valued two-layer networks under rank and related nondegeneracy assumptions. The work of [ATV21] develops tensor-based algorithms for scalar-output ReLU networks with biases and general coefficients. In contrast, we focus on small prediction error rather than parameter recovery, with complexity independent of separation, rank, condition-number, or any other parameters requiring strong assumptions about the configuration of the hidden weight vectors.

Computational lower bounds. For general linear combinations of ReLUs, [DKKZ20, GGJ+20] establish super-polynomial lower bounds in the correlational statistical-query (CSQ) model [BF02, Fel08]. This is a restriction of the Statistical Query (SQ) model [Kea93], in which the learner accesses the target through estimates of correlations with query functions; see also [BFJ+94, VW19]. For networks with two hidden layers, [CGKM22] establishes general SQ and cryptographic hardness even in the noiseless setting. More recently, [LZZ24] proves cryptographic hardness for one-hidden-layer ReLU networks with both positive and negative coefficients and biases, in the presence of small independent Gaussian label noise. Whether general linear combinations of ReLUs without biases can be learned in fully-polynomial time under Gaussian inputs remains open. Since our algorithm can be implemented in the CSQ model, the CSQ lower bounds above rule out an immediate extension to general coefficients.

[2] Here we focus on learning from random labeled examples. If query access to the function is allowed, [CKM21] gives a learner for one-hidden-layer ReLU networks with general coefficients and biases whose query complexity and running time are polynomial in dimension, the number of hidden units, inverse accuracy, and bounds on the weights and biases.

1.4 Concurrent Work and AI Use

All ideas and technical arguments in this paper are the intellectual ownership of the authors, who have worked on developing them over the past 2+ years. In the process of writing the paper, OpenAI GPT models were used to improve the presentation (including generating Figures 1 and 2), audit the proofs, and simplify some of the statements. We note that a very recent paper [Zhe26] appeared on the arXiv, claiming essentially the same main result. Rather curiously, that work is quoting some of the exact same core techniques as our work, but is lacking in most of the proof details. The paper [Zhe26] further uses some of the same non-standard terminology appearing in our work (such as "refine," "prune," "cover," etc.) and even some of the same universal constants chosen lazily by the authors to be smaller than needed. As a result, we were forced to rush posting of this work, before we were entirely satisfied with its presentation, but with complete technical details included.

1.5 Paper Organization

The paper is structured as follows: In Section 2, we record the notation and background required in our technical sections. Sections 3 to 5 provide a detailed description and analysis of our algorithm.

2 Notation and Preliminaries

Basic Notation. For $n \in \mathbb{Z}_+$, we denote $[n] \coloneqq \{1, \ldots, n\}$. We use bold lowercase letters for vectors and bold uppercase letters for matrices and tensors. For $\mathbf{x} \in \mathbb{R}^d$ and $i \in [d]$, $(\mathbf{x})_i$ denotes the ith coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2 \coloneqq (\sum_{i=1}^d (\mathbf{x})_i^2)^{1/2}$ denotes the ℓ_2 norm of $\mathbf{x}$. We write $\mathbf{x} \cdot \mathbf{y}$ for the standard inner product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\theta(\mathbf{x}, \mathbf{y}) \coloneqq \arccos(\mathbf{x} \cdot \mathbf{y}/(\|\mathbf{x}\|_2 \|\mathbf{y}\|_2))$ when $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{y} \neq \mathbf{0}$. We denote $\theta_\pm(\mathbf{x}, \mathbf{y}) \coloneqq \min\{\theta(\mathbf{x}, \mathbf{y}), \pi - \theta(\mathbf{x}, \mathbf{y})\}$.

For unit vectors $\mathbf{u}, \mathbf{v}$, we use $\mathbf{u}^{\perp \mathbf{v}}$ to denote the component of $\mathbf{u}$ orthogonal to $\mathbf{v}$, i.e., $\mathbf{u}^{\perp \mathbf{v}} = (\mathbf{I} - \mathbf{v}\mathbf{v}^\top)\mathbf{u}$. For any vector $\mathbf{u} \neq \mathbf{0}$, we use $\widehat{\mathbf{u}}$ to denote its normalization to unit ℓ_2 norm. Given $\mathbf{v} \in \mathbb{R}^d$, we use $\mathbf{v}^\perp$ to denote the linear subspace of vectors orthogonal to $\mathbf{v}$, i.e., $\mathbf{v}^\perp = \{\mathbf{u} \in \mathbb{R}^d : \mathbf{u} \perp \mathbf{v}\}$. We denote the projection from $\mathbb{R}^d$ to $\mathbf{v}^\perp$ by $\mathbf{P}_{\perp \mathbf{v}} = \mathbf{I} - \mathbf{v}\mathbf{v}^\top$. More generally, for a linear subspace $\mathbb{F}$ of $\mathbb{R}^d$, the projection from $\mathbb{R}^d$ to $\mathbb{F}$ is denoted by $\mathbf{P}_{\mathbb{F}}$.

Throughout the paper, $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$. We use $\mathbb{1}\{\mathcal{E}\}$ to denote the indicator of an event $\mathcal{E}$. $\sigma(z) \coloneqq \max\{0, z\}$. We write $E \gtrsim F$ for two nonnegative expressions E and F to denote that there exists some universal constant $c > 0$ (independent of the variables or parameters on which E and F depend) such that $E \geq cF$. The expression $E \lesssim F$ is similarly defined. We use the notation $g(t) \leq \text{poly}(t)$ for a quantity $t \geq 1$ to indicate that there exist constants $c, C > 0$ such that $g(t) \leq Ct^c$. Similarly, we use $g(t) \geq \text{poly}(t)$ for a quantity $t < 1$ to denote that there exist constants $c, C > 0$ such that $g(t) \geq Ct^c$. Finally, we use standard asymptotic notation $O(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$, and $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$ (the latter two suppress polylogarithmic factors in the argument).

Complex Space. We use $\mathbb{C}^d$ to denote the d-dimensional complex vector space. For $\mathbf{z} \in \mathbb{C}^d$, $\mathbf{z} = \mathbf{x} + \mathrm{i}\mathbf{y}$ where i is the imaginary unit, and $\bar{\mathbf{z}} = \mathbf{x} - \mathrm{i}\mathbf{y}$ is the complex conjugate of $\mathbf{z}$. We use $\text{Re}(\mathbf{z})$ to denote the real part of $\mathbf{z}$ and $\text{Im}(\mathbf{z})$ to denote its imaginary part. For a complex number $z = a + \mathrm{i}b$, recall that $|z|^2 = a^2 + b^2 = z\bar{z}$.

Matrices and Tensors. For a matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, we use $\|\mathbf{M}\|_2$ to denote its spectral/operator norm. For a symmetric matrix $\mathbf{M}$, $\lambda_i(\mathbf{M})$ denotes its ith largest eigenvalue and $\mathbf{r}_i(\mathbf{M})$ denotes the corresponding (unit) eigenvector, $i \in [d]$. $\text{tr}(\mathbf{M})$ denotes the trace of a matrix. Our analysis makes use of the following standard fact.

Fact 2.1 (Wedin's Theorem)**.** *For any pair of symmetric matrices $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{d \times d}$, if $\lambda_1(\mathbf{M}_1) - \lambda_2(\mathbf{M}_1) - \|\mathbf{M}_1 - \mathbf{M}_2\|_2 > 0$, then*

$$\sin(\theta(\mathbf{r}_1(\mathbf{M}_1), \mathbf{r}_1(\mathbf{M}_2))) \leq \frac{\|\mathbf{M}_1 - \mathbf{M}_2\|_2}{\lambda_1(\mathbf{M}_1) - \lambda_2(\mathbf{M}_1) - \|\mathbf{M}_1 - \mathbf{M}_2\|_2}.$$

A complex tensor $\mathbf{W} \in (\mathbb{C}^d)^{\otimes p}$ of dimension d and order p is a multilinear map defined by a p-index array with complex entries $(\mathbf{W})_\alpha$, where $\alpha = (i_1, \ldots, i_p)$ with $i_1, \ldots, i_p \in [d]$. A tensor $\mathbf{W}$ is symmetric

if $(\mathbf{W})_\alpha = (\mathbf{W})_{\alpha_\pi}$, where $\alpha_\pi = (i_{\pi(1)}, \ldots, i_{\pi(p)})$, for any permutation $\pi : [p] \to [p]$. For complex tensors $\mathbf{W}, \mathbf{T}$ of dimension d and order p, we denote by $\langle \mathbf{W}, \mathbf{T} \rangle$ their entrywise inner product:

$$\langle \mathbf{W}, \mathbf{T} \rangle = \sum_{i_1,\ldots,i_p} (\overline{\mathbf{W}})_{i_1,\ldots,i_p} (\mathbf{T})_{i_1,\ldots,i_p} \ .$$

For a complex tensor $\mathbf{W}$, let $\|\mathbf{W}\|_F^2 = \langle \overline{\mathbf{W}}, \mathbf{W} \rangle = \sum_{i_1,\ldots,i_p} |\mathbf{W}_{i_1,\ldots,i_p}|^2$ denote its Frobenius norm. Let $\mathbb{U}, \mathbb{V}$ be finite-dimensional complex inner-product spaces. The Hermitian adjoint of a complex-linear map $\mathbf{W} : \mathbb{U} \to \mathbb{V}$ is the unique complex-linear map $\mathbf{W}^* : \mathbb{V} \to \mathbb{U}$ satisfying $\langle \mathbf{W}(\mathbf{A}), \mathbf{B} \rangle = \langle \mathbf{A}, \mathbf{W}^*(\mathbf{B}) \rangle$ for any $\mathbf{A} \in \mathbb{U}$ and $\mathbf{B} \in \mathbb{V}$.

We overload the notation $\langle \cdot, \cdot \rangle$ with partial tensor contraction: let $\mathbf{W} \in (\mathbb{C}^d)^{\otimes(p+m)}$, $\mathbf{T} \in (\mathbb{C}^d)^{\otimes p}$, $\langle \mathbf{W}, \mathbf{T} \rangle$ is a tensor in $(\mathbb{C}^d)^{\otimes m}$ defined by

$$(\langle \overline{\mathbf{W}}, \mathbf{T} \rangle)_{j_1,\ldots,j_m} = \sum_{i_1,\ldots,i_p} (\overline{\mathbf{W}})_{i_1,\ldots,i_p,j_1,\ldots,j_m} (\mathbf{T})_{i_1,\ldots,i_p}.$$

Let $\mathrm{Sym}(\mathbf{W})$ be the symmetrization operator:

$$(\mathrm{Sym}(\mathbf{W}))_{i_1,\ldots,i_p} = \frac{1}{p!} \sum_{\text{permutation } \pi \text{ of } [p]} (\mathbf{W})_{i_{\pi(1)},\ldots,i_{\pi(p)}}.$$

Identify an order-p tensor $\mathbf{T}$ with the vector $\mathbf{t} = \mathrm{vec}(\mathbf{T})$ of its entries. Its covariance is the Hermitian positive semidefinite matrix $\mathbf{Cov}(\mathbf{T}) = \mathbf{E}[(\mathbf{t} - \mathbf{E}[\mathbf{t}])(\mathbf{t} - \mathbf{E}[\mathbf{t}])^*]$, and its uncentered second moment is $\mathbf{E}[\mathbf{t}\mathbf{t}^*]$. We say that its covariance is bounded above by V if $\mathbf{Cov}(\mathbf{T}) \preceq V\mathbf{I}$, which is equivalent to $\sup_{\|\mathbf{W}\|_F=1} \mathbf{E}[|\langle \mathbf{W}, \mathbf{T} - \mathbf{E}[\mathbf{T}] \rangle|^2] \leq V$. For real-valued tensors, these definitions reduce to the usual real covariance and second-moment matrices after flattening.

Gaussian Space and Hermite Analysis. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the d-dimensional Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. We will primarily use the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, denoted simply by $\mathcal{N}$. The standard Gaussian distribution on a linear subspace $\mathbb{F}$ of $\mathbb{R}^d$ is denoted by $\widetilde{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_\mathbb{F})$ and defined via $\widetilde{\mathbf{x}} = \mathbf{P}_\mathbb{F}\mathbf{x}$, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The Gaussian L_2 inner product is $\langle f, g \rangle \coloneqq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})g(\mathbf{x})]$. For $f \in L_2(\mathcal{N})$, define $\|f\|_{L_2} \coloneqq (\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})^2])^{1/2}$.

Fact 2.2 (Gaussian Hypercontractivity; see, e.g., [O'D14])**.** *For any real polynomial $f(z)$ of degree at most p and real $q \geq 2$, it holds that $(\mathbf{E}_{z\sim\mathcal{N}}[|f(z)|^q])^{1/q} \leq (q-1)^{p/2}(\mathbf{E}_{z\sim\mathcal{N}}[f(z)^2])^{1/2}$.*

We now proceed to define Hermite polynomials and the associated orthonormal basis they define. In one dimension, define the normalized probabilists' Hermite polynomial

$$\mathrm{he}_p(z) \ \coloneqq \ \frac{1}{\sqrt{p!}}(-1)^p \exp(z^2/2) \frac{\mathrm{d}^p}{\mathrm{d}z^p} \exp(-z^2/2). \tag{5}$$

The family $\{\mathrm{he}_p\}_{p\geq 0}$ is an orthonormal basis for $L_2(\mathcal{N}(0,1))$. Thus, any $f : \mathbb{R} \to \mathbb{R}$ with $f \in L_2(\mathcal{N}(0,1))$ has a unique Hermite expansion

$$f(z) = \sum_{p\geq 0} \widehat{f}_p \, \mathrm{he}_p(z), \qquad \widehat{f}_p \coloneqq \mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}[f(z)\mathrm{he}_p(z)].$$

We make use of a number of basic properties of Hermite polynomials summarized below.

Fact 2.3 (Basic Hermite Facts, see, e.g., [Bog98, O'D14])**.** *For normalized Hermite polynomials he_p, as defined in (5), the following hold:*

1. *(Differentiation) $(\mathrm{he}_p(z))' = \sqrt{p}\,\mathrm{he}_{p-1}(z)$, for $p \geq 1$.*
2. *(Degree Identity) $z\mathrm{he}_p(z) = \sqrt{p+1}\,\mathrm{he}_{p+1}(z) + \sqrt{p}\,\mathrm{he}_{p-1}(z)$, for $p \geq 2$.*
3. *(Smoothing) Let $f \in L_2(\mathcal{N}(0,1))$, $\rho \in [0,1]$ and denote $\mathrm{T}_\rho(f)(x) \coloneqq \mathbf{E}_{z\sim\mathcal{N}(0,1)}[f(\rho x + \sqrt{1-\rho^2}z)]$. Then the smoothed function $\mathrm{T}_\rho(f)(x)$ has Hermite expansion:*

$$\mathrm{T}_\rho(f)(x) = \sum_{p\geq 0} \rho^p \widehat{f}_p \mathrm{he}_p(x).$$

4. (Hermite expansion in one specific direction) Let $f \in L_2(\mathcal{N})$ and $\mathbf{u}, \mathbf{v}$ be unit vectors in $\mathbb{R}^d$. Then:

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[f(\mathbf{v}\cdot\mathbf{x})\mathrm{he}_p(\mathbf{u}\cdot\mathbf{x})] = \widehat{f}_p \cos^p(\theta(\mathbf{v},\mathbf{u})).$$

Throughout the paper, we denote by $\widehat{\sigma}_i$ the ith Hermite coefficient of the ReLU function σ.

These Hermite coefficients admit the following characterization.

Fact 2.4 (Hermite expansion of ReLU; see, e.g., [GGK20, DK24])**.** *For $\sigma(z) \coloneqq \max\{0, z\}$, we have $\sigma(z) = \sum_{p=0}^{\infty} \widehat{\sigma}_p \mathrm{he}_p(z)$, where $\widehat{\sigma}_0 = 1/\sqrt{2\pi}$, $\widehat{\sigma}_1 = 1/2$, and*

$$\widehat{\sigma}_{2p-1} = 0 \quad \textit{for all } p \geq 2, \qquad \widehat{\sigma}_{2p} = \frac{\mathrm{he}_{2p-2}(0)}{\sqrt{2\pi(2p)(2p-1)}} \quad \textit{for all } p \geq 1.$$

In particular, $|\widehat{\sigma}_{2p}| = \Theta(p^{-5/4})$ and $\widehat{\sigma}_{2p}^2 = \Theta(p^{-5/2})$ for $p \geq 1$. Moreover, if we formally differentiate the Hermite series, we obtain $\sigma'(z) = \sum_{p\geq 0} \widehat{\sigma}'_p \mathrm{he}_p(z)$ with $\widehat{\sigma}'_p = \widehat{\sigma}_{p+1}\sqrt{p+1}$.

Hermite polynomials can be extended to Hermite tensors in $\mathbb{R}^d$. Throughout this paper, we will use the following normalized Hermite tensor.

Definition 2.5 (Normalized Probabilist's Hermite Tensor)**.** *We define the normalized order-p Hermite tensor in d dimensions as follows: for any $i_1, \ldots, i_p \in [d]$,*

$$(\mathbf{He}_p(\mathbf{x}))_{i_1,\ldots,i_p} \coloneqq \frac{1}{\sqrt{p!}} \sum_{\substack{\textit{Partitions } P \textit{ of } [p] \\ \textit{into sets of size 1 and 2}}} \prod_{\substack{\textit{all size-2 sets } \{a,b\} \\ \textit{in partition } P}} -(\mathbf{I}_d)_{i_a,i_b} \prod_{\substack{\textit{all size-1 sets } \{c\} \\ \textit{in partition } P}} (\mathbf{x})_{i_c} \,. \tag{6}$$

We gather here important facts about Hermite tensors that are used in our analysis.

Fact 2.6 (Properties of Normalized Hermite Tensor)**.** *Let $\mathbf{He}_p$ be the pth-order normalized Hermite tensor. The following holds:*

1. $\mathbf{He}_p(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes p})/\sqrt{p!}]$, *for all* $\mathbf{x} \in \mathbb{R}^d$.
2. *Let $f \in L_2(\mathcal{N}(0,1))$. For any $\mathbf{v} \in \mathbb{S}^{d-1}$, $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[f(\mathbf{v}\cdot\mathbf{z})\mathbf{He}_2(\mathbf{z})] = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[f''(\mathbf{v}\cdot\mathbf{z})]\mathbf{v}\mathbf{v}^\top$.*

3 Optimization via Coarse Direction Recovery, Refinement, and Pruning

In this section, we provide the full technical details of our analysis, assuming access to a coarse direction recovery (CDR) procedure as an oracle; the implementation of CDR is presented in Section 4. Our oracle access to CDR is formally stated as the following assumption.

Assumption 1 (Coarse Direction Recovery)**.** *Let $d, k \in \mathbb{Z}_+$ and $\lambda, \delta \in (0,1)$. Given access to a Gaussian sum of k ReLUs distribution on $\mathbb{R}^d \times \mathbb{R}$, a Coarse Direction Recovery algorithm $\mathsf{CDR}(\mathcal{D}, \lambda, \delta)$ draws $N_{\mathsf{CDR}}(d,k,\lambda,\delta)$ i.i.d. samples from $\mathcal{D}$, runs in time $T_{\mathsf{CDR}}(d,k,\lambda,\delta)$, and with probability at least $1-\delta$ returns a list $\mathcal{L} \subseteq \mathbb{S}^{d-1}$ of size at most $L_{\mathsf{CDR}}(d,k,\lambda,\delta)$ such that: for every λ-heavy component $i \in [k]$, there exists $\mathbf{r}_i \in \mathcal{L}$ satisfying $|\mathbf{r}_i \cdot \mathbf{w}_i| \geq 0.99$.*

At a high level, the CDR procedure postulated in Assumption 1 guarantees that if we had i.i.d. samples from a Gaussian sum of k ReLUs distribution $\mathcal{D}$, then we could efficiently compute a not-too-large list $\mathcal{L}$ of "candidate" unit vectors such that for each λ-heavy hidden weight vector $\mathbf{w}_i$ there exists a candidate $\mathbf{r} \in \mathcal{L}$ that is "close" to $\mathbf{w}_i$ (as measured by the nontrivial correlation $|\mathbf{r}\cdot\mathbf{w}_i| \geq 0.99$). The list may also contain other vectors with no specific guarantee. Given the existence of such an algorithm that runs in polynomial time (which we establish in Section 4), our main result in this section is a polynomial-time learner of an unknown sum of ReLUs, as stated in the following theorem.

Algorithm 1 $\mathsf{LearnSumReLUs}(d, k, \epsilon; \mathcal{D})$

Input: $k, \epsilon > 0$, and access to a Gaussian sum of k ReLUs distribution $\mathcal{D}$.
Output: a hypothesis $\widehat{y}$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\widehat{y}(\mathbf{x}) - y)^2] \leq \epsilon$ with probability at least $2/3$.

$\epsilon_0 \leftarrow \sqrt{\epsilon}/(4k)$, $\delta \leftarrow 1/100$ ▷ ϵ_0: error threshold of Parameters; δ: failure probability

$\widehat{\mathcal{V}} \leftarrow \mathsf{CoverAndPrune}(\epsilon_0, \delta; \mathcal{D})$ ▷ Returns a candidate list by repeatedly calling the CDR Subroutine.

Let $\widehat{\mathcal{V}} = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ where $m = |\widehat{\mathcal{V}}|$. ▷ Enumerate the elements

Draw $\Theta(dk^2/\epsilon^2)$ samples from $\mathcal{D}$ to construct an estimator $\widehat{\mathbf{g}}$ of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}]$.

Draw $\Theta(k^{10}/\epsilon^2)$ samples from $\mathcal{D}$ to formulate the empirical distribution $\widehat{\mathcal{D}}$.

$\widehat{\boldsymbol{\alpha}} \leftarrow \mathrm{argmin}_{\boldsymbol{\alpha}\in\{\alpha_i\geq 0:\sum \alpha_i \leq k\}} \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[(y - \widehat{\mathbf{g}}\cdot\mathbf{x} - \sum_{i=1}^m \frac{\alpha_i}{2}|\mathbf{v}_i\cdot\mathbf{x}|)^2]$.
▷ Convex Quadratic Program

Return: $\widehat{y}(\mathbf{x}) = \widehat{\mathbf{g}}\cdot\mathbf{x} + \sum_{i\in[m]}(\widehat{\alpha}_i/2)|\mathbf{v}_i\cdot\mathbf{x}|$.

Theorem 3.1 (Direction Covering Algorithm)**.** *Let $\epsilon \in (0,1)$, $k, d \in \mathbb{Z}_+$, and $\mathcal{D}$ be a Gaussian sum of k ReLUs distribution on $\mathbb{R}^d \times \mathbb{R}$. Given access to a procedure* CDR *satisfying Assumption 1, there exists an algorithm ($\mathsf{LearnSumReLUs}(d, k, \epsilon; \mathcal{D})$) that repeatedly calls* CDR *with parameters $\lambda = \mathrm{poly}(\epsilon/k)$ and $\delta = \mathrm{poly}(\epsilon/k)$ and attains the following performance guarantee: it draws $N = \widetilde{O}(d\,\mathrm{poly}(k/\epsilon) + L_{\mathsf{CDR}}^3(N_{\mathsf{CDR}} + L_{\mathsf{CDR}}^3)\mathrm{poly}(k/\epsilon))$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(d, N, T_{\mathsf{CDR}})$ time and, with probability at least $2/3$, outputs a hypothesis $\widehat{y}(\mathbf{x})$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \widehat{y}(\mathbf{x}))^2] \leq \epsilon$.*

The pseudocode for our main algorithm is displayed below.

The body of this section is devoted to the proof of Theorem 3.1. In Section 3.1, we show that the main subroutine of our algorithm, $\mathsf{CoverAndPrune}(\epsilon_0, \delta; \mathcal{D})$, successfully covers all hidden weight vectors $\mathbf{w}_i$ with non-trivial coefficients. Then in Section 3.2, we present the proof of Theorem 3.1.

3.1 Learning λ-Heavy Hidden Weight Vectors

Learning λ-heavy (i.e., nontrivial-coefficient) hidden weight vector described in this subsection is the technical core of the paper. We begin this section by first arguing that the CDR procedure can be used to construct updates to candidate vectors maintained as estimates of the hidden vectors $\mathbf{w}_i$. In particular, in Section 3.1.1, we argue that, given a fixed vector $\mathbf{v} \in \mathbb{S}^{d-1}$ and a simple to implement rejection sampling procedure (used for selecting samples from $\mathcal{D}$), CDR procedure can be used to construct a polynomially-sized list $\mathcal{L}(\mathbf{v})$ that contains constant-correlation estimates of $\mathbf{w}_i^{\perp\mathbf{v}}$ for λ-heavy ReLU components. On input $\mathbf{v}$, our $\mathsf{Refine}(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$ procedure uses CDR to construct a list $\mathcal{V}$ of gradient descent-style updates to $\mathbf{v}$ such that for each λ-heavy $\mathbf{w}_i$, there exists at least one updated vector that contracts the distance between $\mathbf{v}$ and $\mathbf{w}_i$ by a small constant factor. At the end of $\mathsf{Refine}(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$, only a small subset of updated candidate vectors in the output list $\mathcal{V}$ is useful to algorithm progress, that is, to improving the estimation of hidden λ-heavy vectors. The next step in the algorithm is thus to prune the list, after $\mathsf{Refine}(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$ is applied to all candidate vectors maintained by the algorithm. As a first coarse pruning step, Section 3.1.2 introduces a testing procedure that can be used to identify elements of $\mathcal{V}$ farther from all λ-heavy $\mathbf{w}_i$ than a prespecified threshold. As a final step, Section 3.1.3 introduces $\mathsf{Prune}(\mathcal{V}, r)$ procedure that prunes the list further so that only up to k candidate vectors are retained by the algorithm. A complete algorithm statement for learning λ-heavy hidden vectors is then provided in $\mathsf{CoverAndPrune}(\epsilon_0, \delta; \mathcal{D})$ and the main result of the subsection (Proposition 3.10) is proved.

3.1.1 Boosting Hidden Vector Estimates

We begin by arguing that, given any $\mathbf{v} \in \mathbb{S}^{d-1}$, the CDR procedure can be called to obtain constant-correlation estimates of $\mathbf{w}_i^{\perp\mathbf{v}}$, for any λ-heavy component $i \in [k]$. These estimates are then used to boost the estimation of $\mathbf{w}_i$, by treating estimates of $\mathbf{w}_i^{\perp}$ as gradient descent directions from $\mathbf{v}$. Of course, it is not immediately clear that CDR can be applied here, as Assumption 1 requires samples from a Gaussian sum of k ReLUs distribution with $\widehat{\mathbf{w}_i^{\perp\mathbf{v}}} := \mathbf{w}_i^{\perp\mathbf{v}}/\|\mathbf{w}_i^{\perp\mathbf{v}}\|_2$ as the hidden directions.

For intuition on how CDR could be applied to estimating $\mathbf{w}_i^{\perp\mathbf{v}}$ and boosting the estimate of $\mathbf{w}_i$, recall first that a sample $(\mathbf{x}, y)$ drawn from our distribution $\mathcal{D}$ satisfies $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $y = \sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$. Fix a unit candidate $\mathbf{v}$ and a target $\mathbf{w}_1$. Given $\mathbf{v}$, we can express $\mathbf{w}_1$ as $\mathbf{w}_1 = \cos(\theta_1)\,\mathbf{v} + \sin(\theta_1)\,\widehat{\mathbf{w}_1^{\perp\mathbf{v}}}$, where $\theta_1 = \theta(\mathbf{w}_1, \mathbf{v})$. Consider the samples $(\mathbf{x}, y) \sim \mathcal{D}$ that lie in a band of width γ around the origin, namely satisfy $|\mathbf{v} \cdot \mathbf{x}| \leq \gamma$. Then, it is possible to argue that $\sigma(\mathbf{w}_1 \cdot \mathbf{x}) \approx \sigma(\sin(\theta_1)\widehat{\mathbf{w}_1^{\perp\mathbf{v}}}) + O(\gamma)$. Thus, when $\gamma \to 0$, for such samples we have that $y \approx \sum_{i=1}^k \beta_i \sin(\theta(\mathbf{w}_i, \mathbf{v}))\sigma(\widehat{\mathbf{w}_i^{\perp\mathbf{v}}} \cdot \mathbf{x})$, which is another sum of ReLUs with coefficients $b_i = \beta_i \sin(\theta(\mathbf{w}_i, \mathbf{v}))$ and with hidden weight vectors $\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}$ lying in the subspace $\mathbf{v}^\perp$. Moreover, the component of $\mathbf{x}$ that lies in the subspace $\mathbf{v}^\perp$ follows the standard normal distribution (the rest is irrelevant, as y depends only on $\mathbf{w}_i^{\perp\mathbf{v}} \cdot \mathbf{x} = \mathbf{w}_i^{\perp\mathbf{v}} \cdot \mathbf{x}^{\perp\mathbf{v}}$). This means that we can apply CDR to this new instance (obtained via conditioning) to obtain a new list of unit vectors $\mathcal{L}$ with the following guarantee: there exists $\mathbf{r} \in \mathcal{L}$ such that $|\mathbf{r} \cdot \widehat{\mathbf{w}_1^{\perp\mathbf{v}}}| \geq 0.99$. Then, updating $\mathbf{v}$ along the direction $\mathbf{r}$ with a carefully chosen step-size, we obtain a new vector $\mathbf{v}'$ such that $|\mathbf{v}' \cdot \mathbf{w}_1| \gg |\mathbf{v} \cdot \mathbf{w}_1|$, boosting the estimation of $\mathbf{w}_1$. We now formalize this intuition.

Rejection Sampling. For technical reasons, instead of using a "hard" band, i.e., conditioning on $|\mathbf{v} \cdot \mathbf{x}| \leq \gamma$ for some $\gamma > 0$, we will use a soft-conditioning method based on rejection sampling. Soft-conditioning is more convenient because the marginal distribution remains a standard Gaussian while ensuring that the retained samples define an instance of the same learning problem (with different parameters). Formally, we use the following rejection sampling procedure:

Fact 3.2 (see, e.g., Lemma 4.7 in [DKS18])**.** *Given parameters $\mathbf{v} \in \mathbb{S}^{d-1}$ and $\gamma \in (0, 1)$, the $(\gamma, \mathbf{v})$-rejection procedure $\mathrm{RS}_{\gamma,\mathbf{v}}$ is defined as follows: A vector $\mathbf{x} \in \mathbb{R}^d$ is accepted by $\mathrm{RS}_{\gamma,\mathbf{v}}$ with probability*

$$a_{\gamma,\mathbf{v}}(\mathbf{x}) \coloneqq \exp\left(-\frac{\gamma^{-2} - 1}{2}(\mathbf{v} \cdot \mathbf{x})^2\right).$$

Let $\mathbf{x}_\gamma$ denote the sample $\mathbf{x}$ conditioned on being accepted by $\mathrm{RS}_{\gamma,\mathbf{v}}$. Then, for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, we have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\mathbf{x}$ is accepted$] = \gamma$ and $\mathbf{x}_\gamma \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_{\gamma,\mathbf{v}})$ with $\mathbf{A}_{\gamma,\mathbf{v}} \coloneqq \mathbf{I}_d - (1-\gamma^2)\mathbf{v}\mathbf{v}^\top$. Furthermore, for any $\delta > 0$, in order to obtain N i.i.d. samples from $\mathbf{x}_\gamma$ with probability at least $1 - \delta$, we need $O(N \log(N/\delta)/\gamma)$ samples from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

For concreteness of discussion, we introduce the following notation and terminology for the conditional distribution of samples following the rejection sampling procedure from Fact 3.2.

Definition 3.3 ($(\gamma, \mathbf{v})$-Rejection Sampling Distribution)**.** *Given a distribution $\mathcal{D}$ supported on $\mathbb{R}^d \times \mathbb{R}$ and samples $(\mathbf{x}, y) \sim \mathcal{D}$, consider the rejection sampling procedure from Fact 3.2 applied to the covariates $\mathbf{x}$. Whenever $\mathbf{x}$ is accepted, accept the whole labeled example $(\mathbf{x}, y)$, and let $\mathbf{z} = \mathbf{A}_{\gamma,\mathbf{v}}^{-1/2}\mathbf{x}$. The distribution of transformed accepted samples $(\mathbf{z}, y)$ is then called the $(\gamma, \mathbf{v})$-rejection sampling distribution obtained from $\mathcal{D}$, and it is denoted by $\mathcal{D}^{\gamma,\mathbf{v}}$.*

We now argue that $\mathcal{D}^{\gamma,\mathbf{v}}$ itself defines an instance of a sum of k ReLUs under Gaussian marginals.

Fact 3.4 (Properties of Rejection Sampling)**.** *Fix parameters $\mathbf{v} \in \mathbb{S}^{d-1}$ and $\gamma \in (0, 1)$. Let $(\mathbf{x}, y) \sim \mathcal{D}$ be such that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $y(\mathbf{x}) = \sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$ with $\|\mathbf{w}_i\|_2 = 1$ and $\beta_i \in [0, 1]$ for $i \in [k]$. Let $(\mathbf{x}_\gamma, y_\gamma)$ be the sample accepted by $\mathrm{RS}_{\gamma,\mathbf{v}}$. Let $\mathbf{z} \coloneqq \mathbf{A}_{\gamma,\mathbf{v}}^{-1/2}\mathbf{x}_\gamma$, where $\mathbf{A}_{\gamma,\mathbf{v}} \coloneqq \mathbf{I}_d - (1-\gamma^2)\mathbf{v}\mathbf{v}^\top$. Then, the following statements both hold:*

1. *$\mathbf{z}$ follows the standard normal distribution, i.e., $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.*
2. *$\mathbf{E}[y_\gamma \mid \mathbf{z}] = \sum_{i=1}^k \widetilde{\beta}_i \sigma(\widetilde{\mathbf{w}}_i \cdot \mathbf{z})$, where $\overline{\mathbf{w}}_i = \mathbf{w}_i^{\perp\mathbf{v}} + \gamma(\mathbf{v} \cdot \mathbf{w}_i)\mathbf{v}$, $\widetilde{\mathbf{w}}_i = \overline{\mathbf{w}}_i / \|\overline{\mathbf{w}}_i\|_2$, and $\widetilde{\beta}_i = \beta_i \|\overline{\mathbf{w}}_i\|_2 \leq \beta_i$.*

Proof. The acceptance decision depends only on $\mathbf{x}$. Since $\mathbf{x}_\gamma = \mathbf{A}_{\gamma,\mathbf{v}}^{1/2}\mathbf{z}$, we have

$$\mathbf{E}[y_\gamma \mid \mathbf{z}] = \sum_{i=1}^k \beta_i \sigma\left(\mathbf{w}_i^\top \mathbf{A}_{\gamma,\mathbf{v}}^{1/2}\mathbf{z}\right) = \sum_{i=1}^k \beta_i \sigma\left(\left(\mathbf{A}_{\gamma,\mathbf{v}}^{1/2}\mathbf{w}_i\right)^\top \mathbf{z}\right).$$

We can write $\mathbf{w}_i = \mathbf{w}_i^{\perp\mathbf{v}} + (\mathbf{v}\cdot\mathbf{w}_i)\mathbf{v}$. Therefore, $\mathbf{A}_{\gamma,\mathbf{v}}^{1/2}\mathbf{w}_i = \mathbf{w}_i^{\perp\mathbf{v}} + \gamma(\mathbf{v}\cdot\mathbf{w}_i)\mathbf{v} = \overline{\mathbf{w}}_i$. Using positive homogeneity of ReLU, namely $\sigma(\overline{\mathbf{w}}_i\cdot\mathbf{z}) = \|\overline{\mathbf{w}}_i\|_2\sigma(\widetilde{\mathbf{w}}_i\cdot\mathbf{z})$, we get

$$\mathbf{E}[y_\gamma \mid \mathbf{z}] = \sum_{i=1}^{k}\beta_i\|\overline{\mathbf{w}}_i\|_2\sigma(\widetilde{\mathbf{w}}_i\cdot\mathbf{z}) = \sum_{i=1}^{k}\widetilde{\beta}_i\sigma(\widetilde{\mathbf{w}}_i\cdot\mathbf{z}).$$

Finally, $\widetilde{\beta}_i \le \beta_i$ because $\|\overline{\mathbf{w}}_i\|_2 \le \|\mathbf{w}_i\|_2$, as $\gamma\in(0,1)$. □

With Fact 3.4 on hand, we now formally argue that CDR can be applied to estimate $\mathbf{w}_i^{\perp\mathbf{v}}$.

Proposition 3.5 (Estimating Update Directions)**.** *Fix a unit vector $\mathbf{v}\in\mathbb{R}^d$ and let $\lambda,\delta\in(0,1)$. Let $\mathcal{D}$ be a sum of k ReLUs under Gaussian marginals and, given $\gamma=\lambda/100$, let $\mathcal{D}^{\gamma,\mathbf{v}}$ be the $(\gamma,\mathbf{v})$-rejection sampling distribution obtained from $\mathcal{D}$. Then, procedure $\mathsf{CDR}(\mathcal{D}^{\gamma,\mathbf{v}},\lambda,\delta)$ that satisfies Assumption 1 applied to i.i.d. samples from $\mathcal{D}^{\gamma,\mathbf{v}}$ returns a list $\mathcal{L}$ such that with probability at least $1-\delta$, for any $\mathbf{w}_i$ with $\beta_i\sin(\theta_\pm(\mathbf{w}_i,\mathbf{v}))\ge\lambda$, there exists a unit vector $\mathbf{r}_i\in\mathcal{L}$ with $|\widehat{\mathbf{r}_i^{\perp\mathbf{v}}}\cdot\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}|\ge 9/10$.*

Proof. Denote the projective angle between $\mathbf{w}_i$ and $\mathbf{v}$ by $\theta_i=\theta_\pm(\mathbf{w}_i,\mathbf{v})$. From Fact 3.4, we have that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}^{\gamma,\mathbf{v}}}[y\mid\mathbf{x}] = \sum_{i=1}^{k}\widetilde{\beta}_i\sigma(\widetilde{\mathbf{w}}_i\cdot\mathbf{z})\ ,$$

with $\overline{\mathbf{w}}_i=\mathbf{w}_i^{\perp\mathbf{v}}+\gamma(\mathbf{v}\cdot\mathbf{w}_i)\mathbf{v}$, $\widetilde{\mathbf{w}}_i=\overline{\mathbf{w}}_i/\|\overline{\mathbf{w}}_i\|_2$, and $\widetilde{\beta}_i=\beta_i\|\overline{\mathbf{w}}_i\|_2$. Observe that

$$\begin{aligned}\|\widetilde{\mathbf{w}}_i-\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}\|_2^2 &= \|\widetilde{\mathbf{w}}_i\|_2^2+\|\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}\|_2^2-2\widetilde{\mathbf{w}}_i^\top\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}\\ &= 2\left(1-\frac{\|\mathbf{w}_i^{\perp\mathbf{v}}\|_2}{\|\mathbf{w}_i^{\perp\mathbf{v}}+\gamma(\mathbf{v}\cdot\mathbf{w}_i)\mathbf{v}\|_2}\right)\\ &= 2\left(1-\frac{1}{(1+\gamma^2(\mathbf{v}\cdot\mathbf{w}_i)^2/\|\mathbf{w}_i^{\perp\mathbf{v}}\|_2^2)^{1/2}}\right)\le\frac{\gamma^2(\mathbf{v}\cdot\mathbf{w}_i)^2}{\|\mathbf{w}_i^{\perp\mathbf{v}}\|_2^2}\ ,\end{aligned}\tag{7}$$

where the second equality uses that both $\widetilde{\mathbf{w}}_i,\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}$ are unit vectors and applies their definitions to simplify the inner product, while the last inequality uses $1/(1+z^2)^{1/2}\ge 1-z^2/2$ (this can be seen as $\varphi(x)=1/(1+x)^{1/2}$, $x\ge0$, is convex, by applying $\varphi(x)\ge\varphi(0)+\varphi'(0)x$ and plugging in $x=z^2$). By using that $\beta_i\sin\theta_i\ge\lambda$ from our assumptions, we have that $1/\sin(\theta_i)\le\beta_i/\lambda\le1/\lambda$, as $\beta_i\in(0,1]$. Combining with (7), we thus get that

$$\|\widetilde{\mathbf{w}}_i-\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}\|_2\le\frac{\gamma|\mathbf{v}\cdot\mathbf{w}_i|}{\left\|\mathbf{w}_i^{\perp\mathbf{v}}\right\|_2}=\frac{\gamma|\cos(\theta_i)|}{|\sin(\theta_i)|}\le\gamma/\lambda\le1/100.$$

Therefore, we have that $\mathsf{CDR}(\mathcal{D}^{\gamma,\mathbf{v}},\lambda,\delta)$ returns a list $\mathcal{L}$ containing $\mathbf{r}_i\in\mathcal{L}$ with $|\mathbf{r}_i\cdot\widetilde{\mathbf{w}}_i|\ge9.1/10$ for any $\mathbf{w}_i$ satisfying $\beta_i\sin(\theta_i)\ge\lambda$. Since, by the triangle inequality and Cauchy-Schwarz,

$$|\mathbf{r}_i\cdot\widetilde{\mathbf{w}}_i|\le|\mathbf{r}_i\cdot\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}|+\|\mathbf{r}_i\|_2\|\widetilde{\mathbf{w}}_i-\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}\|\le|\mathbf{r}_i\cdot\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}|+1/100,$$

we get that $|\widehat{\mathbf{r}_i^{\perp\mathbf{v}}}\cdot\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}|\ge|\mathbf{r}_i\cdot\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}|\ge9.1/10-1/100=9/10$. □

Since the rejection sampling procedure works by (approximately) localizing the marginal distribution of $\mathbf{x}$ to a thin band $|\mathbf{v}\cdot\mathbf{x}|\le\gamma$, it is typically referred to as "localization". The localization argument in Proposition 3.5 allows us to obtain components of the hidden weight vectors that our algorithm potentially has not yet discovered. The subroutine $\mathsf{Refine}(\mathbf{v},r,\epsilon_0,\delta_{\mathsf{CDR}};\mathcal{D})$ formalizes the updating procedure to obtain vectors closer to the target ones.

In Steps (5) through (11), we guarantee that we have created a new set that contains some vectors that are closer than before to some of the hidden vectors, by a multiplicative factor 3/5. We note that for technical reasons (Proposition 3.10 requires a large decrease between iterations), we require the update to decrease the angle by a factor larger than 3/5; for this purpose, in $\mathsf{Refine}(\mathbf{v},r,\epsilon_0,\delta_{\mathsf{CDR}};\mathcal{D})$, we boost this procedure by applying three update steps instead of one.

Algorithm 2 $\mathsf{Refine}(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$

Input: unit vector $\mathbf{v} \in \mathbb{S}^{d-1}$, radius upper bound $r \in (\epsilon_0, \pi/2]$, accuracy $\epsilon_0 > 0$, failure probability $\delta_{\mathsf{CDR}} \in (0,1)$, i.i.d. access to $\mathcal{D}$.
Output: A list of vectors $\mathcal{V}$ satisfying the guarantees of Lemma 3.6.

$\chi \leftarrow 3/5$, $\mathcal{V} \leftarrow \{\mathbf{v}\}$, $\lambda_{\mathsf{CDR}} \leftarrow \epsilon_0 \sin(\chi^3 \epsilon_0)$, $\gamma \leftarrow \lambda_{\mathsf{CDR}}/100$. ▷ λ_{CDR}: coefficient threshold; γ: rejection sampling width.
for $j = 0, 1, 2$ **do**
  $\eta_j \leftarrow \tan(4\chi^j r/5)$.
  for all $\mathbf{v} \in \mathcal{V}$ **do**
    Construct $\mathcal{D}^{\gamma,\mathbf{v}}$ by applying the rejection sampling procedure $\mathrm{RS}_{\gamma,\mathbf{v}}$ (Fact 3.2).
    $\mathcal{L}_0(\mathbf{v}) \leftarrow \mathsf{CDR}(\mathcal{D}^{\gamma,\mathbf{v}}, \lambda_{\mathsf{CDR}}, \delta_{\mathsf{CDR}})$ ▷ Apply the CDR procedure that satisfies Assumption 1.
    $\mathcal{L}(\mathbf{v}) \leftarrow \{\widehat{\mathbf{q}^{\perp \mathbf{v}}} : \mathbf{q} \in \mathcal{L}_0(\mathbf{v})\}$. ▷ Project the vectors obtained from the initialization procedure to the tangent space of $\mathbf{v}$ then normalize to unit norm (Proposition 3.5).
    for $\mathbf{r} \in \mathcal{L}(\mathbf{v})$ **do**
      Add $\mathrm{proj}_{\mathbb{S}^{d-1}}(\mathbf{v} + \eta_j \mathbf{r})$ and $\mathrm{proj}_{\mathbb{S}^{d-1}}(\mathbf{v} - \eta_j \mathbf{r})$ to $\mathcal{V}$. ▷ Gradient descent/Refinement step
Return: $\mathcal{V}$

The following lemma establishes that our $\mathsf{Refine}(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$ subroutine contracts the distance between the input candidate vector $\mathbf{v}$ and the λ_{CDR}-heavy hidden vectors $\mathbf{w}_i$, and bounds its sample and computational complexities.

Lemma 3.6 (Properties of Refine)**.** *Fix the input parameters to* $\mathsf{Refine}(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$*:* $\mathbf{v} \in \mathbb{S}^{d-1}$, $r \in [\epsilon_0, \pi/2)$, $\epsilon_0 \in (0, 1/2)$, $\delta_{\mathsf{CDR}} \in (0,1)$*, and a Gaussian sum of* k *ReLUs distribution* $\mathcal{D}$*, and let all other algorithm parameters be specified as in the algorithm statement. Suppose that for some* $i \in [k]$*,* $\beta_i \geq \epsilon_0$ *and* $\theta_\pm(\mathbf{v}, \mathbf{w}_i) \leq r$*. Then, with probability at least* $1 - O(L_{\mathsf{CDR}}^3 \delta_{\mathsf{CDR}})$*, the list* $\mathcal{V}$ *returned by* $\mathsf{Refine}(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$ *contains a vector* $\mathbf{v}_i \in \mathcal{V}$ *satisfying*

$$\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \leq \chi^3 r.$$

Moreover, $\mathsf{Refine}(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$ *draws* $N_{\mathsf{Refine}} = \widetilde{O}(L_{\mathsf{CDR}}^3 N_{\mathsf{CDR}} \log(1/\delta_{\mathsf{CDR}})/\gamma)$ *samples from* $\mathcal{D}$*, runs in time* $d\,\mathrm{poly}(N_{\mathsf{Refine}}, T_{\mathsf{CDR}})$ *and the size of the output list* $\mathcal{V}$ *is bounded by* $|\mathcal{V}| = O(L_{\mathsf{CDR}}^3)$*, where* $L_{\mathsf{CDR}}, N_{\mathsf{CDR}}, T_{\mathsf{CDR}}$ *are the number of vectors returned, the sample complexity, and the runtime of* $\mathsf{CDR}(\mathcal{D}^{\gamma,\mathbf{v}}; \lambda_{\mathsf{CDR}}, \delta_{\mathsf{CDR}})$*, respectively.*

Proof. Recall that $\chi := 3/5$. If $\theta_\pm(\mathbf{v}, \mathbf{w}_i) \leq \chi^3 r$, then $\mathbf{v}$ already satisfies the condition. Otherwise, since we have that $\beta_i \geq \epsilon_0$ and $r \geq \epsilon_0$, we get that $\beta_i \sin\theta_\pm(\mathbf{v}, \mathbf{w}_i) \geq \beta_i \sin(\chi^3 r) \geq \epsilon_0 \sin(\chi^3 \epsilon_0) = \lambda_{\mathsf{CDR}}$. We argue that after each step $j \in \{0, 1, 2\}$, there exists a vector $\mathbf{v}_i \in \mathcal{V}$ such that $\chi^{j+1} r \geq \theta_\pm(\mathbf{v}_i, \mathbf{w}_i)$. Before the algorithm starts, the input vector $\mathbf{v}$ satisfies $\theta_\pm(\mathbf{v}, \mathbf{w}_i) \leq r$, by our assumption. Since $\beta_i \sin\theta_\pm(\mathbf{v}, \mathbf{w}_i) \geq \lambda_{\mathsf{CDR}}$, Proposition 3.5 implies that running CDR with N_{CDR} samples from $\mathcal{D}^{\gamma,\mathbf{v}}$ returns a list $\mathcal{L}$ of unit vectors that contains a vector $\mathbf{r}$ such that $\mathbf{r} \perp \mathbf{v}$ and

$$|\mathbf{r} \cdot \widehat{\mathbf{w}_i^{\perp \mathbf{v}}}| \geq 0.9 \,.$$

The main insight is that $\mathbf{r}$ defines a direction that can improve the estimate of the hidden direction $\mathbf{w}_i$. In particular, as argued in the next claim, a projected gradient descent-style update of a candidate vector $\mathbf{v}$ along either $\mathbf{r}$ or $-\mathbf{r}$ leads to a new unit vector that is better aligned with (and thus closer to) $\mathbf{w}_i$, unless the vector $\mathbf{v}$ was already well-aligned with (i.e., sufficiently close to) $\mathbf{w}_i$.

Claim 3.7 (One-step improvement in projective angle)**.** *Let* $0 < r \leq \pi/2$*. Let* $\mathbf{u}, \mathbf{w} \in \mathbb{S}^{d-1}$ *be such that* $\theta_\pm(\mathbf{u}, \mathbf{w}) \leq r$*. Let* $\mathbf{r} \in \mathbf{u}^\perp$ *be a unit vector such that* $|\mathbf{r} \cdot \widehat{\mathbf{w}^{\perp \mathbf{u}}}| \geq 0.9$*. Let* $\phi := \frac{4r}{5}, \eta := \tan\phi$*, and define*

$$\mathbf{u}_+ := \mathrm{proj}_{\mathbb{S}^{d-1}}(\mathbf{u} + \eta \mathbf{r}), \qquad \mathbf{u}_- := \mathrm{proj}_{\mathbb{S}^{d-1}}(\mathbf{u} - \eta \mathbf{r}).$$

Then $\min_{\mathbf{z} \in \{\mathbf{u}, \mathbf{u}_+, \mathbf{u}_-\}} \theta_\pm(\mathbf{z}, \mathbf{w}) \leq 3r/5 = \chi r$.

Proof of Claim 3.7. Let $\theta := \theta_{\pm}(\mathbf{u}, \mathbf{w})$. Assume without loss of generality that $\theta(\mathbf{u}, \mathbf{w}) \leq \pi/2$. If $\theta \leq 3r/5$, then the vector $\mathbf{u}$ already satisfies the conclusion. Therefore, we assume throughout the proof that $3r/5 < \theta \leq r$.

Let $\mathbf{q} = \widehat{\mathbf{w}^{\perp \mathbf{u}}}$ and note that $\mathbf{q} \cdot \mathbf{u} = 0$ and $\mathbf{w} = \cos\theta\, \mathbf{u} + \sin\theta\, \mathbf{q}$. Recall that $|\mathbf{r} \cdot \mathbf{q}| \geq 0.9$. Assume without loss of generality that $\mathbf{r} \cdot \mathbf{q} \geq 0.9$ and consider $\mathbf{u}' = \mathrm{proj}_{\mathbb{S}^{d-1}}(\mathbf{u} + \eta \mathbf{r})$, as the other case is captured by the update rule $\mathbf{u}_-$ with exactly the same analysis. Letting $\delta = \theta(\mathbf{r}, \mathbf{q})$, then $\delta \leq \arccos(0.9) < 1/2$. Consider the ideal update, i.e., the update we could take if we knew exactly the unknown part $\mathbf{q}$:

$$\mathbf{z} := \cos\phi\, \mathbf{u} + \sin\phi\, \mathbf{q}\ .$$

Note that the angle between $\mathbf{u}$ and $\mathbf{z}$ is ϕ, and the angle between $\mathbf{u}$ and $\mathbf{w}$ is θ. Therefore

$$\theta(\mathbf{z}, \mathbf{w}) = |\theta - \phi| \leq |\theta - 4r/5| \leq r/5\ , \tag{8}$$

where we used that $\phi = 4r/5$ and $r \geq \theta \geq 3r/5$. Recall that $\mathbf{u}'$, the output of the update, was defined by $\mathbf{u}' = \mathrm{proj}_{\mathbb{S}^{d-1}}(\mathbf{u} + \eta \mathbf{r})$, and observe that

$$\mathbf{u}' = \frac{\mathbf{u} + \tan\phi \mathbf{r}}{\sqrt{1 + \tan^2\phi}} = \cos\phi\, \mathbf{u} + \sin\phi\, \mathbf{r}\ .$$

Next we bound $\theta(\mathbf{u}', \mathbf{w})$. For that, we need to bound $\theta(\mathbf{u}', \mathbf{z})$. Note that $\|\mathbf{u}' - \mathbf{z}\|_2 = 2\sin(\theta(\mathbf{u}', \mathbf{z})/2)$. Furthermore, by construction, $\|\mathbf{u}' - \mathbf{z}\|_2 = \sin\phi \|\mathbf{q} - \mathbf{r}\|_2 = 2\sin\phi\sin(\delta/2)$. Therefore, we have that $\sin(\theta(\mathbf{u}', \mathbf{z})/2) \leq \sin\phi\sin(\delta/2)$. Noting that the sine function is concave on $(0, \pi/2)$, from Jensen's inequality we have $\sin((\delta/2)\,\phi) \geq (\delta/2)\sin\phi + (1 - \delta/2)\sin(0)$. Using that $\sin(\delta/2) \leq \delta/2$, we get that $\sin\phi\sin(\delta/2) \leq \delta/2\sin(\phi) \leq \sin(\phi\delta/2)$. Therefore, from monotonicity, we get that

$$\theta(\mathbf{u}', \mathbf{z}) \leq \delta\phi \leq 2r/5\ . \tag{9}$$

Since $\theta(\mathbf{u}, \mathbf{v})$ is a metric on the unit sphere, by the triangle inequality, and using (8) and (9), we get

$$\theta(\mathbf{u}', \mathbf{w}) \leq \theta(\mathbf{u}', \mathbf{z}) + \theta(\mathbf{z}, \mathbf{w}) \leq \frac{2r}{5} + \frac{r}{5} = \frac{3r}{5}\ .$$

Since $\mathbf{u}' \in \{\mathbf{u}_+, \mathbf{u}_-\}$, this gives $\min_{\mathbf{z} \in \{\mathbf{u}, \mathbf{u}_+, \mathbf{u}_-\}} \theta_{\pm}(\mathbf{z}, \mathbf{w}) \leq \frac{3r}{5}$, completing the proof of Claim 3.7. □

By Claim 3.7, the list $\mathcal{V}$ contains a vector $\mathbf{v}_i$ such that $\chi r \geq \theta_{\pm}(\mathbf{v}_i, \mathbf{w}_i)$. Hence, after step $j = 0$, we have $\chi r \geq \theta_{\pm}(\mathbf{v}_i, \mathbf{w}_i)$ for some vector $\mathbf{v}_i \in \mathcal{V}$. This is the base case of the induction argument.

Now at step $j = \{1, 2\}$, by induction, we know that before updating, there exists a vector $\mathbf{v}_i \in \mathcal{V}$ such that $\chi^j r \geq \theta_{\pm}(\mathbf{v}_i, \mathbf{w}_i)$. Suppose $\theta_{\pm}(\mathbf{v}_i, \mathbf{w}_i) \geq \chi^{j+1} r$, which implies $\beta_i \sin\theta_{\pm}(\mathbf{v}_i, \mathbf{w}_i) \geq \lambda_{\mathsf{CDR}}$. Then, using Proposition 3.5 again, the CDR procedure returns a list $\mathcal{L}(\mathbf{v}_i)$ of unit vectors that contains a vector $\mathbf{r}$ such that $\mathbf{r} \perp \mathbf{v}_i$ and $|\mathbf{r} \cdot \widehat{\mathbf{w}_i^{\perp \mathbf{v}_i}}| \geq 0.9$. Then Claim 3.7 implies that after updating $\mathbf{v}_i$, inside the list $\mathcal{V}$ there is a vector $\mathbf{v}_i'$ so that $\chi^{j+1} r \geq \theta_{\pm}(\mathbf{v}_i', \mathbf{w}_i)$. Therefore, at step $j = 2$, the algorithm produces a vector $\mathbf{v} \in \mathcal{V}$ such that $\theta_{\pm}(\mathbf{v}, \mathbf{w}_i) \leq \chi^3 r$.

At each call, CDR uses N_{CDR} samples from $\mathcal{D}^{\gamma, \mathbf{v}}$. By Fact 3.2, with probability at least $1 - \delta_{\mathsf{CDR}}$, these N_{CDR} samples from $\mathcal{D}^{\gamma, \mathbf{v}}$ can be obtained by drawing $O(N_{\mathsf{CDR}} \log(N_{\mathsf{CDR}}/\delta_{\mathsf{CDR}})/\gamma)$ samples from the original distribution $\mathcal{D}$, and passing through the $(\gamma, \mathbf{v})$-rejection procedure $\mathrm{RS}_{\gamma, \mathbf{v}}$. Conditioned on successfully obtaining these N_{CDR} from $\mathcal{D}^{\gamma, \mathbf{v}}$, the CDR procedure succeeds with probability at least $1 - \delta_{\mathsf{CDR}}$, by Assumption 1 and Proposition 3.5. Therefore, the total number of samples required by CDR is $O(N_{\mathsf{CDR}} \log(N_{\mathsf{CDR}}/\delta_{\mathsf{CDR}})/\gamma)$, and the probability of success for one call of CDR procedure is at least $(1 - \delta_{\mathsf{CDR}})^2 \geq 1 - 2\delta_{\mathsf{CDR}}$. Recall that for each query vector $\mathbf{v} \in \mathcal{V}$, CDR returns a list of $|\mathcal{L}(\mathbf{v})| \leq L_{\mathsf{CDR}}$ vectors. Therefore, the CDR gets called $T = (2|\mathcal{L}|)^3 = O(L_{\mathsf{CDR}}^3)$ times.

Thus, in total, $\mathsf{Refine}(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$ uses $T N_{\mathsf{CDR}} = O(L_{\mathsf{CDR}}^3 N_{\mathsf{CDR}} \log(N_{\mathsf{CDR}}/\delta_{\mathsf{CDR}})/\gamma)$ samples from $\mathcal{D}$, and succeeds with probability $1 - 2T\delta_{\mathsf{CDR}} = 1 - O(L_{\mathsf{CDR}}^3 \delta_{\mathsf{CDR}})$. The size of $\mathcal{V}$ is $|\mathcal{V}| = O(L_{\mathsf{CDR}}^3)$. This completes the proof of Lemma 3.6. □

3.1.2 Closeness Tester

The Refine procedure ($\mathsf{Refine}(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$) returns a large list $\mathcal{V}$ of vectors with the guarantee that for each λ-heavy hidden weight vector $\mathbf{w}_i$, there is at least one vector $\mathbf{v} \in \mathcal{V}$ that is sufficiently close,

i.e., $\theta_\pm(\mathbf{w}_i, \mathbf{v}) \le r$ for some $r > 0$ (Lemma 3.6). However, Refine branches over all possible updates, because it has no information on which hidden weight vector the input $\mathbf{v}$ approximates nor which update directions from CDR are useful for improving the approximation. As a result, the output list may contain many irrelevant vectors that provide no useful approximation of any of the hidden weight vectors. To eliminate irrelevant candidate vectors, we design a tester that distinguishes between $\mathbf{v} \in \mathcal{V}$ that are r-close to at least one heavy hidden weight vector versus $(1+\tau)r$-far from all hidden weight vectors, for constant $\tau > 0$.

For an even degree p, we use the following moment test (see Fact 2.3):

$$a_p(\mathbf{v}) := \frac{1}{\widehat{\sigma}_p} \mathbf{E}[y \mathrm{he}_p(\mathbf{v} \cdot \mathbf{x})] = \sum_i \beta_i |\mathbf{v} \cdot \mathbf{w}_i|^p.$$

The main intuition behind the tester is that the small value of $a_p(\mathbf{v})$ identifies vectors $\mathbf{v}$ that are far from all hidden directions $\mathbf{w}_i$. To see this, recall first that $|\mathbf{v} \cdot \mathbf{w}_i| = \cos(\theta_\pm(\mathbf{w}_i, \mathbf{v}))$. If $\mathbf{v}$ is r-close to at least one ϵ_0-heavy $\mathbf{w}_i$ (i.e., if $\theta_\pm(\mathbf{w}_i, \mathbf{v}) \le r$), then $a_p(\mathbf{v}) \ge \epsilon_0 \cos^p(r)$. On the other hand, if $\theta_\pm(\mathbf{w}_i, \mathbf{v}) > (1+\tau)r$ for all $i \in [k]$, then $|\mathbf{v} \cdot \mathbf{w}_i| = \cos(\theta_\pm(\mathbf{w}_i, \mathbf{v})) \le (1 - \tau r^2/2)\cos(r)$ for all $i \in [k]$, and so $a_p(\mathbf{v}) \le k(1-\tau r^2)^p \cos^p(r)$. Choosing p sufficiently large ($p \gg \log(k/\epsilon_0)/(\tau r^2)$) then ensures $a_p(\mathbf{v}) \ll \epsilon_0 \cos^p(r)$, so the two cases can be separated by the tester.

Algorithm 3 $\mathsf{Test}(\mathbf{v}, r, \epsilon_0, \tau, \delta_{\mathsf{Test}}; \mathcal{D})$

Input: Unit vector $\mathbf{v}$, $\epsilon_0, \delta_{\mathsf{Test}} > 0$, $\tau > 0$, upper bound r, sample access to $\mathcal{D}$.
Output: TRUE if there exists $i \in [k]$ such that $\theta_\pm(\mathbf{w}_i, \mathbf{v}) \le r$ and $\beta_i \ge \epsilon_0$; FALSE if for all $i \in [k]$: $\theta_\pm(\mathbf{w}_i, \mathbf{v}) > (1+\tau)r$.
$p \leftarrow 2\lceil (1/(\tau r^2)) \log(k/\epsilon_0) \rceil$.
Draw $N = \mathrm{poly}(k, 1/\epsilon_0)(1/\delta_{\mathsf{Test}})$ samples from $\mathcal{D}$ to formulate the empirical distribution $\widehat{\mathcal{D}}$.
$\widehat{a}_p \leftarrow (1/\widehat{\sigma}_p) \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[y \mathrm{he}_p(\mathbf{v} \cdot \mathbf{x})]$.
if $\widehat{a}_p \ge (7/8)\epsilon_0 \cos^p(r)$ **then**
    Return: TRUE.
else
    Return: FALSE.

Lemma 3.8 (Smallest Angle Tester)**.** *Let $\epsilon_0 \in (0, 1/2]$, $r \ge \epsilon_0/(5k)$, let $\tau > 0$ be a small absolute constant, and define $p = 2\lceil (1/(\tau r^2)) \log(k/\epsilon_0) \rceil$. Using $N = \mathrm{poly}(k, 1/\epsilon_0)(1/\delta_{\mathsf{Test}})$ samples and $O(dN)$ time with probability at least $1 - \delta_{\mathsf{Test}}$, $\mathsf{Test}(\mathbf{v}, r, \epsilon_0, \tau, \delta_{\mathsf{Test}}; \mathcal{D})$ has the following performance guarantee:*

1. *(Completeness) If there exists $i \in [k]$ such that $\theta_\pm(\mathbf{w}_i, \mathbf{v}) \le r$ and $\beta_i \ge \epsilon_0$, then $\mathsf{Test}(\mathbf{v}, r, \epsilon_0, \tau, \delta_{\mathsf{Test}}; \mathcal{D})$ returns* TRUE.
2. *(Soundness) If $\min_{i\in[k]} \theta_\pm(\mathbf{w}_i, \mathbf{v}) > (1+\tau)r$, then $\mathsf{Test}(\mathbf{v}, r, \epsilon_0, \tau, \delta_{\mathsf{Test}}; \mathcal{D})$ returns* FALSE.

Proof. Denote $\theta_i := \theta_\pm(\mathbf{w}_i, \mathbf{v})$ and suppose without loss of generality that $\theta_1 = \min_{i\in[k]} \theta_i$ and $\beta_1 \ge \epsilon_0$. For even degree p, by Fact 2.3, we have that

$$a_p := \frac{1}{\widehat{\sigma}_p} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y \mathrm{he}_p(\mathbf{v} \cdot \mathbf{x})] = \frac{1}{\widehat{\sigma}_p} \sum_{i=1}^k \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x}) \mathrm{he}_p(\mathbf{v} \cdot \mathbf{x})] = \sum_{i=1}^k \beta_i \cos^p \theta_i,$$

where $\widehat{\sigma}_p$ is the pth Hermite coefficient of the activation σ. We first prove the following simple claim, establishing a bound between the empirical estimate $\widehat{a}_p$ and the population mean a_p. Note that in Claim 3.9, the number of samples required for our tester has polynomial dependence on $1/\delta_{\mathsf{Test}}$. This can be replaced by logarithmic dependence using median-of-means estimation [NY83] instead of the simple empirical estimator used in $\mathsf{Test}(\mathbf{v}, r, \epsilon_0, \tau, \delta_{\mathsf{Test}}; \mathcal{D})$. Yet, for our purposes, a polynomial dependence is sufficient.

Claim 3.9. *With probability at least $1 - \delta_{\mathsf{Test}}$, the empirical estimator $\widehat{a}_p$ of a_p that uses $N = \mathrm{poly}(k, 1/\epsilon_0)(1/\delta_{\mathsf{Test}})$ samples satisfies $|\widehat{a}_p - a_p| \le (1/8)\epsilon_0 \cos^p(r)$.*

Proof of Claim 3.9. Note that $\widehat{a}_p$ is an unbiased estimator of a_p. The proof of the claim follows by bounding the variance of the estimate and applying Chebyshev's inequality.

We start by noting that the variance of $y\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})$ is bounded by:

$$\mathbf{E}[y^2\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})^2] \le k\sum_{i=1}^{k}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\sigma(\mathbf{w}_i\cdot\mathbf{x})^2\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})^2] \le k\sum_{i=1}^{k}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{w}_i\cdot\mathbf{x})^2\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})^2]\,. \tag{10}$$

Using the standard Hermite identity $z\mathrm{he}_p(z) = \sqrt{p+1}\mathrm{he}_{p+1}(z) + \sqrt{p}\mathrm{he}_{p-1}(z)$ (see Fact 2.3) and the orthogonality of Hermite polynomials, we get that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{v}\cdot\mathbf{x})^2\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})^2] = (p+1)\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\mathrm{he}_{p+1}(\mathbf{v}\cdot\mathbf{x})^2] + p\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\mathrm{he}_{p-1}(\mathbf{v}\cdot\mathbf{x})^2] = 2p+1.$$

Therefore, by (10), since $p \ge 2$, we have:

$$\begin{aligned}
\mathbf{E}[y^2\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})^2] &\le k\sum_{i=1}^{k}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\cos\theta(\mathbf{w}_i,\mathbf{v})\mathbf{v}\cdot\mathbf{x} + \sin\theta(\mathbf{w}_i,\mathbf{v})\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}\cdot\mathbf{x})^2\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})^2] \\
&\le k\sum_{i=1}^{k}(2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{v}\cdot\mathbf{x})^2\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})^2] + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\widehat{\mathbf{w}_i^{\perp\mathbf{v}}}\cdot\mathbf{x})^2]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})^2]) \\
&= 2k\sum_{i=1}^{k}((2p+1)+1) \le 6k^2p.
\end{aligned}$$

Since p is an even integer, $\widehat{\sigma}_p \ne 0$. From Chebyshev's inequality, we get

$$\mathbf{Pr}[|\widehat{a}_p - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(1/\widehat{\sigma}_p)y\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})]| \ge t] \le \frac{\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y^2\mathrm{he}_p(\mathbf{v}\cdot\mathbf{x})^2]}{\widehat{\sigma}_p^2Nt^2} \le \frac{6k^2p}{N\widehat{\sigma}_p^2t^2}.$$

Choosing $t = (1/8)\epsilon_0\cos^p(r)$ and $N = (6k^2p)/(\widehat{\sigma}_p^2t^2\delta_{\mathsf{Test}}) \lesssim k^2p/(\widehat{\sigma}_p^2\epsilon_0^2\cos^{2p}(r)\delta_{\mathsf{Test}})$ then leads to $|\widehat{a}_p - a_p| \le (1/8)\epsilon_0\cos^p(r)$ with probability at least $1-\delta_{\mathsf{Test}}$.

Recall that $\widehat{\sigma}_p^2 \gtrsim p^{-3}$ (see Fact 2.4) and the elementary identity $\cos\theta = \exp(-\int_0^\theta \tan(s)\,\mathrm{d}s) \ge \exp(-\theta^2)$. Therefore, when τ is chosen as an absolute positive constant, plugging in our choice of $p = 2\lceil(1/(\tau r^2))\log(k/\epsilon_0)\rceil$, and the assumption $r \ge \epsilon_0/(5k)$, we get that

$$N \lesssim \left(\frac{1}{\delta_{\mathsf{Test}}}\right)\left(\frac{k}{\epsilon_0}\right)^2\frac{p}{\widehat{\sigma}_p^2}\exp(2pr^2) \lesssim \left(\frac{1}{\delta_{\mathsf{Test}}}\right)p^4\left(\frac{k}{\epsilon_0}\right)^{2+4/\tau} \lesssim \left(\frac{\log^4(k/\epsilon_0)}{\delta_{\mathsf{Test}}}\right)\left(\frac{k}{\epsilon_0}\right)^{10+4/\tau}.$$

This completes the proof of Claim 3.9. □

We henceforth condition on the event of Claim 3.9 and we are ready to prove the two statements of the lemma, i.e., the completeness and the soundness of the tester. Note that a direct calculation yields for $\theta_i > r$:

$$\frac{\cos^p\theta_i}{\cos^p(r)} = \exp\left(-p\int_r^{\theta_i}\tan(s)\,\mathrm{d}s\right) \le \exp\left(-\frac{p(\theta_i^2-r^2)}{2}\right).$$

Completeness. Assume first that $\theta_1 \le r$ and $\beta_1 \ge \epsilon_0$. Then, we have:

$$a_p = \sum_{i=1}^{k}\beta_i\cos^p\theta_i \ge \beta_1\cos^p\theta_1 \ge \epsilon_0\cos^p(r).$$

By Claim 3.9, this implies $\widehat{a}_p \ge (7/8)\epsilon_0\cos^p(r)$. Thus, when $\theta_1 \le r$ and $\beta_1 \ge \epsilon_0$, Algorithm 3 would return TRUE.

Soundness. Now assume $(1+\tau)r < \theta_1 \le \theta_i$ for all $i \in [k]$. Then, we have

$$a_p = \sum_{i=1}^{k}\beta_i\cos^p\theta_i \le \sum_{i=1}^{k}\cos^p\theta_i \le ke^{-p(\theta_i^2-r^2)/2}\cos^p(r) \le ke^{-p(\tau r^2)}\cos^p(r) \le \frac{1}{2}\epsilon_0\cos^p(r),$$

where the last inequality is due to our choice of $p = 2\lceil(1/(\tau r^2))\log(k/\epsilon_0)\rceil$. Similarly, by Claim 3.9, this implies $\widehat{a}_p \le (5/8)\epsilon_0\cos^p(r)$. Thus, when no angle θ_i is smaller than $(1+\tau)r$, Algorithm 3 would return FALSE. □

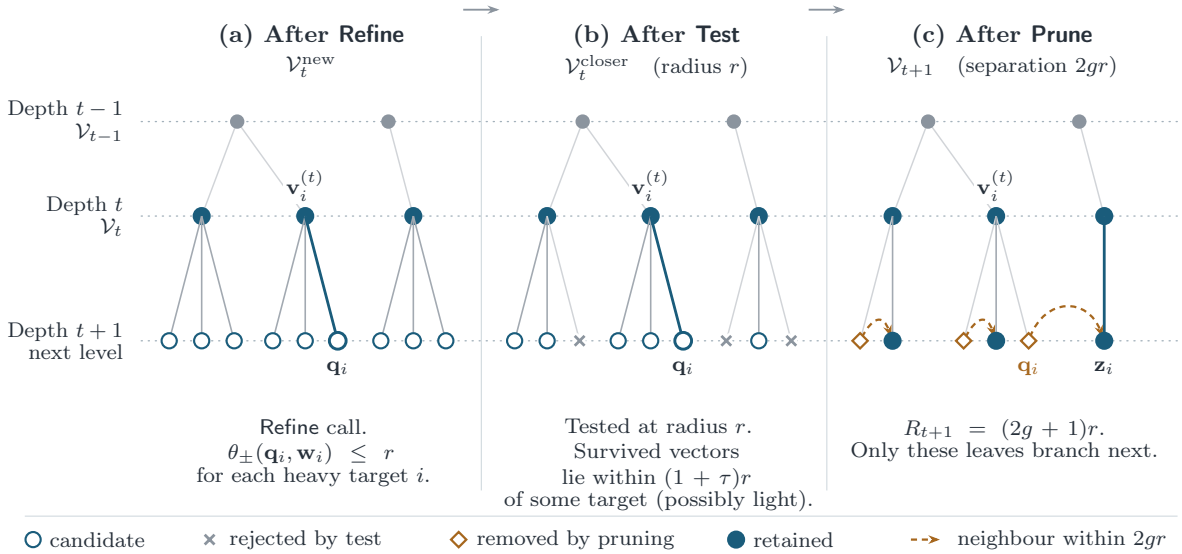


Figure 2: **List of candidate vectors over the different stages of** CoverAndPrune. **Left**: after Refine is called, each $\mathbf{v} \in \mathcal{V}_t$ produces a large list of new candidate vectors; all produced lists combined contain, for each λ-heavy $\mathbf{w}_i$, at least one candidate $\mathbf{v}_i$ with $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \leq r$. **Center**: Test procedure is used to prune all candidates $\mathbf{v}$ far from all $\mathbf{w}_i$, that is, all $\mathbf{v}$ such that $\theta_\pm(\mathbf{v}, \mathbf{w}_i) > (1+\tau)r$, $\forall i \in [k]$. **Right**: Prune procedure makes a single iterative pass through the candidate vectors that survive after Test, forming the output list $\mathcal{V}_{t+1}$; it starts with an empty list and adds a new iterate if and only if it is at least r-far from all the candidates already in the list.

3.1.3 Covering and Pruning

The approach described thus far, based on the Refine subroutine and the Test procedure, provides a concrete way of constructing lists of candidate vectors that progressively improve approximations of ϵ_0-heavy hidden weight vectors. In particular, suppose that the algorithm starts with a list of candidates $\mathcal{V}_t$ such that for each ϵ_0-heavy $\mathbf{w}_i$, $i \in [k]$, there exists at least one $\mathbf{v}_i \in \mathcal{V}_t$ such that $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \leq R_t$. We say in this case that $\mathbf{w}_i$ is covered by $\mathbf{v}_i$ at resolution R_t.

A call to Refine for each $\mathbf{v}_i \in \mathcal{V}_t$ creates a new list of candidate vectors, within which only one candidate vector is guaranteed to land significantly closer to $\mathbf{w}_i$, with $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \leq \chi^3 R_t$. However, it is unclear how to identify such a candidate. What is possible to identify are the candidate vectors that are far from all hidden weight vectors; this is where our Test procedure is used. Thus, Refine and Test can be used jointly to construct a list $\mathcal{V}_t'$ such that all ϵ_0-heavy $\mathbf{w}_i$, $i \in [k]$, are covered at resolution slightly larger than $\chi^3 R_t$ (as testing increases the resolution slightly), but still a constant factor smaller than R_t. Since the resolution improves by a constant factor, repeating this combination of Refine and Test geometrically reduces the resolution and leads to covering all ϵ_0-heavy hidden weight vectors at resolution ϵ/k within order-$\log(k/\epsilon)$ repetitions. Unfortunately, the described approach does not run in polynomial-time, since there is no guarantee that the number of candidate vectors in our list covering the heavy hidden weight vectors is small; as a result, as described, the list $\mathcal{V}_t$ could grow super-polynomially.

To prevent such a size blow up, we additionally introduce a simple pruning procedure (Prune$(\mathcal{V}, r)$). The procedure takes as input the list $\mathcal{V}_t'$ produced after Refine and Test are used. It then creates $\mathcal{V}_{t+1}$, starting from an empty list as follows. It traverses $\mathcal{V}_t'$ element by element, and adds the traversed element to $\mathcal{V}_{t+1}$ if and only if it is sufficiently far from all elements already in $\mathcal{V}_{t+1}$. The intuition behind this method is that if two vectors $\mathbf{v}, \mathbf{u}$ are very close to each other, then both of them have similar correlation with the hidden weight vectors. Thus, we can remove one of them without substantially altering the coverage of the hidden weight vectors. The whole sequence of steps following Refine, Test, and Prune is illustrated in Figure 2.

We argue that the last pruning step (Prune$(\mathcal{V}, r)$) ensures that all heavy hidden weight vectors

remain covered by vectors in $\mathcal{V}_{t+1}$ at a slightly larger resolution $R_{t+1} \le \rho R_t$, $\rho \in (0,1)$, but only up to k candidate vectors are added to $\mathcal{V}_{t+1}$. More broadly, the main procedure of this section, $\mathsf{CoverAndPrune}(\epsilon_0, \delta; \mathcal{D})$, guarantees the following: in each iteration $t \ge 0$ of $\mathsf{CoverAndPrune}$, the following invariant holds:

$$\text{for all } \epsilon_0\text{-heavy } \mathbf{w}_i, \text{ there exists } \mathbf{v}_i^{(t)} \in \mathcal{V}_t \text{ such that } \theta_\pm(\mathbf{v}_i^{(t)}, \mathbf{w}_i) \le R_t, \text{ and } |\mathcal{V}_t| \le k. \tag{Invariant$_t$}$$

Algorithm 4 $\mathsf{Prune}(\mathcal{V}, r)$

Input: A finite set $\mathcal{V} \subseteq \mathbb{S}^{d-1}$ and a separation radius $r > 0$.
Output: A finite set $\mathcal{V}_{\text{output}} \subseteq \mathbb{S}^{d-1}$ such that every pair of elements is separated by r in $\theta_\pm$.

$\mathcal{V}_{\text{output}} \leftarrow \emptyset$.
for $\mathbf{v} \in \mathcal{V}$ **do**
  if $\theta_\pm(\mathbf{v}, \mathbf{v}') > r$ for every $\mathbf{v}' \in \mathcal{V}_{\text{output}}$ **then**
    $\mathcal{V}_{\text{output}} \leftarrow \mathcal{V}_{\text{output}} \cup \{\mathbf{v}\}$.
Return: $\mathcal{V}_{\text{output}}$.

Algorithm 5 $\mathsf{CoverAndPrune}(\epsilon_0, \delta; \mathcal{D})$

Input: Coefficient threshold $\epsilon_0 > 0$, failure probability $\delta \in (0,1)$, sample access to $\mathcal{D}$.
Output: A list $\widehat{\mathcal{V}}$ of unit vectors satisfying Proposition 3.10.

Tester accuracy $\tau \leftarrow 1/200$, ratio $g \leftarrow 1 + 2\tau$, radius shrinking rate $\chi \leftarrow (3/5)$.
Initial radius $r_0 \leftarrow \arccos(9/10)$, error threshold $\lambda_{\mathsf{CDR}} \leftarrow \epsilon_0 \sin(\chi^3 \epsilon_0)$.
Relaxed initial radius $R_0 \leftarrow (2g+1) r_0$, relaxed radius shrinking rate $\rho \leftarrow (2g+1)\chi^3$.
Number of iterations $T \leftarrow \lceil \frac{\log(R_0/\epsilon_0)}{\log(1/\rho)} \rceil$.
Failure probability of immediate steps $\delta' \leftarrow O(\delta/(TkL^3_{\mathsf{CDR}}))$
$\mathcal{L} \leftarrow \mathsf{CDR}(\mathcal{D}, \lambda_{\mathsf{CDR}}, \delta')$. ▷ Apply the Coarse Direction Recovery procedure.
$\mathcal{L}' \leftarrow \{\mathbf{v} \in \mathcal{L} : \mathsf{Test}(\mathbf{v}, r_0, \epsilon_0, \tau, \delta'; \mathcal{D}) = \mathsf{TRUE}\}$.
$\mathcal{V}_0 \leftarrow \mathsf{Prune}(\mathcal{L}', 2gr_0)$. ▷ Initialized vector set satisfying the induction (Invariant$_t$).
for $t = 0, 1, \ldots, T-1$ **do**
  $r_{t+1} \leftarrow \chi^3 R_t$ and $\mathcal{V}_t^{\text{new}} \leftarrow \mathcal{V}_t$.
  for $\mathbf{v} \in \mathcal{V}_t$ **do**
    $\mathcal{V}_t^{\text{new}} \leftarrow \mathcal{V}_t^{\text{new}} \cup \mathsf{Refine}(\mathbf{v}, R_t, \epsilon_0, \delta'; \mathcal{D})$. ▷ "Gradient Descent/Refinement" step; Vectors in $\mathcal{V}_t^{\text{new}}$ have smaller distance to the hidden weight vectors compared to $\mathcal{V}_t$
  $\mathcal{V}_t^{\text{closer}} \leftarrow \{\mathbf{v} \in \mathcal{V}_t^{\text{new}} : \mathsf{Test}(\mathbf{v}, r_{t+1}, \epsilon_0, \tau, \delta'; \mathcal{D}) = \mathsf{TRUE}\}$. ▷ Discard candidate vectors that do not cover hidden weight vectors.
  $\mathcal{V}_{t+1} \leftarrow \mathsf{Prune}(\mathcal{V}_t^{\text{closer}}, 2gr_{t+1})$. ▷ Discard candidate vectors covering the same hidden weight vectors.
  $R_{t+1} \leftarrow (2g+1) r_{t+1}$. ▷ $\mathcal{V}_{t+1}$ satisfies (Invariant$_t$) at $t+1$ with parameter R_{t+1}.
Return: $\widehat{\mathcal{V}} \leftarrow \mathcal{V}_T$.

We briefly discuss how $\mathsf{CoverAndPrune}(\epsilon_0, \delta; \mathcal{D})$ preserves the induction invariant at each step. The refinement subroutine in Line 14 generates new vectors that are closer to the hidden weights, but also produces many redundant and unwanted candidates. In order to guarantee that the vectors we keep cover the hidden weight vectors within R_t, we use our testing routine from the previous subsection ($\mathsf{Test}(\mathbf{v}, r, \epsilon_0, \tau, \delta_{\mathsf{Test}}; \mathcal{D})$). This procedure, applied in Line 15, allows us to discard candidate vectors that

do not have any hidden weight vector within radius R_t (in $\theta_\pm$). Then in Line 16, the Pruning subroutine removes redundant candidates, since two well-separated vectors cannot both be near the same hidden weight vector. By the triangle inequality, after the pruning step, we guarantee that each candidate vector covers hidden weight vectors within a relaxed radius, $R_{t+1} = (2g+1)r_{t+1} = \rho R_t$, with $\rho < 1$. Thus, as a result, at most k candidates remain in $\mathcal{V}_{t+1}$, and every hidden weight with non-negligible coefficient is covered by a candidate vector within radius R_{t+1}, i.e., the invariant (Invariant$_t$) is satisfied after each update.

Proposition 3.10 (Covering and Pruning)**.** *Assume access to an algorithm* CDR *that satisfies Assumption 1. Given parameters* $\epsilon_0, \delta > 0$, CoverAndPrune$(\epsilon_0, \delta; \mathcal{D})$ *draws*

$$N = \widetilde{O}(L_{\mathsf{CDR}}^3(L_{\mathsf{CDR}}^3 + N_{\mathsf{CDR}}))\,\mathrm{poly}(k, 1/\epsilon_0, 1/\delta)$$

i.i.d. samples from $\mathcal{D}$*, runs in* $\mathrm{poly}(d, N, T_{\mathsf{CDR}})$ *time, and returns a list* $\widehat{\mathcal{V}}$ *such that with probability at least* $1-\delta$*, the following holds:*

1. $|\widehat{\mathcal{V}}| \le k$*; and*
2. *For all* ϵ_0*-heavy* $\mathbf{w}_i$*, there exists* $\widehat{\mathbf{v}} \in \widehat{\mathcal{V}}$ *such that* $\theta_\pm(\widehat{\mathbf{v}}, \mathbf{w}_i) \le \epsilon_0$.

Proof. Let g, χ, r_0, R_0, and ρ be defined as in CoverAndPrune$(\epsilon_0, \delta; \mathcal{D})$. For iteration t, let

$$R_{t+1} = \rho R_t = (2g+1)\chi^3 R_t = (2g+1)r_{t+1}.$$

By Assumption 1, we have that running the CDR routine on $\mathcal{D}$ returns a list of unit vectors $\mathcal{L}$ such that for each heavy $\mathbf{w}_i$, $i \in [k]$, there exists $\mathbf{v}_i \in \mathcal{L}$ satisfying $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \le \arccos(9/10) = r_0$. After Line 9, the algorithm CoverAndPrune$(\epsilon_0, \delta; \mathcal{D})$ keeps only the vectors that pass the smallest-angle test. These are the candidate vectors that cover the heavy hidden weight vectors within radius r_0, i.e.,

$$\mathcal{L}' \coloneqq \{\mathbf{v} \in \mathcal{L} : \mathsf{Test}(\mathbf{v}, r_0, \epsilon_0, \tau, \delta_{\mathsf{Test}}; \mathcal{D}) = \mathsf{TRUE}\},$$

and then in Line 10 prunes them to obtain the initial list $\mathcal{V}_0$:

$$\mathcal{V}_0 \coloneqq \mathsf{Prune}(\mathcal{L}', 2gr_0).$$

Our goal is to show that for each $t \in [T]$, (Invariant$_t$) holds. We prove this claim by induction on t.

Claim 3.11 (Induction Base)**.** *At* $t = 0$*,* (Invariant$_t$) *condition is satisfied.*

Proof. By Proposition 3.5, the call to CDR in Line 8 gives a list $\mathcal{L}$ such that for every ϵ_0-heavy $\mathbf{w}_i$, there exists $\mathbf{v}_i \in \mathcal{L}$ with the property that

$$\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \le r_0.$$

Further, after the testing step in Line 9, for any ϵ_0-heavy $\mathbf{w}_i$, any vector $\mathbf{v}_i \in \mathcal{L}$ such that $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \le r_0$ will be kept in $\mathcal{L}'$, by Lemma 3.8. If $\mathbf{v}_i$ survives pruning in Line 10, it covers $\mathbf{w}_i$ with radius r_0 (i.e., $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \le r_0$). If, on the other hand, $\mathbf{v}_i$ is discarded, then there must exist some other vector $\mathbf{v}_i' \in \mathcal{V}_0$ that satisfies

$$\theta_\pm(\mathbf{v}_i', \mathbf{v}_i) \le 2gr_0.$$

Hence, by the triangle inequality, we get that

$$\theta_\pm(\mathbf{v}_i', \mathbf{w}_i) \le \theta_\pm(\mathbf{v}_i', \mathbf{v}_i) + \theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \le 2gr_0 + r_0 = R_0.$$

This establishes the second part of the claim. It remains to bound the size of the list.

Every vector in $\mathcal{V}_0$ passed the smallest-angle test with cutoff r_0. By Lemma 3.8, every such vector is within $(1+\tau)r_0$ of some hidden weight vector $\mathbf{w}_j$ with respect to $\theta_\pm$. If two retained vectors $\mathbf{v}_j, \mathbf{v}_j'$ cover the same $\mathbf{w}_j$ with radius $(1+\tau)r_0$, then, since $g = 1 + 2\tau$, we would have

$$\theta_\pm(\mathbf{v}_j', \mathbf{v}_j) \le \theta_\pm(\mathbf{v}_j', \mathbf{w}_j) + \theta_\pm(\mathbf{w}_j, \mathbf{v}_j) \le 2(1+\tau)r_0 < 2gr_0$$

contradicting the pruning rule in Line 10. Therefore, we have $|\mathcal{V}_0| \le k$, completing the proof of Claim 3.11. □

Now it remains to show that the invariant holds for all $t > 0$. This is done in the following claim.

Claim 3.12 (Inductive Step)**.** *Assume that* (Invariant$_t$) *holds at step t. Then, conditioned on successful* Refine *and* Test *procedures in Lines 14 and 15, it also holds at step $t+1$.*

Proof. Fix an ϵ_0-heavy $\mathbf{w}_i$ and choose $\mathbf{v} \in \mathcal{V}_t$ such that $\theta_\pm(\mathbf{v}, \mathbf{w}_i) \le R_t$. Such a $\mathbf{v}$ is guaranteed to exist because (Invariant$_t$) holds for step t. In Line 14 of CoverAndPrune$(\epsilon_0, \delta; \mathcal{D})$, the Refine subroutine (Refine$(\mathbf{v}, r, \epsilon_0, \delta_{\mathsf{CDR}}; \mathcal{D})$) is called with radius R_t. Therefore, by Lemma 3.6, it gives a list $\mathcal{V}_t^{\text{new}}$ that contains a unit vector $\mathbf{v}' \in \mathcal{V}_t^{\text{new}}$ such that $\theta_\pm(\mathbf{v}', \mathbf{w}_i) \le \chi^3 R_t = r_{t+1}$ (since $\mathbf{w}_i$ is ϵ_0-heavy, $\beta_i \ge \epsilon_0$). Thus, by Lemma 3.8, $\mathbf{v}'$ will pass the test in Line 15, i.e., Test$(\mathbf{v}, r_{t+1}, \epsilon_0, \tau, \delta'; \mathcal{D})$, and $\mathbf{v}' \in \mathcal{V}_t^{\text{closer}}$.

It remains to show that after pruning (Line 16), $\mathcal{V}_{t+1}$ covers all ϵ_0-heavy $\mathbf{w}_i$'s at resolution R_{t+1}. Recall that, in Line 16, Prune$(\mathcal{V}, r)$ is applied to $\mathcal{V} = \mathcal{V}_t^{\text{closer}}$ with separation radius $r = 2gr_{t+1}$. Fix any ϵ_0-heavy $\mathbf{w}_i$ and let $\mathbf{v}_i \in \mathcal{V}_t^{\text{closer}}$ be such that $\theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \le r_{t+1} = \chi^3 R_t$ (we have argued above that this vector will pass the smallest-angle test). If $\mathbf{v}_i \in \mathcal{V}_{t+1}$, it covers $\mathbf{w}_i$ at scale $\chi^3 R_t \le R_{t+1}$. If it is discarded, then some retained vector $\mathbf{v}_i' \in \mathcal{V}_{t+1}$ satisfies

$$\theta_\pm(\mathbf{v}_i', \mathbf{v}_i) \le 2gr_{t+1}.$$

Hence,

$$\theta_\pm(\mathbf{v}_i', \mathbf{w}_i) \le \theta_\pm(\mathbf{v}_i', \mathbf{z}_i) + \theta_\pm(\mathbf{v}_i, \mathbf{w}_i) \le (2g+1)r_{t+1} = R_{t+1}\ .$$

It remains to bound $|\mathcal{V}_{t+1}|$. Similar to the base argument, since every vector in $\mathcal{V}_{t+1}$ passed the smallest-angle test in Line 15, by Lemma 3.8, for every $\mathbf{v}_j \in \mathcal{V}_{t+1}$ there is a hidden weight vector $\mathbf{w}_j$ such that $\theta_\pm(\mathbf{v}_j, \mathbf{w}_j) \le (1+\tau)r_{t+1}$. For the sake of contradiction, suppose there were two distinct vectors $\mathbf{v}_j, \mathbf{v}_j'$ both covering $\mathbf{w}_j$ retained in $\mathcal{V}_{t+1}$. Then, by the triangle inequality, $\theta_\pm(\mathbf{v}_j', \mathbf{v}_j) \le 2(1+\tau)r_{t+1} < 2gr_{t+1}$, indicating that one of them will not pass the pruning step in Line 16. This provides the desired contradiction and establishes that $|\mathcal{V}_{t+1}| \le k$. □

To complete the proof of Proposition 3.10, it remains to bound the sample and computational complexities. Note that $R_t = \rho^t R_0$. Choosing

$$T = \left\lceil \frac{\log(R_0/\epsilon_0)}{\log(1/\rho)} \right\rceil,$$

we have $R_T \le \epsilon_0$. Thus, for all ϵ_0-heavy $\mathbf{w}_i$, there exists $\widehat{\mathbf{v}}_i \in \mathcal{V}_T$ such that $\theta_\pm(\widehat{\mathbf{v}}_i, \mathbf{w}_i) \le \epsilon_0$.

Let L_{CDR} be the uniform upper bound on the length of the list returned by the CDR subroutine, N_{CDR} be the corresponding sample complexity, and δ_{CDR} be the CDR subroutine's failure probability.

At each iteration $t \in [T]$, the smallest angle tester (Test$(\mathbf{v}, r, \epsilon_0, \tau, \delta_{\mathsf{Test}}; \mathcal{D})$) is called

$$|\mathcal{V}_t^{\text{new}}| = |\mathcal{V}^t| \times |\text{number of vectors returned by } \mathsf{Refine}| = O(kL_{\mathsf{CDR}}^3)$$

many times, where we used the fact that

$$\text{number of vectors returned by } \mathsf{Refine} = O(L_{\mathsf{CDR}}^3)$$

from Lemma 3.6. Therefore, by a union bound over the success events of the tester routine, the probability that all Test calls succeed at step t is $1 - O(kL_{\mathsf{CDR}}^3 \delta_{\mathsf{Test}})$.

Meanwhile, at step t, the Refine subroutine is called $|\mathcal{V}_t| \le k$ times. Therefore, by a union bound over all the calls of Refine for all T rounds and by Lemma 3.6, with probability at least $1 - O(TkL_{\mathsf{CDR}}^3 \delta_{\mathsf{CDR}})$, the Refine routine successfully returns a list $\mathcal{V}_t^{\text{new}}$ containing new candidate vectors that are closer to the hidden weight vectors compared to the ones in $\mathcal{V}_t$, at each step $t \in [T]$. Thus, by setting $\delta_{\mathsf{CDR}} = \delta' = O(\delta/(TkL_{\mathsf{CDR}}^3))$ and $\delta_{\mathsf{Test}} = \delta' = O(\delta/(TkL_{\mathsf{CDR}}^3))$, the subroutine CoverAndPrune$(\epsilon_0, \delta; \mathcal{D})$ succeeds with probability at least $1 - O(TkL_{\mathsf{CDR}}^3(\delta_{\mathsf{CDR}} + \delta_{\mathsf{Test}})) \ge 1 - \delta$.

By Lemma 3.8, it follows that the tester Test succeeds with $(1/\delta_{\mathsf{Test}})\text{poly}(k/\epsilon_0) = \text{poly}(k/\epsilon_0)TkL_{\mathsf{CDR}}^3/\delta$ samples for each run. By Lemma 3.6, the Refine subroutine requires $\widetilde{O}(L_{\mathsf{CDR}}^3 N_{\mathsf{CDR}} \log(1/\delta_{\mathsf{CDR}})/\gamma)$ samples from $\mathcal{D}$, where $\gamma = \lambda_{\mathsf{CDR}}/100 = \Theta(\epsilon_0^2)$. Therefore, in total, we need

$$\begin{aligned} N &\gtrsim \text{ number of } \mathsf{Test} \text{ runs} \times N_{\mathsf{Test}} + \text{number of } \mathsf{Refine} \text{ runs } \times N_{\mathsf{Refine}} \\ &\gtrsim TkL_{\mathsf{CDR}}^3 \times \text{poly}(k/\epsilon_0)TkL_{\mathsf{CDR}}^3/\delta + Tk \times \text{poly}(1/\epsilon_0)L_{\mathsf{CDR}}^3 N_{\mathsf{CDR}} \log(1/\delta_{\mathsf{CDR}}) \\ &\gtrsim \log^2(1/\epsilon_0)\text{poly}(k/\epsilon_0, 1/\delta)L_{\mathsf{CDR}}^3(L_{\mathsf{CDR}}^3 + N_{\mathsf{CDR}}) \end{aligned}$$

samples. By Lemma 3.6, Refine has runtime $\text{poly}(d, 1/\epsilon_0, L_{\mathsf{CDR}}, N_{\mathsf{CDR}}, T_{\mathsf{CDR}})$ and Test has runtime $O(dN_{\mathsf{Test}})$. Thus, in summary, CoverAndPrune$(\epsilon_0, \delta; \mathcal{D})$ has runtime $\text{poly}(d, k/\epsilon_0, 1/\delta, L_{\mathsf{CDR}}, N_{\mathsf{CDR}}, T_{\mathsf{CDR}}) = \text{poly}(d, N, T_{\mathsf{CDR}})$. This completes the proof of Proposition 3.10. □

3.2 Proof of Theorem 3.1

We are now ready to prove the main result of this section.

Proof of Theorem 3.1. Let $\epsilon_0 = \sqrt{\epsilon}/(4k)$. Run $\mathsf{CoverAndPrune}(\epsilon_0, \delta; \mathcal{D})$ with failure probability 1/9. By Proposition 3.10, with probability at least 8/9, it returns $\widehat{\mathcal{V}} = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$, where $m \le k$, such that for every i with $\beta_i \ge \epsilon_0$ there exists $\mathbf{v} \in \widehat{\mathcal{V}}$ such that $\theta_{\pm}(\mathbf{w}_i, \mathbf{v}) \le \epsilon_0$. We condition on this event. Assume without loss of generality that all the weights $\beta_i \ge \epsilon_0$, as otherwise, ignoring these ReLUs will only increase the L_2^2 error by ϵ. Indeed, we have the following sequence of inequalities:

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \widehat{y}(\mathbf{x}))^2] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left(\sum_{i:\beta_i\ge\epsilon_0} \beta_i\sigma(\mathbf{w}_i\cdot\mathbf{x}) - \widehat{y}(\mathbf{x}) + \sum_{i:\beta_i\le\epsilon_0} \beta_i\sigma(\mathbf{w}_i\cdot\mathbf{x})\right)^2\right] \\
&\le 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left(\sum_{i:\beta_i\ge\epsilon_0} \beta_i\sigma(\mathbf{w}_i\cdot\mathbf{x}) - \widehat{y}(\mathbf{x})\right)^2\right] + 2\,\mathbf{E}\left[\left(\sum_{i:\beta_i\le\epsilon_0} \beta_i\sigma(\mathbf{w}_i\cdot\mathbf{x})\right)^2\right] \\
&\le 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[(y'(\mathbf{x}) - \widehat{y}(\mathbf{x}))^2] + 2k\sum_{i:\beta_i\le\epsilon_0}^{k} \beta_i^2\,\mathbf{E}[(\sigma(\mathbf{w}_i\cdot\mathbf{x}))^2] \\
&\le 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[(y'(\mathbf{x}) - \widehat{y}(\mathbf{x}))^2] + \epsilon\,,
\end{aligned}$$

where $y'(\mathbf{x}) = \sum_{i:\beta_i\ge\epsilon_0} \beta_i\sigma(\mathbf{w}_i\cdot\mathbf{x})$.

Observe that, by standard concentration arguments, with $N_{\mathbf{g}} = O(d\mathrm{poly}(k)/\epsilon^2)$ samples we can get an estimate $\widehat{\mathbf{g}}$ of the vector $\mathbf{g} = \mathbf{E}[y\mathbf{x}]$ up to ℓ_2 error $\epsilon/16$ with probability at least 8/9. Since $\mathbf{E}[\sigma(\mathbf{w}\cdot\mathbf{x})\mathbf{x}] = \mathbf{w}/2$ for every unit vector $\mathbf{w}$, we have that

$$y - \mathbf{g}\cdot\mathbf{x} = \frac{1}{2}\sum_{i=1}^{k} \beta_i|\mathbf{w}_i\cdot\mathbf{x}|.$$

Let $\sigma_{\mathrm{abs}}(z) = \sigma(z) - z/2 = |z|/2$. Then

$$\left(\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \widehat{\mathbf{g}}\cdot\mathbf{x} - \sum_{j=1}^{k}\beta_j\sigma_{\mathrm{abs}}(\mathbf{v}_j\cdot\mathbf{x}))^2]\right)^{1/2} \le \frac{\sqrt{\epsilon}}{4} + 12\sum\beta_i \min_{s_i\in\{\pm1\}}\|\mathbf{w}_i - s_i\mathbf{v}_i\|_2 \le \frac{\sqrt{\epsilon}}{4} + \frac{k\epsilon_0}{4} \le \frac{\sqrt{\epsilon}}{2}.$$

For $\boldsymbol{\alpha} \in [0,k]^m$ with $\sum\alpha_i \le k$, write

$$\mathcal{R}(\boldsymbol{\alpha}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left(y - \widehat{\mathbf{g}}\cdot\mathbf{x} - \sum_{j=1}^{m}\alpha_j\sigma_{\mathrm{abs}}(\mathbf{v}_j\cdot\mathbf{x})\right)^2\right]$$

Setting $\mathbf{Z} = (y(\mathbf{x}) - \widehat{\mathbf{g}}\cdot\mathbf{x}, \sigma_{\mathrm{abs}}(\mathbf{v}_1\cdot\mathbf{x}), \ldots, \sigma_{\mathrm{abs}}(\mathbf{v}_m\cdot\mathbf{x}))$, we have that for any coordinate of $\mathbf{Z}\mathbf{Z}^\top$, $\mathbf{E}[(\mathbf{Z}\mathbf{Z}^\top)_{i,j}^2] \lesssim k^4$. Construct an empirical distribution $\widehat{\mathcal{D}}$ of $\mathcal{D}$ using $N_{\mathrm{quad}} = k^{10}/(\epsilon^2\delta)$ samples. Then, by Chebyshev's inequality and a union bound over all indices i,j, we have that

$$\mathbf{Pr}\left(\max_{i,j}|(\mathop{\mathbf{E}}_{\widehat{\mathcal{D}}}[\mathbf{Z}\mathbf{Z}^\top] - \mathbf{E}[\mathbf{Z}\mathbf{Z}^\top])_{i,j}| > \frac{\epsilon}{(k+1)^2}\right) \lesssim \frac{k^6(k+1)^4}{N_{\mathrm{quad}}\epsilon^2} \lesssim \delta.$$

Let $\overline{\boldsymbol{\alpha}} = (1, \alpha_1, \ldots, \alpha_m)$. Therefore, if the event inside the probability above is satisfied, we have that

$$\begin{aligned}
|\widehat{\mathcal{R}}(\boldsymbol{\alpha}) - \mathcal{R}(\boldsymbol{\alpha})| &= |\overline{\boldsymbol{\alpha}}^\top(\mathop{\mathbf{E}}_{\widehat{\mathcal{D}}}[\mathbf{Z}\mathbf{Z}^\top] - \mathbf{E}[\mathbf{Z}\mathbf{Z}^\top])\overline{\boldsymbol{\alpha}}| \\
&\le \max_{i,j}|(\mathop{\mathbf{E}}_{\widehat{\mathcal{D}}}[\mathbf{Z}\mathbf{Z}^\top] - \mathbf{E}[\mathbf{Z}\mathbf{Z}^\top])_{i,j}|\|\overline{\boldsymbol{\alpha}}\|_1^2 \le \frac{\epsilon}{(k+1)^2}(m+1)^2 \le \epsilon.
\end{aligned}$$

Minimizing over $\boldsymbol{\alpha}$ is a convex program and the empirical minimizer using $N_{\mathrm{quad}} + N_{\mathbf{g}} = d\,\mathrm{poly}(k/\epsilon)$ samples satisfies

$$\mathcal{R}(\widehat{\boldsymbol{\alpha}}) \lesssim \mathcal{R}(\boldsymbol{\alpha}^*) + \epsilon \lesssim \epsilon\,.$$

By Proposition 3.10, the procedure $\mathsf{CoverAndPrune}(\epsilon_0, \delta; \mathcal{D})$ uses $N_{\mathrm{opt}} = \widetilde{O}(\mathrm{poly}(k/\epsilon)L_{\mathsf{CDR}}^3(L_{\mathsf{CDR}}^3 + N_{\mathsf{CDR}}))$ samples (recall that we have set $\epsilon_0 = \sqrt{\epsilon}/(4k)$ and $\delta = 1/9$). Thus, the total number of samples required for $\mathsf{LearnSumReLUs}(d, k, \epsilon; \mathcal{D})$ is $N = \widetilde{O}(d\,\mathrm{poly}(k/\epsilon) + \mathrm{poly}(k/\epsilon)L_{\mathsf{CDR}}^3(L_{\mathsf{CDR}}^3 + N_{\mathsf{CDR}}))$. The runtime of the algorithm is $\mathrm{poly}(d, N, T_{\mathsf{CDR}})$. This completes the proof. □

4 Efficient Coarse Direction Recovery

In this section, we show that the CDR procedure postulated in Assumption 1 can be implemented in polynomial time. At a high-level, an algorithm that satisfies Assumption 1 needs to guarantee the following: Given sample access to a distribution $(\mathbf{x}, y) \sim \mathcal{D}$ such that (i) the $\mathbf{x}$-marginal is the standard Gaussian, i.e., $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and (ii) $y(\mathbf{x}) = \sum_{i=1}^{k} \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$, with σ being the ReLU activation, $b_i \in (0, 1]$ and $\|\mathbf{w}_i\|_2 = 1$, the algorithm returns a list $\mathcal{L}$ of unit vectors such that for each target weight vector $\mathbf{w}_i$, there exists $\mathbf{v} \in \mathcal{L}$ that correlates non-trivially with $\mathbf{w}_i$, i.e., $|\mathbf{w}_i \cdot \mathbf{v}| = 1 - c$, for a small constant $c > 0$. To this end, we prove that the top eigenvector of a carefully selected matrix (see Equation (M)) aligns well with one of the directions $\mathbf{w}_i$.

Our main result in this section is the following polynomial-time implementation of the CDR subroutine. The proof of Proposition 4.1 is deferred to Section 4.3.

Proposition 4.1 (Polynomial-Time CDR)**.** *Let $k, d \in \mathbb{Z}_+$ and $\lambda, \delta \in (0, 1)$. Let $\mathcal{D}$ be a Gaussian sum of k ReLUs. Then, there is an algorithm (CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$) that draws $N = \mathrm{poly}(d, k, 1/\lambda, 1/\delta)$ i.i.d. samples from $\mathcal{D}$, runs in $T = \mathrm{poly}(d, N)$ time, and outputs a list $\mathcal{L}$ of unit vectors of size $|\mathcal{L}| = \mathrm{poly}(k/\lambda)\log(1/\delta)$ such that with probability at least $1 - \delta$, for any $\mathbf{w}_i$, $i \in [k]$, with $\beta_i \geq \lambda$, there exists a vector $\mathbf{r}_i \in \mathcal{L}$ such that $|\mathbf{r}_i \cdot \mathbf{w}_i| \geq 0.99$.*

We begin this section by formally arguing that the guarantee of the CDR subroutine can be reduced to finding the top eigenvector of a carefully chosen matrix (M) (Section 4.1). Then, in Section 4.2, we give a simple estimation procedure for this matrix that uses a quasi-polynomial number of samples from $\mathcal{D}$. Section 4.2 highlights why more sophisticated techniques are needed for estimation and motivates the techniques based on implicit tensor computation that we employ in Section 4.3. Specifically, in Section 4.3, we present an algorithm (that uses techniques from implicit tensor computation) to estimate the matrix (M) with polynomially many samples and in polynomial time. Using this algorithm, we then prove Proposition 4.1. Throughout this section, we take without loss of generality that λ is sufficiently smaller than one (namely, $\lambda \leq \sqrt{\xi}/10$, where ξ as absolute constant). This only impacts the constants suppressed by the big-Oh notation.

4.1 Low-Degree Filtering: Revealing Hidden Directions

As discussed in Section 1.2.1, our implementation of CDR is based on the properties of Hermite tensors of ReLU functions; namely, the fact that degree-p moments of the ReLU sums are proportional to rank-tensors (recall (2)). For technical reasons, instead of working with the tensor $\mathbf{Q}_{2p}$ introduced in (3) in Section 1.2.1, we use a closely related matrix that is likewise proportional to $\sum_{i=1}^{k} \beta_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p} \mathbf{w}_i \mathbf{w}_i^\top$

$$\begin{aligned} \mathbf{M}_{2p+2}(\mathbf{u}) &\overset{(i)}{:=} \frac{1}{Z_p(\mathcal{U})} \left\langle \frac{1}{\widehat{\sigma}_{2p+2}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [y \mathbf{He}_{2p+2}(\mathbf{x})],\, \mathbf{u}^{\otimes p} \otimes \bar{\mathbf{u}}^{\otimes p} \right\rangle \\ &\overset{(ii)}{=} \frac{1}{Z_p(\mathcal{U})} \sum_{i=1}^{k} \beta_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p} \mathbf{w}_i \mathbf{w}_i^\top \,, \end{aligned} \tag{M}$$

where $Z_p(\mathcal{U}) > 0$ is a fixed parameter depending on the sampling distribution $\mathcal{U}$ from which $\mathbf{u}$ is drawn and on the degree p. The probing vector $\mathbf{u}$ can be chosen from a complex distribution, where $\bar{\mathbf{u}}$ is its complex conjugate. When $\mathbf{u}$ is a real vector, we have $\bar{\mathbf{u}} = \mathbf{u}$, and in this case, the tensor contraction in (i) becomes $\langle \mathbf{E}[y \mathbf{He}_{2p+2}(\mathbf{x})], \mathbf{u}^{\otimes 2p} \rangle$. Intuitively, the tensor $\mathbf{u}^{\otimes p} \otimes \bar{\mathbf{u}}^{\otimes p} \in (\mathbb{C}^d)^{\otimes 2p}$ can be viewed as a "filter" that projects the signal tensor $\mathbf{E}[y \mathbf{He}_{2p+2}(\mathbf{x})]$ to the direction of $\mathbf{u}$. Choosing the order p and the distribution $\mathcal{U}$ properly, we can guarantee that with large probability, only the target vector $\mathbf{w}_i$, $i \in [k]$ will come through the filter—as discussed in Section 1.2.1. To see why (ii) is true, recall that by the definition of Hermite tensors and the definition of our learning problem (Definition 1.2), we have $\mathbf{E}[y \mathbf{He}_{2p+2}(\mathbf{x})] = \sum_{i=1}^{k} \widehat{\sigma}_{2p+2} \beta_i \mathbf{w}_i^{\otimes 2p+2}$, which yields (ii) after contraction with the filter. The matrix M can be estimated via samples from $\mathcal{D}$. Naively, one can estimate the population tensor $\mathbf{E}[y \mathbf{He}_{2p+2}(\mathbf{x})]$ using $d^{O(p)}$ samples, then perform contraction with the filter. However, since the target matrix only concerns the projection to the probing vector $\mathbf{u}$, estimating the tensor $\mathbf{E}[y \mathbf{He}_{2p+2}(\mathbf{x})]$ accurately in all the directions is unnecessary. In Section 4.2, we show that M can be estimated approximately via a polynomial; then in Section 4.3 we give an algorithm that estimates the contraction

$\langle \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{He}_{2p+2}(\mathbf{x})], \mathbf{u}^{\otimes p} \otimes \bar{\mathbf{u}}^{\otimes p}\rangle$ as a whole using polynomially many samples, avoiding a separate estimation of the tensor $\mathbf{E}[y\mathbf{He}_{2p+2}(\mathbf{x})]$.

In the following, we show that the CDR subroutine can be reduced to estimating the top eigenvector of the empirical estimator of the matrix M.

Lemma 4.2. *Let $\xi \in (0, 1/300]$ be a small absolute constant. Let $\mathbf{w}_i$ be unit vectors in $\mathbb{R}^d$, $\beta_i \in [0,1]$, $i \in [k]$, and $\lambda \in (0, \sqrt{\xi}/10]$ be less than a sufficiently small constant. Let $p = \lceil \log(k/\lambda)/\log(1+\xi)\rceil$. Consider any $i^* \in [k]$ such that $\beta_{i^*} \geq \lambda$, and denote $\mathcal{A}_{i^*}(\xi) = \{i \in [k] : |\mathbf{w}_i \cdot \mathbf{w}_{i^*}| \leq 1-\xi\}$. Draw and fix $\mathbf{u}$ from the distribution $\mathcal{U}$. Let $\mathbf{M}_{2p+2}(\mathbf{u})$ be the matrix defined in (M). Define the following "correlation" and "separation" events on $\mathcal{U}$:*

$$\mathcal{E}_{i^*}^{\mathrm{Cor}} = \{\mathbf{u} : |\mathbf{w}_{i^*} \cdot \mathbf{u}| \geq (Z_p(\mathcal{U}))^{1/(2p)}\},\ \mathcal{E}_{i^*}^{\mathrm{Sep}} = \{\mathbf{u} : |\mathbf{w}_{i^*} \cdot \mathbf{u}| \geq (1+\xi)|\mathbf{w}_j \cdot \mathbf{u}|, \forall j \in \mathcal{A}_{i^*}(\xi)\}. \tag{11}$$

Suppose $\mathbf{u} \in \mathcal{E}_{i^*}^{\mathrm{Cor}} \cap \mathcal{E}_{i^*}^{\mathrm{Sep}}$. Let $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})$ be any real matrix such that $\|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 \leq (\sqrt{\xi}/3)\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2$. Then, the top eigenvector $\widehat{\mathbf{r}}_1$ of $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})$ satisfies $|\widehat{\mathbf{r}}_1 \cdot \mathbf{w}_{i^*}| \geq 0.99$.

Proof. The high-level idea of the proof is that under the assumptions of Lemma 4.2, the target matrix $\mathbf{M}_{2p+2}(\mathbf{u})$ is approximately rank-1 and proportional to $\mathbf{w}_{i^*}\mathbf{w}_{i^*}^\top$. Let us define

$$W_A = \frac{1}{Z_p(\mathcal{U})} \sum_{i \in \mathcal{A}_{i^*}(\xi)} \beta_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p},\ W_B = \frac{1}{Z_p(\mathcal{U})} \sum_{i \notin \mathcal{A}_{i^*}(\xi)} \beta_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p}, \text{ and } \mathbf{M}_{2p+2}^B(\mathbf{u}) = W_B \mathbf{w}_{i^*}\mathbf{w}_{i^*}^\top. \tag{12}$$

Observe that $\mathrm{tr}(\mathbf{M}_{2p+2}(\mathbf{u})) = W_A + W_B$. We establish the following:

Claim 4.3. *Under the assumptions of Lemma 4.2, it holds:*

$$W_A \leq (\lambda/k) W_B \text{ and } \|\mathbf{M}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}^B(\mathbf{u})\|_2 \leq (\lambda/k + \sqrt{2\xi})\|\mathbf{M}_{2p+2}^B(\mathbf{u})\|_2. \tag{13}$$

Proof of Claim 4.3. Since $i^* \notin \mathcal{A}_{i^*}(\xi)$, $\beta_{i^*} \geq \lambda$, and $|\mathbf{w}_{i^*} \cdot \mathbf{u}|^{2p} \geq Z_p(\mathcal{U})$ for $\mathbf{u} \in \mathcal{E}_{i^*}^{\mathrm{Cor}}$, the weight W_B is large, namely $W_B \geq \beta_{i^*}|\mathbf{w}_{i^*} \cdot \mathbf{u}|^{2p}/Z_p(\mathcal{U}) \geq \lambda$. On the other hand, for any $i \in \mathcal{A}_{i^*}(\xi)$, the probing vector $\mathbf{u}$ satisfies $|\mathbf{w}_i \cdot \mathbf{u}| \leq |\mathbf{w}_{i^*} \cdot \mathbf{u}|/(1+\xi)$ (because $\mathbf{u} \in \mathcal{E}_{i^*}^{\mathrm{Sep}}$). Hence, using the assumption $\beta_i \in (0,1]$ and our choice of p, the weight W_A is small. Specifically, we have

$$\begin{aligned} W_A &\leq \frac{1}{Z_p(\mathcal{U})} \sum_{i \in \mathcal{A}_{i^*}(\xi)} \frac{|\mathbf{w}_{i^*} \cdot \mathbf{u}|^{2p}}{(1+\xi)^{2p}} \leq \frac{|\mathbf{w}_{i^*} \cdot \mathbf{u}|^{2p}}{Z_p(\mathcal{U})} k \exp(-(2p)\log(1+\xi)) \\ &\leq \frac{\lambda^2}{k} \frac{|\mathbf{w}_{i^*} \cdot \mathbf{u}|^{2p}}{Z_p(\mathcal{U})} \leq \frac{\lambda}{k} \frac{\beta_{i^*}|\mathbf{w}_{i^*} \cdot \mathbf{u}|^{2p}}{Z_p(\mathcal{U})} \leq \frac{\lambda}{k} W_B. \end{aligned}$$

Therefore, we have

$$\begin{aligned} &\|\mathbf{M}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}^B(\mathbf{u})\|_2 \\ &= \left\| \frac{1}{Z_p(\mathcal{U})} \sum_{i \in \mathcal{A}_{i^*}(\xi)} \beta_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p} \mathbf{w}_i \mathbf{w}_i^\top + \frac{1}{Z_p(\mathcal{U})} \sum_{i \notin \mathcal{A}_{i^*}(\xi)} \beta_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p} (\mathbf{w}_i \mathbf{w}_i^\top - \mathbf{w}_{i^*}\mathbf{w}_{i^*}^\top) \right\|_2 \\ &\leq \frac{1}{Z_p(\mathcal{U})} \sum_{i \in \mathcal{A}_{i^*}(\xi)} \beta_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p} \|\mathbf{w}_i \mathbf{w}_i^\top\|_2 + \frac{1}{Z_p(\mathcal{U})} \sum_{i \notin \mathcal{A}_{i^*}(\xi)} \beta_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p} \|\mathbf{w}_i \mathbf{w}_i^\top - \mathbf{w}_{i^*}\mathbf{w}_{i^*}^\top\|_2 \end{aligned}$$

Observe that $\|\mathbf{v}\mathbf{v}^\top - \mathbf{w}\mathbf{w}^\top\|_2 \leq \sqrt{1-(\mathbf{v}\cdot\mathbf{w})^2}$ for any unit vectors $\mathbf{v}, \mathbf{w}$. By the definition of the index set $\mathcal{A}_{i^*}(\xi)$, for any $i \notin \mathcal{A}_{i^*}(\xi)$, it holds $|\mathbf{w}_i \cdot \mathbf{w}_{i^*}| \geq 1-\xi$, therefore we have $\|\mathbf{w}_i\mathbf{w}_i^\top - \mathbf{w}_{i^*}\mathbf{w}_{i^*}^\top\|_2 \leq \sqrt{2\xi}$. Thus, we conclude that:

$$\|\mathbf{M}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}^B(\mathbf{u})\|_2 \leq W_A + \sqrt{2\xi} W_B \leq (\lambda/k + \sqrt{2\xi}) W_B\ ,$$

completing the proof of Claim 4.3. □

It remains to prove that $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})$ is close to $\mathbf{M}^B_{2p+2}(\mathbf{u})$ in spectral norm, which would imply that the top-eigenvector of $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})$ aligns with $\mathbf{w}_{i^*}$. As $\lambda \in (0, \sqrt{\xi}/10]$, it holds that $\lambda/k \le \sqrt{\xi}/10$. By the triangle inequality and Claim 4.3, we have

$$\begin{aligned}\|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}^B_{2p+2}(\mathbf{u})\|_2 &\le \|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 + \|\mathbf{M}_{2p+2}(\mathbf{u}) - \mathbf{M}^B_{2p+2}(\mathbf{u})\|_2 \\ &\le \left(\frac{\sqrt{\xi}}{10} + \sqrt{2\xi}\right) W_B + \frac{\sqrt{\xi}}{3}\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2.\end{aligned}$$

In addition, since $\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2 \le W_A + W_B$ and $W_A \le (\lambda/k)W_B \le (\sqrt{\xi}/10)W_B$, we have $\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2 \le (1 + (\sqrt{\xi}/10))W_B$. Thus, we conclude that

$$\|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}^B_{2p+2}(\mathbf{u})\|_2 \le \left(\frac{\sqrt{\xi}}{10} + \sqrt{2\xi} + \frac{\sqrt{\xi}}{3}\left(1 + \frac{\sqrt{\xi}}{10}\right)\right) W_B.$$

Denote the top (unit) eigenvector of $\widehat{\mathbf{M}}_{2p+2}$ by $\widehat{\mathbf{r}}_1$ and the top eigenvector of $\mathbf{M}^B_{2p+2}(\mathbf{u})$ by $\mathbf{r}_1$; note that $\mathbf{r}_1 = \mathbf{w}_{i^*}$. Then, using Wedin's inequality (Fact 2.1), with $\xi \le 1/300$, we have

$$\begin{aligned}\sin(\theta(\widehat{\mathbf{r}}_1, \mathbf{r}_1)) &\le \frac{\|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}^B_{2p+2}(\mathbf{u})\|_2}{\|\mathbf{M}^B_{2p+2}(\mathbf{u})\|_2 - \|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}^B_{2p+2}(\mathbf{u})\|_2} \\ &\le \frac{(\frac{\sqrt{\xi}}{10} + \sqrt{2\xi} + \frac{\sqrt{\xi}}{3}(1 + \frac{\sqrt{\xi}}{10}))W_B}{W_B - (\frac{\sqrt{\xi}}{10} + \sqrt{2\xi} + \frac{\sqrt{\xi}}{3}(1 + \frac{\sqrt{\xi}}{10}))W_B} \le \sqrt{1 - 0.99^2}\,.\end{aligned}$$

Thus, $|\widehat{\mathbf{r}}_1 \cdot \mathbf{w}_{i^*}| \ge 0.99$, completing the proof of Lemma 4.2. □

What remains are the following steps: (1) pick a suitable distribution of $\mathbf{u}$ such that the event $\mathbf{u} \in \mathcal{E}^{\mathrm{Cor}}_{i^*} \cap \mathcal{E}^{\mathrm{Sep}}_{i^*}$ happens with probability at least $\mathrm{poly}(\lambda/k)$; (2) approximate the matrix $\mathbf{M}_{2p+2}(\mathbf{u})$ efficiently using samples from $(\mathbf{x}, y) \sim \mathcal{D}$. In Section 4.2, we choose $\mathcal{U}$ to be the uniform distribution on the sphere, obtaining quasi-polynomial sample complexity; then in Section 4.3, we choose $\mathcal{U}$ to be a complex distribution, and by applying implicit tensor computation techniques, we argue that $\mathbf{M}_{2p+2}$ can be estimated efficiently, using $\mathrm{poly}(k/\lambda)$ samples.

4.2 Warm-up: Quasi-Polynomial Time CDR Algorithm

Given a distribution $(\mathbf{x}, y) \sim \mathcal{D}$ such that $y(\mathbf{x}) = \sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$ and $\mathbf{x} \sin \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, a simple choice of the probing vector is to sample from the uniform distribution on the unit sphere in a subspace containing $\mathrm{Span}(\mathbf{w}_1, \ldots, \mathbf{w}_k)$. The latter can be estimated efficiently via the following standard result.

Fact 4.4 (see, e.g., Lemma 3.4 of [DKKZ20])**.** *Let $(\mathbf{x}, y) \sim \mathcal{D}$ be the distribution on $\mathbb{R}^d \times \mathbb{R}$ such that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $y(\mathbf{x}) = \sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$, $\beta_i \in [0, 1]$, $\|\mathbf{w}_i\|_2 = 1$, and σ is the ReLU activation. Let $\widehat{\sigma}_2$be the second- order Hermite coefficient of σ. Let $\eta > 0$. Then, using $N = \mathrm{poly}(d, k, 1/\eta)$ samples $(\mathbf{x}^{(i)}, y^{(i)})$ drawn from $\mathcal{D}$, it holds with high probability that:*

$$\left\| \frac{1}{\sqrt{2}N} \sum_{i=1}^N y^{(i)} (\mathbf{x}^{(i)} \mathbf{x}^{(i)\top} - \mathbf{I}_d) - \sum_{j=1}^k |\widehat{\sigma}_2| \beta_j \mathbf{w}_j \mathbf{w}_j^\top \right\|_2 \le \eta\,.$$

Let $\widehat{\mathbb{W}}$ be the subspace spanned by the top-k vectors of the empirical matrix constructed in Fact 4.4. Then, Fact 4.4 implies that for any $\mathbf{w}_i$, $i \in [k]$, it holds $\|\mathbf{w}_i - \mathbf{P}_{\widehat{\mathbb{W}}} \mathbf{w}_i\|_2 \le \sqrt{2\eta/\beta_i}$ (with high probability, otherwise, ascribe to the failure event of the algorithm). For any λ-heavy weights $\mathbf{w}_{i^*}$ with coefficients $\beta_{i^*} \ge \lambda$, choosing $\eta = \mathrm{poly}(\lambda/k)$ yields that

$$\mathrm{dist}(\widehat{\mathbb{W}}, \mathrm{Span}(\{\mathbf{w}_{i^*} : \beta_{i^*} \ge \lambda\})) \le \lambda.$$

Therefore, for a probing vector $\mathbf{u}$ sampled uniformly from $\widehat{\mathbb{W}} \cap \mathbb{S}^{d-1}$, the probability that it satisfies the correlation event $\mathcal{E}^{\mathrm{Cor}}_{i^*}$ and the separation $\mathcal{E}^{\mathrm{Sep}}_{i^*}$ is almost the same as if sampled from $\mathbb{W} \cap \mathbb{S}^{d-1}$, with only differences in the absolute constants. Thus, to simplify the analysis, in the remainder of this section, we assume access to the subspace $\mathbb{W} = \mathrm{Span}(\mathbf{w}_1, \ldots, \mathbf{w}_k)$ with dimension m and we sample the probing vectors $\mathbf{u}$ from $\mathcal{U} = \mathrm{Unif}(\mathbb{W} \cap \mathbb{S}^{d-1})$.

Remark 4.5 (Range of parameters.)**.** Throughout this section, we assume that the dimension m of the subspace $\mathbb{W} = \mathrm{Span}(\mathbf{w}_1, \ldots, \mathbf{w}_k)$ satisfies $m \gtrsim \log(k)$. Indeed, the CDR procedure for small m admits a simpler approach. More generally, consider the regime $m \lesssim \log(k)$. In this case, we can obtain the desired CDR guarantee by sampling random directions from the m-dimensional subspace $\mathbb{W}$. For any fixed unit vector in $\mathbb{W}$, a uniformly random unit vector has correlation at least 0.99 with probability $2^{-O(m)}$. By taking $q = 2^{O(m)} \log(k/\delta) \leq \mathrm{poly}(k, 1/\epsilon) \log(1/\delta)$ random vectors $\mathbf{u}_1, \ldots, \mathbf{u}_q$, with probability at least $1 - \delta$, we have that $\max_{j \in [q]} |\mathbf{u}_j \cdot \mathbf{w}_i| \geq 0.99$ for all $i \in [k]$.

The pseudocode of our algorithm (CDRviaSpectral$(\mathcal{D}, \lambda, \delta)$) is given below.

Algorithm 6 CDRviaSpectral$(\mathcal{D}, \lambda, \delta)$

Input: A subspace $\mathbb{W}$, access to a distribution of a sum of k ReLUs with Gaussian marginals $(\mathbf{x}, y) \sim \mathcal{D}$, positive parameters $\lambda, \delta \in (0, 1)$.
Output: A list of vectors satisfying Assumption 1.

Fix an absolute constant $\xi \in (0, 1/300]$, let $p \leftarrow \lceil \log(k/\lambda)/\log(1+\xi) \rceil$, $\lambda \leftarrow \min\{\lambda, \sqrt{\xi}/10\}$.
Let $\widehat{\sigma}_{2p+2}$ be the $(2p+2)^{\mathrm{th}}$ Hermite coefficient of ReLU and $\widehat{\sigma}''_{2p} \leftarrow \widehat{\sigma}_{2p+2}\sqrt{(2p+1)(2p+2)}$.
$Z_p(\mathcal{U}) \leftarrow k^{-p}$, $\mathcal{L} \leftarrow \emptyset$.
Draw a set $\widehat{\mathcal{U}}$ of $\mathrm{poly}(k)\log(1/\delta)$ probing vectors from $\mathcal{U} = \mathrm{Unif}(\mathbb{S}^{d-1} \cap \mathbb{W})$.
for $\mathbf{u} \in \widehat{\mathcal{U}}$ **do**
    Draw $N = d^2\,\mathrm{poly}(p) k^{O(\log(k/\lambda))}$ samples from $\mathcal{D}$ and construct empirical distribution $\widehat{\mathcal{D}}$.
    $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) \leftarrow (1/(\widehat{\sigma}''_{2p} Z_p(\mathcal{U}))) \, \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[y \mathrm{he}_{2p}(\mathbf{u}\cdot\mathbf{x})(\mathbf{x}^{\perp_{\mathbf{u}}}\mathbf{x}^{\perp_{\mathbf{u}}\top} - \mathbf{I}_{\perp_{\mathbf{u}}})]$.
    Find the top eigenvector of $\mathbf{r}_1 = \mathbf{r}_1(\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}))$, $\mathcal{L} \leftarrow \mathcal{L} \cup \{\pm \mathbf{r}_1\}$.
Return: $\mathcal{L}$.

Our main result for this subsection is the following:

Proposition 4.6 (Quasi-Polynomial-Time Implementation of CDR)**.** *Let $\delta \in (0,1)$, $\xi \in (0, 1/300]$ be an absolute constant and let $\lambda \in (0, \sqrt{\xi}/10]$. Let $(\mathbf{x}, y) \sim \mathcal{D}$ be a distribution such that $y(\mathbf{x}) = \sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$ where σ is the ReLU activation, $\beta_i \in [0,1]$, $\|\mathbf{w}_i\|_2 = 1$, and $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Let $p \geq \lceil \log(k/\lambda)/\log(1+\xi) \rceil$. Then,* CDRviaSpectral$(\mathcal{D}, \lambda, \delta)$ *draws $N = d^2\, k^{O(\log(k/\lambda))} \log(1/\delta)$ samples from $\mathcal{D}$, runs in $T = \mathrm{poly}(d, N)$ time, and outputs a list $\mathcal{L}$ of unit vectors of size $|\mathcal{L}| = \mathrm{poly}(k)\log(1/\delta)$ that with probability at least $1 - \delta$ satisfies the following guarantee: for any $\mathbf{w}_i$, $i \in [k]$ with $\beta_i \geq \lambda$, there exists $\mathbf{r}_i \in \mathcal{L}$ such that $|\mathbf{r}_i \cdot \mathbf{w}_i| \geq 0.99$.*

To prove Proposition 4.6, we need to bound from below the probability of the event $\mathcal{E}^{\mathrm{Cor}}_{i^*} \cap \mathcal{E}^{\mathrm{Sep}}_{i^*}$ for a fixed index $i^* \in [k]$, defined in (11). Intuitively, when the probing vector $\mathbf{u}$ is sampled uniformly from the sphere, with high probability the values $|\mathbf{u} \cdot \mathbf{w}_{i^*}|$ and $|\mathbf{u} \cdot \mathbf{w}_j|$ would be separated when the vectors $\mathbf{w}_{i^*}$ and $\mathbf{w}_j$ are away from each other. This is the separation event $\mathcal{E}^{\mathrm{Sep}}_{i^*}$. At the same time, for the correlation event, $\mathcal{E}^{\mathrm{Cor}}_{i^*}$, with high probability, $|\mathbf{u} \cdot \mathbf{w}_{i^*}|$ cannot be too small and would be concentrated near $1/\sqrt{k}$. Therefore, the probability that $\mathcal{E}^{\mathrm{Cor}}_{i^*} \cap \mathcal{E}^{\mathrm{Sep}}_{i^*}$ holds is bounded from below. We then determine the number of samples required to estimate the target matrix M. Finally, we prove Proposition 4.6 by invoking Lemma 4.2, as outlined previously. The proof of Proposition 4.6 is deferred to the end of this subsection.

Determining the Probability of the Probing Event. We first determine the probability of event $\mathcal{E}^{\mathrm{Cor}}_{i^*} \cap \mathcal{E}^{\mathrm{Sep}}_{i^*}$ under the distribution $\mathcal{U} = \mathrm{Unif}(\mathbb{W} \cap \mathbb{S}^{d-1})$. In fact, we prove a slightly stronger result: with probability at least $\mathrm{poly}(1/k)$ in addition to $\mathbf{u} \in \mathcal{E}^{\mathrm{Cor}}_{i^*} \cap \mathcal{E}^{\mathrm{Sep}}_{i^*}$, $\mathbf{u}$ also satisfies a "near-orthogonal" event $\mathcal{E}^{\mathrm{Ort}}$. The "near-orthogonal" event $\mathcal{E}^{\mathrm{Ort}}$ states that $\mathbf{u}$ satisfies $|\mathbf{u} \cdot \mathbf{w}_i| \leq c$ for a sufficiently small absolute constant c for all hidden weight vectors $\mathbf{w}_i$. This event is needed in estimating the target matrix $\mathbf{M}_{2p+2}$ to high accuracy, as will become apparent in the proof of Proposition 4.8.

Lemma 4.7. *Let parameters $p, Z_p(\mathcal{U})$ and constant ξ be as prescribed in Proposition 4.6. Let $\mathbf{w}_1, \ldots, \mathbf{w}_k$ be k unit vectors and $\mathbb{W} = \mathrm{Span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k\} \subset \mathbb{R}^d$ be a linear subspace of dimension $m \leq k$. Let $\mathbf{u} \sim \mathcal{U}$ where $\mathcal{U} = \mathrm{Unif}(\mathbb{W} \cap \mathbb{S}^{d-1})$. Fix (any) index $i^* \in [k]$ and let $\mathcal{A}_{i^*}(\xi) = \{i \in [k] : |\mathbf{w}_i \cdot \mathbf{w}_{i^*}| \leq 1 - \xi\}$. Define the "correlation" and "separation" events on $\mathcal{U}$ as in Lemma 4.2:*

$$\mathcal{E}^{\mathrm{Cor}}_{i^*} = \{\mathbf{u} : |\mathbf{w}_{i^*} \cdot \mathbf{u}| \geq (Z_p(\mathcal{U}))^{1/(2p)}\}, \; \mathcal{E}^{\mathrm{Sep}}_{i^*} = \{\mathbf{u} : |\mathbf{w}_{i^*} \cdot \mathbf{u}| \geq (1+\xi)|\mathbf{w}_j \cdot \mathbf{u}|, \forall j \in \mathcal{A}_{i^*}(\xi)\}.$$

In addition, let the "near-orthogonal" event on $\mathcal{U}$ be:

$$\mathcal{E}^{\mathrm{Ort}} = \{\mathbf{u} : |\mathbf{u} \cdot \mathbf{w}_i| \leq \sqrt{\xi}/20, \forall i \in [k]\}.$$

Then, $\mathbf{Pr}_{\mathbf{u}\sim\mathcal{U}}[\mathcal{E}^{\mathrm{Cor}}_{i^*} \cap \mathcal{E}^{\mathrm{Sep}}_{i^*} \cap \mathcal{E}^{\mathrm{Ort}}] \geq \mathrm{poly}(1/k)$.

Proof. Let $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{\mathbb{W}})$ and $Z_j \coloneqq \mathbf{p} \cdot \mathbf{w}_j$. Then, by definition, $\mathbf{u} = \mathbf{p}/\|\mathbf{p}\|_2$ and $\mathbf{u} \cdot \mathbf{w}_i = \mathbf{p} \cdot \mathbf{w}_i/\|\mathbf{p}\|_2$. The roadmap of the proof is the following: we first show that $\mathcal{E}^{\mathrm{Sep}}_{i^*} \cap \{t \leq |\mathbf{p} \cdot \mathbf{w}_{i^*}| \leq 2t\}$ happens with probability at least $\mathrm{poly}(1/k)$ for some $t \lesssim \sqrt{\log(k)}$ then we argue that $\mathbf{Pr}[\|\mathbf{p}\|_2 \in (\sqrt{m/2}, \sqrt{m})] \geq 1 - \mathrm{poly}(1/m)$. Since we have assumed without loss of generality (see Remark 4.5) that the dimension m of the subspace $\mathbb{W}$ satisfies $m \gtrsim \log(k)$, applying a union bound on these events gives a lower bound on $\mathbf{Pr}[\mathcal{E}^{\mathrm{Sep}}_{i^*} \cap \{t/\sqrt{m} \leq |\mathbf{u} \cdot \mathbf{w}_{i^*}| \leq 2t/\sqrt{m/2}\}] = \mathbf{Pr}[\mathcal{E}^{\mathrm{Cor}}_{i^*} \cap \mathcal{E}^{\mathrm{Sep}}_{i^*} \cap \mathcal{E}^{\mathrm{Ort}}]$, completing the proof.

For each $j \in \mathcal{A}_{i^*}(\xi)$, let $\rho_{i^*,j} = \mathbf{w}_{i^*} \cdot \mathbf{w}_j = \cos(\theta(\mathbf{w}_{i^*}, \mathbf{w}_j))$. By the definition of $\mathcal{A}_{i^*}(\xi)$, it holds $|\rho_{i^*,j}| \leq \rho \coloneqq 1 - \xi$. Then, we can write Z_j as

$$Z_j = \mathbf{p} \cdot \cos(\theta(\mathbf{w}_{i^*}, \mathbf{w}_j))\mathbf{w}_{i^*} + \mathbf{p} \cdot \sin(\theta(\mathbf{w}_{i^*}, \mathbf{w}_j))\widehat{\mathbf{w}_{i^*}^{\perp \mathbf{w}_j}} = \rho_{i^*,j} Z_{i^*} + \sqrt{1 - \rho_{i^*,j}^2} G_j \ ,$$

where G_j is a standard normal random variable. By the definition of $\mathbf{u}$, it holds $\mathbf{u} \cdot \mathbf{w}_i = \mathbf{p} \cdot \mathbf{w}_i/\|\mathbf{p}\|_2$, and hence the event $|\mathbf{u} \cdot \mathbf{w}_j| \leq |\mathbf{u} \cdot \mathbf{w}_{i^*}|/(1+\xi)$ is equivalent to $|Z_j| \leq |Z_{i^*}|/(1+\xi)$. Conditional on $Z_{i^*} = z$, consider the event that $\mathcal{E}^{\mathrm{Sep}}_{i^*}$ is violated, i.e., there exists some Z_j, $j \in \mathcal{A}_{i^*}(\xi)$, such that $|Z_j| \geq |Z_{i^*}|/(1+\xi)$. Note that $|Z_j| \leq |\rho_{i^*,j}||z| + |G_j|$ and using that $|\rho_{i^*,j}| \leq 1 - \xi$ and rearranging, we get that on a violation, $|G_j| \geq |z|\xi^2/(1+\xi)$. Therefore, from a union bound, we have

$$\begin{aligned}\mathbf{Pr}[\cup_{j \in \mathcal{A}_{i^*}(\xi)}\{|Z_j| \geq |Z_{i^*}|/(1+\xi)\} \mid Z_{i^*} = z] &\leq \sum_{j \in \mathcal{A}_{i^*}(\xi)} \mathbf{Pr}[|G_j| \geq |z|\xi^2/(1+\xi) \mid Z_{i^*} = z] \\ &\leq k \max_{j \in \mathcal{A}_{i^*}(\xi)} \mathbf{Pr}[|G_j| \geq |z|\xi^2/(1+\xi) \mid Z_{i^*} = z] \ .\end{aligned}$$

Using that $\mathbf{Pr}[|G_j| \geq q] \leq 2\exp(-q^2/2)$, we get that

$$\mathbf{Pr}[\cup_{j \in \mathcal{A}_{i^*}(\xi)}\{|Z_j| \geq |Z_{i^*}|/(1+\xi)\} \mid Z_{i^*} = z] \leq 2k\exp(-z^2\xi^4/(2(1+\xi)^2)) \ .$$

Let $C_\xi = 4(1+\xi)^2/\xi^4$. We get that for any $z \geq t \coloneqq \sqrt{C_\xi \log(2k)}$, it holds

$$\mathbf{Pr}[\cup_{j \in \mathcal{A}_{i^*}(\xi)}\{|Z_j| \geq |Z_{i^*}|/(1+\xi)\} \mid |Z_{i^*}| = z] \leq 1/2 \ .$$

Therefore, the probability of no violation conditioned on $|Z_{i^*}| = Z \geq t$ is at least

$$\mathbf{Pr}[\cap_{j \in \mathcal{A}_{i^*}(\xi)}\{|Z_j| \leq |Z_{i^*}|/(1+\xi)\} \mid |Z_{i^*}| = z] \geq 1/2.$$

Furthermore, by the anti-concentration inequality for standard Gaussians, we have:

$$\mathbf{Pr}[2t \geq |Z_{i^*}| \geq t] \geq \frac{1}{\sqrt{2\pi}t} e^{-t^2/2} \geq e^{-C_\xi \log(2k)} = (2k)^{-C_\xi}.$$

Therefore, we get that ($t = \sqrt{C_\xi \log(2k)}$):

$$\begin{aligned}\mathbf{Pr}[\mathcal{E}^{\mathrm{Sep}}_{i^*} \cap \{2t \geq |Z_{i^*}| \geq t\}] &= \mathbf{Pr}[\cap_{j \in \mathcal{A}_{i^*}(\xi)}\{|Z_j| \leq |Z_{i^*}|/(1+\xi)\} \cap \{2t \geq |Z_{i^*}| \geq t\}] \\ &\geq \mathbf{Pr}[\cap_{j \in \mathcal{A}_{i^*}(\xi)}\{|Z_j| \leq |Z_{i^*}|/(1+\xi)\} \mid t \leq |Z_{i^*}| \leq 2t]\, \mathbf{Pr}[2t \geq |Z_{i^*}| \geq t] \\ &\geq (1/2)(2k)^{-C_\xi} \ . \end{aligned} \tag{14}$$

Next observe that the random variable $\|\mathbf{p}\|_2^2$ is a sum of m squares of independent Gaussians (as $\mathbb{W}$ is an m-dimensional subspace). Therefore, $\|\mathbf{p}\|_2^2$ follows the χ^2_m distribution. From the concentration of χ^2_m, we have that

$$\mathbf{Pr}[\|\mathbf{p}\|_2 \notin (\sqrt{m/2}, \sqrt{2m})] \leq 2\exp(-m/16) \ .$$

It is without loss of generality that m is greater than a large absolute constant (Remark 4.5), and we have $m \geq 17C_\xi \log(2k)$. Thus, we obtain

$$\mathbf{Pr}[\|\mathbf{p}\|_2 \notin (\sqrt{m/2}, \sqrt{2m})] \leq (1/8)(2k)^{-C_\xi} \ . \tag{15}$$

Finally, since Z_j, $j \in [k]$, are standard Gaussian random variables, by the Gaussian concentration inequality and a union bound, we obtain:

$$\mathbf{Pr}\left[\max_{j\in[k]} |Z_j| \geq 4t\right] \leq 2k\exp(-8t^2) = 2k\exp(-8C_\xi \log(2k)) = (2k)^{-8C_\xi+1} \leq (1/8)(2k)^{-C_\xi}. \tag{16}$$

Combining (14), (15) and (16), since $\mathbf{Pr}[A \cap B \cap C] \geq 1 - \mathbf{Pr}[A^c] - \mathbf{Pr}[B^c] - \mathbf{Pr}[C^c]$, with

$$A = \mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \{2t \geq |Z_{i^*}| \geq t\},\ B = \|\mathbf{p}\|_2 \in (\sqrt{m/2}, \sqrt{2m}),\ C = \max_{j\in[k]} |Z_j| \geq 4t\ ,$$

we get that

$$\mathbf{Pr}[\mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \{2t \geq |Z_{i^*}| \geq t\} \cap \{\|\mathbf{p}\|_2 \in (\sqrt{m/2}, \sqrt{2m})\} \cap \{\max_{j\in[k]} |Z_j| \leq 4t\}] \geq (1/8)(2k)^{-C_\xi}\ .$$

Therefore, plugging in the value $t = \sqrt{C_\xi \log(2k)}$, on this event, we have

$$|\mathbf{u}\cdot\mathbf{w}_{i^*}| \geq \sqrt{\frac{C_\xi \log(2k)}{2m}} \geq 1/\sqrt{k}\ ,$$

and since $m \geq 2 \times 80^2 (C_\xi/\xi)\log(2k)$, we also have

$$|\mathbf{u}\cdot\mathbf{w}_j| \leq \frac{4t}{\sqrt{m/2}} \leq \frac{4\sqrt{2C_\xi \log(2k)}}{\sqrt{m}} \leq \frac{\sqrt{\xi}}{20}, \forall j \in [k].$$

Since $(Z_p(\mathcal{U}))^{1/(2p)} = 1/\sqrt{k}$, we conclude that $\mathbf{Pr}_{\mathbf{u}\sim\mathcal{U}}[\mathcal{E}_{i^*}^{\mathrm{Cor}} \cap \mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \mathcal{E}^{\mathrm{Ort}}] \geq \mathrm{poly}(1/k)$. □

Estimating the Target Matrix in Quasi-Polynomial Time. We now show that, with quasi-polynomial sample and computational complexity, we can approximate the target matrix $\mathbf{M}_{2p+2}(\mathbf{u})$ (M) in operator norm to high accuracy.

Proposition 4.8 (Sample complexity for the spectral method)**.** *Let $0 < \xi \leq 1/300$ be an absolute constant and let $0 < \lambda \leq \sqrt{\xi}/10$. Let $(\mathbf{x}, y) \sim \mathcal{D}$ be a sum of k ReLUs under Gaussian marginals. Let $p = \lceil \log(k/\lambda)/\log(1+\xi)\rceil$, $Z_p(\mathcal{U}) = k^{-p}$ and $\widehat{\sigma}''_{2p} = \widehat{\sigma}_{2p+2}\sqrt{(2p+1)(2p+2)}$. Fix $i^* \in [k]$ and suppose the unit probing vector $\mathbf{u}$ satisfies the "correlation", "separation" and "near orthogonality" events defined in Lemma 4.7: $\mathbf{u} \in \mathcal{E}_{i^*}^{\mathrm{Cor}} \cap \mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \mathcal{E}^{\mathrm{Ort}}$. Draw $N = d^2\,\mathrm{poly}(k, p, 1/\lambda)k^{O(\log(k/\lambda))}$ i.i.d. samples $(\mathbf{x}, y) \sim \mathcal{D}$ and construct the empirical distribution $\widehat{\mathcal{D}}$. Then, for*

$$\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}} = \frac{1}{\widehat{\sigma}''_{2p} Z_p(\mathcal{U})} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}} \left[y\mathrm{he}_{2p}(\mathbf{u}\cdot\mathbf{x})(\mathbf{x}^{\perp\mathbf{u}}\mathbf{x}^{\perp\mathbf{u}\top} - \mathbf{I}_{\perp\mathbf{u}})\right]\ ,$$

with probability at least $2/3$, it holds $\|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 \leq (\sqrt{\xi}/3)\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2$.

Proof. The high-level idea of the proof is that the target matrix $\mathbf{M}_{2p+2}(\mathbf{u})$ defined in (M) can be approximately written as the following expectation.

Claim 4.9. *Given any unit vector $\mathbf{u}$, define $\mathbf{M}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}} = \mathbf{P}_{\perp\mathbf{u}}\mathbf{M}_{2p+2}(\mathbf{u})\mathbf{P}_{\perp\mathbf{u}}$. Then,*

$$\mathbf{M}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}} = \frac{1}{\widehat{\sigma}''_{2p} Z_p(\mathcal{U})} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[y\mathrm{he}_{2p}(\mathbf{u}\cdot\mathbf{x})\mathbf{He}_2^{\mathbf{u}^\perp}(\mathbf{P}_{\perp\mathbf{u}}\mathbf{x})\right].$$

Proof of Claim 4.9. By our definition, the distribution $(\mathbf{x}, y) \sim \mathcal{D}$ satisfies $y(\mathbf{x}) = \sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$ and $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$. Therefore, denoting $\theta_{i\mathbf{u}} := \theta(\mathbf{w}_i, \mathbf{u})$, we have that for any unit vector $\mathbf{u} \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[y\mathrm{he}_{2p}(\mathbf{u}\cdot\mathbf{x})\mathbf{He}_2^{\mathbf{u}^\perp}(\mathbf{P}_{\perp\mathbf{u}}\mathbf{x})\right] \\
&= \sum_{i=1}^k \beta_i \mathop{\mathbf{E}}_{\mathbf{u}\cdot\mathbf{x}} \left[\mathop{\mathbf{E}}_{\mathbf{x}^{\perp\mathbf{u}}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_{\mathbf{u}^\perp})} \left[\sigma(\cos(\theta_{i\mathbf{u}})\widehat{\mathbf{w}_i^{\perp\mathbf{u}}}\cdot\mathbf{x}^{\perp\mathbf{u}})\mathbf{He}_2^{\mathbf{u}^\perp}(\mathbf{x}^{\perp\mathbf{u}}) \,\middle|\, \mathbf{u}\cdot\mathbf{x}\right] \mathrm{he}_{2p}(\mathbf{u}\cdot\mathbf{x})\right] \\
&\overset{(i)}{=} \sum_{i=1}^k \beta_i \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \left[\sigma''(\cos(\theta_{i\mathbf{u}})\mathbf{u}\cdot\mathbf{x} + \sin(\theta_{i\mathbf{u}})\widehat{\mathbf{w}_i^{\perp\mathbf{u}}}\cdot\mathbf{x})\mathrm{he}_{2p}(\mathbf{u}\cdot\mathbf{x})\right] \sin^2(\theta_{i\mathbf{u}})\widehat{\mathbf{w}_i^{\perp\mathbf{u}}}\widehat{\mathbf{w}_i^{\perp\mathbf{u}}}^\top \\
&\overset{(ii)}{=} \sum_{i=1}^k \beta_i \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \left[\mathrm{T}_{\cos(\theta_{i\mathbf{u}})}(\sigma'')(\mathbf{u}\cdot\mathbf{x})\mathrm{he}_{2p}(\mathbf{u}\cdot\mathbf{x})\right]\mathbf{w}_i^{\perp\mathbf{u}}\mathbf{w}_i^{\perp\mathbf{u}\top},
\end{aligned} \tag{17}$$

where in (i) we used Fact 2.6, by which $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[f(\mathbf{r}\cdot\mathbf{z})\mathbf{He}_2(\mathbf{z})] = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[f''(\mathbf{r}\cdot\mathbf{z})]\mathbf{r}\mathbf{r}^\top$ for any unit vector $\mathbf{r}$, and in (ii) $\mathrm{T}_\rho(\sigma'')$, $\rho\in[0,1]$, is the smoothing operator defined in Fact 2.3. By Fact 2.3, Part 2, we know that σ'' has Hermite expansion $\sigma''(z) = \sum_{i\geq 0} c_i''\mathrm{he}_i(z)$, where $\widehat{\sigma}_i'' = \widehat{\sigma}_{i+2}\sqrt{(i+1)(i+2)}$. Thus, applying Fact 2.3, Part 3 to (17) leads to:

$$\frac{1}{\widehat{\sigma}_{2p}'' Z_p(\mathcal{U})}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y\mathrm{he}_{2p}(\mathbf{u}\cdot\mathbf{x})\mathbf{He}_2^{\mathbf{u}^\perp}(\mathbf{P}_{\perp\mathbf{u}}\mathbf{x})\right] = \frac{1}{Z_p(\mathcal{U})}\sum_{i=1}^k \beta_i|\mathbf{w}_i\cdot\mathbf{u}|^{2p}\mathbf{w}_i^{\perp\mathbf{u}}\mathbf{w}_i^{\perp\mathbf{u}\top}$$
$$= \mathbf{P}_{\perp\mathbf{u}}\mathbf{M}_{2p+2}(\mathbf{u})\mathbf{P}_{\perp\mathbf{u}} =: \mathbf{M}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}}\,,$$

completing the proof of Claim 4.9. □

We next show that when $\mathbf{u}$ is a unit vector such that the event $\mathbf{u}\in\mathcal{E}_{i^*}^{\mathrm{Cor}}\cap\mathcal{E}_{i^*}^{\mathrm{Sep}}\cap\mathcal{E}^{\mathrm{Ort}}$ holds for λ-heavy hidden weight $\mathbf{w}_{i^*}$, then the spectral difference $\|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})\|_2$ is small.

Claim 4.10. *Under the assumptions of Proposition 4.8, it holds*

$$\|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 \leq \frac{\sqrt{\xi}}{8}\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2 \quad \textit{and} \quad \|\mathbf{M}_{2p+2}(\mathbf{u})\|_2 \geq \lambda\,.$$

Proof of Claim 4.10. Note that for any unit vector $\mathbf{w}_i$, by the triangle inequality, it holds

$$\|\mathbf{w}_i\mathbf{w}_i^\top - \mathbf{w}_i^{\perp\mathbf{u}}\mathbf{w}_i^{\perp\mathbf{u}\top}\|_2 \leq \|(\mathbf{w}_i - \mathbf{w}_i^{\perp\mathbf{u}})\mathbf{w}_i^\top\|_2 + \|\mathbf{w}_i^{\perp\mathbf{u}}(\mathbf{w}_i - \mathbf{w}_i^{\perp\mathbf{u}})\|_2 \leq 2\|\mathbf{w}_i - \mathbf{w}_i^{\perp\mathbf{u}}\|_2 = 2|\mathbf{u}\cdot\mathbf{w}_i|.$$

Since $\mathbf{u}$ also satisfies the near-orthogonal event $\mathcal{E}^{\mathrm{Ort}}$, i.e, for all $i\in[k]$ it holds $|\mathbf{u}\cdot\mathbf{w}_i|\leq\sqrt{\xi}/20$, we have $\|\mathbf{w}_i\mathbf{w}_i^\top - \mathbf{w}_i^{\perp\mathbf{u}}\mathbf{w}_i^{\perp\mathbf{u}\top}\|_2 \leq \sqrt{\xi}/10$. This implies the following inequality

$$\|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 \leq \frac{1}{Z_p(\mathcal{U})}\sum_{i=1}^k \beta_i|\mathbf{w}_i\cdot\mathbf{u}|^{2p}\|\mathbf{w}_i\mathbf{w}_i^\top - \mathbf{w}_i^{\perp\mathbf{u}}\mathbf{w}_i^{\perp\mathbf{u}\top}\|_2 \leq \frac{\sqrt{\xi}}{10}\sum_{i=1}^k \beta_i|\mathbf{w}_i\cdot\mathbf{u}|^{2p}.$$

Recall that in (12) we defined: $\mathcal{A}_{i^*}(\xi) = \{i\in[k] : |\mathbf{w}_i\cdot\mathbf{w}_{i^*}|\leq 1-\xi\}$, and

$$W_A = \frac{1}{Z_p(\mathcal{U})}\sum_{i\in\mathcal{A}_{i^*}(\xi)}\beta_i|\mathbf{w}_i\cdot\mathbf{u}|^{2p},\ W_B = \frac{1}{Z_p(\mathcal{U})}\sum_{i\notin\mathcal{A}_{i^*}(\xi)}\beta_i|\mathbf{w}_i\cdot\mathbf{u}|^{2p},$$

so that $W_A + W_B = (1/Z_p(\mathcal{U}))\sum_{i=1}^k \beta_i|\mathbf{w}_i\cdot\mathbf{u}|^{2p}$. Then in Claim 4.3, equation (13), we proved that when $\mathbf{u}\in\mathcal{E}_{i^*}^{\mathrm{Cor}}\cap\mathcal{E}_{i^*}^{\mathrm{Sep}}$, it holds $W_A\leq(\lambda/k)W_B$. Thus, we conclude that

$$\|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 \leq \frac{\sqrt{\xi}}{10}(W_A + W_B) \leq \frac{\sqrt{\xi}(\lambda/k+1)W_B}{10}.$$

On the other hand, since $\beta_i\geq 0$ for all $i\in[k]$, we have:

$$\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2 \geq \mathbf{w}_{i^*}^\top\mathbf{M}_{2p+2}(\mathbf{u})\mathbf{w}_{i^*} \geq \frac{1}{Z_p(\mathcal{U})}\sum_{i\notin\mathcal{A}_{i^*}(\xi)}\beta_i|\mathbf{w}_i\cdot\mathbf{u}|^{2p}(\mathbf{w}_i\cdot\mathbf{w}_{i^*})^2 \geq (1-\xi)^2 W_B.$$

Therefore, with the choice $\xi\in(0,1/300]$ and $\lambda\leq\sqrt{\xi}/10$, we conclude that:

$$\|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 \leq \frac{\sqrt{\xi}(\lambda/k+1)}{10(1-\xi^2)}\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2 \leq \frac{\sqrt{\xi}}{8}\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2.$$

Finally, note that $\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2 \geq \beta_{i^*}|\mathbf{w}_{i^*}\cdot\mathbf{u}|^{2p}/Z_p(\mathcal{U})$, since $\mathbf{w}_{i^*}$ is a λ-heavy hidden weight and $\mathbf{u}$ satisfies $\mathcal{E}_{i^*}^{\mathrm{Cor}}$, it holds that $\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2\geq\lambda$. This completes the proof of Claim 4.10. □

Claim 4.10 implies that an accurate estimation of the matrix $\mathbf{M}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}}$ approximates the target matrix from (M) in operator norm.

The rest of the proof is devoted to determining how many samples are required to estimate $\mathbf{M}_{2p+2}(\mathbf{u})^{\perp\mathbf{u}}$. We show that:

Lemma 4.11. *Using $N \gtrsim \frac{d^2 3^{2p} k^2}{(\widehat{\sigma}''_{2p})^2 \lambda^2 Z_p(\mathcal{U})^2}$ samples, with probability at least 2/3, it holds*

$$\left\|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\right\|_2 \leq \frac{\sqrt{\xi}}{5} \|\mathbf{M}_{2p+2}(\mathbf{u})\|_2.$$

Proof of Lemma 4.11. Set $Q = \widehat{\sigma}''_{2p} Z_p(\mathcal{U})$, and let $\mathbf{X}_i = Q^{-1} y^{(i)} \mathrm{he}_{2p}(\mathbf{u} \cdot \mathbf{x}^{(i)})(\mathbf{x}^{(i)\perp \mathbf{u}} \mathbf{x}^{(i)\perp \mathbf{u}\top} - \mathbf{I}_{\perp \mathbf{u}})$ and note that $\mathbf{E}[\mathbf{X}_i] = \mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}$ by Claim 4.9. By Chebyshev inequality, we have that

$$\mathbf{Pr}\left[\left\|\frac{1}{N}\sum \mathbf{X}_i - \mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\right\|_2 \geq t\|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\|_2\right] \leq \frac{\mathbf{E}[\|\frac{1}{N}\sum \mathbf{X}_i - \mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\|_2^2]}{t^2 \|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\|_2^2}.$$

By the independence of samples, $\mathbf{E}[\|\frac{1}{N}\sum \mathbf{X}_i - \mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\|_2^2] \leq \frac{2}{N}\mathbf{E}[\|\mathbf{X}_i\|_F^2]$. In addition, by the definition of the labels, we know that $y = \sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$. Furthermore, by the independence between $\mathbf{u} \cdot \mathbf{x}$ and $\mathbf{x}^{\perp \mathbf{u}}$ and by Cauchy–Schwarz inequality, the second-order moment of $\mathbf{X}_i$ is bounded by

$$\begin{aligned}
\mathbf{E}[\|\mathbf{X}_i\|_F^2] &\leq Q^{-2}\,\mathbf{E}\left[y^2 \mathrm{he}_{2p}^2(\mathbf{u}\cdot\mathbf{x})\|(\mathbf{x}^{\perp \mathbf{u}}\mathbf{x}^{\perp \mathbf{u}\top} - \mathbf{I}_{\perp \mathbf{u}})\|_F^2\right] \\
&= \frac{1}{Q^2}\,\mathbf{E}\left[\Big(\sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})\Big)^2 \mathrm{he}_{2p}^2(\mathbf{u}\cdot\mathbf{x})\|(\mathbf{x}^{\perp \mathbf{u}}\mathbf{x}^{\perp \mathbf{u}\top} - \mathbf{I}_{\perp \mathbf{u}})\|_F^2\right] \\
&\leq \frac{1}{Q^2}\sqrt{\mathbf{E}\left[\Big(\sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})\Big)^4\right]\mathbf{E}[\mathrm{he}_{2p}^4(\mathbf{u}\cdot\mathbf{x})\|(\mathbf{x}^{\perp \mathbf{u}}\mathbf{x}^{\perp \mathbf{u}\top} - \mathbf{I}_{\perp \mathbf{u}})\|_F^4]} \\
&= \frac{1}{Q^2}\sqrt{\mathbf{E}\left[\Big(\sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})\Big)^4\right]\mathbf{E}[\mathrm{he}_{2p}^4(\mathbf{u}\cdot\mathbf{x})]\,\mathbf{E}[\|(\mathbf{x}^{\perp \mathbf{u}}\mathbf{x}^{\perp \mathbf{u}\top} - \mathbf{I}_{\perp \mathbf{u}})\|_F^4]}\,.
\end{aligned}$$

The fourth moment of y is bounded above by $\mathbf{E}[(\sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x}))^4] \lesssim k^4$. In addition, using Fact 2.2, we have that $\mathbf{E}[\mathrm{he}_{2p}^4(\mathbf{u}\cdot\mathbf{x})] \leq 3^{4p}$. Further, $\mathbf{E}[\|(\mathbf{x}^{\perp \mathbf{u}}\mathbf{x}^{\perp \mathbf{u}\top} - \mathbf{I}_{\perp \mathbf{u}})\|_F^4] \leq d^4$. Therefore, we have:

$$\mathbf{E}[\|\mathbf{X}_i\|_F^2] \lesssim \frac{d^2 3^{2p} k^2}{Q^2}.$$

Plugging the upper bound on the second moment $\mathbf{E}[\|\mathbf{X}_i\|_2^2]$ back into Chebyshev inequality yields

$$\mathbf{Pr}\left[\left\|\frac{1}{N}\sum \mathbf{X}_i - \mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\right\|_2 \geq t\|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\|_2\right] \lesssim \frac{d^2 3^{2p}\sqrt{k^4}}{Q^2 N t^2 \|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\|_2^2}.$$

When $\xi \in (0, 1/300]$, by Claim 4.10 and using the triangle inequality, it holds that

$$\|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\|_2 \geq (1 - \sqrt{\xi}/8)\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2 \geq (1/3)\lambda.$$

Thus, choosing $t \lesssim \sqrt{\xi}$, if the sample size satisfies

$$N \gtrsim \frac{d^2 3^{2p} k^2}{Q^2 \lambda^2} = \frac{d^2 3^{2p} k^2}{(\widehat{\sigma}''_{2p})^2 \lambda^2 Z_p(\mathcal{U})^2},$$

we have that with probability at least 2/3,

$$\left\|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\right\|_2 \leq \frac{\sqrt{\xi}\lambda}{10} \leq \frac{\sqrt{\xi}}{10}\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2,$$

where in the last inequality we used the fact that conditioned on $\mathbf{u} \in \mathcal{E}_{i^*}^{\mathrm{Cor}} \cap \mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \mathcal{E}^{\mathrm{Ort}}$, it holds $\|\mathbf{M}_{2p+2}(\mathbf{u})\|_2 \geq \lambda$ by Claim 4.10. This completes the proof of Lemma 4.11. □

Now combining Claim 4.10, Lemma 4.11 and using the triangle inequality yields:

$$\begin{aligned}
\left\|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})\right\|_2 &\leq \left\|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}}\right\|_2 + \left\|\mathbf{M}_{2p+2}(\mathbf{u})^{\perp \mathbf{u}} - \mathbf{M}_{2p+2}(\mathbf{u})\right\|_2 \\
&\leq \frac{\sqrt{\xi}}{3}\left\|\mathbf{M}_{2p+2}(\mathbf{u})\right\|_2.
\end{aligned}$$

Since $p = \lceil \log(k/\lambda)/\log(1+\xi) \rceil$ and $Z_p(\mathcal{U}) = k^{-p} = k^{-O(\log(k/\lambda))}$, $\widehat{\sigma}_{2p+2} \geq 1/(2p+2)^3$ (by Fact 2.4), by Lemma 4.11 it follows that the number of samples required is

$$N = d^2 \operatorname{poly}(p) k^{O(\log(k/\lambda))}.$$

The probability of success is at least 2/3, conditioned on the event $\mathbf{u} \in \mathcal{E}_{i^*}^{\mathrm{Cor}} \cap \mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \mathcal{E}^{\mathrm{Ort}}$. □

Proof of the Main Result of the Subsection: Quasi-Polynomial Complexity CDR. We are now ready to prove Proposition 4.6, which we restate below:

Proposition 4.6 (Quasi-Polynomial-Time Implementation of CDR)**.** *Let $\delta \in (0,1)$, $\xi \in (0, 1/300]$ be an absolute constant and let $\lambda \in (0, \sqrt{\xi}/10]$. Let $(\mathbf{x}, y) \sim \mathcal{D}$ be a distribution such that $y(\mathbf{x}) = \sum_{i=1}^k \beta_i \sigma(\mathbf{w}_i \cdot \mathbf{x})$ where σ is the ReLU activation, $\beta_i \in [0,1]$, $\|\mathbf{w}_i\|_2 = 1$, and $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Let $p \geq \lceil \log(k/\lambda)/\log(1+\xi) \rceil$. Then, CDRviaSpectral$(\mathcal{D}, \lambda, \delta)$ draws $N = d^2\, k^{O(\log(k/\lambda))} \log(1/\delta)$ samples from $\mathcal{D}$, runs in $T = \operatorname{poly}(d, N)$ time, and outputs a list $\mathcal{L}$ of unit vectors of size $|\mathcal{L}| = \operatorname{poly}(k)\log(1/\delta)$ that with probability at least $1-\delta$ satisfies the following guarantee: for any $\mathbf{w}_i$, $i \in [k]$ with $\beta_i \geq \lambda$, there exists $\mathbf{r}_i \in \mathcal{L}$ such that $|\mathbf{r}_i \cdot \mathbf{w}_i| \geq 0.99$.*

Proof of Proposition 4.6. Fix any $i^* \in [k]$ such that $\beta_{i^*} \geq \lambda$. By Lemma 4.7, we have that $\mathbf{Pr}[\mathcal{E}_{i^*}^{\mathrm{Cor}} \cap \mathcal{E}_{i^*}^{\mathrm{Sep}}] \geq \mathbf{Pr}[\mathcal{E}_{i^*}^{\mathrm{Cor}} \cap \mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \mathcal{E}^{\mathrm{Ort}}] \geq \operatorname{poly}(1/k)$. Combining Lemma 4.2 and Proposition 4.8, for a probing vector $\mathbf{u}$ satisfying $\mathcal{E}_{i^*}^{\mathrm{Cor}} \cap \mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \mathcal{E}^{\mathrm{Ort}}$, using $N = d^2 k^{O(\log(k/\lambda))}$ samples, the empirical matrix $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})$ from Line 9 of CDRviaSpectral$(\mathcal{D}, \lambda, \delta)$ satisfies $|\mathbf{r}_1(\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})) \cdot \mathbf{w}_{i^*}| \geq 0.99$ with probability at least 2/3. Thus, for any fixed $i^* \in [k]$ with $\beta_{i^*} \geq \lambda$, the probability that the top eigenvector of the empirical matrix $\mathbf{r}_{i^*} = \mathbf{r}_1(\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}))$ satisfies $|\mathbf{r}_{i^*} \cdot \mathbf{w}_{i^*}| \geq 0.99$ is at least $\operatorname{poly}(1/k)$. Hence, by a union bound and a Chernoff bound, drawing $\operatorname{poly}(k)\log(1/\delta)$ probing vectors and constructing $\operatorname{poly}(k)\log(1/\delta)$ empirical matrices ensures that with probability at least $1-\delta$, for any $i^* \in [k]$ with $\beta_{i^*} \geq \lambda$, the list $\mathcal{L}$ returned by CDRviaSpectral$(\mathcal{D}, \lambda, \delta)$ contains a unit vector $\mathbf{r}_{i^*}$ such that $|\mathbf{r}_{i^*} \cdot \mathbf{w}_{i^*}| \geq 0.99$. This completes the proof. □

4.3 Polynomial Time CDR via Implicit Tensor Computation

In this section, we provide a CDR subroutine that uses polynomially many samples and runs in polynomial time. Similar to Section 4.2, the core idea is to estimate the target matrix $\mathbf{M}_{2p+2}(\mathbf{u})$ (M), constructed by projecting the moment tensor $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{He}_{2p+2}(\mathbf{x})]$ to a low-degree filter $\mathbf{u}^{\otimes p} \otimes \bar{\mathbf{u}}^{\otimes p}$:

$$\begin{aligned}\mathbf{M}_{2p+2}(\mathbf{u}) &= \frac{1}{Z_p(\mathcal{U})}\left\langle \frac{1}{\widehat{\sigma}_{2p+2}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{He}_{2p+2}(\mathbf{x})],\, \mathbf{u}^{\otimes p} \otimes \bar{\mathbf{u}}^{\otimes p}\right\rangle \\ &= \frac{1}{Z_p(\mathcal{U})} \sum_{i=1}^k b_i |\mathbf{w}_i \cdot \mathbf{u}|^{2p} \mathbf{w}_i \mathbf{w}_i^\top,\end{aligned}$$

where $\mathbf{u}$ is a random probing vector sampled from some distribution $\mathcal{U}$. Once we obtain an accurate estimation $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})$, we can invoke Lemma 4.2, which states that the top eigenvector of $\mathbf{M}_{2p+2}(\mathbf{u})$ aligns well with the target vector $\mathbf{w}_i$, conditions on $\mathbf{u}$ that are satisfied with inverse-polynomial probability.

To ensure polynomial complexity, we take a significant departure from Section 4.2 in both components of the algorithm: (1) instead of using the uniform distribution on the sphere, we choose $\mathcal{U}$ to be a complex Gaussian distribution: $\mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$, $\mathbf{p}, \mathbf{q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$; and (2) instead of estimating the matrix $\mathbf{M}_{2p+2}(\mathbf{u})$ via low-degree polynomials, we use more advanced techniques from implicit tensor computation [DK25]. Choosing $\mathcal{U}$ to be the complex Gaussian has two main advantages: first, the operator norm of the target matrix $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})$ becomes larger, since $|\mathbf{w}_i \cdot \mathbf{u}|$ is no longer restricted to be smaller than one, as was the case when $\mathbf{u}$ were sampled uniformly on the sphere; second, the covariance of $\mathbf{u}^{\otimes p}$ for $\mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$ is bounded above by $2^p p!$ (cf. Lemma 4.23)—when $p \lesssim k$, this is much smaller compared to $O(k^{p/2})$, the covariance of $\mathbf{u}^{\otimes p}$ for $\mathbf{u} \sim \mathrm{Unif}(\mathbb{S}^{k-1})$. These benefits together ensure that we can apply the implicit tensor computation techniques to estimate the target matrix accurately using only a polynomial number of samples, in polynomial time. Finally, while it is possible to choose other types of distributions $\mathcal{U}$ with similar properties; the complex Gaussian distribution is the most convenient one for our analysis.

The pseudo-code of the polynomial-time CDR subroutine is given in CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$.

Algorithm 7 CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$

Input: Sample access to a sum of k ReLUs with Gaussian marginals $(\mathbf{x}, y) \sim \mathcal{D}$, positive parameters $\lambda, \delta \in (0, 1)$.
Output: A list of unit vectors satisfying Assumption 1.

Fix an absolute constant $\xi \in (0, 1/300]$, $\lambda \leftarrow \min\{\lambda, \sqrt{\xi}/10\}$.
$p \leftarrow \lceil \log(k/\lambda)/\log(1+\xi) \rceil$. ▷ Filter degree
$N_{\mathbf{u}} \leftarrow \mathrm{poly}(k/\lambda)\log(1/\delta)$. ▷ Number of probing vectors/Low-degree filter
$\tau \leftarrow \delta/(10d^2 N_{\mathbf{u}})$. ▷ Failure probability of the immediate steps
$N_0 \leftarrow \mathrm{poly}(d, k/\lambda, 1/\tau, 1/\delta)$. ▷ Number of samples
Let the distribution $\mathcal{I} \sim \mathcal{D}(\mathcal{I})$ be $\mathcal{I} = (\mathbf{x}, y, \mathbf{z})$, where $(\mathbf{x}, y) \sim \mathcal{D}$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.
$\widehat{\Phi}_p \leftarrow$ TensorCompressor$(p, N_0, \mathcal{D}(\mathcal{I}))$
$\mathbf{A}_p \leftarrow$ CompressEmpMoment$(p, N_0, \widehat{\Phi}_p, \mathcal{D}(\mathcal{I}))$
$\mathcal{L} \leftarrow \emptyset$.
for $j = 1, \ldots, N_{\mathbf{u}}$ **do**
  Draw a fresh vector $\mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$, where $\mathbf{p}, \mathbf{q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$
  $\mathbf{B}_p(\mathbf{u}) \leftarrow$ CompressFilter$(p, \widehat{\Phi}_p, \mathbf{u})$
  $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) = \langle \mathbf{A}_p, \mathbf{B}_p(\mathbf{u}) \rangle$ ▷ Empirical matrix, cf. EstM
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{\pm \mathbf{r}_1(\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}))\}$. ▷ Top eigenvectors of the empirical matrix
Return: $\mathcal{L}$.

The whole section is devoted to proving Proposition 4.1, which is restated below. We defer its proof to Section 4.4, after completing all the ingredients.

Proposition 4.1 (Polynomial-Time CDR)**.** *Let $k, d \in \mathbb{Z}_+$ and $\lambda, \delta \in (0, 1)$. Let $\mathcal{D}$ be a Gaussian sum of k ReLUs. Then, there is an algorithm (CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$) that draws $N = \mathrm{poly}(d, k, 1/\lambda, 1/\delta)$ i.i.d. samples from $\mathcal{D}$, runs in $T = \mathrm{poly}(d, N)$ time, and outputs a list $\mathcal{L}$ of unit vectors of size $|\mathcal{L}| = \mathrm{poly}(k/\lambda)\log(1/\delta)$ such that with probability at least $1 - \delta$, for any $\mathbf{w}_i$, $i \in [k]$, with $\beta_i \geq \lambda$, there exists a vector $\mathbf{r}_i \in \mathcal{L}$ such that $|\mathbf{r}_i \cdot \mathbf{w}_i| \geq 0.99$.*

Remark 4.12. Throughout this section, we assume without loss of generality that $k \leq d$. If $k > d$, before running our CDR procedure CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$, set $d' = \max\{d, k\}$. If $d' > d$, augment every example $\mathbf{x}$ by an independent vector $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_{d'-d})$, and use $\mathbf{x}' = (\mathbf{x}, \mathbf{z})$ with the same label. The lifted hidden weights are $\mathbf{w}_i' = (\mathbf{w}_i, 0)$, so the target and the Gaussian model are preserved. Run the CDR procedure (CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$) in dimension d'. If a unit vector $\mathbf{r} \in \mathbb{R}^{d'}$ is returned, retain $\mathbf{P}\,\mathbf{r}/\|\mathbf{P}\,\mathbf{r}\|_2$ when $\mathbf{P}\,\mathbf{r} \neq 0$, where $\mathbf{P}$ projects onto the original d coordinates; discard a zero projection. For every original unit weight $\mathbf{w}_i$, it holds

$$\left| \mathbf{w}_i \cdot \frac{\mathbf{P}\,\mathbf{r}}{\|\mathbf{P}\,\mathbf{r}\|_2} \right| = \frac{|\mathbf{w}_i' \cdot \mathbf{r}|}{\|\mathbf{P}\,\mathbf{r}\|_2} \geq |\mathbf{w}_i' \cdot \mathbf{r}|.$$

Thus, a useful output retains its correlation guarantee after projection, the output list cannot grow, and the cost remains polynomial in d and k. In the tensor analysis below, relabel d' as d, so $d \geq k$.

4.3.1 Determining the Probability of the Probing Event

Our first step is to determine the probability that the signal of the target vectors passes through the random filter—in other words, the probability that the "separation" and "correlation" events $\mathcal{E}_{i^*}^{\mathrm{Sep}}$, $\mathcal{E}_{i^*}^{\mathrm{Cor}}$ (defined in (11)) happen under the complex distribution $\mathcal{U}$. We prove that, with $p = \lceil \log(k/\lambda)/\log(1+\xi) \rceil$ and $Z_p(\mathcal{U}) = p!$, the event $\mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \mathcal{E}_{i^*}^{\mathrm{Cor}}$ happens with probability at least $\mathrm{poly}(\lambda/k)$.

Lemma 4.13 (Probability of Successful Filtration)**.** *Let $\mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$, where $\mathbf{p}, \mathbf{q} \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_d)$. Let $\xi \in (0, 1/300]$ be a small absolute constant, $p = \lceil \log(k/\lambda)/\log(1+\xi) \rceil$ and $Z_p(\mathcal{U}) = p!$. Let $\mathbf{w}_1, \ldots, \mathbf{w}_k$ be k unit vectors. Fixing any $i^* \in [k]$, let $\mathcal{A}_{i^*}(\xi) = \{i \in [k] : |\mathbf{w}_i \cdot \mathbf{w}_{i^*}| \leq 1 - \xi\}$ and let the "correlation" and "separation" events on $\mathcal{U}$ be defined by (11). Then, $\mathbf{Pr}_{\mathbf{u} \sim \mathcal{U}}[\mathcal{E}_{i^*}^{\mathrm{Cor}} \cap \mathcal{E}_{i^*}^{\mathrm{Sep}}] \geq \mathrm{poly}(\lambda/k)$.*

Proof. Without loss of generality, let $i^* = 1$ and fix $\mathbf{w}_1$. Let $C_\xi = \sqrt{12}((1+\xi)/\xi^2)$ and $R = C_\xi \sqrt{\log(k/\lambda)}$. By our choice of parameters, we have $R^2 \geq p$. Furthermore, since $R^{2p} \geq p^p \geq p! = Z_p(\mathcal{U})$, we have

$$\mathcal{E}_1^{\mathrm{Cor}} = \{\mathbf{u} : |\mathbf{w}_1 \cdot \mathbf{u}| \geq (Z_p(\mathcal{U}))^{1/(2p)}\} \supset \{\mathbf{u} : |\mathbf{w}_1 \cdot \mathbf{u}| \geq R\} := \mathcal{E}_1^{\mathrm{Cor},2}.$$

We proceed by bounding the probability of the event $\mathcal{E}_1^{\mathrm{Cor}}$. For any fixed unit vector $\mathbf{r} \in \mathbb{R}^d$, we have $|\mathbf{r} \cdot \mathbf{u}|^2 = (\mathbf{p} \cdot \mathbf{r})^2 + (\mathbf{q} \cdot \mathbf{r})^2 = X_1^2 + X_2^2$, where X_1, X_2 are independent standard Gaussian random variables; therefore, by the anti-concentration inequality for the standard Gaussian distribution, it holds that

$$\begin{aligned}\mathbf{Pr}[|\mathbf{r} \cdot \mathbf{u}| \geq R] = \mathbf{Pr}[|\mathbf{r} \cdot \mathbf{u}|^2 \geq R^2] \geq \mathbf{Pr}[X_1^2 \geq R^2/2, X_2^2 \geq R^2/2] \geq \mathbf{Pr}[|X_1| \geq R/\sqrt{2}]^2 \\ \geq \exp(-cR^2) = \mathrm{poly}(\lambda/k).\end{aligned}$$

Therefore, $\mathbf{Pr}[\mathcal{E}_1^{\mathrm{Cor}}] \geq \mathbf{Pr}[\mathcal{E}_1^{\mathrm{Cor},2}] \geq \mathrm{poly}(\lambda/k)$.

It remains to bound $\mathbf{Pr}[\mathcal{E}_1^{\mathrm{Sep}} | \mathcal{E}_1^{\mathrm{Cor},2}]$. Denote $z_i := \mathbf{u} \cdot \mathbf{w}_i$. Consider the joint distribution of $z_1 = \mathbf{w}_1 \cdot \mathbf{u}$ and $z_i = \mathbf{w}_i \cdot \mathbf{u}$ for any $i \in \mathcal{A}_1(\xi)$. Letting $\cos\theta_i := \cos(\theta_\pm(\mathbf{w}_i, \mathbf{w}_1)) = |\mathbf{w}_i \cdot \mathbf{w}_1|$, we have $z_i = \mathbf{u} \cdot (\cos\theta_i \mathbf{w}_1 + \sin\theta_i \mathbf{s}_i)$, where $\mathbf{s}_i = \mathbf{w}_i^{\perp \mathbf{w}_1} / \|\mathbf{w}_i^{\perp \mathbf{w}_1}\|_2$. Denote for simplicity $g_i = \mathbf{u} \cdot \mathbf{s}_i$. We condition on $\mathcal{E}_1^{\mathrm{Cor},2}$, i.e., $|z_1| = Z \geq R$. For any $i \in \mathcal{A}_1(\xi)$, we have $\cos(\theta_i) = |\mathbf{w}_i \cdot \mathbf{w}_1| \leq 1 - \xi$. By the triangle inequality, whenever $|\mathbf{w}_i \cdot \mathbf{u}| \geq |\mathbf{w}_1 \cdot \mathbf{u}|/(1+\xi)$ and $|\mathbf{w}_1 \cdot \mathbf{u}| = Z$, it holds

$$|g_i| \geq -\frac{Z\cos(\theta_i)}{\sin(\theta_i)} + \frac{Z}{(1+\xi)\sin(\theta_i)} \geq \frac{Z}{\sin(\theta_i)}\left(\frac{1}{1+\xi} - (1-\xi)\right) \geq \frac{Z\xi^2}{1+\xi} = \frac{\sqrt{12}Z}{C_\xi} ,$$

where we used that $\cos(\theta_i) \leq 1 - \xi$. Therefore, we have:

$$\mathbf{Pr}\left[|z_i| \geq \frac{|z_1|}{1+\xi} \middle| |z_1| = Z, Z \geq R\right] \leq \mathbf{Pr}\left[|g_i| \geq \frac{\sqrt{12}Z}{C_\xi} \middle| Z \geq R\right] \leq \mathbf{Pr}\left[|g_i| \geq \frac{\sqrt{12}R}{C_\xi}\right]. \tag{18}$$

Recall that $|g_i|^2 = |(\mathbf{p} + \mathrm{i}\mathbf{q}) \cdot \mathbf{s}_i|^2 = X_1^2 + X_2^2 \sim \chi_2^2$, therefore, by the standard concentration inequality of χ_2^2 random variables [LM00], we have

$$\mathbf{Pr}[|g_i| \geq t] = \mathbf{Pr}[|g_i|^2 \geq t^2] \leq \exp(-t^2/6)$$

when $t^2 \geq 2$. Plugging in $t = \sqrt{12}R/C_\xi = \sqrt{12\log(k/\lambda)}$, and combining with (18), we obtain that:

$$\mathbf{Pr}\left[|z_i| \geq \frac{|z_1|}{1+\xi} \middle| |z_1| \geq R\right] \leq \exp(-2\log(k/\lambda)) = (\lambda/k)^2.$$

Applying a union bound on all such events for $i \in \mathcal{A}_1(\xi)$, we obtain

$$\mathbf{Pr}[\exists i \in \mathcal{A}_1(\xi),\ |z_i| \geq |z_1|/(1+\xi) \mid |z_1| \geq R] \leq \lambda^2/k.$$

Therefore, we get that $\mathbf{Pr}[\mathcal{E}_1^{\mathrm{Sep}} | \mathcal{E}_1^{\mathrm{Cor},2}] \geq 1 - \lambda^2/k$ and hence $\mathbf{Pr}[\mathcal{E}_1^{\mathrm{Sep}} \cap \mathcal{E}_1^{\mathrm{Cor}}] \geq \mathbf{Pr}[\mathcal{E}_1^{\mathrm{Sep}} \cap \mathcal{E}_1^{\mathrm{Cor},2}] \geq \mathrm{poly}(\lambda/k)$, completing the proof of Lemma 4.13. □

4.3.2 Estimating the Target Matrix Using Implicit Tensor Computation

In this section, we show how to use implicit tensor computation to estimate the target matrix (M). Our main result in this context is the following:

Proposition 4.14 (Matrix Estimation Using Implicit Tensor Computation)**.** *Let $\mathcal{D}$ be a Gaussian sum of k ReLUs. Let $p, N_0, N_{\mathbf{u}} \in \mathbb{Z}_+$, $\xi, \lambda, \delta \in (0,1)$ be the parameters prescribed in* CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$*. Let $\mathbf{M}_{2p+2}(\mathbf{u})$ be the matrix defined in (M). Let $\widehat{\mathcal{U}}$ be a set of $N_{\mathbf{u}}$ random probing vectors following the distribution $\mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$, $\mathbf{p}, \mathbf{q} \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_d)$. Then the matrices $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})$ computed in Line 15 of* CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$ *satisfy the following guarantee: with probability at least $1 - \delta/2$ over all the samples and the probing vectors, $\max_{\mathbf{u} \in \widehat{\mathcal{U}}} \|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 \leq \sqrt{\xi}\lambda$.*

The rest of this subsection is devoted to the proof of Proposition 4.14.

Our first observation is that $\mathbf{M}_{2p+2}(\mathbf{u})$ is obtained via contracting the moment tensor $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{He}_{2p+2}(\mathbf{x})]$ with $\mathbf{F}_p(\mathbf{u}) \otimes \mathbf{F}_p(\bar{\mathbf{u}})$, $\mathbf{F}_p(\mathbf{u}) = \mathbf{u}^{\otimes p}/\sqrt{p!}$, where we remind the reader that $\bar{\mathbf{u}}$ is the complex conjugate of the vector $\mathbf{u}$.

Claim 4.15. *Let $\mathbf{M}_{2p+2}(\mathbf{u})$ be the matrix defined by (i) in (M). Let $\mathbf{F}_p(\mathbf{u}) = \mathbf{u}^{\otimes p}/\sqrt{p!}$. Then:*

$$\mathbf{M}_{2p+2}(\mathbf{u}) = \frac{1}{\widehat{\sigma}_{2p+2}} \left\langle \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[y\mathbf{He}_{2p+2}(\mathbf{x}) \right], \mathbf{F}_p(\mathbf{u}) \otimes \mathbf{F}_p(\bar{\mathbf{u}}) \right\rangle.$$

Proof. By the Hermite expansion of the ReLU activation σ, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[y\mathbf{He}_{2p+2}(\mathbf{x}) \right] = \sum_{i=1}^{k} \beta_i \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\sigma(\mathbf{w}_i \cdot \mathbf{x})\mathbf{He}_{2p+2}(\mathbf{x}) \right] = \sum_{i=1}^{k} \beta_i \widehat{\sigma}_{2p+2} \mathbf{w}_i^{\otimes(2p+2)}.$$

Taking the tensor product and noting that $\langle \mathbf{w}_i^{\otimes 2p}, \mathbf{u}^{\otimes p} \otimes \bar{\mathbf{u}}^{\otimes p} \rangle = (\mathbf{w}_i \cdot \mathbf{u})^p (\mathbf{w}_i \cdot \bar{\mathbf{u}})^p = |\mathbf{w}_i \cdot \mathbf{u}|^{2p}$ gives the stated claim. □

Therefore, the problem boils down to estimating the contraction of two tensors $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{He}_{2p+2}(\mathbf{x})]$ and $\mathbf{F}_p(\mathbf{u}) \otimes \mathbf{F}_p(\bar{\mathbf{u}})$ using samples from $\mathcal{D}$. To this end, we apply the implicit tensor technique from [DK25].

Preliminaries for Implicit Tensor Computation. The central idea is the following: Recall that the target matrix $\mathbf{M}_{2p+2}(\mathbf{u})$ is defined by

$$\mathbf{M}_{2p+2}(\mathbf{u}) = \frac{1}{\widehat{\sigma}_{2p+2}} \left\langle \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[y\mathbf{He}_{2p+2}(\mathbf{x}) \right], \mathbf{F}_p(\mathbf{u}) \otimes \mathbf{F}_p(\bar{\mathbf{u}}) \right\rangle = \sum_{i=1}^{k} \langle \mathbf{w}_i^{\otimes p}, \mathbf{F}_p(\mathbf{u}) \rangle \langle \mathbf{w}_i^{\otimes p}, \mathbf{F}_p(\bar{\mathbf{u}}) \rangle \mathbf{w}_i \otimes \mathbf{w}_i.$$

Let $\mathbf{h}_1, \ldots, \mathbf{h}_k \in (\mathbb{R}^d)^{\otimes p}$ be the basis of the subspace $\mathbb{U} \subset (\mathbb{R}^d)^{\otimes p}$ spanned by "vectors" $\mathbf{w}_i^{\otimes p}$, $i \in [k]$. Suppose we have the projection $\Phi_p \in \mathbb{R}^{k \times d^p}$, $\Phi_p^\top = [\mathbf{h}_1, \ldots, \mathbf{h}_k]$ that maps a "vector" from $(\mathbb{R}^d)^{\otimes p}$ to the subspace $\mathbb{U} \subset (\mathbb{R}^d)^{\otimes p}$ followed by an isometry from $\mathbb{U}$ to $\mathbb{R}^k$, then since $\mathbf{w}_i^{\otimes p}$ is a "vector" in $\mathbb{U}$, it holds $\Phi_p^\top \Phi_p(\mathbf{w}_i^{\otimes p}) = \mathbf{w}_i^{\otimes p}$. In addition, since Φ_p is a matrix with real entries, consider its natural complex extension on a complex tensor $\mathbf{F}_p(\mathbf{u})$: $\Phi_p(\mathbf{F}_p(\mathbf{u})) = \Phi_p(\mathrm{Re}(\mathbf{F}_p(\mathbf{u}))) + \mathrm{i}\Phi_p(\mathrm{Im}(\mathbf{F}_p(\mathbf{u})))$. Then its Hermitian adjoint is $\Phi_p^* = \Phi_p^\top$. Therefore, we have

$$\langle \mathbf{w}_i^{\otimes p}, \mathbf{F}_p(\mathbf{u}) \rangle = \langle \Phi_p^\top \Phi_p(\mathbf{w}_i^{\otimes p}), \mathbf{F}_p(\mathbf{u}) \rangle = \langle \Phi_p(\mathbf{w}_i^{\otimes p}), (\Phi_p^\top)^*(\mathbf{F}_p(\mathbf{u})) \rangle = \langle \Phi_p(\mathbf{w}_i^{\otimes p}), \Phi_p(\mathbf{F}_p(\mathbf{u})) \rangle.$$

By the linearity of expectation, this implies that the target matrix $\mathbf{M}_{2p+2}(\mathbf{u})$ can be written as:

$$\begin{aligned}
\mathbf{M}_{2p+2}(\mathbf{u}) &= \sum_{i=1}^{k} \langle \Phi_p(\mathbf{w}_i^{\otimes p}), \Phi_p(\mathbf{F}_p(\mathbf{u})) \rangle \langle \Phi_p(\mathbf{w}_i^{\otimes p}), \Phi_p(\mathbf{F}_p(\bar{\mathbf{u}})) \rangle \mathbf{w}_i \otimes \mathbf{w}_i \\
&= \sum_{i=1}^{k} \left\langle (\Phi_p \otimes \Phi_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d)(\mathbf{w}_i^{\otimes 2p+2}), \Phi_p(\mathbf{F}_p(\mathbf{u})) \otimes \Phi_p(\mathbf{F}_p(\bar{\mathbf{u}})) \right\rangle \\
&= \left\langle \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(\Phi_p \otimes \Phi_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d) \left(\frac{1}{\widehat{\sigma}_{2p+2}} y\mathbf{He}_{2p+2}(\mathbf{x}) \right) \right], \Phi_p(\mathbf{F}_p(\mathbf{u})) \otimes \Phi_p(\mathbf{F}_p(\bar{\mathbf{u}})) \right\rangle. \qquad (19)
\end{aligned}$$

Notably, the right-hand side of the equality above is a product of a $k \times k \times d \times d$ tensor and a $k \times k$ matrix, rather than two higher-order tensors. This implies that the dimension of the estimation problem is greatly reduced from d^{2p} to $k^2 + (dk)^2$, via the "compressor" Φ_p. The main idea of [DK25] is that, when $\sum_{i=1}^{k} \mathbf{w}_i^{\otimes t} = \mathbf{E}_{\mathcal{I}}[\mathbf{S}_t(\mathcal{I})]$ for $t = 2, 4, \ldots, 2p$, where $\mathbf{S}_t(\mathcal{I})$ are recursively-constructed random tensors depending on the random variable $\mathcal{I}$, the compressor Φ_p can be discovered and constructed recursively and efficiently from samples $\mathcal{I}$. Furthermore, Φ_p can also be efficiently applied to other recursively-constructed tensors.

These recursively-constructed tensors are formally defined as tensors generated from *Sequential Tensor Computation (STC) circuits.*

Definition 4.16 (Sequential Tensor Computation [DK25])**.** *A sequential tensor computation (STC) is an arithmetic circuit S_t that takes as input $\mathcal{I}$ a set of scalars and vectors in $\mathbb{C}^d$ and outputs an order-t tensor of dimension d, denoted by $\mathbf{S}_t(\mathcal{I}) \in (\mathbb{C}^d)^{\otimes t}$, for some $t \in \mathbb{Z}_+$, by applying the following operations: scalar multiplication, addition of objects of the same order and dimension, and appending a vector by tensor product (in the last component). When $\mathcal{I}$ is sampled from a distribution $\mathcal{D}(\mathcal{I})$, we denote $\mathbf{S}_t(\mathcal{I}) \sim \mathsf{S}_t(\mathcal{D}(\mathcal{I}))$. The size of S_t is defined as the size of the underlying arithmetic circuit.*

In short, Sequential Tensor Computation (STC) circuits capture a class of tensors that can be constructed via recursive simple operations. One simple example is order-p rank-1 tensors, i.e., tensors of the form $\mathbf{u}^{\otimes p}$, since these tensors can be constructed by recursively adding a new vector in its last component via tensor product. These tensors are exactly the objectives we consider.

Lemma 4.17 (Moment Tensor and Filter Tensor are STCs)**.** *Let $\mathcal{D}$ be a Gaussian sum of k ReLUs, and let $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Let $\mathcal{I} = (\mathbf{x}, y, \mathbf{z})$ and denote the distribution of $\mathcal{I}$ by $\mathcal{D}(\mathcal{I}) = \mathcal{D} \otimes \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. For even $t \in \mathbb{Z}_+$, denote the tth Hermite coefficient of the ReLU activation σ by $\widehat{\sigma}_t$. Then, for any even $t \in \mathbb{Z}_+$, there is a real STC $\mathsf{S}_t(\mathcal{D}(\mathcal{I}))$ of size $O(t)$ that given a vector $\mathcal{I} \sim \mathcal{D}(\mathcal{I})$, outputs:*

$$\mathbf{S}_t(\mathcal{I}) = \mathbf{S}_t(\mathbf{x}, y, \mathbf{z}) \coloneqq \frac{y}{\widehat{\sigma}_t \sqrt{t!}} \mathrm{Re}((\mathbf{x} + \mathrm{i}\mathbf{z})^{\otimes t}). \tag{20}$$

The tensor $\mathbf{S}_t(\mathbf{x}, y, \mathbf{z})$ satisfies

$$\mathop{\mathbf{E}}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)} [\mathbf{S}_t(\mathbf{x}, y, \mathbf{z})] = \frac{1}{\widehat{\sigma}_t} y \mathbf{He}_t(\mathbf{x}), \qquad \mathop{\mathbf{E}}_{\mathcal{I} \sim \mathcal{D}(\mathcal{I})} [\mathbf{S}_t(\mathcal{I})] = \sum_{i=1} \beta_i \mathbf{w}_i^{\otimes t} .$$

Furthermore, let $\mathbf{u} \sim \mathcal{U}, \mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$, $\mathbf{p}, \mathbf{q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Then, for any $p \in \mathbb{Z}_+$, there is a complex STC $\mathsf{F}_p(\mathcal{U})$ of size $O(p)$ that given a vector $\mathbf{u}$, outputs:

$$\mathbf{F}_p(\mathbf{u}) = \frac{\mathbf{u}^{\otimes p}}{\sqrt{p!}} . \tag{21}$$

Proof. We first show that $\mathsf{S}_t(\mathcal{D}(\mathcal{I}))$ is a real STC of size $O(t)$. Let $a = y/(\widehat{\sigma}_t \sqrt{t!})$. Define tensors $\mathbf{R}_j, \mathbf{Im}_j$ such that $\mathbf{R}_j + \mathrm{i}\mathbf{Im}_j = a(\mathbf{x} + \mathrm{i}\mathbf{z})^{\otimes j}$. To construct $j+1$, note that

$$a(\mathbf{x} + \mathrm{i}\mathbf{z})^{\otimes (j+1)} = (\mathbf{R}_j + \mathrm{i}\mathbf{Im}_j) \otimes (\mathbf{x} + \mathrm{i}\mathbf{z}) = (\mathbf{R}_j \otimes \mathbf{x} - \mathbf{Im}_j \otimes \mathbf{z}) + \mathrm{i}(\mathbf{R}_j \otimes \mathbf{z} + \mathbf{Im}_j \otimes \mathbf{x}) .$$

This gives the recursive construction: $\mathbf{R}_{j+1} = \mathbf{R}_j \otimes \mathbf{x} - \mathbf{Im}_j \otimes \mathbf{z}$ and $\mathbf{Im}_{j+1} = \mathbf{R}_j \otimes \mathbf{z} + \mathbf{Im}_j \otimes \mathbf{x}$ with $\mathbf{R}_0 = a$ and $\mathbf{Im}_0 = 0$. Each step uses an STC circuit of size $O(1)$, therefore the real STC $\mathsf{S}_t(\mathcal{D}(\mathcal{I}))$ is of size $O(t)$.

Next, applying Fact 2.6, we have

$$\mathop{\mathbf{E}}_{\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_d)} [\mathbf{S}_t(\mathbf{x}, y, \mathbf{z})] = \frac{y}{\widehat{\sigma}_t} \mathop{\mathbf{E}}_{\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\frac{\mathrm{Re}((\mathbf{x} + \mathrm{i}\mathbf{z})^{\otimes t})}{\sqrt{t!}} \right] = (1/\widehat{\sigma}_t) y \mathbf{He}_t(\mathbf{x}) .$$

Therefore, by the law of total expectation, we have $\mathbf{E}_{\mathcal{I} \sim \mathcal{D}(\mathcal{I})}[\mathbf{S}_t(\mathcal{I})] = (1/\widehat{\sigma}_t) \mathbf{E}_{(\mathbf{x}, y) \sim \mathcal{D}}[y \mathbf{He}_t(\mathbf{x})] = \sum_{i=1}^k \beta_i \mathbf{w}_i^{\otimes t}$, using the definition of Hermite expansions.

For $\mathbf{F}_p(\mathbf{u})$, directly by the definition of STCs, we know that it can be generated by an STC $\mathsf{F}_p(\mathcal{U})$ of size $O(p)$. This completes the proof of Lemma 4.17. □

Combining Claim 4.15 and Lemma 4.17, and using (19), the task of estimating the matrix $\mathbf{M}_{2p+2}(\mathbf{u})$ is reduced to estimating the tensor contraction of these two STCs mentioned above:

$$\begin{aligned} \mathbf{M}_{2p+2}(\mathbf{u}) &= \left\langle \mathop{\mathbf{E}}_{\mathcal{I} \sim \mathcal{D}(\mathcal{I})} [\mathbf{S}_{2p+2}(\mathcal{I})], \mathbf{F}_p(\mathbf{u}) \otimes \mathbf{F}_p(\bar{\mathbf{u}}) \right\rangle \\ &= \left\langle \mathop{\mathbf{E}}_{\mathcal{I} \sim \mathcal{D}(\mathcal{I})} \left[(\Phi_p \otimes \Phi_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d) \Big(\mathbf{S}_{2p+2}(\mathcal{I}) \Big) \right], \Phi_p(\mathbf{F}_p(\mathbf{u})) \otimes \Phi_p(\mathbf{F}_p(\bar{\mathbf{u}})) \right\rangle, \end{aligned} \tag{22}$$

where Φ_p is the "compressor" we are looking for that maps from $(\mathbb{R}^d)^{\otimes p}$ to the subspace spanned by the hidden weight tensors $\mathbf{w}_i^{\otimes p}$. These compressors are formally called (recursive) pseudo-projections:

Definition 4.18 (Recursive pseudo-projections; Definitions 3.4 and 3.6 in [DK25])**.** *A real linear map $\Phi : \mathbb{U} \to \mathbb{V}$ between finite-dimensional inner product spaces $\mathbb{U}$ and $\mathbb{V}$ is a* pseudo-projection *if $\Phi\Phi^\top = \mathbf{I}_{\mathbb{V}}$. A (d, t, k)-recursive pseudo-projection Φ has dimensions $n_0 = 1, n_1, \ldots, n_t = k$ and base pseudo-projections $\phi_j : \mathbb{R}^{n_{j-1}} \otimes \mathbb{R}^d \to \mathbb{R}^{n_j}$ such that*

$$\Phi_0 = \mathbf{I}_d, \qquad \Phi_j = \phi_j(\Phi_{j-1} \otimes \mathbf{I}_d), \quad 1 \le j \le t, \qquad \Phi = \Phi_t.$$

Its order is t and its size is k.

Remark 4.19 (Decomposition of Pseudo-Projections)**.** As proved in [DK25, Lemma 3.5], every pseudo-projection Φ can be written as a projection $\mathbf{P} : \mathbb{U} \to \mathbb{U}'$ (i.e., $\mathbf{P}$ is a linear map satisfying $\mathbf{P}\mathbf{P}^\top = \mathbf{I}$) composed with an isometry $\mathbf{R}$ from $\mathbb{U}'$ to $\mathbb{V}$; and in addition, the composition of pseudo-projections is a pseudo-projection. This implies the recursive map defined above is indeed a pseudo-projection, as tensorization with an identity and composition preserves the equation $\Phi_j \Phi_j^\top = \mathbf{I}_{n_j}$.

Remark 4.20 (Complex Extension of the Pseudo-Projections)**.** The mapping Φ is a linear map with real matrix entries. We also use the same notation for its complex-linear extension, defined by

$$\Phi(\mathbf{W} + \mathrm{i}\mathbf{T}) = \Phi(\mathbf{W}) + \mathrm{i}\Phi(\mathbf{T}), \qquad \mathbf{W}, \mathbf{T} \text{ real tensors in real space } \mathbb{U}.$$

Thus, Φ can be applied to complex tensors. Since the matrix entries of Φ are real, its Hermitian adjoint coincides with its transpose $\Phi^* = \Phi^\top$, and $\Phi\Phi^* = \mathbf{I}_{\mathbb{V}}, \Phi(\overline{\mathbf{W} + \mathrm{i}\mathbf{T}}) = \overline{\Phi(\mathbf{W} + \mathrm{i}\mathbf{T})}$. In the rest of the section, whenever we apply this real linear map Φ to a complex tensor, we are applying its natural complex extension.

Fact 4.21. *Let $\Phi : \mathbb{U} \to \mathbb{V}$ be a real pseudo-projection. The following all hold:*

1. *For any $\mathbf{W} \in \mathbb{V}$, $\|\Phi^\top \mathbf{W}\|_F = \|\mathbf{W}\|_F$.*
2. *For every $\mathbf{T} \in \mathbb{U}$, $\|\Phi\mathbf{T}\|_F = \|\Phi^\top \Phi\mathbf{T}\|_F \le \|\mathbf{T}\|_F$.*
3. *For any $\mathbf{V} \in \mathbb{V}, \mathbf{T} \in \mathbb{U}$, $\langle \Phi\mathbf{T}, \mathbf{V}\rangle = \langle \mathbf{T}, \mathbf{\Phi}^\top \mathbf{V}\rangle$.*

Proof. For Item 1, we have that

$$\|\Phi^\top \mathbf{W}\|_F^2 = \langle \Phi^\top \mathbf{W}, \Phi^\top \mathbf{W}\rangle = \langle \mathbf{W}, \Phi\Phi^\top \mathbf{W}\rangle = \langle \mathbf{W}, \mathbf{W}\rangle = \|\mathbf{W}\|_F^2 .$$

For Item 2, the first part follows by the same argument as Item 1. The second part follows from the fact that every pseudo-projection Φ can be written as a projection $\mathbf{P} : \mathbb{U} \to \mathbb{U}'$ (i.e., $\mathbf{P}$ is a linear map satisfying $\mathbf{P}\mathbf{P}^\top = \mathbf{I}$) composed with an isometry from $\mathbb{U}'$ to $\mathbb{V}$. □

To obtain the pseudo-projection Φ_p, we apply the results from [DK25]. However, the output of the main algorithm in [DK25] is a real-valued inner product between two projected tensors. In our application, we aim to utilize the recursive pseudo-projections explicitly. Due to this small difference, we slightly adapt the main algorithm from [DK25] to our setting, as stated in $\mathsf{TensorCompressor}(p, N_0, \mathcal{D}(\mathcal{I}))$.

Algorithm 8 $\mathsf{TensorCompressor}(p, N_0, \mathcal{D}(\mathcal{I}))$

Input: Parameters $k, d, p, N_0 \in \mathbb{Z}_+$; sample access to $\mathcal{D}(\mathcal{I})$
Output: A recursive pseudo-projection $\widehat{\Phi}_p$, represented by $\phi_1, \ldots, \phi_p$.
Construct STC circuit S_t, $t = 2, 4, \ldots, 2p + 2$, according to Lemma 4.17.
$\widehat{\Phi}_0 \leftarrow \mathbf{I}_d$.
for $r = 1, \ldots, p$ **do**
  $\Psi_r \leftarrow \widehat{\Phi}_{r-1} \otimes I_d$. ▷ $\widehat{\Phi}_{r-1} : (\mathbb{R}^d)^{\otimes(r-1)} \to \mathbb{R}^k$, $\Psi_r : (\mathbb{R}^d)^{\otimes(r-1)} \otimes \mathbb{R}^d \to \mathbb{R}^k \otimes \mathbb{R}^d$
  Draw N_0 fresh samples from $\mathcal{D}(\mathcal{I})$, construct empirical distribution $\widehat{\mathcal{D}}(\mathcal{I})$.
  Let $\mathbf{S}_{2r}(\mathcal{I})$ be the output of S_{2r} on sample $\mathcal{I}$.
  $\mathbf{V}_r \leftarrow \mathbf{E}_{\mathcal{I} \sim \widehat{\mathcal{D}}(\mathcal{I})}[(\Psi_r \otimes \Psi_r)\mathbf{S}_{2r}(\mathcal{I})]$ ▷ Viewed as a $d \times d$ matrix for $r = 1$ and a $dk \times dk$ matrix otherwise.
  $\mathbf{U}_r \leftarrow \mathsf{Top(k)Eigenvectors}[\mathbf{V}_r]$
  $\phi_r \leftarrow \mathbf{U}_r^\top$; $\widehat{\Phi}_r \leftarrow \phi_r \Psi_r$.
return $\widehat{\Phi}_p$.

$\mathsf{TensorCompressor}(p, N_0, \mathcal{D}(\mathcal{I}))$ has the following guarantee:

Proposition 4.22 (Implicit Tensor Computation)**.** *Let $\mathcal{D}$ be a Gaussian sum of k ReLUs and denote $\mathbf{M}_p = \sum_{i=1}^k \beta_i \mathbf{w}_i^{\otimes p}$. Let $\mathcal{I} = (\mathbf{x}, y, \mathbf{z})$ where $(\mathbf{x}, y) \sim \mathcal{D}$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and let $\mathsf{S}_t(\mathcal{D}(\mathcal{I}))$ be the STC defined in Lemma 4.17, Equation (20). Let $p \in \mathbb{Z}_+$. Suppose that for all positive even integers $t \le 2p$, the output tensor $\mathbf{S}_t(\mathcal{I})$, $\mathcal{I} \sim \mathcal{D}_t(\mathcal{I})$, has covariance bounded above by V_{S}, for some $V_{\mathsf{S}} > 0$.*

Then $\mathsf{TensorCompressor}(p, N_0, \mathcal{D}(\mathcal{I}))$ *has the following guarantee: the algorithm draws* pN_0 *samples from* $\mathcal{D}(\mathcal{I})$*, runs in* $\mathrm{poly}(d, p, N_0)$ *time, outputs a* (d, p, k)*-recursive pseudo-projection* $\widehat{\Phi}_p$ *from* $(\mathbb{R}^d)^{\otimes p}$ *to* $\mathbb{R}^k$*, such that with probability at least* $1 - \tau$ *(over the samples drawn from* $\mathcal{D}(\mathcal{I})$*),* $\widehat{\Phi}_p$ *satisfies*

$$\|\widehat{\Phi}_p^\top \widehat{\Phi}_p(\mathbf{M}_p) - \mathbf{M}_p\|_F \le kp\sqrt{\zeta}, \text{ and for each } i \in [k],\ \|\widehat{\Phi}_p^\top \widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) - \mathbf{w}_i^{\otimes p}\|_F \le p\sqrt{\zeta/\beta_i}\ ,$$

where $\zeta = O(pdkV_{\mathsf{S}}/\sqrt{N_0\tau})$.

Proof of Proposition 4.22. Recall we proved that $\mathbf{E}_{\mathcal{I}\sim\mathcal{D}(\mathcal{I})}[\mathbf{S}_t(\mathcal{I})] = \sum_{i=1} \beta_i \mathbf{w}_i^{\otimes t}$ in Lemma 4.17, which satisfies the prerequisites of Proposition 3.2 in [DK25]. Then the proof follows directly from [DK25, Lemma 3.9 and Claim 3.10]. □

Proposition 4.22 states that, using $\mathrm{poly}(d, k, p, 1/\tau, V_{\mathsf{S}})$ samples, we can approximately recover the compressor Φ_p we are looking for: the recursive pseudo-projection $\widehat{\Phi}_p$ will almost project to the subspace spanned by the hidden weight tensors $\mathbf{w}_i^{\otimes p}$ whose coefficients are not too small.

Applying Implicit Tensor Computation for Target Matrix Approximation. To apply $\mathsf{TensorCompressor}(p, N_0, \mathcal{D}(\mathcal{I}))$ and Proposition 4.22, we need to determine the covariance of the STC output $\mathbf{S}_t(\mathcal{I})$ defined in Lemma 4.17. To do so, we establish the following key lemma:

Lemma 4.23. *Let* $\mathcal{D}$ *be a Gaussian sum of* k *ReLUs, and let* $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$*. Let* $\mathcal{I} = (\mathbf{x}, y, \mathbf{z})$ *and denote the distribution of* $\mathcal{I}$ *by* $\mathcal{D}(\mathcal{I}) = \mathcal{D} \otimes \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$*. Let* $\mathbf{S}_t(\mathcal{I})$ *be the output of the STC circuit defined in Lemma 4.17, Equation (20). Then the covariance of* $\mathbf{S}_t(\mathcal{I})$ *is bounded by*

$$\sup_{\|\mathbf{W}\|_F=1} \mathop{\mathbf{E}}_{\mathcal{I}\sim\mathcal{D}(\mathcal{I})}[|\langle \mathbf{S}_t(\mathcal{I}) - \mathop{\mathbf{E}}_{\mathcal{I}\sim\mathcal{D}}[\mathbf{S}_t(\mathcal{I})], \mathbf{W}\rangle|^2] \le V_{\mathsf{S}_t} := k^2 2^{O(t)}/\widehat{\sigma}_t^2\ .$$

Proof. By the definition of tensor covariance and the distribution of the labels y, we have:

$$\begin{aligned}
&\sup_{\|\mathbf{W}\|_F=1} \mathop{\mathbf{E}}_{(\mathbf{x},y,\mathbf{z})\sim\mathcal{D}(\mathcal{I})} \left[\left|\left\langle \frac{y}{\widehat{\sigma}_t\sqrt{t!}}\mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathbf{W}\right\rangle\right|^2\right] \\
&\le \sup_{\|\mathbf{W}\|_F=1} \frac{1}{\widehat{\sigma}_t^2 t!}\sqrt{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y^4] \mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)} \left[\left|\left\langle \mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathbf{W}\right\rangle\right|^4\right]}\ ,
\end{aligned} \tag{23}$$

where the last inequality is by Cauchy-Schwarz. We prove the following claim:

Claim 4.24. *For any* $\mathbf{W}$ *such that* $\|\mathbf{W}\|_F = 1$*, it holds that*

$$\sqrt{\mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)} \left[\left|\left\langle \mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathbf{W}\right\rangle\right|^4\right]} \le 2^{O(t)} t!.$$

Proof. Note that $|\langle(\mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathbf{W}\rangle|^2 = (\langle\mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathrm{Re}(\mathbf{W})\rangle)^2 + (\langle\mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathrm{Im}(\mathbf{W})\rangle)^2$, where $a(\mathbf{x}, \mathbf{z}) = \langle\mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathrm{Re}(\mathbf{W})\rangle$ and $b(\mathbf{x}, \mathbf{z}) = \langle\mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathrm{Im}(\mathbf{W})\rangle$ are both degree-t polynomials. Therefore, by Gaussian Hypercontractivity (Fact 2.2), we have that

$$\|a(\mathbf{x}, \mathbf{z})^2\|_{L_2(\mathcal{N})}^{1/2} = \left(\mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[a(\mathbf{x}, \mathbf{z})^4]\right)^{1/4} \le 3^{t/2}\left(\mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[a(\mathbf{x}, \mathbf{z})^2]\right)^{1/2} = 3^{t/2}\|a(\mathbf{x}, \mathbf{z})\|_{L_2(\mathcal{N})}\ ,$$

and a similar inequality holds for $b(\mathbf{x}, \mathbf{z})$. Therefore, we obtain:

$$\begin{aligned}
\left(\mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)} \left[\left|\left\langle \mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathbf{W}\right\rangle\right|^4\right]\right)^{1/2} &= \left(\mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)} \left[\left(a(\mathbf{x}, \mathbf{z})^2 + b(\mathbf{x}, \mathbf{z})^2\right)^2\right]\right)^{1/2} \\
&= \|a(\mathbf{x}, \mathbf{z})^2 + b(\mathbf{x}, \mathbf{z})^2\|_{L_2(\mathcal{N})} \\
&\le \|a(\mathbf{x}, \mathbf{z})^2\|_{L_2(\mathcal{N})} + \|b(\mathbf{x}, \mathbf{z})^2\|_{L_2(\mathcal{N})} \\
&\le 3^t(\|a(\mathbf{x}, \mathbf{z})\|_{L_2(\mathcal{N})}^2 + \|b(\mathbf{x}, \mathbf{z})\|_{L_2(\mathcal{N})}^2)\ .
\end{aligned}$$

By [DK25, Lemma 4.3], we have that

$$\|a(\mathbf{x},\mathbf{z})\|_{L_2(\mathcal{N})}^2 = \mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left\langle \mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathrm{Re}(\mathbf{W})\right\rangle^2\right]$$
$$\leq \sup_{\|\widetilde{\mathbf{W}}\|_F=1} \mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left|\left\langle \mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \widetilde{\mathbf{W}}\right\rangle\right|^2\right] \leq t!2^{O(t)} .$$

The same inequality also holds for $b(\mathbf{x},\mathbf{z})$. Thus we conclude that:

$$\left(\mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left|\left\langle \mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathbf{W}\right\rangle\right|^4\right]\right)^{1/2} \leq 3^t \mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left|\left\langle \mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathbf{W}\right\rangle\right|^2\right]$$
$$\leq t!\,3^t 2^{O(t)} = t!\,2^{O(t)} .$$

□

Therefore, by plugging $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[(\sum_{i=1}^k \beta_i\sigma(\mathbf{w}_i\cdot\mathbf{x}))^4] \lesssim k^4$ back into (23), we obtain:

$$\sup_{\|\mathbf{W}\|_F=1} \mathop{\mathbf{E}}_{(\mathbf{x},y,\mathbf{z})\sim\mathcal{D}(\mathcal{I})}\left[\left|\left\langle \frac{y}{c_t\sqrt{t!}}\mathrm{Re}((\mathbf{x}+\mathrm{i}\mathbf{z})^{\otimes t}), \mathbf{W}\right\rangle\right|^2\right] \lesssim \frac{k^2 2^{O(t)} t!}{\widehat{\sigma}_t^2 t!} = \frac{k^2 2^{O(t)}}{\widehat{\sigma}_t^2} .$$

This completes the proof of Lemma 4.23. □

Combining Lemma 4.23 with Proposition 4.22 immediately implies that when $p = O(\log(k/\lambda))$ (as was set in CDRviaImplicitTensor$(\mathcal{D},\lambda,\delta)$), using $N_0 = \mathrm{poly}(d,k,1/\tau)2^{O(p)} = \mathrm{poly}(d,k/\lambda,1/\tau)$ samples from $\mathcal{D}$, the pseudo-projection $\widehat{\Phi}_p$ maps a tensor $O(\lambda)$-close to the space of the λ-heavy hidden weight tensors $\mathbf{w}_i^{\otimes p}$ (i.e., $\beta_i \geq \lambda$) with probability at least $1-\tau$. The algorithm TensorCompressor$(p, N_0, \mathcal{D}(\mathcal{I}))$ runs efficiently, in time $\mathrm{poly}(d,p,N_0)$. In short, we can efficiently approximate the compressor Φ_p using $\widehat{\Phi}_p$ to reduce the problem dimension. Recalling from (22) that the target matrix equals

$$\mathbf{M}_{2p+2}(\mathbf{u}) = \left\langle \mathop{\mathbf{E}}_{\mathcal{I}\sim\mathcal{D}(\mathcal{I})}\left[(\Phi_p\otimes\Phi_p\otimes\mathbf{I}_d\otimes\mathbf{I}_d)\left(\mathbf{S}_{2p+2}(\mathcal{I})\right)\right], \Phi_p(\mathbf{F}_p(\mathbf{u}))\otimes\Phi_p(\mathbf{F}_p(\bar{\mathbf{u}}))\right\rangle ,$$

an estimation of $\mathbf{M}_{2p+2}(\mathbf{u})$ can therefore be constructed via:

$$\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) = \left\langle \mathop{\mathbf{E}}_{\mathcal{I}\sim\widehat{\mathcal{D}}(\mathcal{I})}\left[(\widehat{\Phi}_p\otimes\widehat{\Phi}_p\otimes\mathbf{I}_d\otimes\mathbf{I}_d)\left(\mathbf{S}_{2p+2}(\mathcal{I})\right)\right], \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))\otimes\widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}}))\right\rangle, \tag{EstM}$$

where $\widehat{\mathcal{D}}(\mathcal{I})$ is a fresh empirical distribution of $\mathcal{D}(\mathcal{I})$ (independent of the samples used to construct $\widehat{\Phi}_p$). The two tensors on the right-hand side of (EstM) are computed in the following procedures: CompressEmpMoment$(p, N_0, \widehat{\Phi}_p, \mathcal{D}(\mathcal{I}))$ and CompressFilter$(p, \widehat{\Phi}_p, \mathbf{u})$. In CDRviaImplicitTensor$(\mathcal{D},\lambda,\delta)$, Line 15, we compute the empirical matrix $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})$ displayed in (EstM), using the tensors returned by CompressEmpMoment$(p, N_0, \widehat{\Phi}_p, \mathcal{D}(\mathcal{I}))$ and CompressFilter$(p, \widehat{\Phi}_p, \mathbf{u})$.

Algorithm 9 CompressEmpMoment$(p, N_0, \widehat{\Phi}_p, \mathcal{D}(\mathcal{I}))$

Input: Moment degree p, number of samples N_0, tensor compressor $\widehat{\Phi}_p$, sample access to $\mathcal{D}(\mathcal{I})$.
Output: A tensor $\mathbf{A}_p \in \mathbb{R}^{k\times k\times d\times d}$

Draw N_0 fresh samples $\widehat{\mathcal{D}}(\mathcal{I}) = \{\mathcal{I}^{(i)}\}_{i=1}^{N_0}$ from $\mathcal{D}(\mathcal{I})$.
Construct STC circuit S_{2p+2}, according to Lemma 4.17.
Compute $(\widehat{\Phi}_p\otimes\widehat{\Phi}_p\otimes\mathbf{I}_d\otimes\mathbf{I}_d)\mathbf{S}_{2p+2}(\mathcal{I}^{(i)})$, where $\mathbf{S}_{2p+2}(\mathcal{I}^{(i)})$ is the output of S_{2p+2}, using Lemma 4.25.
$\mathbf{A}_p \leftarrow \frac{1}{N_0}\sum_{i=1}^{N_0}(\widehat{\Phi}_p\otimes\widehat{\Phi}_p\otimes\mathbf{I}_d\otimes\mathbf{I}_d)\mathbf{S}_{2p+2}(\mathcal{I}^{(i)})$.
Return: $\mathbf{A}_p$.

Algorithm 10 $\mathsf{CompressFilter}(p, \widehat{\Phi}_p, \mathbf{u})$

1: **Input:** Moment degree p, tensor compressor $\widehat{\Phi}_p$, probing vector $\mathbf{u}$.
2: **Output:** A matrix $\mathbf{B}_p(\mathbf{u}) \in \mathbb{R}^{k \times k}$
3: Construct STC circuits F_p according to Lemma 4.17.
4: Compute $\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))$ and $\widehat{\Phi}_p(\mathbf{F}_p(\overline{\mathbf{u}}))$, where $\mathbf{F}_p(\mathbf{u})$ is the output of F_p, using Lemma 4.25.
5: $\mathbf{B}_p(\mathbf{u}) \leftarrow \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\overline{\mathbf{u}}))$
6: **Return:** $\mathbf{B}_p(\mathbf{u})$.

We next show that these procedures run efficiently.

Lemma 4.25 (Efficient Computation of Tensor Projections)**.** *Let $\mathsf{S}_t(\mathcal{D}(\mathcal{I}))$ and $\mathsf{F}_p(\mathcal{U})$ be the STCs defined in Lemma 4.17, equations (20) and (21) respectively. Let Φ_p be a (d, p, k)-recursive pseudo-projection. Then, for any $\mathbf{F}_p(\mathbf{u}) \sim \mathsf{F}_p(\mathcal{U})$, $\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))$ can be computed in time $\mathrm{poly}(d, k, p)$; and for any $\mathbf{S}_{2p+2}(\mathcal{I}) \sim \mathsf{S}_{2p+2}(\mathcal{D}(\mathcal{I}))$, the tensor $(\widehat{\Phi}_p \otimes \widehat{\Phi}_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d)(\mathbf{S}_{2p+2}(\mathcal{I}))$ can be computed in time $\mathrm{poly}(d, k, p)$.*

Proof. [DK25, Lemma 3.7] states that given an order-t recursive pseudo-projection Ψ along with their defining pseudo-projections θ_j, and an STC $\mathbf{S}$ of order t, dimension d, and size S, the projection $\Psi(\mathbf{S})$ can be computed in time $\mathrm{poly}(d, t, S)$. This immediately implies that $\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))$ can be computed in $\mathrm{poly}(d, p)$ time. Furthermore, according to [DK25, Lemma 3.8], given recursive pseudo-projections $\Phi : (\mathbb{R}^d)^{\otimes t} \to \mathbb{R}^n$ and $\Psi : (\mathbb{R}^d)^{\otimes s} \to \mathbb{R}^m$ along with their defining pseudo-projections ϕ_i, θ_j, one can efficiently construct the recursive pseudo-projection $\Phi \otimes \Psi : (\mathbb{R}^d)^{\otimes (t+s)} \to \mathbb{R}^{nm}$ of size $\max(\mathrm{size}(\Phi), n\,\mathrm{size}(\Psi))$ in time $\mathrm{poly}(d, t, s, \max(\mathrm{size}(\Phi), n\,\mathrm{size}(\Psi)))$. Therefore, the recursive-pseudo projection $\widehat{\Phi}_p \otimes \widehat{\Phi}_p : (\mathbb{R}^d)^{\otimes p} \otimes (\mathbb{R}^d)^{\otimes p} \to \mathbb{R}^k \otimes \mathbb{R}^k$ can be efficiently constructed. Since $\mathbf{I}_d : \mathbb{R}^d \to \mathbb{R}^d$ is also a pseudo-projection, we can similarly efficiently construct the recursive pseudo-projection $(\widehat{\Phi}_p \otimes \widehat{\Phi}_p) \otimes \mathbf{I}_d$, and then construct $((\widehat{\Phi}_p \otimes \widehat{\Phi}_p) \otimes \mathbf{I}_d) \otimes \mathbf{I}_d$. The process takes time $\mathrm{poly}(d, k, p)$. After constructing the pseudo-projection $\widehat{\Phi}_p \otimes \widehat{\Phi}_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d$, since $\mathbf{S}_{2p+2}(\mathcal{I})$ is generated from a size $O(p)$ STC, the projection $(\widehat{\Phi}_p \otimes \widehat{\Phi}_t \otimes \mathbf{I}_d \otimes \mathbf{I}_d)\mathbf{S}_{2p+2}(\mathcal{I})$ can also be computed in $\mathrm{poly}(d, k, p)$ time. □

Main Result of this Subsection: Matrix Estimation Using Implicit Tensor Computation. What remains is to prove that the empirical matrix estimation in (EstM) is close to the target matrix $\mathbf{M}_{2p+2}(\mathbf{u})$. However, we need one last technical piece, which is to bound the covariance of the tensor $\mathbf{F}_p(\mathbf{u})$. This is the place where we crucially use the property of complex Gaussian tensors: in Lemma 4.26, we show that the covariance of the complex Gaussian filter $\mathbf{F}_p(\mathbf{u})$ is bounded above by $2^{O(p)}$, whereas in Lemma 4.27 we prove that were we using a real Gaussian filter, the covariance would be much larger.

To see why a bound on the covariance of $\mathbf{F}_p(\mathbf{u})$ is required, consider the difference between the empirical matrix and the target matrix:

$$\begin{aligned}\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}(\mathbf{u}) &= \left\langle \mathop{\mathbf{E}}_{\mathcal{I} \sim \widehat{\mathcal{D}}(\mathcal{I})} \left[(\widehat{\Phi}_p \otimes \widehat{\Phi}_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d) \left(\mathbf{S}_{2p+2}(\mathcal{I}) \right) \right], \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}})) \right\rangle \\ &\quad - \left\langle \mathop{\mathbf{E}}_{\mathcal{I} \sim \mathcal{D}(\mathcal{I})} \left[\mathbf{S}_{2p+2}(\mathcal{I}) \right], \mathbf{F}_p(\mathbf{u}) \otimes \mathbf{F}_p(\bar{\mathbf{u}}) \right\rangle \\ &= \langle \Delta_1, \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}})) \rangle + \langle \Delta_2, \mathbf{F}_p(\mathbf{u}) \rangle + \langle \Delta_2, \mathbf{F}_p(\bar{\mathbf{u}}) \rangle \,,\end{aligned}$$

where Δ_1, Δ_2, and Δ_3 are error terms (see, e.g., (25) in the proof of Proposition 4.14). Even though Δ_1, Δ_2, and Δ_3 can be tiny, if the spectral norm of $\mathbf{F}_p(\mathbf{u})$ or $\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}}))$ is large, the difference displayed above is not necessarily small. To argue that $\mathbf{F}_p(\mathbf{u})$ and $\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}}))$ have small spectral norm with large probability, we bound their covariance and then apply standard concentration inequalities.

Lemma 4.26. *Let $\mathbf{u} \sim \mathcal{U}, \mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$, $\mathbf{p}, \mathbf{q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and let $\mathbf{F}_p(\mathbf{u})$ be the output of the STC $\mathsf{F}_p(\mathcal{U})$ defined in (21) in Lemma 4.17. Then, we have $\mathbf{E}_{\mathbf{u} \sim \mathcal{U}}[|\langle \mathbf{F}_p(\mathbf{u}), \mathbf{W} \rangle|^2] \le 2^{O(p)} \|\mathbf{W}\|_F^2$, $\mathbf{E}_{\mathbf{u} \sim \mathcal{U}}[|\langle \mathbf{F}_p(\mathbf{u}), \mathbf{W} \rangle|^4] \le 2^{O(p)} \|\mathbf{W}\|_F^4$. In addition, let $\widehat{\Phi}_p$ be the complex extension of a real-valued pseudo-projection from $(\mathbb{R}^d)^{\otimes p}$ to $\mathbb{R}^k$. Then:*

$$\mathop{\mathbf{E}}_{\mathbf{u} \sim \mathcal{U}}[\|\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}}))\|_F^2] \le k^2 2^{O(p)}.$$

Proof. Let $Z(\mathbf{u}) = \langle \mathbf{F}_p(\mathbf{u}), \mathbf{W}\rangle$. The random variables $Z(\mathbf{u})$ and $\mathrm{i}Z(\mathbf{u})$ have the same distribution; to see that, note that for any $\theta \in \mathbb{R}$, we have $\exp(\mathrm{i}\theta)\mathbf{u} = \cos\theta\mathbf{p} - \sin\theta\mathbf{q} + \mathrm{i}(\cos\theta\mathbf{q} + \sin\theta\mathbf{p})$ and both the real part and imaginary part are independent standard normal from orthogonality. Therefore, $Z(\exp(-\mathrm{i}\pi/(2p))\mathbf{u}) = \mathrm{i}Z(\mathbf{u})$ and because $Z(\exp(-\mathrm{i}\pi/(2p))\mathbf{u})$ has the same distribution as $Z(\mathbf{u})$, it holds that $Z(\mathbf{u})$ and $\mathrm{i}(Z(\mathbf{u}))$ have the same distribution. This further indicates that $\mathrm{Re}(Z(\mathbf{u}))$ and $\mathrm{Im}(Z(\mathbf{u}))$ have the same distribution. Similarly, $\mathrm{Re}(\mathbf{F}_p(\mathbf{u}))$ and $\mathrm{Im}(\mathbf{F}_p(\mathbf{u}))$ also have the same distribution. Taking imaginary parts, we conclude that $\mathrm{Re}(Z(\mathbf{u}))$ and $\mathrm{Im}(Z(\mathbf{u}))$ have the same distribution. The same rotation also shows that $\mathrm{Re}(\mathbf{F}_p(\mathbf{u}))$ and $\mathrm{Im}(\mathbf{F}_p(\mathbf{u}))$ have the same distribution. By [DK25, Lemma 4.3], for every real tensor $\mathbf{T}$ of order p, we have

$$\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}\left[\langle \mathrm{Re}(\mathbf{F}_p(\mathbf{u})), \mathbf{T}\rangle^2\right] = \frac{1}{p!}\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}\left[\langle \mathrm{Re}(\mathbf{u}^{\otimes p}), \mathbf{T}\rangle^2\right] \le 2^{O(p)}\|\mathbf{T}\|_F^2.$$

Since $\mathrm{Re}(\mathbf{F}_p(\mathbf{u}))$ and $\mathrm{Im}(\mathbf{F}_p(\mathbf{u}))$ have the same distribution, the same bound holds for $\mathrm{Im}(\mathbf{F}_p(\mathbf{u}))$.

Since the inner product is conjugate-linear in its first argument, we have

$$\mathrm{Re}(Z(\mathbf{u})) = \langle \mathrm{Re}(\mathbf{F}_p(\mathbf{u})), \mathrm{Re}(\mathbf{W})\rangle + \langle \mathrm{Im}(\mathbf{F}_p(\mathbf{u})), \mathrm{Im}(\mathbf{W})\rangle.$$

Using $(a+b)^2 \le 2a^2 + 2b^2$, it follows that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}\left[\mathrm{Re}(Z(\mathbf{u}))^2\right] &\le 2\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}\left[\langle \mathrm{Re}(\mathbf{F}_p(\mathbf{u})), \mathrm{Re}(\mathbf{W})\rangle^2\right] + 2\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}\left[\langle \mathrm{Im}(\mathbf{F}_p(\mathbf{u})), \mathrm{Im}(\mathbf{W})\rangle^2\right]\\ &\le 2^{O(p)}\left(\|\mathrm{Re}(\mathbf{W})\|_F^2 + \|\mathrm{Im}(\mathbf{W})\|_F^2\right) = 2^{O(p)}\|\mathbf{W}\|_F^2.\end{aligned}$$

The inequality also holds for $\mathrm{Im}(Z(\mathbf{u}))$. Therefore,

$$\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}\left[|\langle \mathbf{F}_p(\mathbf{u}), \mathbf{W}\rangle|^2\right] = \mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}\left[\mathrm{Re}(Z(\mathbf{u}))^2 + \mathrm{Im}(Z(\mathbf{u}))^2\right] = 2\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}\left[\mathrm{Re}(Z(\mathbf{u}))^2\right] \le 2^{O(p)}\|\mathbf{W}\|_F^2.$$

For the fourth moment, we apply Claim 4.24. By the definition of $\mathbf{F}_p(\mathbf{u}) = \mathbf{u}^{\otimes p}/\sqrt{p!}$, for every real tensor $\mathbf{T}$ of order p, we have

$$\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}\left[|\langle \mathrm{Re}(\mathbf{F}_p(\mathbf{u})), \mathbf{T}\rangle|^4\right] = \frac{1}{(p!)^2}\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}\left[|\langle \mathrm{Re}(\mathbf{u}^{\otimes p}), \mathbf{T}\rangle|^4\right] \le 2^{O(p)}\|\mathbf{T}\|_F^4.$$

Since $\mathrm{Re}(\mathbf{F}_p(\mathbf{u}))$ and $\mathrm{Im}(\mathbf{F}_p(\mathbf{u}))$ have the same distribution, the same bound holds with $\mathrm{Re}(\mathbf{F}_p(\mathbf{u}))$ replaced by $\mathrm{Im}(\mathbf{F}_p(\mathbf{u}))$. Recalling that $\mathrm{Re}(Z(\mathbf{u})) = \langle \mathrm{Re}(\mathbf{F}_p(\mathbf{u})), \mathrm{Re}(\mathbf{W})\rangle + \langle \mathrm{Im}(\mathbf{F}_p(\mathbf{u})), \mathrm{Im}(\mathbf{W})\rangle$, and using $(a+b)^4 \le 8(a^4+b^4)$, we obtain

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[\mathrm{Re}(Z(\mathbf{u}))^4] &\le 8\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[\langle \mathrm{Re}(\mathbf{F}_p(\mathbf{u})), \mathrm{Re}(\mathbf{W})\rangle^4] + 8\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[\langle \mathrm{Im}(\mathbf{F}_p(\mathbf{u})), \mathrm{Im}(\mathbf{W})\rangle^4]\\ &\le 2^{O(p)}\left(\|\mathrm{Re}(\mathbf{W})\|_F^4 + \|\mathrm{Im}(\mathbf{W})\|_F^4\right) \le 2^{O(p)}\|\mathbf{W}\|_F^4,\end{aligned}$$

where the last inequality uses

$$\|\mathrm{Re}(\mathbf{W})\|_F^4 + \|\mathrm{Im}(\mathbf{W})\|_F^4 \le \left(\|\mathrm{Re}(\mathbf{W})\|_F^2 + \|\mathrm{Im}(\mathbf{W})\|_F^2\right)^2 = \|\mathbf{W}\|_F^4.$$

Finally, using $(a^2+b^2)^2 \le 2a^4+2b^4$ and the fact that $\mathrm{Re}(Z(\mathbf{u}))$ and $\mathrm{Im}(Z(\mathbf{u}))$ have the same distribution, we conclude that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[|\langle \mathbf{F}_p(\mathbf{u}), \mathbf{W}\rangle|^4] &= \mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[(\mathrm{Re}(Z(\mathbf{u}))^2 + \mathrm{Im}(Z(\mathbf{u}))^2)^2] \le 2\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[\mathrm{Re}(Z(\mathbf{u}))^4 + \mathrm{Im}(Z(\mathbf{u}))^4]\\ &= 4\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[\mathrm{Re}(Z(\mathbf{u}))^4] \le 2^{O(p)}\|\mathbf{W}\|_F^4. \qquad (24)\end{aligned}$$

For the second part of the statement, note that since $\widehat{\Phi}_p$ is a map with real entries, we have $\widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}})) = \overline{\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))}$, therefore, it holds that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[\|\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}}))\|_F^2] &\le \mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[\|\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))\|_F^2\|\widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}}))\|_F^2]\\ &= \mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[\|\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))\|_F^2\|\overline{\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))}\|_F^2] = \mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{U}}[\|\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))\|_F^4].\end{aligned}$$

Recall that $\widehat{\Phi}_p$ maps $\mathbf{F}_p(\mathbf{u})$ to a vector in $\mathbb{C}^k$, thus we have $\|\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))\|_2^2 = \sum_{i=1}^k |\langle \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})), \mathbf{e}_i\rangle|^2$. Furthermore, by Jensen's inequality, it holds:

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{u}}[\|\widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))\|_2^4] = \mathop{\mathbf{E}}_{\mathbf{u}}\left[\left(\sum_{i=1}^k |\langle \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})), \mathbf{e}_i\rangle|^2\right)^2\right] &\le \mathop{\mathbf{E}}_{\mathbf{u}}\left[k\sum_{i=1}^k |\langle \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})), \mathbf{e}_i\rangle|^4\right] \\ &= k\sum_{i=1}^k \mathop{\mathbf{E}}_{\mathbf{u}}[|\langle \mathbf{F}_p(\mathbf{u}), \widehat{\Phi}_p^\top(\mathbf{e}_i)\rangle|^4] \le k^2 2^{O(p)}\ ,\end{aligned}$$

where for the last inequality we use that $\widehat{\Phi}_p\widehat{\Phi}_p^\top = \mathbf{I}_k$ (and therefore $\|\widehat{\Phi}_p^\top \mathbf{e}_i\|_F = 1$) and (24). □

For completeness, in the following lemma, we demonstrate the crucial difference between complex and real Gaussian tensors in terms of their covariance bounds, which justifies the use of complex Gaussians in our approach.

Lemma 4.27 (Real vs Complex Gaussian Tensor Powers)**.** *Let $d, p \ge 1$ be integers, let $\mathbf{p}, \mathbf{q} \sim \mathcal{N}(0, \mathbf{I}_d)$, and put $\mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$. Define*

$$\mathbf{F}_p^{\mathrm{R}}(\mathbf{p}) = \frac{\mathbf{p}^{\otimes p}}{\sqrt{(2p-1)!!}}, \qquad \mathbf{F}_p^{\mathrm{C}}(\mathbf{u}) = \frac{\mathbf{u}^{\otimes p}}{\sqrt{2^p p!}} = \mathbf{F}_p(\mathbf{u})/\sqrt{2^p}.$$

Both normalizations give, for every real unit vector $\mathbf{w}$,

$$\mathbf{E}\left[|\langle \mathbf{F}_p^{\mathrm{R}}(\mathbf{p}), \mathbf{w}^{\otimes p}\rangle|^2\right] = \mathbf{E}[|\langle \mathbf{F}_p^{\mathrm{C}}(\mathbf{u}), \mathbf{w}^{\otimes p}\rangle|^2] = 1.$$

Their uncentered second moments in general tensor directions differ as follows.

1. *For every complex order-p tensor $\mathbf{R}$,*

$$\mathbf{E}[|\langle \mathbf{F}_p(\mathbf{u}), \mathbf{R}\rangle|^2] = \|\mathrm{Sym}(\mathbf{R})\|_F^2 \le \|\mathbf{R}\|_F^2.$$

2. *For every even $p \ge 2$, the real tensor $\mathbf{R}_p = (\mathbf{I}_d/\sqrt{d})^{\otimes p/2}$ has $\|\mathbf{R}_p\|_F = 1$ and*

$$\mathbf{E}[\left|\langle \mathbf{F}_p^{\mathrm{R}}(\mathbf{p}), \mathbf{R}_p\rangle\right|^2] = \frac{\prod_{j=0}^{p-1}(d+2j)}{d^{p/2}(2p-1)!!} \ge \frac{d^{p/2}}{(2p)^p}.$$

Proof. Let p be even. Multiplicativity of the Frobenius norm under tensor products gives $\|\mathbf{R}_p\|_F = 1$, and direct contraction gives

$$\langle \mathbf{p}^{\otimes p}, \mathbf{R}_p\rangle = \frac{\|\mathbf{p}\|_2^p}{d^{p/4}}.$$

The radial Gaussian density is proportional to $r^{d-1}e^{-r^2/2}$. Integration by parts gives $\mathbf{E}\,\|\mathbf{p}\|_2^{2s} = (d+2s-2)\,\mathbf{E}\,\|\mathbf{p}\|_2^{2s-2}$ for $s \ge 1$, so $\mathbf{E}\,\|\mathbf{p}\|_2^{2p} = \prod_{j=0}^{p-1}(d+2j)$. Squaring the contraction and dividing by $(2p-1)!!$ proves the equality. Its lower bound follows from $d + 2j \ge d$ and $(2p-1)!! \le (2p)^p$. □

We are now ready to prove the main result of this subsection, Proposition 4.14.

Proposition 4.14 (Matrix Estimation Using Implicit Tensor Computation)**.** *Let $\mathcal{D}$ be a Gaussian sum of k ReLUs. Let $p, N_0, N_{\mathbf{u}} \in \mathbb{Z}_+$, $\xi, \lambda, \delta \in (0,1)$ be the parameters prescribed in* CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$*. Let $\mathbf{M}_{2p+2}(\mathbf{u})$ be the matrix defined in (M). Let $\widehat{\mathcal{U}}$ be a set of $N_{\mathbf{u}}$ random probing vectors following the distribution $\mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$, $\mathbf{p}, \mathbf{q} \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_d)$. Then the matrices $\widehat{\mathbf{M}}_{2p+2}(\mathbf{u})$ computed in Line 15 of* CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$ *satisfy the following guarantee: with probability at least $1 - \delta/2$ over all the samples and the probing vectors, $\max_{\mathbf{u}\in\widehat{\mathcal{U}}} \|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 \le \sqrt{\xi}\lambda$.*

Proof of Proposition 4.14. Let $\mathcal{D}(\mathcal{I})$ be the distribution of $\mathcal{I} = (\mathbf{x}, y, \mathbf{z})$ where $(\mathbf{x}, y) \sim \mathcal{D}$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. By Proposition 4.22, the pseudo-projection $\widehat{\Phi}_p$ returned by TensorCompressor$(p, N_0, \mathcal{D}(\mathcal{I}))$ (see Line 9 in CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$) satisfies: with probability $1 - \tau_0$ over the samples drawn in TensorCompressor$(p, N_0, \mathcal{D}(\mathcal{I}))$, the following event holds

$$\mathcal{E}_0 := \left\{\widehat{\Phi}_p \text{ satisfies } \|\widehat{\Phi}_p^\top \widehat{\Phi}_p(\mathbf{M}_p) - \mathbf{M}_p\|_F \le kp\sqrt{\zeta},\ \|\widehat{\Phi}_p^\top \widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) - \mathbf{w}_i^{\otimes p}\|_F \le p\sqrt{\frac{\zeta}{\beta_i}}, \forall i \in [k]\right\}.$$

where $\zeta = O(pdkV_{\mathsf{S}}/\sqrt{N_0\tau_0})$. By Lemma 4.23, we have that $V_{\mathsf{S}} \le k^2 2^{O(p)}/\widehat{\sigma}_{2p+2}^2$. From now on, we condition on this event $\mathcal{E}_0$.

In CompressEmpMoment$(p, N_0, \widehat{\Phi}_p, \mathcal{D}(\mathcal{I}))$, we draw N_0 new samples from $\mathcal{D}(\mathcal{I})$ and construct a new empirical distribution, which we denote by $\widehat{\mathcal{D}}_{\text{new}}(\mathcal{I})$. Then, using the pseudo-projection $\widehat{\Phi}_p$ returned from TensorCompressor$(p, N_0, \mathcal{D}(\mathcal{I}))$, CompressEmpMoment$(p, N_0, \widehat{\Phi}_p, \mathcal{D}(\mathcal{I}))$ constructs

$$\mathbf{A}_p = (\widehat{\Phi}_p \otimes \widehat{\Phi}_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d)\widehat{\mathbf{S}}_{2p+2}, \quad \widehat{\mathbf{S}}_{2p+2} = \mathop{\mathbf{E}}_{\mathcal{I}\sim\widehat{\mathcal{D}}_{\text{new}}(\mathcal{I})}[\mathbf{S}_{2p+2}(\mathcal{I})].$$

Given a probing vector $\mathbf{u}$ from $\widehat{\mathcal{U}}$ (of $N_{\mathbf{u}}$ probing vectors in total), CompressFilter$(p, \widehat{\Phi}_p, \mathbf{u})$ generates:

$$\mathbf{B}_p(\mathbf{u}) = \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}})).$$

Then in Line 15, we compute the empirical estimation (EstM)

$$\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) = \langle \mathbf{A}_p, \mathbf{B}_p(\mathbf{u})\rangle = \langle (\widehat{\Phi}_p \otimes \widehat{\Phi}_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d)\widehat{\mathbf{S}}_{2p+2}, \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}}))\rangle.$$

Let $a, b \in [d]$, and define

$$(\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}))_{ab} = \left\langle (\widehat{\Phi}_p \otimes \widehat{\Phi}_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d)(\widehat{\mathbf{S}}_{2p+2}), \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}})) \otimes \mathbf{e}_a \otimes \mathbf{e}_b \right\rangle,$$

which is the (a, b)-entry of the empirical matrix.

Lemma 4.28. *Conditioned on $\mathcal{E}_0$, with probability at least $1 - 4\tau_0$ (over the fresh N_0 samples $\widehat{\mathcal{D}}_{\text{new}}(\mathcal{I})$ and $\mathbf{u}$), for any indices $a, b \in [d]$, it holds:*

$$|(\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}))_{ab} - (\mathbf{M}_{2p+2}(\mathbf{u}))_{ab}| \le \frac{dk^3p^2 2^{O(p)}}{|\widehat{\sigma}_{2p+2}|\tau_0^{3/2}N_0^{1/4}}.$$

Proof. For this proof, we use the following notation:

$$\Psi_{2p} = \widehat{\Phi}_p \otimes \widehat{\Phi}_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d \qquad \mathbf{B}_p(\mathbf{u}) = \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}}))$$

First, note that the difference between $(\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}))_{ab}$ and $(\mathbf{M}_{2p+2}(\mathbf{u}))_{ab}$ can be decomposed as:

$$\begin{aligned}
&(\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}))_{ab} - (\mathbf{M}_{2p+2}(\mathbf{u}))_{ab} \\
&= \left\langle \Psi_{2p}(\widehat{\mathbf{S}}_{2p+2}), \mathbf{B}_p(\mathbf{u}) \otimes \mathbf{e}_a \otimes \mathbf{e}_b \right\rangle - \left\langle \sum_{i=1} \beta_i \mathbf{w}_i^{\otimes(2p+2)}, \mathbf{F}_p(\mathbf{u}) \otimes \mathbf{F}_p(\bar{\mathbf{u}}) \otimes \mathbf{e}_a \otimes \mathbf{e}_b \right\rangle \\
&= \underbrace{\left\langle \Psi_{2p}\left(\widehat{\mathbf{S}}_{2p+2} - \mathop{\mathbf{E}}_{\mathcal{I}\sim\mathcal{D}(\mathcal{I})}[\mathbf{S}_{2p+2}(\mathcal{I})]\right), \mathbf{B}_p(\mathbf{u}) \otimes \mathbf{e}_a \otimes \mathbf{e}_b \right\rangle}_{I_1} \\
&\quad + \underbrace{\left\langle \Psi_{2p}\left(\mathop{\mathbf{E}}_{\mathcal{I}\sim\mathcal{D}(\mathcal{I})}[\mathbf{S}_{2p+2}(\mathcal{I})] - \sum_{i=1} \beta_i \mathbf{w}_i^{\otimes(2p+2)}\right), \mathbf{B}_p(\mathbf{u}) \otimes \mathbf{e}_a \otimes \mathbf{e}_b \right\rangle}_{I_2} \\
&\quad + \underbrace{\left\langle \Psi_{2p}\left(\sum_{i=1} \beta_i \mathbf{w}_i^{\otimes(2p+2)}\right), \mathbf{B}_p(\mathbf{u}) \otimes \mathbf{e}_a \otimes \mathbf{e}_b \right\rangle}_{I_3} - \underbrace{\left\langle \sum_{i=1} \beta_i \mathbf{w}_i^{\otimes(2p+2)}, \mathbf{F}_p(\mathbf{u}) \otimes \mathbf{F}_p(\bar{\mathbf{u}}) \otimes \mathbf{e}_a \otimes \mathbf{e}_b \right\rangle}_{I_4}.
\end{aligned} \tag{25}$$

We bound each term respectively.

Claim 4.29. *With probability at least $1 - \tau_{\mathbf{u}}$ over the probing vector $\mathbf{u}$, and probability at least $1 - \tau_{2p+2}$ over the samples from $\widehat{\mathcal{D}}_{\text{new}}(\mathcal{I})$, it holds:*

$$|I_1| \le \frac{dk^2 2^{O(p)}}{|\widehat{\sigma}_{2p+2}|\sqrt{\tau_{2p+2}\tau_{\mathbf{u}}N_0}}.$$

Proof. Let $\mathbf{W} = \widehat{\mathbf{S}}_{2p+2} - \mathbf{E}_{\mathcal{I}\sim\mathcal{D}(\mathcal{I})}[\mathbf{S}_{2p+2}(\mathcal{I})]$. Note that $\Psi_{2p}(\mathbf{W}) \in \mathbb{R}^{k\times k\times d\times d}$, we have

$$\mathbf{E}[\|\Psi_{2p}(\mathbf{W})\|_F^2] = \sum_{i_1,i_2\in[k],i_3,i_4\in[d]} \mathbf{E}\left[\left|\left\langle \mathbf{W},\ \Psi_{2p}^\top(\mathbf{e}_{i_1}\otimes\mathbf{e}_{i_2}\otimes\mathbf{e}_{i_3}\otimes\mathbf{e}_{i_4})\right\rangle\right|^2\right].$$

Note that by Item 1 from Fact 4.21, it holds $\|\Psi_{2p}^\top(\mathbf{e}_{i_1}\otimes\mathbf{e}_{i_2}\otimes\mathbf{e}_{i_3}\otimes\mathbf{e}_{i_4})\|_F = 1$. Thus, by Lemma 4.23, for any $i_1, i_2 \in [k], i_3, i_4 \in [d]$, we have:

$$\begin{aligned}
&\mathbf{E}\left[\left|\left\langle \mathbf{W},\ \Psi_{2p}^\top(\mathbf{e}_{i_1}\otimes\mathbf{e}_{i_2}\otimes\mathbf{e}_{i_3}\otimes\mathbf{e}_{i_4})\right\rangle\right|^2\right] \\
&\le \frac{1}{N_0}\mathop{\mathbf{E}}_{\mathcal{I}\sim\mathcal{D}(\mathcal{I})}\left[\left|\left\langle \mathbf{S}_{2p+2}(\mathcal{I}) - \mathop{\mathbf{E}}_{\mathcal{I}\sim\mathcal{D}(\mathcal{I})}[\mathbf{S}_{2p+2}(\mathcal{I})],\ \Psi_{2p}^\top(\mathbf{e}_{i_1}\otimes\mathbf{e}_{i_2}\otimes\mathbf{e}_{i_3}\otimes\mathbf{e}_{i_4})\right\rangle\right|^2\right] \\
&\le \frac{V_{\mathsf{S}}}{N_0} \le \frac{k^2 2^{O(p)}}{\widehat{\sigma}_{2p+2}^2 N_0}.
\end{aligned}$$

This implies $\mathbf{E}[\|\Psi_{2p}(\mathbf{W})\|_F^2] \le d^2k^4 2^{O(p)}/(\widehat{\sigma}_{2p+2}^2 N_0)$. Therefore, by Chebyshev's inequality, with probability at least $1-\tau_{2p+2}$ over the samples $\widehat{\mathcal{D}}_{\text{new}}(\mathcal{I})$:

$$\|\Psi_{2p}(\mathbf{W})\|_F \le \sqrt{\frac{k^4 d^2 2^{O(p)}}{\widehat{\sigma}_{2p+2}^2 \tau_{2p+2} N_0}} = \frac{k^2 d 2^{O(p)}}{|\widehat{\sigma}_{2p+2}|\sqrt{\tau_{2p+2}N_0}}.$$

Let us denote this event by $\mathcal{E}_{\|\mathbf{W}\|_F}$, and condition on this event:

$$\Pr_{\mathbf{u},\widehat{\mathcal{D}}_{\text{new}}(\mathcal{I})}[|I_1| \le t] \ge \Pr_{\mathbf{u},\widehat{\mathcal{D}}_{\text{new}}(\mathcal{I})}[|I_1| \le t, \mathcal{E}_{\|\mathbf{W}\|_F}] \ge (1-\tau_{2p+2})\left(1 - \Pr_{\mathbf{u}}[|I_1| \ge t \mid \mathcal{E}_{\|\mathbf{W}\|_F}]\right). \tag{26}$$

Using Chebyshev's inequality again with respect to the probing vector $\mathbf{u}$, we have

$$\begin{aligned}
\Pr_{\mathbf{u}}[|I_1| \ge t \mid \mathcal{E}_{\|\mathbf{W}\|_F}] &= \Pr_{\mathbf{u}}[|\langle \Psi_{2p}(\mathbf{W}),\ \mathbf{B}_p(\mathbf{u})\otimes\mathbf{e}_a\otimes\mathbf{e}_b\rangle| \ge t \mid \mathcal{E}_{\|\mathbf{W}\|_F}] \\
&\le \frac{\mathbf{E}[|\langle \Psi_{2p}(\mathbf{W}),\ \mathbf{B}_p(\mathbf{u})\otimes\mathbf{e}_a\otimes\mathbf{e}_b\rangle|^2 \mid \mathcal{E}_{\|\mathbf{W}\|_F}]}{t^2}
\end{aligned} \tag{27}$$

By Lemma 4.26, it holds $\mathbf{E}[\|\mathbf{B}_p(\mathbf{u})\otimes\mathbf{e}_a\otimes\mathbf{e}_b\|_F^2] \le \mathbf{E}[\|\mathbf{B}_p(\mathbf{u})\|_F^2] \le k^2 2^{O(p)}$. Using the fact that $\langle \mathbf{W}, \mathbf{B}_p(\mathbf{u})\rangle \le \|\mathbf{W}\|_F \|\mathbf{B}_p(\mathbf{u})\|_F$, we obtain

$$\begin{aligned}
&\mathbf{E}\left[\left|\left\langle \Psi_{2p}(\mathbf{W}),\ \mathbf{B}_p(\mathbf{u})\otimes\mathbf{e}_a\otimes\mathbf{e}_b\right\rangle\right|^2 \,\middle|\, \mathcal{E}_{\|\mathbf{W}\|_F}\right] \\
&\le \mathbf{E}\left[\left\|\Psi_{2p}(\mathbf{W})\right\|_F^2 \left\|\mathbf{B}_p(\mathbf{u})\otimes\mathbf{e}_a\otimes\mathbf{e}_b\right\|_F^2 \,\middle|\, \mathcal{E}_{\|\mathbf{W}\|_F}\right] \\
&\le \frac{d^2k^4 2^{O(p)}}{\widehat{\sigma}_{2p+2}^2 \tau_{2p+2} N_0}\mathbf{E}\left[\left\|\mathbf{B}_p(\mathbf{u})\otimes\mathbf{e}_a\otimes\mathbf{e}_b\right\|_F^2\right] \le \frac{d^2k^6 2^{O(p)}}{\widehat{\sigma}_{2p+2}^2 \tau_{2p+2} N_0}.
\end{aligned} \tag{28}$$

Therefore, combining (28) and (27) and plugging back into (26), we obtain that with probability at least $(1-\tau_{2p+2})(1-\tau_{\mathbf{u}}) \ge 1 - \tau_{\mathbf{u}} - \tau_{2p+2}$, it holds

$$|I_1| \le \frac{dk^3 2^{O(p)}}{|\widehat{\sigma}_{2p+2}|\sqrt{\tau_{2p+2}\tau_{\mathbf{u}} N_0}}.$$

□

For the term I_2, since $\mathbf{E}_{\mathcal{I}\sim\mathcal{D}(\mathcal{I})}[\mathbf{S}_{2p+2}(\mathcal{I})] = \sum_{i=1}\beta_i \mathbf{w}_i^{\otimes(2p+2)}$ by Lemma 4.23, we have $I_2 = 0$. It remains to bound the last two terms in (25).

Claim 4.30. *With probability at least $1-\tau_{\mathbf{u}}$ we have:*

$$|I_3 - I_4| \le \sqrt{\frac{dk^5 p^3 2^{O(p)}}{\widehat{\sigma}_{2p+2}^2 \tau_{\mathbf{u}}\sqrt{N_0\tau_0}}}.$$

Proof. First, recall that by the property of pseudo-projectors we know that $\widehat{\Phi}_p$ is a composition of a isometry $\mathbf{R}_p$ and a projection operator $\mathbf{P}_p$ that maps $(\mathbb{R}^d)^{\otimes p}$ to a subspace $\mathbb{U}_p$: $\widehat{\Phi}_p = \mathbf{R}_p\mathbf{P}_p$, therefore, for two tensors $\mathbf{T}$ and $\mathbf{W}$, it holds $\langle\widehat{\Phi}_p(\mathbf{T}), \widehat{\Phi}_p(\mathbf{W})\rangle = \langle\mathbf{R}_p(\mathbf{P}_p(\mathbf{T})), \mathbf{R}_p(\mathbf{P}_p(\mathbf{W}))\rangle = \langle\mathbf{P}_p^\top\mathbf{P}_p(\mathbf{T}), \mathbf{W}\rangle$ since $\mathbf{R}_p^\top\mathbf{R}_p = \mathbf{I}_k$ as $\mathbf{R}_p$ is an isometry.

Note that the linear operator Ψ_{2p} acting on $\sum_{i=1}\beta_i\mathbf{w}_i^{\otimes(2p+2)}$ is simply:

$$(\widehat{\Phi}_p \otimes \widehat{\Phi}_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d)\left(\sum_{i=1}\beta_i\mathbf{w}_i^{\otimes(2p+2)}\right) = \sum_{i=1}^k \beta_i\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) \otimes \widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) \otimes \mathbf{w}_i \otimes \mathbf{w}_i,$$

and its tensor product with $\mathbf{B}_p(\mathbf{u}) \otimes \mathbf{e}_a \otimes \mathbf{e}_b$ is

$$\begin{aligned}
I_3 &= \left\langle (\widehat{\Phi}_p \otimes \widehat{\Phi}_p \otimes \mathbf{I}_d \otimes \mathbf{I}_d)\left(\sum_{i=1}\beta_i\mathbf{w}_i^{\otimes(2p+2)}\right), \mathbf{B}_p(\mathbf{u}) \otimes \mathbf{e}_a \otimes \mathbf{e}_b\right\rangle \\
&= \sum_{i=1}^k \beta_i\left\langle \widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) \otimes \widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) \otimes \mathbf{w}_i \otimes \mathbf{w}_i, \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u})) \otimes \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}})) \otimes \mathbf{e}_a \otimes \mathbf{e}_b\right\rangle \\
&= \sum_{i=1}^k \beta_i\left\langle \widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}), \widehat{\Phi}_p(\mathbf{F}_p(\mathbf{u}))\right\rangle\left\langle \widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}), \widehat{\Phi}_p(\mathbf{F}_p(\bar{\mathbf{u}}))\right\rangle (\mathbf{w}_i)_a(\mathbf{w}_i)_b \\
&= \sum_{i=1}^k \beta_i\left\langle \widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}), \mathbf{F}_p(\mathbf{u})\right\rangle\left\langle \widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}), \mathbf{F}_p(\bar{\mathbf{u}})\right\rangle (\mathbf{w}_i)_a(\mathbf{w}_i)_b \\
&= \sum_{i=1}^k \beta_i\left|\langle \widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}), \mathbf{F}_p(\mathbf{u})\rangle\right|^2 (\mathbf{w}_i)_a(\mathbf{w}_i)_b\ ,
\end{aligned}$$

where we used that $\mathbf{F}_p(\bar{\mathbf{u}}) = \overline{\mathbf{F}_p(\mathbf{u})}$ and $(\mathbf{w}_i)_a$ is the a-th entry of the vector $\mathbf{w}_i$. On the other hand, we have

$$I_4 = \left\langle \sum_{i=1}\beta_i\mathbf{w}_i^{\otimes(2p+2)}, \mathbf{F}_p(\mathbf{u}) \otimes \mathbf{F}_p(\bar{\mathbf{u}}) \otimes \mathbf{e}_a \otimes \mathbf{e}_b\right\rangle = \sum_{i=1}^k \beta_i\left|\langle\mathbf{w}_i^{\otimes p}, \mathbf{F}_p(\mathbf{u})\rangle\right|^2(\mathbf{w}_i)_a(\mathbf{w}_i)_b.$$

Using the complex inequality $||x|^2 - |y|^2| \le |x-y|(|x|+|y|)$, the difference between these terms equals:

$$|I_3 - I_4| \le \sum_{i=1}^k \beta_i\left|\langle \widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) - \mathbf{w}_i^{\otimes p}, \mathbf{F}_p(\mathbf{u})\rangle\right|\left(\left|\langle \widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}), \mathbf{F}_p(\mathbf{u})\rangle\right| + \left|\langle\mathbf{w}_i^{\otimes p}, \mathbf{F}_p(\mathbf{u})\rangle\right|\right)|(\mathbf{w}_i)_a||(\mathbf{w}_i)_b|\ .$$

To bound the variance, we use Cauchy–Schwarz and the crude bound $(X^2+Y^2)^2 \le 2(X^4+Y^4)$ to get

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{u}}[|I_3 - I_4|^2] \le 2k\sum_{i=1}^k \beta_i^2\sqrt{\mathop{\mathbf{E}}_{\mathbf{u}}\left[\left|\left\langle \widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) - \mathbf{w}_i^{\otimes p}, \mathbf{F}_p(\mathbf{u})\right\rangle\right|^4\right]\mathop{\mathbf{E}}_{\mathbf{u}}\left[\left|\left\langle \widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}), \mathbf{F}_p(\bar{\mathbf{u}})\right\rangle\right|^4\right]}|(\mathbf{w}_i)_a(\mathbf{w}_i)_b|^2 \\
+ 2k\sum_{i=1}^k \beta_i^2\sqrt{\mathop{\mathbf{E}}_{\mathbf{u}}\left[\left|\left\langle\mathbf{w}_i^{\otimes p}, \mathbf{F}_p(\mathbf{u})\right\rangle\right|^4\right]\mathop{\mathbf{E}}_{\mathbf{u}}\left[\left|\left\langle \widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) - \mathbf{w}_i^{\otimes p}, \mathbf{F}_p(\bar{\mathbf{u}})\right\rangle\right|^4\right]}|(\mathbf{w}_i)_a(\mathbf{w}_i)_b|^2.
\end{aligned}$$

Using Lemma 4.26, since $\|\widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p})\|_F \le \|\mathbf{w}_i^{\otimes p}\|_F = 1$ according to Fact 4.21, we further obtain:

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{u}}[|I_3 - I_4|^2] &\le 2k\sum_{i=1}^k \beta_i^2\Bigg(\sqrt{2^{O(p)}\|\widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) - \mathbf{w}_i^{\otimes p}\|_F^4 \times 2^{O(p)}} \\
&\qquad + \sqrt{2^{O(p)} \times 2^{O(p)}\|\widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) - \mathbf{w}_i^{\otimes p}\|_F^4}\Bigg)|(\mathbf{w}_i)_a(\mathbf{w}_i)_b|^2 \\
&\le 4k\sum_{i=1}^k 2^{O(p)}\beta_i^2\|\widehat{\Phi}_p^\top\widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) - \mathbf{w}_i^{\otimes p}\|_F^2.
\end{aligned}$$

As we have conditioned on the event $\mathcal{E}_0$, we know that for every $i \in [k]$, it holds $\|\widehat{\Phi}_p^\top \widehat{\Phi}_p(\mathbf{w}_i^{\otimes p}) - \mathbf{w}_i^{\otimes p}\|_F \leq p\sqrt{\zeta/\beta_i}$, where $\zeta = O(pdkV_{\mathsf{S}}/\sqrt{N_0\tau_0})$ and $V_{\mathsf{S}} \leq k^2 2^{O(p)}/\widehat{\sigma}_{2p+2}^2$. Thus, as $\beta_i \in (0,1)$, we finally get that conditioned on $\mathcal{E}_0$, it holds

$$\mathop{\mathbf{E}}_{\mathbf{u}}[|I_3 - I_4|^2] \leq 4k \sum_{i=1}^{k} 2^{O(p)} \beta_i^2 p^2 \frac{\zeta}{\beta_i} \leq k^2 p^2 2^{O(p)} \zeta \lesssim \frac{dk^5 p^3 2^{O(p)}}{\widehat{\sigma}_{2p+2}^2 \sqrt{N_0 \tau_0}}.$$

Then, applying Markov's inequality, with probability at least $1 - \tau_{\mathbf{u}}$ we have:

$$|I_3 - I_4| \leq \sqrt{\frac{dk^5 p^3 2^{O(p)}}{\widehat{\sigma}_{2p+2}^2 \tau_{\mathbf{u}} \sqrt{N_0 \tau_0}}}\,.$$

□

In summary, after a union bound on these events, we have that with probability at least $1 - \tau_{2p+2}$ over the fresh samples from $\widehat{\mathcal{D}}_{\text{new}}$ and with probability at least $1 - 3\tau_{\mathbf{u}}$ over the random probing vector $\mathbf{u}$, we have

$$\begin{aligned} |(\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}))_{ab} - (\mathbf{M}_{2p+2}(\mathbf{u}))_{ab}| &\leq |I_1| + |I_2| + |I_3 - I_4| \\ &\leq \frac{dk^2 2^{O(p)}}{|\widehat{\sigma}_{2p+2}| \sqrt{\tau_{2p+2} \tau_{\mathbf{u}} N_0}} + \sqrt{\frac{dk^5 p^3 2^{O(p)}}{\widehat{\sigma}_{2p+2}^2 \tau_{\mathbf{u}} \sqrt{N_0 \tau_0}}}. \end{aligned}$$

After rearranging the terms, we have:

$$|(\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}))_{ab} - (\mathbf{M}_{2p+2}(\mathbf{u}))_{ab}| \lesssim \frac{dk^3 p^2 2^{O(p)}}{|\widehat{\sigma}_{2p+2}| (\tau_{2p+2} \tau_{\mathbf{u}} \tau_0)^{1/2}} \frac{1}{N_0^{1/4}},$$

with probability at least $(1 - \tau_{2p+2})(1 - 3\tau_{\mathbf{u}}) \geq 1 - 3\tau_{\mathbf{u}} - \tau_{2p+2}$ (by the independence between $\mathcal{D}$ and $\mathcal{U}$), conditioned on the event $\mathcal{E}_0$. Letting $\tau_0 = \tau_{2p+2} = \tau_{\mathbf{u}}$ completes the proof of Lemma 4.28. □

By Lemma 4.28, using a union bound over all d^2 pairs of indices a, b, and recalling from Fact 2.4 that $|\widehat{\sigma}_{2p+2}| \gtrsim 1/p^3$, conditioned on the event $\mathcal{E}_0$, we have that with probability at least $1 - 4d^2\tau_0$ over the fresh samples $\widehat{\mathcal{D}}_{\text{new}}(\mathcal{I})$ and the probing vector $\mathbf{u}$, it holds:

$$\|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}(\mathbf{u})\|_F \leq \frac{d \operatorname{poly}(k, 1/\tau_0) 2^{O(p)}}{N_0^{1/4}}.$$

Then, using another union bound over all $N_{\mathbf{u}}$ probing vectors from $\widehat{\mathcal{U}}$, we have that with probability at least $1 - d^2 N_{\mathbf{u}} \tau_0$ (conditioned on $\mathcal{E}_0$), it holds

$$\max_{\mathbf{u} \in \widehat{\mathcal{U}}} \|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}(\mathbf{u})\|_F \leq d^2 \operatorname{poly}(k, p, 1/\tau_0) 2^{O(p)}/N_0^{1/4}. \tag{29}$$

Recall that $\mathcal{E}_0$ happens with probability at least $1 - \tau_0$, thus, (29) holds with probability at least $1 - (d^2 N_{\mathbf{u}} + 1)\tau_0$ over all $N = (p+1)N_0$ samples and the probing vectors set $\widehat{\mathcal{U}}$. Recall that we have chosen $\tau_0 \lesssim \delta/(d^2 N_{\mathbf{u}})$, $p \lesssim \log(k/\lambda)$, $N_{\mathbf{u}} = \operatorname{poly}(k/\lambda)\log(1/\delta)$ in $\mathsf{CDRviaImplicitTensor}(\mathcal{D}, \lambda, \delta)$. Hence, with $N_0 = \operatorname{poly}(d, k, N_{\mathbf{u}}/\delta) 2^{O(p)} = \operatorname{poly}(d, k/\lambda, 1/\delta)$, (29) yields that with probability at least $1 - \delta/2$,

$$\max_{\mathbf{u} \in \widehat{\mathcal{U}}} \|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 \leq \max_{\mathbf{u} \in \widehat{\mathcal{U}}} \|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}(\mathbf{u})\|_F \leq \frac{d^2 \operatorname{poly}(k, p, 1/\tau_0) 2^{O(p)}}{N_0^{1/4}} \leq \sqrt{\xi}\lambda,$$

completing the proof of Proposition 4.14. □

4.4 Proof of the Main Result of the Section: Polynomial Complexity CDR

We are ready to prove Proposition 4.1, which is restated below.

Proposition 4.1 (Polynomial-Time CDR)**.** *Let $k, d \in \mathbb{Z}_+$ and $\lambda, \delta \in (0,1)$. Let $\mathcal{D}$ be a Gaussian sum of k ReLUs. Then, there is an algorithm (CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$) that draws $N = \mathrm{poly}(d, k, 1/\lambda, 1/\delta)$ i.i.d. samples from $\mathcal{D}$, runs in $T = \mathrm{poly}(d, N)$ time, and outputs a list $\mathcal{L}$ of unit vectors of size $|\mathcal{L}| = \mathrm{poly}(k/\lambda)\log(1/\delta)$ such that with probability at least $1-\delta$, for any $\mathbf{w}_i$, $i \in [k]$, with $\beta_i \geq \lambda$, there exists a vector $\mathbf{r}_i \in \mathcal{L}$ such that $|\mathbf{r}_i \cdot \mathbf{w}_i| \geq 0.99$.*

Proof of Proposition 4.1. CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$ uses a total of $(p+1)N_0 = \mathrm{poly}(d, k/\lambda, 1/\delta)$ samples. According to Proposition 4.22, line 9 takes $\mathrm{poly}(d, N_0)$ time. Further, by Lemma 4.25, lines 10 and 14 (algorithms CompressEmpMoment$(p, N_0, \widehat{\Phi}_p, \mathcal{D}(\mathcal{I}))$ and CompressFilter$(p, \widehat{\Phi}_p, \mathbf{u})$) take runtime only $\mathrm{poly}(d, k, p)$. Thus, we conclude that the runtime of CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$ is $\mathrm{poly}(dN_0)$.

It remains to argue the correctness of CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$. Recall that by Lemma 4.13, fixing any $i^* \in [k]$ such that $\beta_{i^*} \geq \lambda$, the "separation" and "correlation" events

$$\mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \mathcal{E}_{i^*}^{\mathrm{Cor}} = \{\mathbf{u} : (1+\xi)|\mathbf{w}_j \cdot \mathbf{u}| \leq |\mathbf{w}_{i^*} \cdot \mathbf{u}|, \forall j \in \mathcal{A}_{i^*}(\xi)\} \cap \{\mathbf{u} : |\mathbf{w}_{i^*} \cdot \mathbf{u}| \geq (Z_p(\mathcal{U}))^{1/(2p)}\}$$

happen with probability at least $\mathrm{poly}(\lambda/k)$ over the choice of the random probing vector $\mathbf{u} = \mathbf{p} + \mathrm{i}\mathbf{q}$, $\mathbf{p}, \mathbf{q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Since the number of probing vectors $N_{\mathbf{u}}$ is chosen to be $N_{\mathbf{u}} = \mathrm{poly}(k/\lambda)\log(k/\delta)$, by the Chernoff bound, we know that drawing a set $\widehat{\mathcal{U}}$ of $N_{\mathbf{u}}$ independent probing vectors $\mathbf{u}$, with probability at least $1 - \delta/2$, it holds that:

$$\text{for every } i^* \in [k] \text{ satisfying } \beta_{i^*} \geq \lambda,\ \exists \mathbf{u} \in \widehat{\mathcal{U}},\ s.t.\ \mathbf{u} \in \mathcal{E}_{i^*}^{\mathrm{Sep}} \cap \mathcal{E}_{i^*}^{\mathrm{Cor}}\ . \tag{30}$$

On the other hand, by Proposition 4.14 we know that under the choice of parameters specified in CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$, with probability at least $1 - \delta/2$ over all the samples from $\mathcal{D}$ and the random vectors $\mathbf{u} \in \widehat{\mathcal{U}}$, it holds

$$\max_{\mathbf{u} \in \widehat{\mathcal{U}}} \|\widehat{\mathbf{M}}_{2p+2}(\mathbf{u}) - \mathbf{M}_{2p+2}(\mathbf{u})\|_2 \leq \sqrt{\xi}\lambda. \tag{31}$$

Thus, the joint probability that these two events, (30) and (31) hold at the same time is at least $1 - \delta$. Invoking Lemma 4.2, we obtain that with probability at least $1 - \delta$, for every $i^* \in [k]$ satisfying $\beta_{i^*} \geq \lambda$, there exists a vector $\mathbf{r}_{i^*}$ in the list returned by Algorithm 7 that satisfies $|\mathbf{r}_{i^*} \cdot \mathbf{w}_{i^*}| \geq 0.99$. Note that the size of the list is $2N_{\mathbf{u}} = \mathrm{poly}(k/\lambda)\log(1/\delta)$. This completes the proof of Proposition 4.1. □

5 Putting Everything Together: Proof of Theorem 1.3

By putting everything together, we obtain the following theorem, which is the formal version of Theorem 1.3.

Theorem 5.1. *Let $d, k \in \mathbb{Z}_+$ and $\epsilon > 0$. Let $(\mathbf{x}, y) \sim \mathcal{D}$ be a distribution supported on $\mathbb{R}^d \times \mathbb{R}$ such that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $y = f(\mathbf{x})$ for an unknown $f \in \mathcal{C}_{d,k}$. LearnSumReLUs$(d, k, \epsilon; \mathcal{D})$, using subroutine CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$ with parameters $\lambda = \mathrm{poly}(\epsilon/k)$, $\delta = \mathrm{poly}(\epsilon/k)$, has the following guarantee: with probability at least $2/3$, it outputs a hypothesis $\widehat{y}(\mathbf{x})$ satisfying $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \widehat{y}(\mathbf{x}))^2] \leq \epsilon$. LearnSumReLUs$(d, k, \epsilon; \mathcal{D})$ uses $\mathrm{poly}(d, k, 1/\epsilon)$ samples from $\mathcal{D}$, and runs in $\mathrm{poly}(d, k, 1/\epsilon)$ time.*

Proof of Theorem 5.1. By Theorem 3.1, using a Coarse Direction Recovery procedure (CDR) procedure satisfying Assumption 1 with parameters $\lambda = \mathrm{poly}(\epsilon/k)$ and $\delta = \mathrm{poly}(\epsilon/k)$, the algorithm LearnSumReLUs$(d, k, \epsilon; \mathcal{D})$ outputs a hypothesis $\widehat{y}(\mathbf{x})$ with error $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \widehat{y}(\mathbf{x}))^2] \leq \epsilon$ with probability at least $2/3$. In addition, LearnSumReLUs$(d, k, \epsilon; \mathcal{D})$ uses

$$N = \widetilde{O}(d\,\mathrm{poly}(k/\epsilon) \,+\, L_{\mathsf{CDR}}^3(N_{\mathsf{CDR}} + L_{\mathsf{CDR}}^3)\mathrm{poly}(k/\epsilon))$$

samples from $\mathcal{D}$ and runs in time $\mathrm{poly}(d, N, T_{\mathsf{CDR}})$, where N_{CDR}, T_{CDR} and L_{CDR} are the sample complexity, runtime and the output list length of the CDR procedure respectively.

On the other hand, in Proposition 4.1, we proved that CDRviaImplicitTensor$(\mathcal{D}, \lambda, \delta)$ is a CDR procedure satisfying Assumption 1, with sample complexity $N_{\mathsf{CDR}} = \mathrm{poly}(d, k, 1/\lambda, 1/\delta)$, runtime $T_{\mathsf{CDR}} = \mathrm{poly}(d, k, 1/\lambda, 1/\delta)$, and list size $L_{\mathsf{CDR}} \leq \mathrm{poly}(k/\lambda)\log(1/\delta)$.

Thus, combining the results from Theorem 3.1 and Proposition 4.1, we obtain that our algorithm $\mathsf{LearnSumReLUs}(d, k, \epsilon; \mathcal{D})$, using a CDR procedure $\mathsf{CDRviaImplicitTensor}(\mathcal{D}, \lambda, \delta)$ with parameters $\lambda = \mathrm{poly}(\epsilon/k)$ and $\delta = \mathrm{poly}(\epsilon/k)$, with probability at least 2/3 outputs a hypothesis $\widehat{y}(\mathbf{x})$ with L_2^2 error $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \widehat{y}(\mathbf{x}))^2] \leq \epsilon$. In addition, considering the sample complexity, runtime, and the output list length of the CDR procedure $\mathsf{CDRviaImplicitTensor}(\mathcal{D}, \lambda, \delta)$ under the choice of parameters $\lambda = \mathrm{poly}(\epsilon/k)$ and $\delta = \mathrm{poly}(\epsilon/k)$, it follows that $\mathsf{LearnSumReLUs}(d, k, \epsilon; \mathcal{D})$ uses $\mathrm{poly}(d, k, 1/\epsilon)$ samples from $\mathcal{D}$, and runs in $\mathrm{poly}(d, k, 1/\epsilon)$ time. This completes the proof of Theorem 5.1. □

References

[ATV21] P. Awasthi, A. Tang, and A. Vijayaraghavan. Efficient algorithms for learning depth-2 neural networks with general ReLU activations. In *Advances in Neural Information Processing Systems*, volume 34, pages 13485–13496, 2021.

[BF02] N. Bshouty and V. Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002.

[BFJ+94] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of the Twenty-Sixth Annual Symposium on Theory of Computing*, pages 253–262, Montreal, Quebec, Canada, 1994. ACM Press.

[BJW19] A. Bakshi, R. Jayaram, and D. P. Woodruff. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory, COLT 2019*, 2019.

[Bog98] V. Bogachev. *Gaussian measures*. Mathematical surveys and monographs, vol. 62, 1998.

[CDG+23] S. Chen, Z. Dou, S. Goel, A. R. Klivans, and R. Meka. Learning narrow one-hidden-layer relu networks. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 5580–5614. PMLR, 2023.

[CGKM22] S. Chen, A. Gollakota, A. R. Klivans, and R. Meka. Hardness of noise-free learning for two-hidden-layer neural networks. In *NeurIPS*, 2022.

[Che21] S. Chen. *Rethinking Algorithm Design for Modern Challenges in Data Science*. PhD thesis, MIT, 2021.

[CKM21] S. Chen, A. R. Klivans, and R. Meka. Efficiently learning one hidden layer relu networks from queries. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 24087–24098, 2021.

[CKM22] S. Chen, A. R. Klivans, and R. Meka. Learning deep relu networks is fixed-parameter tractable. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, 2022.

[CN24] S. Chen and S. Narayanan. A faster and simpler algorithm for learning shallow networks. In *The Thirty Seventh Annual Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 981–994. PMLR, 2024.

[DIKR25] I. Diakonikolas, G. Iakovidis, D. Kane, and L. Ren. Algorithms and SQ lower bounds for robustly learning real-valued multi-index models. In *Advances in Neural Information Processing Systems 38, NeurIPS 2025*, 2025.

[DIKZ25] I. Diakonikolas, G. Iakovidis, D. M. Kane, and N. Zarifis. Robust learning of multi-index models via iterative subspace approximation. In *66th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2025, Sydney, Australia, December 14-17, 2025*, pages 2231–2239. IEEE, 2025.

[DK20] I. Diakonikolas and D. M. Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 184–195. IEEE, 2020.

[DK24] I. Diakonikolas and D. M. Kane. Efficiently learning one-hidden-layer relu networks via Schur polynomials. In *The Thirty Seventh Annual Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1364–1378. PMLR, 2024.

[DK25] I. Diakonikolas and D. M. Kane. Implicit high-order moment tensor estimation and learning latent variable models. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1228–1247. IEEE, 2025.

[DKKZ20] I. Diakonikolas, D. M. Kane, V. Kontonis, and N. Zarifis. Algorithms and SQ lower bounds for pac learning one-hidden-layer ReLU networks. In *Conference on Learning Theory, COLT*, pages 1514–1539. PMLR, 2020.

[DKS18] I. Diakonikolas, D. M. Kane, and A. Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1061–1073, 2018.

[DLB25] A. Damian, J. D. Lee, and J. Bruna. The generative leap: Tight sample complexity for efficiently learning Gaussian multi-index models. In *Advances in Neural Information Processing Systems 38, NeurIPS 2025*, 2025.

[Fel08] V. Feldman. Evolvability from learning algorithms. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 619–628. ACM, 2008.

[GGJ+20] S. Goel, A. Gollakota, Z. Jin, S. Karmalkar, and A. R. Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 3587–3596, 2020.

[GGK20] S. Goel, A. Gollakota, and A. R. Klivans. Statistical-query lower bounds via functional gradients. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[GK19] S. Goel and A. R. Klivans. Learning neural networks with two nonlinear layers in polynomial time. In *Conference on Learning Theory, COLT 2019*, pages 1470–1499, 2019.

[GKLW19] R. Ge, R. Kuditipudi, Z. Li, and X. Wang. Learning two-layer neural networks with symmetric inputs. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[GLM18] R. Ge, J. D. Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[JSA15] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods, 2015.

[Kea93] M. J. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, pages 392–401, 1993.

[LL22] A. Liu and J. Li. Clustering mixtures with almost optimal separation in polynomial time. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1248–1261. ACM, 2022.

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[LZZ24] S. Li, I. Zadik, and M. Zampetakis. On the hardness of learning one hidden layer neural networks. *CoRR*, abs/2410.03477, 2024. Available at https://arxiv.org/abs/2410.03477.

[NY83] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. Wiley-Interscience, Chichester, NY, 1983. Translated from Russian by E. R. Dawson.

[O'D14] R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[SJA16] H. Sedghi, M. Janzamin, and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, pages 1223–1231, 2016.

[VW19] S. Vempala and J. Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and SQ lower bounds. In *Conference on Learning Theory, COLT 2019*, pages 3115–3117, 2019. Full version available at https://arxiv.org/abs/1805.02677.

[WZD26] P. Wang, N. Zarifis, and J. Diakonikolas. Efficient optimization on the Euclidean sphere: Riemannian gradient alignment as a unifying principle. *arXiv preprint arXiv:2609.07723*, 2026.

[WZDD23] P. Wang, N. Zarifis, I. Diakonikolas, and J. Diakonikolas. Robustly learning a single neuron via sharpness. *40th International Conference on Machine Learning*, 2023.

[WZDD25] P. Wang, N. Zarifis, I. Diakonikolas, and J. Diakonikolas. Robustly learning monotone single-index models. In *Advances in Neural Information Processing Systems 38, NeurIPS 2025*, 2025.

[Zhe26] Z. Zheng. Polynomial learning of positive ReLU networks under Gaussian inputs. *arXiv preprint arXiv:2610.25116*, 2026.

[ZLJ16] Y. Zhang, J. D. Lee, and M. I. Jordan. L1-regularized neural networks are improperly learnable in polynomial time. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, pages 993–1001, 2016.

[ZSJ+17] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 4140–4149, 2017.

[ZWDD25] N. Zarifis, P. Wang, I. Diakonikolas, and J. Diakonikolas. Robustly learning monotone generalized linear models via data augmentation. In *Proceedings of the Thirty Eighth Conference on Learning Theory*, volume 291, pages 5921–5990. PMLR, 2025.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il