Recovering Imbalanced Clusters via Gradient-Based Projection Pursuit

Martin Eppert[a,*], Satyaki Mukherjee[b], Debarghya Ghoshdastidar[a]

[a] *Technical University of Munich School of Computation, Information and Technology - I7 Boltzmannstr. 3 85748 Garching b. München Germany*
[b] *National University of Singapore Level 4, Block S17 10 Lower Kent Ridge Road Singapore 119076*

Abstract

Projection Pursuit is a classic exploratory technique for finding "interesting" projections of a dataset. We propose a method for recovering projections containing either Imbalanced Clusters or a Bernoulli-Rademacher distribution using a gradient-based technique to optimize the projection index. As sample complexity is a major limiting factor in Projection Pursuit, we analyze our algorithm's sample complexity within a Planted Vector setting where we can observe that Imbalanced Clusters can be recovered more easily than balanced ones. Additionally, we give a generalized result that works for a variety of data distributions and projection indices. We compare these results to computational lower bounds in the Low-Degree-Polynomial Framework. Finally, we experimentally evaluate our method's applicability to real-world data on FashionMNIST.

Keywords: Gradient-Based Methods, Projection Pursuit, Optimization, Statistical Computational Gap
2020 MSC: Primary 62H12, Secondary 62F12

1. Introduction

Projection Pursuit was introduced in Friedman and Tukey [6] as a method for finding maximally interesting projections, as determined by the histogram of the projected data. The measure of "interestingness" is assessed by a function called the projection index of the data projected onto a subspace. A subspace is then obtained for which the projection index is maximized. One example of such a function would be Kurtosis, as maximizing excess Kurtosis can be a good way of finding non-Gaussian components [9]. The recovery of projections with large Kurtosis is commonplace in Projection Pursuit and related methods [3, 9, 15, 27]. Also, recovery of projections with minimal Kurtosis is common for finding bounded projections, for example, containing two evenly sized clusters [4].

A menagerie of projection indices is commonly used as presented in [10, 13, 14, 20, 21]. The choice of the index depends on multiple considerations. Ideally, we would want an index that is simultaneously considered relevant in the specific domain (e.g. finds subspaces where the data is bimodal) and is also easily optimizable. In the interest of practical problems related to optimization, a common choice is to use a differentiable projection index, which can be optimized using off-the-shelf iterative gradient-based optimization techniques. For example, in the setting of Independent Component Analysis(ICA), the gradient-based FastICA algorithm is commonly used to find maximizers of certain projection indices [26]. ICA is a related method to Projection Pursuit with the goal of identifying a set of independent random variables from a linear combination [9]. Even with this assumption, however, it cannot be guaranteed that a global maximizer of the projection index is recovered within polynomial time. Frequently, it can be observed that a minimum amount of samples is required for algorithms to recover a maximizer of the Projection Index. For instance, Davis et al. [4] considers a model where the generated dataset has a sufficiently "interesting" subspace, i.e., the projection index of the data projected to the signal subspace is sufficiently larger than the same projected to any orthogonal subspace, even just with a limited amount of samples. However, in practice, the landscape of the loss function prevents gradient ascent and other methods, such as spectral methods, from finding the correct subspace in polynomial time if too few samples are present. This is commonly called a statistical-computational gap [5, 8, 15]. This means that there are regimes where the amount of data (n), with respect to the dimension (d), is

*Corresponding author. Email address:martin.eppert@tum.de

statistically enough for Projection Pursuit to work, yet in practice, the subspace is not recoverable using polynomial time algorithms. In Projection Pursuit, this generally manifests in a way where it is possible to certify that a given projection possesses an interesting structure using a small amount of samples (usually $n = \widetilde{O}(d)$) but a significantly larger amount of samples are needed to recover a projection (in polynomial time) which reveals the structure. Beyond this statistical-computational gap, as the number of samples becomes much larger than the ambient dimension, it becomes possible to do Projection Pursuit again. For example, in Davis et al. [4], $n = \Omega(d^2)$ is needed for Projection Pursuit to be usable.

However, in practice, the number of samples available frequently is in a significantly smaller order, thus severely limiting the applicability of Projection Pursuit. Thus, it is crucial to design sample-efficient methods for Projection Pursuit. We consider this problem under the context of recovering a planted vector ("signal") from an otherwise Gaussian subspace. This manifests in a way where if the data is projected in a direction, which we will call the "signal direction," the planted vector is revealed, while in all orthogonal directions, the data follows a Gaussian distribution. In this paper, we consider a setting where the variance is unhelpful for recovering the planted vector, i.e., the planted vector has unit variance. Recovering a planted vector is a generic task related to many problems in machine learning and statistics. For example, recovering a planted vector is a subproblem in recovering sparsely used dictionaries. Here, it is generally assumed that an (ortho-normal) basis exists such that the data represented in this basis is sparse. Projection Pursuit is used to identify projections, which correspond to elements in the dictionary. Bai et al. [1] explores the recovery of an orthogonal dictionary using gradient descent on the ℓ_1-norm. Zhai et al. [27] explores the recovery of an orthogonal dictionary using a generalized power method by optimizing the Kurtosis.

This paper focuses on the performance of gradient-based algorithms for recovering planted vectors. As there has already been a significant effort in the literature to find sample-efficient algorithms for recovering planted vectors that match computational lower bounds, it is also of interest if gradient-based methods are also able to match computational lower bounds. Spectral methods can approximately find projections with either large [7, 15] or small [4] Kurtosis. Davis et al. [4] apply spectral methods to recover a planted vector when it is a balanced mixture of Gaussians and Hopkins et al. [7] consider a spectral method to recover a planted vector which follows a Bernoulli-Gaussian/Rademacher distribution. Gradient-based methods are more commonly used in practice as they are both easier to implement and easier to generalize to other projection indices. Additionally, gradient-based algorithms scale well computationally in terms of the number of samples and dimensionality. Typically, gradient ascent runs in $O(nd)$ steps per iteration. Naive spectral methods based on singular-value or eigenvalue decompositions need at least $O(nd^2)$ steps.[1] Additionally, it is often not easy to extend to more complicated settings such as dictionary learning.

We present a general gradient-based Projection Pursuit algorithm, which we apply to the recovery of Imbalanced Clusters in the planted vector setting, where we use p to denote the probability of a sample being in one cluster over the other. A related setting in which Imbalanced Clusters have already been studied is outlier detection. Loperfido [14] considers an imbalanced mixture of Gaussians, which can be recovered by maximizing Kurtosis. We use the projection index $\phi(x) = \max\{0, x\}^2$ (ReLU2) for which we prove that $n = \widetilde{O}(d^2p^2)$ samples are sufficient for gradient ascent to recover the signal direction. In this setting, as the cluster imbalance becomes larger, the smaller one of the two clusters moves further away from the origin, with the larger one approaching zero. This assumption on the data is similar to the assumptions made in dictionary learning [18, 22] where the distribution of the data on optimal projection is assumed sparse (i.e. is zero with a large probability). This is usually modeled using a Bernoulli-Rademacher and Bernoulli-Gaussian distribution, which are constructed by the product of a Bernoulli random variable with another independent random variable to obtain a distribution that is zero with some fixed probability. Thus, we also apply the algorithm to the recovery of a Bernoulli-Rademacher planted vector using Kurtosis as a Projection Index for which $n = \widetilde{\Omega}(d^3p^4)$ are sufficient. To our knowledge, the best-known algorithm, which uses exclusively gradient-based methods to recover a planted sparse vector, has a sample complexity of $n = \widetilde{\Omega}\left(d^4\right)$ [21].

In the planted vector setting, we analyze what changes to the gradient ascent algorithm can be applied to improve its sample complexity. First, we propose using fresh mini-batches in each iteration of gradient ascent [2]. Suppose we reuse the same dataset for each step of gradient ascent. In that case, more samples are necessary to ensure that the optimization problem is smooth enough so that gradient ascent does not get stuck in local maxima. Using minibatches mitigates this problem but comes at a significant cost, as we require new samples in each iteration. However, we prove

[1] Although in some cases it is possible to improve upon the time complexity of the algorithm by computing eigenvalues using power iteration [7].

that only a few steps are necessary to converge, and thus, the impact of resampling on the sample complexity is small.

Additionally, we suggest initializing gradient ascent with normalized samples from the dataset, similar to a technique used in Spielman et al. [22], where linear programs are used to recover projections that follow a very sparse Bernoulli-Gaussian distribution. Many planted vectors, such as Imbalanced Clusters, have longer tails than the standard Gaussian distribution. Exploiting this, we demonstrate that with reasonable probability, it is possible to find an initialization closer to the signal direction than what could be found using a random initialization. As estimating the gradient for a direction closer to the signal directions generally becomes easier, we require fewer samples to estimate the gradient accurately. Additionally, it is possible to reuse the minibatches for multiple initializations simultaneously.

To study the setting of planted Imbalanced Clusters, we study lower bounds on the sample complexity of recovering the planted vector using any polynomial time algorithm. Our study of the setting where the planted vector contains two clusters, with one being significantly larger than the other, differs from commonly studied planted vector settings in which the planted vector is symmetric, such as the Bernoulli-Rademacher in Mao and Wein [15] and Qu et al. [21]. We prove a computational lower bound close to the sample complexity required for the gradient-based algorithm. For this lower bound in the setting of Imbalanced Clusters, we extend Mao and Wein [15], which uses the framework of Low-Degree-Polynomials [8, 12]. Here, we obtain a lower bound on the sample complexity of $n = \widetilde{\Omega}(d^{1.5}p)$. While the number of samples required by our method is not optimal, it is sufficiently close.

Finally, to motivate the pursuit of imbalanced projections, we show that by applying our algorithm to the FashionMNIST dataset [24], we obtain directions that reveal clusters in the data that correspond to their labels. The utility of these projections is measured by how much information they provide on the classes that are present in the dataset. Especially with the projection index $\phi(x) = \max\{0, x\}^2$, we observe that even with very few samples, it is possible to find projections that reveal the class structure of the dataset by separating one class from the others.

2. Main Results

We use standard asymptotic notation $o(\cdot)$, $O(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$ and $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$ hides logarithmic factors. $\mathbb{S}_{d-1}$ denotes the d-dimensional unit sphere. $\mathcal{U}(\cdot)$ denotes the uniform distribution and $\mathcal{N}(\mu, \sigma^2)$ denotes the standard normal distribution with mean μ and standard deviation σ.

2.1. Setup

We follow the literature on recovery of planted vectors [5, 7, 15], which gives a simplified formulation of the data assumption in Projection Pursuit. Throughout the paper we use n as the number of samples present and d as the dimensionality of the data. In the Planted Vector Setting the model is constructed as follows:

Definition 1 (Planted Vector Setting)**.** *We say* $\mathbf{x} \sim \mathcal{D}_{\mathcal{F}}$ *if,*

$$\mathbf{x} \sim \mathcal{N}(\nu\, \mathbf{u}^*, I_d - \mathbf{u}^*\mathbf{u}^{*\top})$$

Given the random variable $\nu \sim \mathcal{F}$ *for some distribution* $\mathcal{F}$ *and a fixed but unknown direction* $\mathbf{u}^*$*.*

The distribution of ν is generally defined to follow a non-gaussian distribution with unit variance. Later on, we will consider the setting where ν follows a distribution containing either two Imbalanced Clusters or a sparse distribution. If d is large, then if the data is projected in a random direction, it will be approximately Gaussian, but if projected in the direction $\mathbf{u}^*$, the structure of ν can be observed.

2.2. Gradient-Based Algorithm

This section describes Algorithm 1, which is a gradient-based algorithm for optimizing differentiable projection indices. Here, we assume that we are given access to a dataset containing a planted vector as defined in Definition 1. The recovery of the signal direction ($\mathbf{u}^*$) is done by finding an (approximate) solution to the following optimization problem where ψ is the projection index.

$$\hat{\mathbf{u}} = \max_{\mathbf{u} \in \mathbb{S}_{d-1}} \sum_{i=1}^{n} \frac{\psi(\langle \mathbf{X}_i, \mathbf{u} \rangle)}{n}$$

This will be done by performing gradient ascent using a different projection index ϕ using multiple initializations. Then ψ is used to pick the best direction $\hat{\mathbf{u}}$.

Two key ideas are used in the design of the algorithm. The first idea is to use multiple initializations from the dataset by using $\mathbf{u}_i = \frac{\mathbf{X}_i}{\|\mathbf{X}\|_2}$ as initialization inspired by Qu et al. [21], Spielman et al. [22] Intuitively, initializing closer to the planted vector allows for a more accurate estimation of the gradient, which allows a decrease in sample complexity. If the distribution of the planted vector takes on larger values, it is possible to find an initialization with a larger inner product to the signal direction than by using uniformly random samples. For example if $\mathbb{P}\left[\nu = \sqrt{\frac{1}{p}}\right] = p$ then with probability p we have an initialization for which $\langle \mathbf{u}, \mathbf{u}^* \rangle \approx \sqrt{\frac{1}{dp}}$. Instead of choosing $\mathbf{u}$ uniformly at random, we only have $\langle \mathbf{u}, \mathbf{u}^* \rangle \approx \sqrt{\frac{1}{d}}$, which can be much worse in the case when p is sufficiently small. Without knowing which samples have large values in the signal direction, this method comes at the price of having to guess many samples as initialization. This initialization scheme is shown in Algorithm 1.

The second idea is to use minibatches to avoid overfitting. As previously noted, this comes at the cost of needing new samples for each step of gradient ascent. Thus speeding up convergence is necessary to decrease the sample complexity of the algorithm. We do this by using the Riemannian gradient $(I_d - \mathbf{u}\mathbf{u}^\top)\frac{\partial}{\partial \mathbf{u}}\left(\sum_{i=1}^{n} \frac{\phi(\langle \mathbf{X}_i, \mathbf{u} \rangle)}{n}\right)$ instead of the gradient itself, which allows us to decrease the number of steps needed to converge.

Using the Riemannian gradient itself also has a downside. If $\langle \mathbf{u}, \mathbf{u}^* \rangle$ becomes sufficiently large, we cannot guarantee that a gradient step does not degrade the current estimate of the direction. Thus, we use a schedule for the learning rate η, decreasing η once we are close to convergence. The algorithm is presented in Algorithm 1, which calls a subroutine described in Algorithm 2. Algorithm 2 runs Riemannian gradient ascent with a large learning rate η_1 and then extracts the solution with the largest value for the projection index, ensuring that a good solution is found. A second run of Algorithm 2 with a smaller learning rate η_2 is used to fine-tune the projection to find a close estimate of the signal direction $\mathbf{u}^*$. As we use multiple initializations, we end up with multiple estimates of the signal direction. Additionally, we are unsure if an estimate may have diverged during gradient ascent. Thus, we use the second projection index $\psi(\cdot)$ to pick the best estimate of the signal direction.

Algorithm 1 Two-Step Gradient Ascent Algorithm

function TWO_STEP_GRADIENT_ASCENT($\mathbf{X}, n, n_{init}, s, \eta_1, \eta_2$)
  for $j = 1 ... n_{init}$ **do**
    $\mathbf{u}_j \leftarrow \frac{\mathbf{X}_j}{\|\mathbf{X}_j\|_2}$
  end for
  $\hat{\mathbf{u}} \leftarrow$ GRADIENT_ASCENT($\{\mathbf{X}_i\}_{i=n_{init}+1}^{n_{init}+ns}, \mathbf{u}, \eta_1, s$)
  $\hat{\mathbf{u}} \leftarrow$ GRADIENT_ASCENT($\{\mathbf{X}_i\}_{i=n_{init}+ns+1}^{n_{init}+2ns}, \hat{\mathbf{u}}, \eta_2, s$)
  return $\arg\max_{\hat{\mathbf{u}} \in \{\hat{\mathbf{u}}_j | j \in [n_{init}]\}} \sum_{k=1}^{n} \frac{\psi(\langle \mathbf{X}_k, \hat{\mathbf{u}} \rangle)}{n}$
end function

Algorithm 2 Gradient Ascent

function GRADIENT_ASCENT($\mathbf{X}, \mathbf{u}, \eta, s$)
  for $i = 0...(s-1)$ **do**
    Choose $\bar{\mathbf{X}} \leftarrow \{\mathbf{X}_k\}_{k=ni}^{n(i+1)}$
    for $j = 1...n_{init}$ **do**
      Calculate $\mathbf{g} \leftarrow \left(I_d - \mathbf{u}_{i,j}\mathbf{u}_{i,j}^\top\right) \frac{\partial}{\partial \mathbf{u}_{i,j}} \left(\sum_{k=1}^n \frac{\phi(\langle \bar{\mathbf{X}}_k, \mathbf{u}_{i,j}\rangle)}{n}\right)$
      Update $\bar{\mathbf{u}}_{i,j} \leftarrow \mathbf{u}_{i,j} + \eta\mathbf{g}$
      Renormalize $\mathbf{u}_{i+1,j} \leftarrow \frac{\bar{\mathbf{u}}_{i+1,j}}{\|\bar{\mathbf{u}}_{i+1,j}\|_2}$
    end for
  end for
  for $j = 1...n_{init}$ **do**
    $\hat{i} \leftarrow \arg\max_{i\in[s]} \sum_{k=1}^n \frac{\psi(\langle \mathbf{X}_k, \mathbf{u}_{i,j}\rangle)}{n}$
    $\hat{\mathbf{u}}_j \leftarrow \mathbf{u}_{\hat{i},j}$
  end for
  return $\hat{\mathbf{u}}$
end function

2.3. Sample Complexity Bounds

Next, we will highlight a method of studying the sample complexity of the Gradient Ascent Subroutine (Algorithm 2) when specified towards a planted vector distribution and a projection index. We state three assumptions that have to be fulfilled by the setting and the projection index to demonstrate convergence. This analysis can then be applied to both uses of Algorithm 2 to complete the analysis of Algorithm 1. Later, in Subection 2.4 and Subection 2.5, we show that Algorithm 1 can recover planted vectors with close to optimal sample complexity.

Lemma 1 gives a convergence result for arbitrary $\phi(\cdot), \psi(\cdot)$ and a planted vector distribution. In order to apply Lemma 1, we have to demonstrate the following Preconditions hold.

Precondition 1 of Lemma 1 guarantees that at least one initialization is close enough to the signal direction, providing a good starting point for our algorithm. Precondition 2 ensures that the gradient estimates are sufficiently accurate for each step. This also ensures that renormalization does not decrease $\langle \mathbf{u}, \mathbf{u}^*\rangle$. Finally, Precondition 3 ensures that the projection index $\psi(\cdot)$ can be used to (sample-)efficiently test if an initialization has converged. In most cases, choosing $\psi = \phi$ is completely sufficient, but choosing a convenient ψ can oftentimes drastically ease the analysis. This is necessary in the last step of the algorithm to select a converged estimate.

In the following we will use $g_{\mathbf{u}}(\mathbf{x}) := (I - \mathbf{u}\mathbf{u}^\top)\frac{\partial \phi(\langle \mathbf{x}, \mathbf{u}\rangle)}{\partial \mathbf{u}}$ for simplicity.

Lemma 1. *Let* $\mathbf{X} \sim \mathcal{D}^n$. *For* $s = \Omega(\log(d))$ *steps,* $\delta \geq 0$, $1 > b > b - \delta > a > 0$, $\delta > 0$ *and* $n > 0$, *if*

1. *Given* $\mathbf{u}_{0,i}$ *for* $i \in [n_{init}]$ *then with probability at least* $1 - o(1)$

$$\max_i \langle \mathbf{u}_i, \mathbf{u}^*\rangle \geq a$$

2. *For an arbitrary constant* $c_0 > 0$, *if* $\langle \mathbf{u}, \mathbf{u}^*\rangle \in (a, b)$ *then*

$$\frac{\langle \mathbf{u}, \mathbf{u}^*\rangle + \eta\left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^*\right\rangle}{\sqrt{1 + \eta^2 \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}}} \geq (1 + c_0)\langle \mathbf{u}, \mathbf{u}^*\rangle$$

   with probability at least $1 - O\left(\frac{1}{s}\right)$.

3. *There exists a threshold* t *where for all* $\mathbf{u} \in \mathbb{S}_{d-1}$ *if* $\langle \mathbf{u}^*, \mathbf{u}\rangle \geq b$ *then* $\sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \mathbf{u}\rangle)}{n} \geq t$ *and if* $\langle \mathbf{u}, \mathbf{u}^*\rangle \leq b - \delta$ *then* $\sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \mathbf{u}\rangle)}{n} \leq t$, *with probability at least* $1 - o(1)$.

Then, Algorithm 2 returns $\hat{\mathbf{u}}$ *such that* $\max_{i\in[n_{init}]} \langle \hat{\mathbf{u}}_i, \mathbf{u}^*\rangle \geq b - \delta$ *with a total of* $\widetilde{O}(n)$ *samples steps with probability at least* $1 - o(1)$.

The proof of Lemma 1 can be found in Appendix A. Thus, to analyze the performance of Algorithm 1, we apply Lemma 1 once for each execution of Algorithm 2.

2.4. Application of Lemma 1 to Imbalanced Clusters

Here we will focus on a distribution $\mathcal{B}(p)$ containing two Imbalanced Clusters with an imbalance parameter p.

Definition 2 (Imbalanced Clusters)**.** *We say $\nu \sim \mathcal{B}(p)$, with $p \in (0,1)$, if*

$$\nu = \begin{cases} \sqrt{(1-p)/p}, & \text{with probability } p \\ -\sqrt{p/(1-p)}, & \text{with probability } (1-p) \end{cases}$$

The cluster centers are chosen so that the mean is zero and the variance is one. Due to the data having unit variance, methods such as PCA cannot recover the planted vector and need Projection Pursuit methods. Note that for smaller p, the first cluster moves further away from the origin, and the second cluster shifts closer to the origin. This behaves similarly to the Bernoulli-Rademacher distribution. Here, we will be interested in the parameter $p \in (\frac{1}{\sqrt{d}}, \frac{1}{2})$. For larger $p > \frac{1}{2}$, the same results follow by symmetry with the notable exception of $p = \frac{1}{2}$ where the clusters are perfectly balanced. We choose to use the projection index $\phi(x) = \max\{0, x\}^2$. In Theorem 1 we demonstrate bounds on the sample complexity of Algorithm 1 using $\phi(x) = \psi(x) = \max\{0, x\}^2$.

Theorem 1. *For arbitrary $\beta > 0$ and $p \in \left(\frac{1}{2}, \frac{1}{\sqrt{d}}\right)$ there exist $\eta_1 = \Omega\left(\sqrt{d}p\right)$, $\eta_2 = \Theta(1)$, $s = \Theta(\log(d))$, $n_{init} = \Omega(1/p)$ and $n = \widetilde{\Theta}\left(d^2p^2\right)$ such that for sufficiently large d and sufficiently small p, Projection Pursuit using Algorithm 1 with $\mathbf{X} \sim \mathcal{D}^n_{\mathcal{B}(p)}$ and a Projection Index $\phi(x) = \max\{0, x\}^2$ will output $\hat{\mathbf{u}}$ such that $\langle \hat{\mathbf{u}}, \mathbf{u}^* \rangle \geq 1 - \beta$ with probability at least $1 - o(1)$ utilizing a total of $\widetilde{\Theta}\left(d^2p^2\right)$ samples.*

The proof of Theorem 1 can be found in Appendix B.

2.5. Application of Lemma 1 to Bernoulli-Rademacher Planted Vectors

Other commonly studied settings are the Bernoulli-Rademacher and Bernoulli-Gaussian settings [7, 15, 27]. These are both sparse distributions, i.e., are 0 with probability $1-p$, thus are of particular interest in compressed sensing [22, 27]. We will prove that a Bernoulli-Rademacher planted vector can be recovered using gradient-based techniques.

Definition 3 (Bernoulli-Rademacher [22])**.** *We say $\nu \sim \mathcal{BR}(p)$, with $p \in (0,1)$, if*

$$\nu = \begin{cases} \sqrt{1/p}, & \text{with probability } p/2 \\ -\sqrt{1/p}, & \text{with probability } p/2 \\ 0, & \text{with probability } (1-p) \end{cases}$$

We demonstrate, that Algorithm 1 using the projection index $\phi(x) = x^4$ can recover the planted vector using $n = \widetilde{O}(d^3p^4)$ samples.

Theorem 2. *For arbitrary $\beta > 0$ there exist $\eta_1 = \Omega\left(dp^2\right), \eta_2 = \Theta(1), s = \Omega(\log(d)), n_{init} = \Theta(1/p)$ and $n = \widetilde{\Theta}\left(d^3p^4\right)$ such that for sufficiently large $d > 0$ and sufficiently small $p > 0$, Projection Pursuit using Algorithm 1 with $\mathbf{X} \sim \mathcal{D}^n_{\mathcal{BR}(p)}$, $\phi(x) = x^4$ and $\psi(x) = -|x|$ will output $\hat{\mathbf{u}}$ such that $\langle \hat{\mathbf{u}}, \mathbf{u}^* \rangle \geq 1 - \beta$ with probability at least $1 - o(1)$ utilizing a total of $\widetilde{\Theta}\left(d^3p^4\right)$ samples.*

The proof can be found in Appendix C.

3. Statistical Computational Lower Bounds of the Planted Vector Setting

Here, we study whether gradient-based methods are optimal in the sense of matching computational lower bounds. For this, we compare the sample complexity of gradient-based methods to computational lower bounds, which assess the minimum sample complexity required in order for any *computationally efficient* algorithm (i.e. computable in polynomial time) to recover the planted vector, as defined in Definition 1. As this is not tractable for such a general class of algorithm, there have been rigorous results in more limited settings such as lower bounds for the statistical

query model [11], sum of squares hierarchies [16] and the Low Degree Polynomial Framework [8, 12]. Here, we will focus on the framework of Low Degree Polynomials.

Generally, the Low Degree Polynomial Framework uses Low Degree Polynomials as a surrogate for efficiently computable algorithms to determine whether an efficient algorithm exists for deciding a hypothesis testing problem. We will utilize this to obtain lower bounds on the sample complexity of efficiently computable tests. Bounds can be obtained using the optimality of the likelihood ratio test [12]. Let $L := \frac{d\mathcal{H}_1}{d\mathcal{H}_0}$ be the likelihood ratio. Neyman and Pearson [17] shows that thresholding the likelihood ratio L is an optimal test and thus allows reasoning about computational lower bounds. The Low Degree Polynomial Framework focuses on the degree-D likelihood ratio $L_d^{\leq D}$ which is defined as the likelihood ratio projected onto the subspace of polynomials of degree at most D, where D is low, i.e. logarithmic in the size of the problem. By demonstrating that a likelihood ratio test using $L_d^{\leq D}$ fails, we can demonstrate that no polynomial of degree $\leq D$ can be used to construct a test to distinguish $\mathcal{H}_1$ and $\mathcal{H}_0$. This is summarized in the following conjecture.

Conjecture (Low Degree Conjecture [8])**.** *For "sufficiently nice" sequences of probability measures $\mathcal{H}_0$ and $\mathcal{H}_1$, if there exists $\epsilon > 0$ and degree $D \geq \log(d)^{1+\epsilon}$ for which $\|L_d^{\leq D}\|$ remains bounded as $d \to \infty$, then there is no polynomial-time algorithm f for which if $\mathbf{X} \sim \mathcal{H}_0$ then $f(\mathbf{X}) = \mathcal{H}_0$ and if $\mathbf{X} \sim \mathcal{H}_1$ then $f(\mathbf{X}) = \mathcal{H}_1$ with high probability.*

3.1. Planted Vectors with Imbalanced Clusters

To our knowledge, computational lower bounds have not been studied for the planted vector with Imbalanced Clusters setting. To obtain lower bounds on the sample complexity needed to recover a close estimate of the planted vector in polynomial time, we follow a three-step procedure following the method used in Mao and Wein [15]. First, we formulate a hypothesis testing problem in Problem 1, which tests between a Gaussian distribution and the Planted Vector distribution as defined in Definition 1. Then, we demonstrate computational lower bounds on Problem 1 in the Low Degree Polynomial Framework. Finally, we extend the computational lower bounds to the estimation problem of finding a direction $\hat{\mathbf{u}}$ close to the signal direction $\mathbf{u}^*$ such that $\langle \hat{\mathbf{u}}, \mathbf{u}^* \rangle \geq 1 - \beta$. This is done by reducing the estimation problem to Problem 1.

Problem 1. *Let ν be a distribution over $\mathbb{R}$. Define the following null and planted distributions:*

- *Under $\mathcal{H}_0$, observe i.i.d. samples $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}(0, I_d)^n$.*
- *Under $\mathcal{H}_1$, first draw $\mathbf{u}^*$ uniformly from $\mathbb{S}_{d-1}$ and i.i.d. $\nu_1, \cdots, \nu_n$. Conditional on $\mathbf{u}^*$ and $\{\mathbf{x}_i\}$, draw independent samples $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ where $\mathbf{X}_i \sim \mathcal{N}(\nu_i \mathbf{u}^*, I_n - \mathbf{u}^* \mathbf{u}^{*\top})$. Note that this is equivalent to $\mathcal{D}_{\mathcal{F}}$ in Definition 1.*

Suppose that we observe the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with rows $\mathbf{X}_1^\top, \ldots, \mathbf{X}_n^\top$. We aim to test between the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$.

In the following, we will show evidence of a computational statistical gap in the Planted Vector Setting Definition 1. Specifically, we will demonstrate that for $n = \widetilde{O}(d^{1.5} p)$ the Low Degree Likelihood ratio stays bounded and thus, according to the Low Degree Conjecture , no polynomial time algorithm can test Problem 1.

Theorem 3. *For an instance of Problem 1 with $\nu \sim \mathcal{B}(p)^n$ and $n = \widetilde{O}(d^{1.5} p)$ the Low Degree Likelihood Ratio stays bounded for degree $D = \log(d)^{1+\epsilon}$ for $\epsilon > 0$.*

$$\left\| L_d^{\leq D} \right\|_2^2 \leq 2$$

Finally, we reduce the estimation problem to Problem 1. To do this, we have to demonstrate that it is possible to construct a test for Problem 1 if we have access to an estimate $\hat{\mathbf{u}}$ for $\mathbf{u}^*$ such that $\langle \hat{\mathbf{u}}, \mathbf{u}^* \rangle \geq 1 - \beta$. Corollary 1 shows it is possible to construct such a test for Problem 1 if $n = \Omega(d)$.

Corollary 1. *For all $\hat{\mathbf{u}}$ for which $\langle \hat{\mathbf{u}}, \mathbf{u}^* \rangle \geq 1 - \beta$ for sufficiently small $\beta > 0$ and $\frac{1}{\sqrt{d}} \leq p < \frac{1}{2}$. Define the test Ψ:*

$$\Psi := \begin{cases} \mathcal{H}_0 & \text{if } \sum_{i=1}^n \frac{\phi(\langle \mathbf{X}_i, \hat{\mathbf{u}} \rangle)}{n} < t \\ \mathcal{H}_1 & \text{if } \sum_{i=1}^n \frac{\phi(\langle \mathbf{X}_i, \hat{\mathbf{u}} \rangle)}{n} \geq t \end{cases}$$

With $\phi(x) = \max\{0, x\}^2$. There exists a threshold t such that

$$\mathbb{P}_{\mathcal{H}_0}\{\Psi = \mathcal{H}_1\} + \mathbb{P}_{\mathcal{H}_1}\{\Psi = \mathcal{H}_0\} \leq \exp\left(-\Theta\left(\frac{n}{d}\right)\right)$$

In the regime where $n < d+1$, we refer to Zadik et al. [25], demonstrating the statistical impossibility of estimation in this regime. Thus, combining Corollary 1 and Theorem 3 yields the result that if the Low Degree Conjecture is true, no polynomial time can estimate the planted vector. To our knowledge, currently, no efficient algorithm exists which can recover the signal direction with $n = o(d^2p^2)$ samples.

Remark 1. *Dudeja and Hsu [5] gives results on the failure of Low-Degree Polynomials in the Planted Vector Setting. If for $i \in 1, \cdots, k-1$ the moments of ν and the standard normal distribution match a bound of $n \ll d^{k/2}\lambda^{-2}$, where λ is the signal to noise ratio defined as*

$$\left|\mathbb{E}[\nu^k] - \mathbb{E}[Z^k]\right| = \lambda \quad Z \sim \mathcal{N}(0,1).$$

In the case of $\nu \sim \mathcal{B}(p)$ for $k = 3$ we obtain $\lambda \geq \frac{2}{\sqrt{p}}$. Here, we note that the bound is not applicable to our setting as by choosing p sufficiently small, Assumption 2 cannot be fulfilled anymore. Thus, for completeness, we give the same bound of $n = \widetilde{O}(d^{1.5}p)$ which is valid for $p \in \left(\frac{1}{\sqrt{p}}, \frac{1}{2}\right)$.

3.2. Bernoulli Rademacher Planted Vectors

The Bernoulli Rademacher setting has been thoroughly studied in Mao and Wein [15]. Here, the failure of Low Degree Polynomials when $n = \widetilde{O}(d^2p^2)$ is demonstrated. This lower bound is known to be tight, as a spectral algorithm can recover a Bernoulli-Rademacher planted vector with $n = \widetilde{\Theta}(d^2p^2)$ samples, which is tight, as there exist spectral methods which can recover the planted vector when $n = \widetilde{\Omega}(d^2p^2)$ samples [7, 15]. Our gradient-based method has a sample complexity of $n = \widetilde{O}(d^3p^4)$, which is larger than what can be achieved using spectral methods. In the case where $p = \frac{1}{\sqrt{d}}$, this matches the bounds obtained using spectral methods.

Remark 2. *Zadik et al. [25] gives an algorithm to recover planted vectors using LLL-basis reduction. This algorithm does not exhibit the statistical to computational gap. This algorithm is only applicable in a very restrictive setting, as it is required that the planted vector can only take on a set of discrete values. This can be avoided by considering a setting with a small amount of noise. E.g., considering a hierarchical setting where $\mathcal{B}(p')$ with $p' \sim \mathcal{U}([\frac{p}{2}, p])$ is a simple counterexample, in which our analysis still works and where the algorithm discussed in Zadik et al. [25] fails due to the lack of robustness to noise.*

4. Experiments

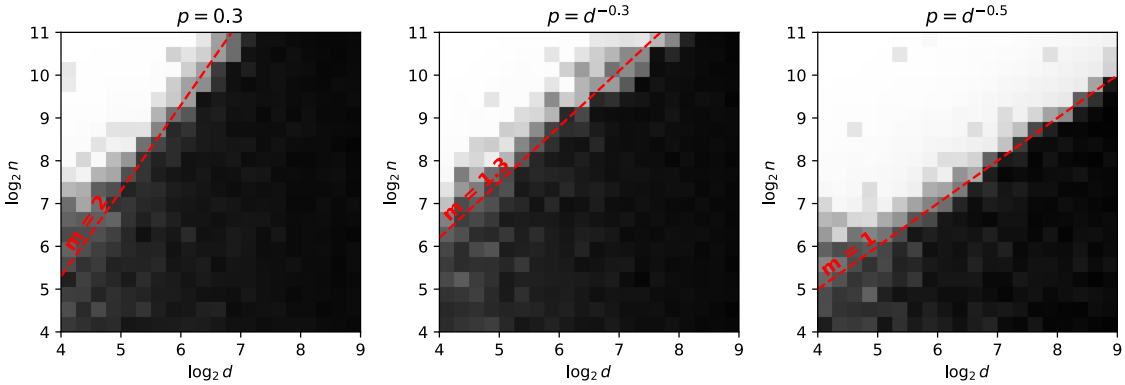
4.1. Experiments with Synthetic Data

In the following, we will validate the findings in Theorem 1 and Theorem 2 by running Algorithm 1 on synthetic datasets. In the Imbalanced Clusters setting, we will be choosing the dimension $d \in [16, 512]$ and the cluster imbalance for $p \in \{d^{-0.5}, d^{-0.3}, 0.3\}$ with $n \in [16, 2048]$. The algorithm is executed for $s = 2\log_2 d$ steps and $\eta_1 = \sqrt{d}p$ and $\eta_2 = 0.5$. The results are plotted in Figure 1a. In the Bernoulli-Rademacher setting we will be choosing the dimension $d \in [16, 256]$ and the cluster imbalance as either $p \in \{d^{-0.5}, d^{-0.5}, 0.3\}$ with $n \in [16, 4096]$. The algorithm is executed for $s = 2\log_2 d$ steps and $\eta_1 = \sqrt{d}p$ and $\eta_2 = 0.5$. The results are plotted in Figure 1b. To evaluate, we plot the average value of $\langle \hat{\mathbf{u}}, \mathbf{u}^* \rangle$ over 30 independently sampled datasets for each d, n, p.

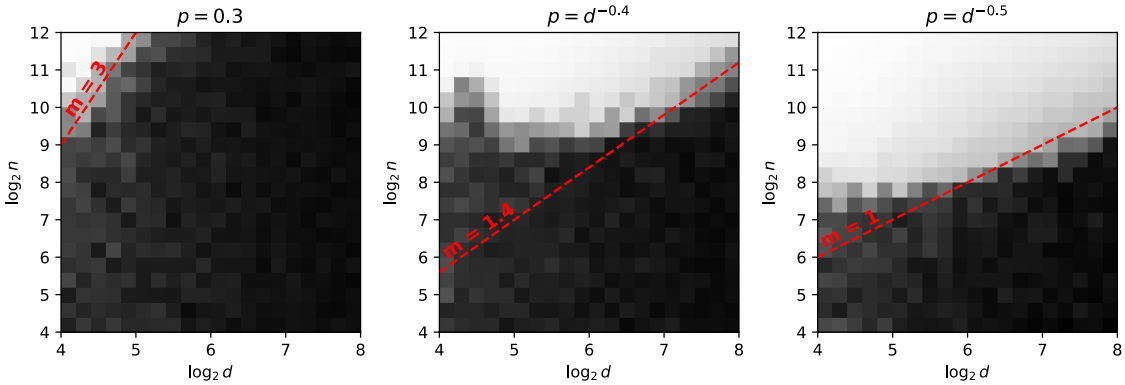
4.2. Comparison to Other Methods

Here, we will apply Algorithm 1 to the dataset with the following projection indices.

Abbreviation	$\phi(x)$	$\psi(x)$
ReLU2	$\max\{0, x\}^2$	$\max\{0, x\}^2$
Kurtosis	x^4	$-\lvert x\rvert$
Abs [21]	$-\lvert x\rvert$	$-\lvert x\rvert$
AbsMax [4]	$\lvert x\rvert$	$\lvert x\rvert$
Skewness [19]	x^3	x^3

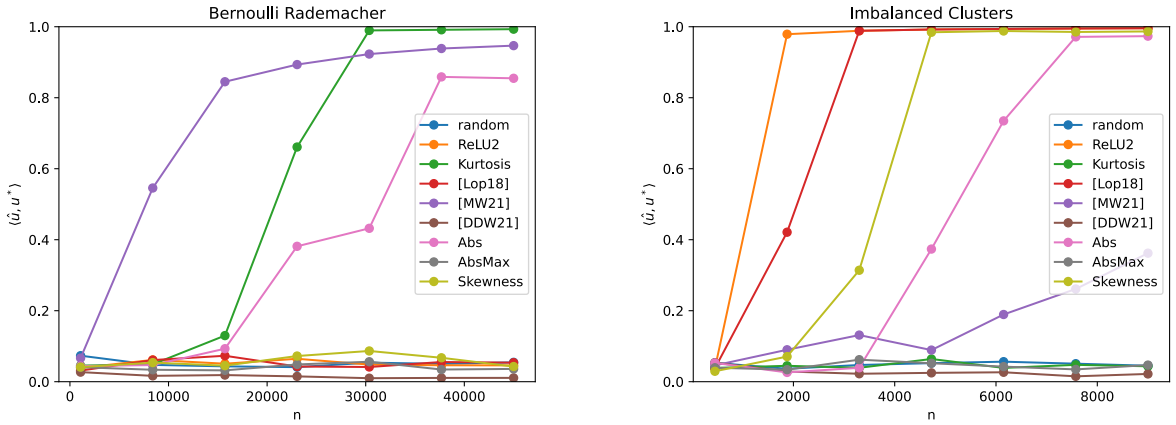


(a) Projection pursuit of Imbalanced Clusters using Algorithm 1 with $\phi(x) = \psi(x) = \max\{0, x\}^2$. The red lines of slopes $(2, 1.3, 1)$ roughly highlight the phase transition.



(b) Projection pursuit of a sparse Bernoulli-Rademacher distribution using Algorithm 1 with $\phi(x) = x^4$ and $\psi(x) = -|x|$. The red lines of slopes $(3, 1.8, 1)$ roughly highlight the phase transition.

Fig. 1: Phase transitions in Projection Pursuit using gradient-based algorithms. The horizontal and vertical axes correspond to $\log_2 d$ and $\log_2 n$. Each pixel shows the average value of the absolute inner product between the predicted direction and the signal direction, where white corresponds to 1 and black to 0.



(a) Bernoulli-Rademacher

(b) Imbalanced Clusters

Fig. 2: Comparison of different methods in the planted vector setting. We plot the average inner product between the signal direction and the recovered direction by each algorithm over 30 datasets.

ReLU2 and **Kurtosis** correspond to the projection indices we study in Theorem 1 and Theorem 2. Additionally, we test the projection indices *Abs*, which corresponds to the projection index used in Qu et al. [21], *AbsMax*, which corresponds to the objective in Davis et al. [4] and *Skewness* as proposed in Paajarvi and LeBlanc [19]. Additionally, we compare the gradient-based methods to three spectral methods. Loperfido [13] introduces a method called MaxSkew for finding directions of large skewness(abbreviated by [Lop18]). Mao and Wein [15] introduces a method for recovering planted sparse vectors(abbreviated by [MW21]). Davis et al. [4] uses a spectral method for identifying directions separating two balanced clusters(abbreviated by [DDW21]).

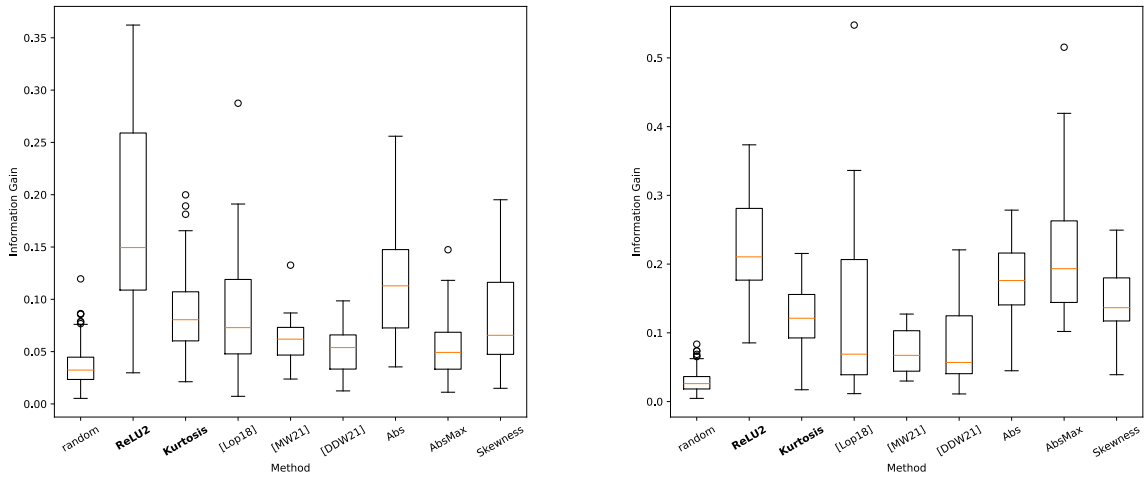
Especially if only a few samples are present, using minibatches seems to restrict the performance of the proposed algorithm. Thus, choosing to subsample the dataset with replacement tends to be beneficial, which we will do for the following experiments.

Here, we compare the previously mentioned methods in the planted vector setting with a Bernoulli Rademacher planted vector in Figure 2a and with an Imbalanced Clusters planted vector in Figure 2b with $d = 300$, $p = 0.1$. For Algorithm 1 we use $n_{init} = 400$ initializations. In Figure 2 we can observe that most algorithms only perform well on one of both settings.

4.3. Experiments with FashionMNIST

We also compare the algorithms by comparing their performance on a small subsample of FashionMNIST [24] with 600 samples and on the whole training dataset with 60000 samples. We use $n_{init} = 500$ initializations and choose the 30 directions with the largest value for the projection index. For spectral methods, we run the spectral method 30 times while removing the recovered directions from the dataset. In order to compare the performance of the different methods, we evaluate how well a single projection can help to predict the labels of the images. This will be evaluated using the Information Gain $\mathrm{IG}(Y, A) = \mathrm{H}(Y) - \mathrm{H}(Y|A)$. Where Y are the labels assigned to the instances in FashionMNIST. Here we choose the indicator $A = I_{\langle \mathbf{u}, \mathbf{x} \rangle > t}$, where t is an automatically chosen threshold for each index to empirically maximize information gain on the training dataset. The final information gain is evaluated on a holdout dataset. In Figure 3, the information gain is plotted for each method. Here, we can observe that the ReLU2 projection index still performs well even if only a few samples(600) are present. For example, AbsMax can only be optimized reliably with a large amount(60000) of samples. Additionally, it can be observed that the best direction found by the MaxSkew method([Lop18]) performs well with small and large amounts of data.

For demonstration purposes, we also show histograms of the data projected onto the recovered projections in Figure 4a and Figure 4b. In Figure 4a, it can be observed that the found projections reveal two Imbalanced Clusters where the smaller cluster contains samples of mostly one class. Figure 4b shows that the Kurtosis projection index



(a) $n = 600$ **(b)** $n = 60000$

Fig. 3: A comparison of projection pursuit approaches on fashionMNIST. We plot the achieved information gain using projections produced by different projection indices.

recovers projections for which many samples are projected close to zero while a few are projected away from zero similar to a Bernoulli-Rademacher distribution.

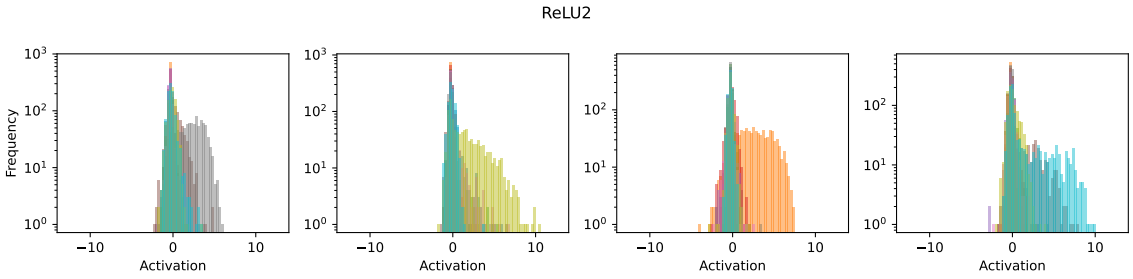
Figure 4a provides insight into why discovering imbalanced projections effectively recovers label-separating projections in classification tasks such as Fashion-MNIST. We can reasonably assume that the data of a classification task follows a cluster structure. If the data is projected in the direction of one cluster center, then we will expect that the samples from the other clusters will collapse close to 0, while the samples from the chosen cluster will move out to one side. This effect can also be observed in Figure 4a. Consequently, projecting the data along the direction of a particular cluster center naturally yields an imbalanced histogram, with samples from the corresponding cluster positioned towards one extreme.

5. Conclusion

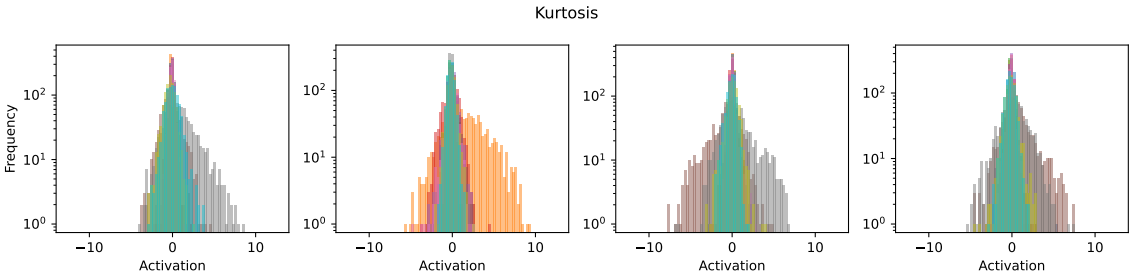
We consider the performance of gradient-based algorithms for projection pursuit in the planted vector setting where we study their sample complexity. Specifically, we consider the setting where the planted vector follows a distribution containing two clusters of imbalanced size or a Bernoulli-Rademacher distribution. In the former setting, Low-Degree Polynomials give a lower bound of $n = \widetilde{\Theta}(d^{1.5}p)$ and gradient-based methods can recover the signal direction provably with $n = \widetilde{\Omega}(d^2p^2)$ thus presenting a gap of a factor of $p\sqrt{d}$ which increases as p increases. It is currently unknown whether an algorithm exists which matches the sample complexity of the lower bound in the Low Degree Polynomial Framework. In the latter setting, $n = \widetilde{\Omega}(d^3p^4)$ samples are sufficient and there exist spectral algorithms matching the computational lower bounds of $n = \widetilde{\Omega}(d^2p^2)$ samples. Although there still exists a gap between gradient-based methods and computational lower bounds, we can observe that in both settings, if the distribution is very imbalanced/sparse, gradient-based methods match computational lower bounds closely. Finally, we demonstrate the performance of our gradient-based algorithm on the FashionMNIST dataset.

6. Acknowledgements

This work has been supported by the German Research Foundation (DFG) through DFG-ANR PRCI “ASCAI” (GH 257/3-1).



(a) 4 recovered projection using the ReLU2 projection index



(b) 4 recovered projection using the Kurtosis projection index

Fig. 4: Histograms of projections recovered by Algorithm 1 on FashionMNIST using $n = 600$ samples. Classes are plotted using different colors.

References

[1] Y. Bai, Q. Jiang, J. Sun, Subgradient descent learns orthogonal dictionaries, 2019.

[2] D. P. Bertsekas, Incremental least squares methods and the extended kalman filter, SIAM Journal on Optimization 6 (1996) 807–822.

[3] M. Breaban, H. Luchian, Outlier detection with nonlinear projection pursuit, International Journal of Computers Communications and Control 8 (2012) 30.

[4] D. Davis, M. Diaz, K. Wang, Clustering a mixture of gaussians with unknown covariance, arXiv preprint arXiv:2110.01602 (2021).

[5] R. Dudeja, D. Hsu, Statistical-computational trade-offs in tensor pca and related problems via communication complexity, 2024.

[6] J. Friedman, J. Tukey, A projection pursuit algorithm for exploratory data analysis, IEEE Transactions on Computers C-23 (1974) 881–890.

[7] S. B. Hopkins, T. Schramm, J. Shi, D. Steurer, Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors, 2016.

[8] S. B. K. Hopkins, Statistical inference and the sum of squares method, Ph.D. thesis, Cornell University, 2018.

[9] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Computation 9 (1997) 1483–1492.

[10] M. C. Jones, R. Sibson, What is projection pursuit?, Journal of the Royal Statistical Society. Series A (General) 150 (1987) 1–37.

[11] M. Kearns, Efficient noise-tolerant learning from statistical queries, in: Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, STOC '93, Association for Computing Machinery, New York, NY, USA, 1993, p. 392–401.

[12] D. Kunisky, A. S. Wein, A. S. Bandeira, Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio, 2019.

[13] N. Loperfido, Skewness-based projection pursuit: A computational approach, Computational Statistics and Data Analysis 120 (2018) 42–57.

[14] N. Loperfido, Kurtosis-based projection pursuit for outlier detection in financial time series, The European Journal of Finance 26 (2020) 142–164.

[15] C. Mao, A. S. Wein, Optimal spectral recovery of a planted vector in a subspace, 2022.

[16] R. Meka, A. Potechin, A. Wigderson, Sum-of-squares lower bounds for planted clique, in: Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, STOC '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 87–96.

[17] J. Neyman, E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, On the Problem of the Most Efficient Tests of Statistical Hypotheses, Springer New York, New York, NY, 1992, pp. 73–108.

[18] B. Olshausen, D. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–9.

[19] P. Paajarvi, J. LeBlanc, Skewness maximization for impulsive sources in blind deconvolution, Report - Helsinki University of Technology, Signal Processing Laboratory, volume 46, pp. 304– 307.

[20] J.-X. Pan, W.-K. Fung, K.-T. Fang, Multiple outlier detection in multivariate data using projection pursuit techniques, Journal of Statistical Planning and Inference 83 (2000) 153–167.

[21] Q. Qu, J. Sun, J. Wright, Finding a sparse vector in a subspace: Linear sparsity using alternating directions, CoRR abs/1412.4659 (2014).

[22] D. A. Spielman, H. Wang, J. Wright, Exact recovery of sparsely-used dictionaries, 2012.

[23] R. Vershynin, High-Dimensional Probability, 2020.

[24] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.
[25] I. Zadik, M. J. Song, A. S. Wein, J. Bruna, Lattice-based methods surpass sum-of-squares in clustering, 2022.
[26] V. Zarzoso, P. Comon, Comparative speed analysis of fastica, in: M. E. Davies, C. J. James, S. A. Abdallah, M. D. Plumbley (Eds.), Independent Component Analysis and Signal Separation, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 293–300.
[27] Y. Zhai, Z. Yang, Z. Liao, J. Wright, Y. Ma, Complete dictionary learning via l4-norm maximization over the orthogonal group, J. Mach. Learn. Res. 21 (2020).

Appendix A. Proof of Lemma 1

Proof. Start off by choosing an initialization $j \in [n_{init}]$ for which $\left\langle \mathbf{u}_{0,j}, \mathbf{u}^* \right\rangle \geq a$ which can be assumed to exist with probability at least $1 - o(1)$. By Precondition 2 we know that for each step $\left\langle \mathbf{u}_{t+1,j}, \mathbf{u}^* \right\rangle \geq (1 + c_0)\left\langle \mathbf{u}_{t,j}, \mathbf{u}^* \right\rangle$ if $\left\langle \mathbf{u}_{t,j}, \mathbf{u}^* \right\rangle < b$ with proabability at least $1 - o\left(\frac{1}{s}\right)$. By applying the union bound over all steps we can obtain that $s = \log_{1+c_0}\left(\frac{b}{a}\right) = O(\log(d))$ steps are sufficient, such that at least one direction $\mathbf{u}_{i,j}$ is encountered for which $\left\langle \mathbf{u}_{i,j}, \mathbf{u}^* \right\rangle \geq b$. By Precondition 3 we can observe that $\hat{\mathbf{u}} = \arg\max_{\mathbf{u} \in \{\mathbf{u}_{i,j} | i \in [s], j \in [n_{init}]\}} \sum_{k=1}^{n} \frac{\phi(\langle \mathbf{X}_k, \mathbf{u} \rangle)}{n}$ has $\langle \hat{\mathbf{u}}, \mathbf{u}^* \rangle \geq b - \delta$ finishing the proof.

□

Appendix B. Proofs in Subection 2.4

In this section use $\phi(x) = \psi(x) = \max\{0, x\}^2$. For simplicity we also define $\mu_1 = \sqrt{\frac{1-p}{p}}$ and $\mu_2 = -\sqrt{\frac{p}{1-p}}$.

Lemma 2. *Given* $\mathbf{u} \sim \mathcal{D}_{\mathcal{B}(p)}$ *for* $p \in (1/\sqrt{d}, 1/2)$ *with probability at least* $\Theta(p)$

$$\frac{\langle \mathbf{u}, \mathbf{u}^* \rangle}{\|\mathbf{u}\|_2} \geq \Theta\left(\frac{1}{\sqrt{pd}}\right)$$

Proof. We know that $\langle \mathbf{u}, \mathbf{u}^* \rangle = \sqrt{\frac{1-p}{p}} \geq \sqrt{\frac{1}{2p}}$ with probability p. Conditioned on $\langle \mathbf{u}^*, \mathbf{u} \rangle = \sqrt{\frac{1-p}{p}}$ we have $\mathbf{u} \sim \mathcal{N}\left(\sqrt{\frac{1-p}{p}}\mathbf{u}^*, I_n - \mathbf{u}^*\mathbf{u}^{*\top}\right)$. By Markovs inequality we have $\left\|(I_d - \mathbf{u}^*\mathbf{u}^{*\top})\mathbf{u}\right\|_2 \leq \Theta\left(\sqrt{d}\right)$ with constant probability. The lemma follows by combining the previous two results.

□

Lemma 3. *For any* $\delta \in (0, 1)$ *and* $\mathbf{X} \sim \mathcal{D}^n_{\mathcal{B}(p)}$ *have*

$$\mathbb{P}\left[\left|\left\langle \sum_{i=1}^{n} \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle - \left\langle \mathbb{E}\left[g_{\mathbf{u}}(\mathbf{X})\right], \mathbf{u}^* \right\rangle\right| \geq \delta\right] \leq \exp\left(-\Theta\left(n\delta^2 \min\left\{1, \frac{p}{\langle \mathbf{u}, \mathbf{u}^* \rangle^2}\right\}\right)\right)$$

Proof. For simplicity we will split $\langle g_{\mathbf{u}}(\mathbf{x}), \mathbf{u}^* \rangle$ into terms t_1, t_2 such that

$$\langle g_{\mathbf{u}}(\mathbf{x}), \mathbf{u}^* \rangle = \underbrace{\max\{0, \langle \mathbf{x}, \mathbf{u} \rangle\} \langle \mathbf{x}, \mathbf{u}^* \rangle}_{=t_1(\mathbf{x})} - \underbrace{\max\{0, \langle \mathbf{x}, \mathbf{u} \rangle\} \langle \mathbf{u}, \mathbf{u}^* \rangle \langle \mathbf{x}, \mathbf{u} \rangle}_{=t_2(\mathbf{x})}$$

Here we use the Sub-Gaussian norm $\|\cdot\|_{\psi_2}$ and Sub-Exponential norm $\|\cdot\|_{\psi_1}$ as defined in [23].

Next we will choose a fixed sample $\hat{\nu} \sim \mathcal{B}(p)$ such that $\langle \mathbf{X}_i, \mathbf{u}^* \rangle = \hat{\nu}_i$ and define the *empirical* balance as $\hat{p} = \sum_{i=1}^{n} \frac{\mathbb{1}_{\hat{\nu}=\mu_1}}{n}$. For now we will assume $(1 - \bar{\delta})p \leq \hat{p} \leq (1 + \bar{\delta})p$.

Bounding $\sum_{i=1}^{n} t_1(\mathbf{X}_i)$. First note that if the random variable Z is Sub-Gaussian then there exists a constant $C_{centering}$ such that $\|Z - \mathbb{E}[Z]\|_{\psi_2} \leq C_{centering} \|Z\|_{\psi_2}$. Also note that $\|\max\{c, Z\}\|_{\psi_2} \leq \|Z\|_{\psi_2}$.

Thus $\sum_{i=1}^{n} \left\| \frac{t_1(\mathbf{X}_i)}{n} - \mathbb{E}\left[\frac{t_1(\mathbf{X}_i)}{n}\right] \right\|^2_{\psi_2} \leq \hat{p}\frac{\mu_1^2}{n} + (1 - \hat{p})\frac{\mu_2^2}{n} \overset{(1)}{\leq} \Theta\left(\frac{1}{n}\right)$. With (1) following from $\hat{p} \leq (1 + \bar{\delta})p$. Thus

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} \frac{t_1(\mathbf{X}_i)}{n} - \mathbb{E}\left[\sum_{i=1}^{n} \frac{t_1(\mathbf{X}_i)}{n}\right]\right| \leq \delta\right] \geq 1 - \exp\left(-\Theta\left(n\delta^2\right)\right)$$

Bounding $\sum_{i=1}^n t_2(\mathbf{X}_i)$. Note that if the random variable Z is Sub-Exponential then there exists a constant $C_{centering}$ such that $\|Z - \mathbb{E}[Z]\|_{\psi_1} \le C_{centering}\|Z\|_{\psi_1}$. Additionally note if $\|Z - \mathbb{E}[Z]\|_{\psi_2} = K$ and $\mathbb{E}[Z] < 0$ then $\|\max\{0, Z\}\|_{\psi_2} \le K$.

Thus $\sum_{i=1}^n \left\| \frac{t_2(\mathbf{X}_i)}{n} - \mathbb{E}\left[\sum_{i=1}^n \frac{t_2(\mathbf{X}_i)}{n}\right]\right\|_{\psi_1}^2 \le \hat{p}\frac{\mu_1^4}{n^3} + (1-\hat{p})\frac{\mu_2^4}{n^3} \le \Theta\left(\frac{1}{n^2}\right)$ for $n > d$ and $\max_{i\in[n]} \|t_2(\mathbf{X}_i)\|_{\psi_1} \le \Theta\left(\frac{1}{n}\right)$. Thus we can bound by Bernstein's inequality for sufficiently large $n \ge \frac{1}{\delta}$

$$\mathbb{P}\left[\left|\sum_{i=1}^n \frac{t_2(\mathbf{X}_i)}{n} - \mathbb{E}\left[\sum_{i=1}^n \frac{t_2(\mathbf{X}_i)}{n}\right]\right| \le \delta\right] \ge 1 - \exp\left(-\Theta\left(\min\left\{\frac{\delta^2}{\sum_{i=1}^n \|t_2(\mathbf{X}_i)\|_{\psi_1}^2}, \frac{\delta}{\max_{i\in[n]}\|t_2(\mathbf{X}_i)\|_{\psi_1}}\right\}\right)\right)$$
$$\ge 1 - \exp\{-\Theta(n\delta)\}$$

Now what is left to do is to bound $\langle g_{\mathbf{u}}(\mathbf{x}), \mathbf{u}^*\rangle$ with a change in $\hat{p}$.

$$\left|\mathbb{E}\left[\sum_{i=1}^n \langle g_{\mathbf{u}}(\mathbf{x}), \mathbf{u}^*\rangle\right] - \mathbb{E}\left[\sum_{i=1}^n \langle g_{\mathbf{u}}(\mathbf{x}), \mathbf{u}^*\rangle \middle| \hat{p} = (1-\bar{\delta})p\right]\right| \le \Theta\left(\langle \mathbf{u}, \mathbf{u}^*\rangle\bar{\delta}\right)$$

By applying the Chernoff bound for the Binomial distribution we obtain

$$\mathbb{P}[(1-\bar{\delta})p \le \hat{p} \le (1+\bar{\delta})p] \ge 1 - 2\exp\left(-\frac{\bar{\delta}^2 np}{3}\right)$$

for all $\bar{\delta} \in (0,1)$. The lemma follows by choosing $\bar{\delta} = \Theta\left(\frac{\delta}{\langle \mathbf{u}, \mathbf{u}^*\rangle}\right)$ and applying the union bound.

□

Lemma 4. *For arbitrary $\beta > 0$ there exist constants $t, p_0 > 0$ such that with $\mathbf{X} \sim \mathcal{D}_{\mathcal{B}(p)}$ for $p < p_0$ we have for $\langle \mathbf{u}, \mathbf{u}^*\rangle(\mu_1 - \mu_2) < t$*

$$\mathbb{E}[\langle g_{\mathbf{u}}(\mathbf{X}), \mathbf{u}^*\rangle] \ge \Theta\left(\frac{\langle \mathbf{u}, \mathbf{u}^*\rangle^2}{\sqrt{p}}\right)$$

and for $\frac{t}{(\mu_1-\mu_2)} \le \langle \mathbf{u}, \mathbf{u}^*\rangle \le 1 - \beta$

$$\mathbb{E}[\langle g_{\mathbf{u}}(\mathbf{X}_i), \mathbf{u}^*\rangle] \ge \Theta\left(\langle \mathbf{u}, \mathbf{u}^*\rangle\right)$$

Proof. For simplicity abbreviate $a_1 = \langle \mathbf{u}, \mathbf{u}^*\rangle$ and $a_2 = \sqrt{1 - \langle \mathbf{u}, \mathbf{u}^*\rangle^2}$. Define $\mathbf{e}_2 = \frac{\mathbf{u} - a_1\mathbf{u}^*}{a_2}$. Since $\langle \mathbf{x}, \mathbf{e}_2\rangle \sim \mathcal{N}(0,1)$ we know $\mathbb{E}[\max\{0, \langle \mathbf{x}, \mathbf{e}_2\rangle\}^2] = \frac{1}{2}$ and $\mathbb{E}[\max\{0, \langle \mathbf{x}, \mathbf{e}_2\rangle\}] = \sqrt{\frac{1}{2\pi}}$ by applying the expectation of the half normal distribution. and $f_{\langle \mathbf{x}, \mathbf{e}_2\rangle}(y) \ge \underline{p}$ for $y \in [-t_0, t_0]$.

Abbreviate $f_i(x) = f_{\langle \mathbf{x}, \mathbf{e}_2\rangle}\left(\frac{x - \mu_i a_1}{a_2}\right) = f_{\mathcal{N}(0,1)}\left(\frac{x-\mu_i a_1}{a_2}\right)$. By applying Lemma 20 have

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_p}[\langle g_{\mathbf{u}}(\mathbf{x}), \mathbf{u}^*\rangle] = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_p}(\langle \mathbf{x}, \mathbf{u}\rangle\langle \mathbf{x}, \mathbf{u}^*\rangle - \langle \mathbf{x}, \mathbf{u}\rangle^2 a_1)$$
$$= \int_0^\infty p(x\mu_1 - x^2 a_1)f_1(x) + (1-p)(x\mu_2 - x^2 a_1)f_2(x)dx$$
$$= \int_0^{(\mu_1-\mu_2)a_1} p\mu_1 x f_1(x)dx + \sqrt{(1-p)p}\int_0^\infty (x + (\mu_1-\mu_2)a_1 - x)\, f_2(x)dx - a_1\int_0^\infty x^2(pf_1(x) + (1-p)f_2(x))dx$$

We continue by bounding each term.

Case 1. If $(\mu_1 - \mu_2)a_1 \le t_0$.

$$\int_0^{(\mu_1-\mu_2)a_1} p\mu_1 x f_1(x)dx \ge \frac{\underline{p}p\mu_1}{2}((\mu_1-\mu_2)a_1)^2 \ge \frac{\underline{p}a_1^2}{2}\left(\frac{(1-p)^{1.5}}{\sqrt{p}} + \frac{p^{1.5}}{\sqrt{1-p}} - 2\sqrt{p(1-p)}\right) \ge \Theta\left(\frac{a_1^2}{\sqrt{p}}\right)$$

Case 2. If $(\mu_1 - \mu_2)a_1 > t_0$.

$$\begin{aligned} p\mu_1 \int_0^{(\mu_1-\mu_2)a_1} x f_1(x)dx &\geq \sqrt{(1-p)p} \int_{(\mu_1-\mu_2)a_1-t_0}^{(\mu_1-\mu_2)a_1} x f_1(x)dx \\ &\geq \underline{p}\sqrt{(1-p)p}\,((\mu_1-\mu_2)a_1)^2 - ((\mu_1-\mu_2)a_1 - t_0)^2 \\ &\geq \underline{p}\sqrt{(1-p)p}\left(2t_0(\mu_1-\mu_2)a_1 - t_0^2\right) \\ &\geq \underline{p}t_0\sqrt{(1-p)p}(\mu_1-\mu_2)a_1 \\ &\geq \underline{p}t_0 a_1 \end{aligned}$$

For $p < p_0$ the following terms can be bounded as such

$$\sqrt{(1-p)p}\int_0^{\infty}((\mu_1-\mu_2)a_1)\,f_2(x)dx = a_1\int_0^{\infty} f_2(x)dx \geq a_1\max\left\{0, \frac{1}{2} - \frac{2}{\sqrt{\pi}}\frac{\sqrt{\frac{p}{1-p}}a_1}{a_2}\right\}$$

and

$$\begin{aligned} &a_1\int_0^{\infty} x^2(pf_1(x) + (1-p)f_2(x))dx \\ \leq &a_1\left(p\int_0^{\mu_1 a_1} x^2 f_1(x)dx + p\int_0^{\infty}(x+\mu_1 a_1)^2 f_{\langle \mathbf{x},\mathbf{e}_2\rangle}\left(\frac{x}{a_2}\right)dx + \frac{(1-p)(1-a_1^2)}{2}\right) \\ = &a_1\left(p\int_0^{\mu_1 a_1} x^2 f_1(x)dx + p\int_0^{\infty} x^2 f_{\langle \mathbf{x},\mathbf{e}_2\rangle}\left(\frac{x}{a_2}\right)dx + p\int_0^{\infty} 2x\mu_1 a_1 f_{\langle \mathbf{x},\mathbf{e}_2\rangle}\left(\frac{x}{a_2}\right)dx + p\,(\mu_1 a_1)^2\int_0^{\infty} f_{\langle \mathbf{x},\mathbf{e}_2\rangle}\left(\frac{x}{a_2}\right)dx + \frac{(1-p)(1-a_1^2)}{2}\right) \\ \leq &a_1\left[p\left(\frac{2}{3}(\mu_1 a_1)^3 + \frac{(1-a_1^2)}{2} + 2\sqrt{\frac{1}{2\pi}}\mu_1 a_1 + (\mu_1 a_1)^2\frac{1}{2}\right) + \frac{(1-p)(1-a_1^2)}{2}\right] \\ \leq &\frac{a_1 a_2^2}{2} + \Theta\left(\sqrt{p}a_1^2\right) \end{aligned}$$

Thus in Case 1 have.

$$\mathbb{E}[\langle g_{\mathbf{u}}(\mathbf{x}), \mathbf{u}^*\rangle] \geq \Theta\left(\frac{a_1^2}{\sqrt{p}}\right) + \frac{a_1}{a_2}\left(\frac{1}{2} - 2\frac{\sqrt{\frac{p}{1-p}}a_1}{a_2}\right) - \frac{a_1 a_2^2}{2} - \Theta\left(\sqrt{p}a_1^2\right) \geq \Theta\left(\frac{a_1^2}{\sqrt{p}}\right)$$

And analogously in Case 2.

$$\mathbb{E}[\langle g_{\mathbf{u}}(\mathbf{x}), \mathbf{u}^*\rangle] \geq \Theta(a_1)$$

□

Lemma 5. *For $n = \Omega(d)$ have*

$$\mathbb{P}\left[\max_{\hat{\mathbf{u}}\in\mathbb{S}_{d-1}}\left|\sum_{i=1}^{n}\frac{\psi(\langle \mathbf{X}_i, \hat{\mathbf{u}}\rangle)}{n} - \mathbb{E}[\psi(\langle \mathbf{X}, \hat{\mathbf{u}}\rangle)]\right| \geq \delta\right] \leq \exp\left(-\Theta\left(\frac{n\delta^2}{d\log(1/\delta)}\right)\right)$$

Proof. First make the distinction into $\mathbf{Y}_j \sim \mathcal{D}_{p,j}^{n_j}$ for $j \in \{1,2\}$, where $\sum_{j\in\{1,2\}} n_j = n$ We will first note that $\mathbf{A}_j(\mathbf{u}) = \psi(\mathbf{Y}_j\mathbf{u}) - \mathbb{E}[\psi(\mathbf{Y}_j\mathbf{u})]$ is subexponential and thus $\left\| \sum_{j\in\{1,2\}} \sum_{i=1}^{n_j} \frac{\mathbf{A}_{j,i}(\mathbf{u})}{n_j} \right\|_{\psi_1} \leq \Theta\left(\frac{1}{n}\right)$. Thus we can obtain

$$\mathbb{P}\left[\left| \sum_{j\in\{1,2\}} \sum_{i=1}^{n_j} \frac{\mathbf{A}_{j,i}(\mathbf{u})}{n} \right| \geq \delta_c\right] \leq \exp\left(-\Theta\left(\delta_c^2 n\right)\right)$$

Let $\mathcal{N}_{\epsilon,d}$ be the minimum size ϵ-Net of the d-dimensional unit sphere. Thus we know $\left|\mathcal{N}_{\epsilon,d}\right| \leq \left(\frac{3}{\epsilon}\right)^d$.

$$\mathbb{P}\left[\max_{\mathbf{u}\in\mathcal{N}_{\epsilon,d}} \left| \sum_{j\in\{1,2\}} \sum_{i=1}^{n_j} \frac{\mathbf{A}_{j,i}(\mathbf{u})}{n} \right| \geq \delta_c\right] \leq 2\exp\left(d\log\left(\frac{3}{\epsilon}\right) - \Theta\left(\delta_c^2 n\right)\right)$$

Next we bound the maximum deviation for $\mathbf{u}$ for $\|\mathbf{u}' - \mathbf{u}\|_2 < \epsilon$

$$\left|\sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \mathbf{u}\rangle)}{n} - \sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \mathbf{u}'\rangle)}{n}\right| \leq \left(\sqrt{\sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \mathbf{u}\rangle)}{n}} + \sqrt{\sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \mathbf{u}'\rangle)}{n}}\right)\left|\sqrt{\sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \mathbf{u}\rangle)}{n}} - \sqrt{\sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \mathbf{u}'\rangle)}{n}}\right|$$
$$\leq \frac{2\|\mathbf{X}\|_{op}\|\mathbf{X}(\mathbf{u}-\mathbf{u}')\|_2}{n} \leq \frac{2\epsilon\|\mathbf{X}\|_{op}^2}{n}$$

We continue by bounding for all $\mathbf{u} \in \mathbb{S}_{d-1}$ by applying the triangle inequality

$$\left| \sum_{j\in\{1,2\}} \sum_{i=1}^{n_j} \frac{\mathbf{A}_{j,i}(\mathbf{u})}{n} \right| \leq \max_{\mathbf{u}'\in\mathcal{N}_{\epsilon,d}} \left| \sum_{j\in\{1,2\}} \sum_{i=1}^{n_j} \frac{\mathbf{A}_{j,i}(\mathbf{u})}{n} \right| + \frac{2\epsilon\|\mathbf{X}\|_{op}^2}{n}$$

Finally we notice that $\mathbf{X}$ can be decomposed into the union of $\mathbf{Y}_1, \mathbf{Y}_2$ with $n_1 + n_2 = n$, where $n_1 \sim \text{Binomial}(n, p)$. Thus by the Chernoff bound for the Binomial distribution we have

$$\mathbb{P}\left[\left|\mathbb{E}[\psi(\langle \mathbf{X}, \mathbf{u}\rangle)] - \sum_{j\in\{1,2\}} \frac{n_j}{n}\mathbb{E}\left[\psi\left(\left\langle \mathbf{Y}_j, \mathbf{u}\right\rangle\right)\right]\right| \geq \delta_p\right] \leq 2\exp\left(-\Theta\left(\delta_p^2 np\right)\right)$$

Thus we can prove the Theorem by choosing $\epsilon = \Theta(\delta)$ for $n > d > 1$ and $\delta < 1$. The second to last step follows bounding $\|\mathbf{X}\|_{op} \leq \left\|\mathbf{X}(\mathbf{u}^*\mathbf{u}^{*\top})\right\|_{op} + \left\|\mathbf{X}(I - \mathbf{u}^*\mathbf{u}^{*\top})\right\|_{op} \leq \Theta\left(\sqrt{n} + t\right)$ with probability $1 - 2\exp(-t^2)$ [23].

$$\begin{aligned}
&\mathbb{P}\left[\max_{\mathbf{u}\in\mathbb{S}_{d-1}} \left|\sum_{i=1}^n \frac{\psi(\mathbf{X}_i\mathbf{u})}{n} - \mathbb{E}[\psi(\mathbf{X}\mathbf{u})]\right| \geq \delta\right] \\
\leq &\mathbb{P}\left[\max_{\mathbf{u}\in\mathcal{N}_{\epsilon,d}} \left| \sum_{j\in\{1,2\}} \sum_{i=1}^{n_j} \frac{\mathbf{A}_{j,i}}{n} \right| \geq \frac{\delta}{3}\right] + \mathbb{P}\left[\frac{2\epsilon\|\mathbf{X}\|_{op}}{n} \geq \frac{\delta}{3}\right] + \mathbb{P}\left[\left|\mathbb{E}[\psi(\langle \mathbf{X}, \mathbf{u}\rangle)] - \sum_{j\in\{1,2\}} \mathbb{E}\left[\psi\left(\left\langle \mathbf{Y}_j, \mathbf{u}\right\rangle\right)\right]\right| \geq \frac{\delta}{3}\right] \\
\leq &\exp\left(d\log\left(\frac{3}{\epsilon}\right) - \Theta\left(\delta^2 n\right)\right) + 2\exp(-n) + \exp\left(-\Theta\left(\delta^2 np\right)\right) \\
\leq &\exp\left(-\Theta\left(\frac{n\delta^2}{d\log(1/\delta)}\right)\right)
\end{aligned}$$

□

Lemma 6 (Testing)**.** *For any $\langle \mathbf{u}, \mathbf{u}^*\rangle > \langle \mathbf{u}', \mathbf{u}^*\rangle \geq a_{min} > 0$ there exists $\bar{p} > 0$ such that there exists a threshold t such that with probability at least $1 - \exp\left(-\Theta\left(\frac{n\Delta^2}{d\log(1/\Delta)}\right)\right)$ where $\Delta = \langle \mathbf{u}, \mathbf{u}^*\rangle - \langle \mathbf{u}', \mathbf{u}^*\rangle$ we have*

$$\sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \mathbf{u}\rangle)}{n} \geq t \text{ and } \sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \mathbf{u}'\rangle)}{n} \leq t$$

Proof. We begin by bounding the expectation of the score function.

$$\begin{aligned}\mathbb{E}[\phi(\langle \mathbf{X}, \mathbf{u}\rangle)] =& p\int_0^\infty \left(x^2 f_{\mathcal{N}(0,1)}\left(\frac{x - \langle \mathbf{u}, \mathbf{u}^*\rangle\sqrt{(1-p)/p}}{\sqrt{1-\langle \mathbf{u}, \mathbf{u}^*\rangle^2}}\right)dx\right) + (1-p)\int_0^\infty \left(x^2 f_{\mathcal{N}(0,1)}\left(\frac{x - \langle \mathbf{u}, \mathbf{u}^*\rangle\sqrt{p/(1-p)}}{\sqrt{1-\langle \mathbf{u}, \mathbf{u}^*\rangle^2}}\right)dx\right)\\ =&(1-p)\langle \mathbf{u}, \mathbf{u}^*\rangle^2 + p(1-\langle \mathbf{u}, \mathbf{u}^*\rangle^2) - p\int_{-\infty}^0 \left(x^2 f_{\mathcal{N}(0,1)}\left(\frac{x - \langle \mathbf{u}, \mathbf{u}^*\rangle\sqrt{(1-p)/p}}{\sqrt{1-\langle \mathbf{u}, \mathbf{u}^*\rangle^2}}\right)dx\right)\\ &+ (1-p)\int_0^\infty \left(x^2 f_{\mathcal{N}(0,1)}\left(\frac{\langle \mathbf{u}, \mathbf{u}^*\rangle\sqrt{p/(1-p)}}{\sqrt{1-\langle \mathbf{u}, \mathbf{u}^*\rangle^2}}\right)dx\right)\end{aligned}$$

Thus for each $c_1 > 0$ there exists a $\bar{p} > 0$ such that for all $p \le \bar{p}$ and $\langle \mathbf{u}, \mathbf{u}^*\rangle \ge l$

$$\langle \mathbf{u}, \mathbf{u}^*\rangle^2(1-p) + (1-\langle \mathbf{u}, \mathbf{u}^*\rangle^2)\left(\frac{1+p}{2} - c_1\right) \le \mathbb{E}[\phi(\langle \mathbf{X}, \mathbf{u}\rangle)] \le \langle \mathbf{u}, \mathbf{u}^*\rangle^2(1-p) + (1-\langle \mathbf{u}, \mathbf{u}^*\rangle^2)\frac{1+p}{2}$$

and for $\langle \mathbf{u}, \mathbf{u}^*\rangle \le 0$ have $\mathbb{E}[\phi(\langle \mathbf{X}, \hat{\mathbf{u}}\rangle)] \le \frac{1}{2}$.

Thus we can bound the difference in expectation for $\mathbf{u}$ and $\mathbf{u}'$

$$\mathbb{E}[\phi(\langle \mathbf{X}, \mathbf{u}\rangle)] - \mathbb{E}[\phi(\langle \mathbf{X}, \mathbf{u}'\rangle)] \ge (\langle \mathbf{u}, \mathbf{u}^*\rangle^2 - \langle \mathbf{u}', \mathbf{u}^*\rangle^2)\frac{1-3p}{2} - c_1(1-\langle \mathbf{u}, \mathbf{u}^*\rangle^2)$$

Thus $\mathbf{u}$ and $\mathbf{u}'$ there exists a $\bar{p} > 0$ such that $\mathbb{E}[\phi(\langle \mathbf{X}, \mathbf{u}\rangle)] - \mathbb{E}[\phi(\langle \mathbf{X}, \mathbf{u}'\rangle)] > 0$ By Lemma 5 we know that $\left|\sum_{i=1}^n \frac{\phi(\langle \mathbf{X}_i, \mathbf{u}\rangle)}{n} - \mathbb{E}[\phi(\langle \mathbf{X}, \mathbf{u}\rangle)]\right| \le \delta$ with probability at least $1 - \exp\left(-\Theta\left(\frac{n\delta^2}{d\log(1/\delta)}\right)\right)$. The lemma follows by choosing $\delta = \frac{|\mathbb{E}[\phi(\langle \mathbf{X}, \mathbf{u}\rangle)] - \mathbb{E}[\phi(\langle \mathbf{X}, \mathbf{u}'\rangle)]|}{2}$ and t accordingly. □

Appendix B.1. Proof of Theorem 1

As previously discussed we will split the proof of Theorem 1 into parts for each execution of Algorithm 2.

Proof. Combining Lemma 4 and Lemma 3 and choosing $\delta = \Theta\left(\min\left\{\frac{\langle \mathbf{u}, \mathbf{u}^*\rangle^2}{\sqrt{p}}, \langle \mathbf{u}, \mathbf{u}^*\rangle\right\}\right)$ we obtain

$$\sum_{i=1}^n \left\langle \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^*\right\rangle \ge \Theta\left(\min\left\{\frac{\langle \mathbf{u}, \mathbf{u}^*\rangle^2}{\sqrt{p}}, \langle \mathbf{u}, \mathbf{u}^*\rangle\right\}\right)$$

with probability at least $1 - \Theta\left(\frac{1}{s}\right)$ if $\langle \mathbf{u}, \mathbf{u}^*\rangle \ge a = \Theta\left(\frac{1}{\sqrt{dp}}\right)$ and $n = \widetilde{\Omega}\left(d^2p^2\right)$. Next notice that with probability at least $1 - \Theta\left(\frac{1}{s}\right)$.

$$\left\|\frac{\phi'(\mathbf{X}\mathbf{u})}{n}\right\|_2 \le \left\|\frac{\mathbf{X}\mathbf{u}}{n}\right\|_2 \le \left\|\frac{(I_d - \mathbf{u}^*\mathbf{u}^{*\top})\mathbf{X}}{n}\right\|_{op} + \|\mathbf{X}\mathbf{u}^*\|_2 \le \Theta\left(\frac{1}{\sqrt{n}}\right)$$

By utilizing that $(I_d - \mathbf{u}^*\mathbf{u}^{*\top})\mathbf{X}$ is a gaussian random matrix and that $\|\mathbf{X}\mathbf{u}^*\|_2^2$ follows a binomial(scaled) distribution. Thus we can bound using Lemma 17

$$\left\|\frac{\phi'(\mathbf{X}\mathbf{u})^\top \mathbf{X}(I_d - \mathbf{u}^*\mathbf{u}^{*\top} - \mathbf{e}_2\mathbf{e}_2^\top)}{n}\right\|_2 \le \Theta\left(\sqrt{\frac{d}{n}}\right)$$

Thus by applying Lemma 16 we obtain probability $1 - \Theta\left(\frac{1}{s}\right)$

$$\left\|\frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}\right\|_2 \le \sqrt{\left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^*\right\rangle^2\left(1 + \frac{\langle \mathbf{u}, \mathbf{u}^*\rangle^2}{1-\langle \mathbf{u}, \mathbf{u}^*\rangle^2}\right) + \Theta\left(\frac{d}{n}\right)} \tag{B.1}$$

First Execution of Algorithm 2 *There exist parameters t_1 and $\eta_1 = \Omega(\sqrt{d}p)$ such that for $a_1 = \Theta\left(\frac{1}{\sqrt{pd}}\right)$ there exists a constant $b_1 \in (0,1)$ such that the first execution of Algorithm 2 fulfills the criteria of Lemma 1 for $n = \widetilde{O}(d^2p^2)$.*

By Lemma 2 we have $\langle \mathbf{u}, \mathbf{u}^* \rangle \geq \Theta\left(\frac{1}{\sqrt{pd}}\right)$ and thus fulfilling Precondition 1.

Using (B.1) we can bound

$$\left\| \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2 \leq \max\left\{ 2\sqrt{\left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle^2 \left(1 + \frac{\langle \mathbf{u}, \mathbf{u}^* \rangle^2}{1 - \langle \mathbf{u}, \mathbf{u}^* \rangle^2}\right)}, \Theta\left(\sqrt{\frac{d}{n}}\right) \right\} \tag{B.2}$$

Thus for all $\langle \mathbf{u}, \mathbf{u}^* \rangle \in (a_1, b_1)$ and $n = \widetilde{\Omega}\left(d^2p^2\right)$ we have

$$\langle \mathbf{u}, \mathbf{u}^* \rangle^{-1} \sum_{i=1}^n \left\langle \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle \geq \left\| \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2$$

This can be verified by using the bound (B.2). By applying Lemma 18 we can verify that Precondition 2 is fulfilled. Finally, using Lemma 6 shows that Precondition 3 is fulfilled.

Second Execution of Algorithm 2 *There exist parameters t_2 and η_2 such that for $a_2 = b_1 - \delta > 0$ and $b_2 = 1 - \beta$ for some $\epsilon > 0$ and $\delta > 0$ the second execution of Algorithm 2 fulfills the criteria of Lemma 1 for $n = \widetilde{O}(d^2p^2)$.*

As a result of the first execution of Algorithm 2 there exists a $\mathbf{u}$ such that $\langle \mathbf{u}, \mathbf{u}^* \rangle \geq b_1 - \delta$ and thus Precondition 1 is fulfilled. For all $\mathbf{u}$ for which $\langle \mathbf{u}, \mathbf{u}^* \rangle \in (a, b)$ there exists a constant choice for $\eta_2 > 0$ such that $\left\| \eta_2 \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2^2 \leq \eta_2 \langle \mathbf{u}, \mathbf{u}^* \rangle^{-1} \left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle$. By applying Lemma 19 we can show Precondition 2 is fulfilled. Finally, Precondition 3 can be fulfilled by Lemma 6.

□

Appendix C. Proofs in Subection 2.5

In this section use $\phi(x) = x^4$ and $\psi(x) = -|x|$.

Lemma 7. *Have $p < \frac{1}{3}$. Then*

$$\mathbb{E}[\langle g_{\mathbf{u}}(\mathbf{X}), \mathbf{u}^* \rangle] \geq \Omega\left(\frac{\langle \mathbf{u}, \mathbf{u}^* \rangle^3}{p}\right)$$

Proof. Define $\mu_0 = 0$ and $\mu_i = \frac{i}{\sqrt{p}}$ and $p_0 = 1 - p$ and $p_i = \frac{p}{2}$ for $i \in \{1, -1\}$. Thus

$$\begin{aligned}
&\mathbb{E}[\langle g_{\mathbf{u}}(\mathbf{X}), \mathbf{u}^* \rangle] \\
&= 4 \sum_{i \in \{-1,0,1\}} p_i \left(3\langle \mathbf{u}, \mathbf{u}^* \rangle (1 - \langle \mathbf{u}, \mathbf{u}^* \rangle^2)^2 (1 - \mu_i^2) + \mu_i^2 \langle \mathbf{u}, \mathbf{u}^* \rangle^3 (1 - \langle \mathbf{u}, \mathbf{u}^* \rangle^2)(\mu_i^2 - 3)\right) \\
&= 4(1 - \langle \mathbf{u}, \mathbf{u}^* \rangle^2) \langle \mathbf{u}, \mathbf{u}^* \rangle^3 \left(\frac{1}{p} - 3\right) \\
&\geq \Omega\left(\frac{\langle \mathbf{u}, \mathbf{u}^* \rangle^3}{p}\right)
\end{aligned}$$

□

Lemma 8. *For $\delta \in (0,1)$ have*

$$\mathbb{P}\left[\left|\left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle - \left\langle \mathbb{E}\left[g_{\mathbf{u}}(\mathbf{X})\right], \mathbf{u}^* \right\rangle\right| \geq \delta\right] \leq \exp\left(-\Theta\left(\frac{n\delta^2}{\log(n\log(s))^2 \left(1 + \max\left\{\langle \mathbf{u}, \mathbf{u}^* \rangle^4 p^{-1}, \langle \mathbf{u}, \mathbf{u}^* \rangle^8 p^{-2}\right\}\right)}\right)\right)$$

Proof. Choose an ortho-normal basis $\mathbf{E} = (\mathbf{u}^*, \mathbf{e}_2, ..., \mathbf{e}_d) \in \mathbb{R}^{d\times d}$ such that $\mathbf{u} = \langle \mathbf{u}, \mathbf{u}^*\rangle \mathbf{u}^* + \sqrt{1 - \langle \mathbf{u}, \mathbf{u}^*\rangle^2}\mathbf{e}_2$. Next we will condition on $\hat{\nu}_i = \langle \mathbf{X}_i, \mathbf{u}^*\rangle$ for all i where $\hat{\nu} \in \{-\sqrt{1/p}, 0, \sqrt{1/p}\}$. For convenience we define $\hat{p} = \sum_{i=1}^n \frac{I_{\hat{\nu}_i \neq 0}}{n}$ as well as $n_j = \sum_{i=1}^n I_{\hat{\nu}_i = j\sqrt{1/p}}$ with $j \in \{-1, 0, 1\}$. Additionally define $\mathbf{Y}_j = \left(\mathbf{X} \middle| \langle \mathbf{X}, \mathbf{u}^*\rangle = j\sqrt{\frac{1}{p}}\right)$.

Next observe that $\max_{i\in[n]} |\langle \mathbf{X}_i, \mathbf{e}_2\rangle| \leq \Theta\left(\log(n\log(s))\right)$ with probability at least $1 - \Theta\left(\frac{1}{s}\right)$ by applying the union bound. Thus for all j we have with probability at least $1 - \Theta\left(\frac{1}{s}\right)$

$$\left|\mathbb{E}\left[\left\langle g_{\mathbf{u}}(\mathbf{Y}_j), \mathbf{u}^*\right\rangle\right] - \left\langle g_{\mathbf{u}}(\mathbf{Y}_j), \mathbf{u}^*\right\rangle\right| \leq \Theta\left(\max\left\{1, \langle \mathbf{u}, \mathbf{u}^*\rangle^2\left(j\frac{1}{\sqrt{d}}\right)^2, \langle \mathbf{u}, \mathbf{u}^*\rangle^4\left(j\frac{1}{\sqrt{d}}\right)^3\right\}\log(n\log(s))\right)$$

by By Hoeffdings inequality we have

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^n \frac{\langle g_{\mathbf{u}}(\mathbf{X}_i), \mathbf{u}^*\rangle}{n} - \sum_{j\in\{-1,0,1\}} \frac{n_j\mathbb{E}\left[\left\langle g_{\mathbf{u}}(\mathbf{Y}_j), \mathbf{u}^*\right\rangle\right]}{n}\right| \geq \delta\right) &\leq \exp\left(-\Theta\left(\frac{\delta^2}{\log(n\log(s))^2\sum_{i=1}^n \frac{\max\left\{1, \langle \mathbf{u},\mathbf{u}^*\rangle^4 v_i^4, \langle \mathbf{u},\mathbf{u}^*\rangle^8 v_i^6\right\}}{n^2}}\right)\right) \\ &\leq \exp\left(-\Theta\left(\frac{n\delta^2}{\log(n\log(s))^2\left(1 + \hat{p}\max\left\{\langle \mathbf{u}, \mathbf{u}^*\rangle^4 p^{-2}, \langle \mathbf{u}, \mathbf{u}^*\rangle^8 p^{-3}\right\}\right)}\right)\right)\end{aligned}$$

By applying the Chernoff bound for the Binomial distribution we obtain

$$\mathbb{P}\left[\left|\sum_{j\in\{-1,0,1\}} \frac{n_j\mathbb{E}\left[\left\langle g_{\mathbf{u}}(\mathbf{Y}_j), \mathbf{u}^*\right\rangle\right]}{n} - \mathbb{E}\left[\sum_{i=1}^n \langle g_{\mathbf{u}}(\mathbf{X}_i), \mathbf{u}^*\rangle\right]\right| \geq \delta\right] \leq \exp\left(-\Theta\left(\frac{np\delta^2}{\max\left\{\langle \mathbf{u}, \mathbf{u}^*\rangle, \frac{\langle \mathbf{u},\mathbf{u}^*\rangle^3}{p}\right\}^2}\right)\right)$$

Finally we can combine all bounds to obtain the full statement. For any constant $\bar{\delta} \in (0, 1)$ have

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{i=1}^n \frac{\langle g_{\mathbf{u}}(\mathbf{X}_i), \mathbf{u}^*\rangle}{n} - \mathbb{E}[\langle g_{\mathbf{u}}(\mathbf{X}), \mathbf{u}^*\rangle]\right| \geq \delta\right) \\ &\leq \mathbb{P}\left(\left|\sum_{i=1}^n \frac{\langle g_{\mathbf{u}}(\mathbf{X}_i), \mathbf{u}^*\rangle}{n} - \sum_{j\in\{-1,0,1\}} \frac{n_j\mathbb{E}\left[\left\langle g_{\mathbf{u}}(\mathbf{Y}_j), \mathbf{u}^*\right\rangle\right]}{n}\right| \geq \frac{\delta}{2}\middle| |\hat{p} - p| \leq p\bar{\delta}\right) + \mathbb{P}\left(|\hat{p} - p| \geq p\bar{\delta}\right) \\ &\qquad + \mathbb{P}\left[\left|\sum_{j\in\{-1,0,1\}} \frac{n_j\mathbb{E}\left[\left\langle g_{\mathbf{u}}(\mathbf{Y}_j), \mathbf{u}^*\right\rangle\right]}{n} - \mathbb{E}\left[\langle g_{\mathbf{u}}(\mathbf{X}), \mathbf{u}^*\rangle\right]\right| \geq \frac{\delta}{2}\right] \\ &\leq \exp\left(-\Theta\left(\frac{n\delta^2}{\log(n\log(s))^2\left(1 + p\max\left\{\langle \mathbf{u}, \mathbf{u}^*\rangle^4 p^{-2}, \langle \mathbf{u}, \mathbf{u}^*\rangle^8 p^{-3}\right\}\right)}\right)\right)\end{aligned}$$

□

Lemma 9. *For $n = \Omega(d)$ have*

$$\mathbb{P}\left[\max_{\hat{\mathbf{u}}\in\mathbb{S}_{d-1}}\left|\sum_{i=1}^n \frac{\psi(\langle \mathbf{X}_i, \hat{\mathbf{u}}\rangle)}{n} - \mathbb{E}[\psi(\langle \mathbf{X}, \hat{\mathbf{u}}\rangle)]\right| \geq \delta\right] \leq \exp\left(-\Theta\left(\frac{n\delta^2}{d}\right)\right)$$

Proof. First make the distinction into $\mathbf{Y}_j \sim \mathcal{N}\left(\frac{j\mathbf{u}^*}{\sqrt{p}}, I_d - \mathbf{u}^*\mathbf{u}^{*\top}\right)^{n_j}$ for $j \in \{-1, 0, 1\}$, where $\sum_{j\in\{-1,0,1\}} n_j = n$. We will first note that $\mathbf{A}_j(\mathbf{u}) = \psi(\mathbf{Y}_j\mathbf{u}) - \mathbb{E}[\psi(\mathbf{Y}_j\mathbf{u})]$ is subgaussian and thus

$$\mathbb{P}\left[\left|\sum_{j\in\{-1,0,1\}}\sum_{i=1}^{n_j} \frac{\mathbf{A}_{j,i}(\mathbf{u})}{n}\right| \geq \delta_c\right] \leq \exp\left(-\Theta\left(\delta_c^2 n\right)\right)$$

Let $\mathcal{N}_{\epsilon,d}$ be the minimum size ϵ-Net of the d-dimensional unit sphere. Thus we know $\left|\mathcal{N}_{\epsilon,d}\right| \leq \left(\frac{3}{\epsilon}\right)^d$.

$$\mathbb{P}\left[\max_{\mathbf{u}\in\mathcal{N}_{\epsilon,d}}\left|\sum_{j\in\{-1,0,1\}}\sum_{i=1}^{n_j}\frac{\mathbf{A}_{j,i}(\mathbf{u})}{n}\right| \geq \delta_c\right] \leq \exp\left(d\log\left(\frac{3}{\epsilon}\right) - \Theta\left(\delta_c^2 n\right)\right)$$

First we bound the maximum deviation for $\mathbf{u}'$ for $\|\mathbf{u}' - \mathbf{u}\|_2 < \epsilon$

$$\begin{aligned}\left|\sum_{i=1}^{n}\frac{\psi(\langle\mathbf{X}_i,\mathbf{u}\rangle)}{n} - \sum_{i=1}^{n}\frac{\psi(\langle\mathbf{X}_i,\mathbf{u}'\rangle)}{n}\right| &\leq \frac{\|\mathbf{X}(\mathbf{u}-\mathbf{u}')\|_1}{n} \leq \frac{\|\mathbf{X}(\mathbf{u}-\mathbf{u}')\|_2}{\sqrt{n}} \\ &\leq \frac{\epsilon\|X\|_{op}}{\sqrt{n}}\end{aligned}$$

We continue by bounding for all $\mathbf{u} \in \mathbb{S}_{d-1}$ by applying the triangle inequality

$$\left|\sum_{j\in\{-1,0,1\}}\sum_{i=1}^{n_j}\frac{\mathbf{A}_{j,i}(\mathbf{u})}{n}\right| \leq \max_{\mathbf{u}'\in\mathcal{N}_{\epsilon,d}}\left|\sum_{j\in\{-1,0,1\}}\sum_{i=1}^{n_j}\frac{\mathbf{A}_{j,i}(\mathbf{u}')}{n}\right| + \frac{\epsilon\|X\|_{op}}{\sqrt{n}}$$

Finally we notice that $\mathbf{X}$ can be decomposed into the union of $\mathbf{Y}_{-1}, \mathbf{Y}_0, \mathbf{Y}_1$. Thus by applying the Chernoff bound for the Binonmial Distribution we obtain

$$\mathbb{P}\left[\left|\mathbb{E}[\psi(\langle\mathbf{X},\mathbf{u}\rangle)] - \sum_{j\in\{-1,0,1\}}\mathbb{E}\left[\psi\left(\left\langle\mathbf{Y}_j,\mathbf{u}\right\rangle\right)\right]\right| \geq \frac{\delta_p}{\sqrt{p}}\right] \leq 2\exp\left(-\Theta\left(\delta_p^2 np\right)\right)$$

Thus we can prove the Theorem by choosing $\epsilon = \Theta(\delta)$ for $n > d > 1$ and $\delta < 1$. The second to last step follows bounding $\|\mathbf{X}\|_{op} \leq \left\|\mathbf{X}(\mathbf{u}^*\mathbf{u}^{*\top})\right\|_{op} + \left\|\mathbf{X}(I - \mathbf{u}^*\mathbf{u}^{*\top})\right\|_{op} \leq \Theta\left(\sqrt{n} + t\right)$ with probability $1 - 2\exp(-t^2)$ [23].

$$\begin{aligned}&\mathbb{P}\left[\max_{\mathbf{u}\in\mathbb{S}_{d-1}}\left|\sum_{i=1}^{n}\frac{\psi(\mathbf{X}_i\mathbf{u})}{n} - \mathbb{E}[\psi(\mathbf{X}\mathbf{u})]\right| \geq \delta\right] \\ \leq&\mathbb{P}\left[\max_{\mathbf{u}\in\mathcal{N}_{\epsilon,d}}\left|\sum_{j\in\{-1,0,1\}}\sum_{i=1}^{n_j}\frac{\mathbf{A}_{j,i}}{n}\right| \geq \frac{\delta}{3}\right] + \mathbb{P}\left[\frac{\epsilon\|X\|_{op}}{\sqrt{n}} \geq \frac{\delta}{3}\right] + \mathbb{P}\left[\left|\mathbb{E}[\psi(\langle\mathbf{X},\mathbf{u}\rangle)] - \sum_{j\in\{-1,0,1\}}\mathbb{E}\left[\psi\left(\left\langle\mathbf{Y}_j,\mathbf{u}\right\rangle\right)\right]\right| \geq \frac{\delta}{3}\right] \\ \leq&\exp\left(d\log\left(\frac{3}{\epsilon}\right) - \Theta\left(\delta^2 n\right)\right) + 2\exp(-n) + \exp\left(-\Theta\left(\delta^2 np^2\right)\right) \\ \leq&2\exp\left(-\Theta\left(\frac{n\delta^2}{d}\right)\right)\end{aligned}$$

□

Lemma 10. *For all $\langle\mathbf{u},\mathbf{u}^*\rangle > \langle\mathbf{u}',\mathbf{u}^*\rangle$ and $\delta > 0$. there exists a threshold t such that*

$$\sum_{i=1}^{n}\frac{\psi(\langle\mathbf{X}_i,\mathbf{u}\rangle)}{n} \geq t \text{ and } \sum_{i=1}^{n}\frac{\psi(\langle\mathbf{X}_i,\mathbf{u}'\rangle)}{n} \leq t$$

with probability at least $1 - 2\exp\left(-\Theta\left(\frac{n\delta^2}{d}\right)\right)$.

Proof. By applying the expectation of the half normal distribution we obtain the following upper and lower bounds.

$$-\sqrt{\frac{2}{\pi}}\sqrt{1-\langle\mathbf{u},\mathbf{u}^*\rangle^2}(1-p) - \sqrt{p}\langle\mathbf{u},\mathbf{u}^*\rangle^2 \geq \mathbb{E}[\psi(\langle\mathbf{X},\mathbf{u}\rangle)] \geq -\sqrt{\frac{2}{\pi}}\sqrt{1-\langle\mathbf{u},\mathbf{u}^*\rangle^2}(1-p) - \sqrt{p}$$

Thus if for all $\mathbf{u}$

$$\left|\sum_{i=1}^{n}\frac{\psi(\langle\mathbf{X}_i,\mathbf{u}\rangle)}{n}-\mathbb{E}[\psi(\langle\mathbf{X},\mathbf{u}\rangle)]\right|\leq\delta$$

then

$$-\sqrt{\frac{2}{\pi}}\sqrt{1-\langle\mathbf{u},\mathbf{u}^*\rangle^2}(1-p)-\sqrt{p}-\delta>-\sqrt{\frac{2}{\pi}}\sqrt{1-\langle\mathbf{u}',\mathbf{u}^*\rangle^2}(1-p)-\sqrt{p}\langle\mathbf{u}',\mathbf{u}^*\rangle^2+\delta$$

indicates the existence of the threshold. The Lemma follows by applying Lemma 9 and reordering terms. □

Lemma 11. *With probability at least* $1-\Theta\left(\frac{1}{s}\right)-\exp\left(-\Theta\left(\frac{np}{\log(ns)}\right)\right)$ *have*

$$\left\|\frac{\phi'(\mathbf{X}\mathbf{u})}{n}\right\|_2\leq\Theta\left(\frac{1+\frac{\langle\mathbf{u},\mathbf{u}^*\rangle^3}{p}}{\sqrt{n}}\right)$$

Proof. Choose an ortho-normal basis $\mathbf{E}=(\mathbf{u}^*,\mathbf{e}_2,...,\mathbf{e}_d)\in\mathbb{R}^{d\times d}$ such that $\mathbf{u}=\langle\mathbf{u},\mathbf{u}^*\rangle\mathbf{u}^*+\sqrt{1-\langle\mathbf{u},\mathbf{u}^*\rangle^2}\mathbf{e}_2$.

$$\begin{aligned}\mathbb{E}[\phi'(\langle\mathbf{X},\mathbf{u}\rangle)^2]&=\mathbb{E}[\langle\mathbf{X},\mathbf{u}\rangle^6]\\&=\mathbb{E}\left[\left(\langle\mathbf{u},\mathbf{u}^*\rangle\langle\mathbf{u}^*,\mathbf{X}\rangle+\sqrt{1-\langle\mathbf{u},\mathbf{u}^*\rangle^2}\langle\mathbf{e}_2,\mathbf{X}\rangle\right)^6\right]\\&\leq\Theta\left(1+\mathbb{E}\left[\langle\mathbf{u},\mathbf{u}^*\rangle\langle\mathbf{u}^*,\mathbf{X}\rangle)^6\right]\right)\\&\leq\Theta\left(1+\frac{\langle\mathbf{u},\mathbf{u}^*\rangle^6}{p^2}\right)\end{aligned}$$

For convenience we define $\hat{p}=\sum_{i=1}^{n}\frac{I_{\langle\mathbf{X}_i,\mathbf{u}^*\rangle\neq0}}{n}$ as well as $n_j=\sum_{i=1}^{n}I_{\langle\mathbf{X}_i,\mathbf{u}^*\rangle=j\sqrt{1/p}}$ with $j\in\{-1,0,1\}$. Additionally define $\mathbf{Y}_j=\left(\mathbf{X}\middle|\langle\mathbf{X},\mathbf{u}^*\rangle=j\sqrt{\frac{1}{p}}\right)$. For $\bar{\delta}\in(0,1)$

$$\begin{aligned}&\mathbb{P}\left[\left|\sum_{i=1}^{n}\frac{\phi'(\langle\mathbf{X}_i,\mathbf{u}\rangle)^2}{n}-\mathbb{E}\left[\phi'(\langle\mathbf{X},\mathbf{u}\rangle)^2\right]\right|\geq\delta\right]\\&\leq\mathbb{P}\left[\left|\sum_{i=1}^{n}\frac{\phi'(\langle\mathbf{X}_i,\mathbf{u}\rangle)^2}{n}-\sum_{j\in\{-1,0,1\}}\frac{n_j}{n}\mathbb{E}\left[\phi'\left(\left\langle\mathbf{Y}_j,\mathbf{u}\right\rangle\right)^2\right]\right|\geq\frac{\delta}{2}\;\middle|\;|\hat{p}-p|\leq p\bar{\delta}\wedge\max_{i\in[n]}\langle\mathbf{X}_i,\mathbf{e}_2\rangle\leq\Theta\left(\sqrt{\log(ns)}\right)\right]+\mathbb{P}\left[|\hat{p}-p|\geq p\bar{\delta}\right]\\&\quad+\mathbb{P}\left[\max_{i\in[n]}\langle\mathbf{X}_i,\mathbf{e}_2\rangle\geq\Theta\left(\sqrt{\log(ns)}\right)\right]+\mathbb{P}\left[\left|\left(\sum_{j\in\{-1,0,1\}}\frac{n_j}{n}\mathbb{E}\left[\phi'\left(\left\langle\mathbf{Y}_j,\mathbf{u}\right\rangle\right)^2\right]\right)-\mathbb{E}\left[\phi'(\langle\mathbf{X},\mathbf{u}\rangle)^2\right]\right|\geq\frac{\delta}{2}\right]\\&\leq\exp\left(-\Theta\left(\frac{n\delta^2}{\log(ns)\left(1+\frac{\langle\mathbf{u},\mathbf{u}^*\rangle^6}{p^2}\right)^2}\right)\right)+2\exp\left(-\frac{\bar{\delta}^2np}{3}\right)+2\exp\left(\log(n)-\Theta\left(\log(ns)\right)\right)+\exp\left(-\Theta\left(\frac{\delta^2np}{\left(1+\frac{\langle\mathbf{u},\mathbf{u}^*\rangle^6}{p^2}\right)^2}\right)\right)\qquad\text{(C.1)}\\&\leq\Theta\left(\frac{1}{s}\right)+\exp\left(-\Theta\left(\frac{\delta^2np}{\log(ns)\left(1+\frac{\langle\mathbf{u},\mathbf{u}^*\rangle^6}{p^2}\right)^2}\right)\right)\end{aligned}$$

Here (C.1) follows by applying Hoeffdings inequality and the Chernoff bound for the Binomial distribution. The last term in (C.1) follows by bounding the difference $\mathbb{E}\left[\phi'\left(\langle\mathbf{Y}_1,\mathbf{u}\rangle\right)^2\right]-\mathbb{E}\left[\phi'\left(\langle\mathbf{Y}_0,\mathbf{u}\rangle\right)^2\right]=\mathbb{E}\left[\phi'\left(\langle\mathbf{Y}_{-1},\mathbf{u}\rangle\right)^2\right]-\mathbb{E}\left[\phi'\left(\langle\mathbf{Y}_0,\mathbf{u}\rangle\right)^2\right]\leq\Theta\left(1+\frac{\langle\mathbf{u},\mathbf{u}^*\rangle^6}{p^2}\right)$ and also applying the Chernoff bound.Choosing $\delta=\Theta\left(1+\frac{\langle\mathbf{u},\mathbf{u}^*\rangle^6}{p^2}\right)$ yields

$$\mathbb{E}[\phi'(\langle\mathbf{X},\mathbf{u}\rangle)^2]\leq\Theta\left(1+\frac{\langle\mathbf{u},\mathbf{u}^*\rangle^6}{p^2}\right)$$

With probability at least $1-\Theta\left(\frac{1}{s}\right)-\exp\left(-\Theta\left(\frac{np}{\log(ns)}\right)\right)$. The lemma follows by computing the norm.

□

Appendix C.1. Proof of Theorem 2

Proof. As discussed we will split the proof of Theorem 2 into parts for each execution of Algorithm 2.

First Execution of Algorithm 2 *There exist parameters t_1 and $\eta_1 = \Omega(dp^2)$ such that for $a_1 = \Theta\left(\frac{1}{\sqrt{pd}}\right)$ and some constant $b_1 \in (0,1)$ the first execution of Algorithm 2 fulfills the criteria of Lemma 1 for $n = \widetilde{O}(d^3p^4)$.*

By the same argument as in Lemma 2 we have $\langle \mathbf{u}, \mathbf{u}^* \rangle \geq \Theta\left(\frac{1}{\sqrt{pd}}\right)$ with sufficient large probability, thus fulfilling Precondition 1.

Using Lemma 7 and applying the concentration result of Lemma 8 we obtain that

$$\sum_{i=1}^{n} \left\langle \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle = \Omega\left(\frac{\langle \mathbf{u}, \mathbf{u}^* \rangle^3}{p}\right) \tag{C.2}$$

with probability at least $1 - \exp\left(-\Theta\left(\frac{n(\langle \mathbf{u},\mathbf{u}^*\rangle^3 p^{-1})^2}{\log(n\log(s))^2\left(1+\max\left\{\langle \mathbf{u},\mathbf{u}^*\rangle^4 p^{-1}, \langle \mathbf{u},\mathbf{u}^*\rangle^8 p^{-2}\right\}\right)}\right)\right)$.

Next bound with probability at least $1 - \Theta\left(\frac{1}{s}\right) - \exp\left(-\Theta\left(\frac{np}{\log(ns)}\right)\right)$

$$\left\| \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2 \leq \sqrt{\left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle^2 \left(1 + \frac{\langle \mathbf{u}, \mathbf{u}^* \rangle^2}{1 - \langle \mathbf{u}, \mathbf{u}^* \rangle^2}\right) + \left\| \frac{\phi'(\mathbf{X}\mathbf{u})^\top \mathbf{X}(I_d - \mathbf{u}^*\mathbf{u}^{*\top} - \mathbf{e}_2\mathbf{e}_2^\top)}{n} \right\|_2^2} \tag{C.3}$$

$$\leq \sqrt{\left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle^2 \left(1 + \frac{\langle \mathbf{u}, \mathbf{u}^* \rangle^2}{1 - \langle \mathbf{u}, \mathbf{u}^* \rangle^2}\right) + \Theta\left(\left(1 + \frac{\langle \mathbf{u}, \mathbf{u}^* \rangle^6}{p^2}\right)\frac{d}{n}\right)} \tag{C.4}$$

$$\leq \max\left\{ 2\sqrt{\left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle^2 \left(1 + \frac{\langle \mathbf{u}, \mathbf{u}^* \rangle^2}{1 - \langle \mathbf{u}, \mathbf{u}^* \rangle^2}\right)}, 2\sqrt{\Theta\left(\left(1 + \frac{\langle \mathbf{u}, \mathbf{u}^* \rangle^6}{p^2}\right)\frac{d}{n}\right)} \right\} \tag{C.5}$$

Here (C.3) follows by by decomposing the norm. (C.4) follows by applying Lemma 11 and Lemma 17. Finally using $\eta_1 = \Omega\left(p^2 d\right)$ we apply Lemma 18 to demonstrate that Precondition 2 is fulfilled if $n = \widetilde{\Omega}\left(d^3p^4\right)$.

Finally Lemma 10 shows that Precondition 3 is fulfilled.

Second Execution of Algorithm 2 *There exist parameters t_2 and $\eta_2 = \Omega(dp^2)$ such that for $a_2 = b_1 - \delta$ and $b_2 = 1 - \beta$ for some $3 > \epsilon > 0$ the first execution of Algorithm 2 fulfills the criteria of Lemma 1 for $n = \widetilde{\Omega}\left(d^3p^4\right)$.*

Precondition 1 is fulfilled by the first execution of Algorithm 2.

Using (C.5) have $\left\| \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2 = O\left(\frac{1}{p}\right)$ if $|\langle \mathbf{u}, \mathbf{u}^* \rangle| \leq 1-\beta$. Choosing $\eta_2 = \Theta(1) > 0$ yields that $\eta_2 \langle \mathbf{u}, \mathbf{u}^* \rangle^{-1} \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle \geq \left\| \eta_2 \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2^2$. Thus applying Lemma 19 we obtain that Precondition 2 is satisfied if $n = \widetilde{\Omega}\left(d^3p^4\right)$.

Finally Lemma 10 shows that Precondition 3 is fulfilled.

□

Appendix D. Proofs in Section 3

Appendix D.1. Proof of Corollary 1

Proof. First note we have $\mathbb{E}_{\mathcal{H}_0}[\max\{0, \mathbf{X}\}^2] = \mathbb{E}_{\mathcal{H}_1}[\max\{0, \mathbf{X}\}^2$ if $\langle \hat{\mathbf{u}}, \mathbf{u}^* \rangle = 0$. Thus we can apply Lemma 6 to obtain the result. □

Lemma 12 is nearly equlvalent to Lemma 6.7 in Mao and Wein [15].

Lemma 12. *For $\alpha \in \mathbb{N}^n$, let $|\alpha| = \sum_{i=1}^n \alpha_i = \|\alpha\|_1$, and let $\|\alpha\|_0$ be the size of the support of α. For $m \in [d]$, define a set*

$$\mathcal{A}(k, m) := \{\alpha \in \mathbb{N}^n : |\alpha| = k,\ \|\alpha\|_0 = m,\ \alpha_i \in \{0\} \cup \{3, 4, \dots\} \text{ for all } i \in [n]\}. \tag{D.1}$$

Then we have $|\mathcal{A}(k, m)| \leq n^m k^k$.

Proof. $|\mathcal{A}(k,m)| \leq \binom{n}{k} k^k$. □

Lemma 13. *Have the Imbalanced Clusters RV* $\mathbf{X} \sim \mathcal{D}_\nu(p)$

$$\mathbb{E}[\hat{h}_0(\mathbf{X})] = 1,\ \mathbb{E}[\hat{h}_1(\mathbf{X})] = 0,\ \mathbb{E}[\hat{h}_2(\mathbf{X})] = 0$$

And given $p < 0.5$ *have for* $k \geq 3$

$$|\mathbb{E}[\hat{h}_k(\mathbf{X})]| \leq k^{k/2} p^{1-k/2}$$

Proof. For $p < 0.5$ have $\mathbb{E}[\mathbf{X}^k] \leq 2p^{1-k/2}$. Thus we can bound

$$|\mathbb{E}[h_k(\mathbf{X})]| = \frac{1}{\sqrt{k!}} \left(\sum_{i=0}^{k} c_i \mathbb{E}[\mathbf{X}^i] \right) \leq |\mathbb{E}[\mathbf{X}^i]| \sqrt{k!}$$

. With the last inequality following from $\sum_{i=0}^{k} |c_i| \leq k!$. □

Lemma 14 (Mao and Wein [15])**.** *Consider the distribution* $\mathcal{H}_1$ *in Problem 1 and suppose the first D moments of* ν *are finite. For* $\alpha \in \mathcal{N}^N$*, let* $|\alpha| := \sum_{i=1}^{N} \alpha_i$*. Then*

$$\left\| L_d^{\leq D} \right\|_2^2 = \sum_{d=0}^{D} \mathbb{E}[\langle \mathbf{u}, \mathbf{u}' \rangle^d] \sum_{\substack{\alpha \in \mathcal{N}^N \\ |\alpha| = d}} \prod_{i=1}^{N} \left(\mathbb{E}_{x \sim \nu}[h_{\alpha_i}(x)] \right)^2 \tag{D.2}$$

where $\mathbf{u}$ *and* $\mathbf{u}'$ *are drawn independently from* $\mathcal{U}$*.*

Lemma 15 (Mao and Wein [15])**.** *Let* $\mathbf{u}$ *and* $\mathbf{u}'$ *be independent uniform random vectors on the unit sphere in* $\mathbb{R}^n$*. For* $k \in \mathbb{N}$*, if k is odd, then* $\mathbb{E}[\langle \mathbf{u}, \mathbf{u}' \rangle^d] = 0$*, and if k is even, then*

$$\mathbb{E}[\langle \mathbf{u}, \mathbf{u}' \rangle^k] \leq (k/n)^{k/2}$$

Appendix D.2. Proof of Theorem 3

The proof of Theorem 3 closely follows the proof of Theorem 4.5 in Mao and Wein [15].

Proof. We will use $\mathcal{A}(k,m)$ as defined in Lemma 12 and note that for $\alpha \in \mathcal{A}(k,m)$, we obtain that $\alpha \geq 3$ and thus $m \leq \lfloor k/3 \rfloor$ using Lemma 13

$$\sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| = k}} \prod_{i=1}^{n} \left(\mathbb{E}[h_{\alpha_i}(x)] \right)^2 = \sum_{m=1}^{\lfloor k/3 \rfloor} \sum_{\alpha \in \mathcal{A}(k,m)} \prod_{i=1}^{n} \left(\mathbb{E}[h_{\alpha_i}(x)] \right)^2 \leq \sum_{m=1}^{\lfloor k/3 \rfloor} |\mathcal{A}(k,m)| \prod_{i \in [n],\, \alpha_i \neq 0} \alpha_i^{2\alpha_i} p^{2-\alpha_i} \leq \sum_{m=1}^{\lfloor k/3 \rfloor} n^m k^{3k} p^{2m-k}$$

By applying the closed form of the geometric series we obtain

$$\sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| = k}} \prod_{i=1}^{n} \left(\mathbb{E}[h_{\alpha_i}(x)] \right)^2 \leq k^{3k} n p^{2-k} \frac{(np^2)^{\lfloor k/3 \rfloor} - 1}{np^2 - 1} \leq k^{3k} n p^{2-k} \frac{(np^2)^{k/3}}{\frac{1}{2} np^2} = 2k^{3k} n^{k/3} p^{-k/3}.$$

This combined with Lemma 15 gives

$$\mathbb{E}[\langle u, u' \rangle^k] \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| = k}} \prod_{i=1}^{n} \left(\mathbb{E}[h_{\alpha_i}(x)] \right)^2 \leq (k/d)^{k/2} \cdot 2k^{3k} n^{k/3} p^{-k/3} = 2 \left(\frac{k^{10.5} n}{d^{3/2} p} \right)^{k/3}.$$

Finally, combining this with Lemma 14, we obtain

$$\left\| L_d^{\leq D} \right\|_2^2 = \sum_{k=0}^{D} \mathbb{E}[\langle u, u' \rangle^k] \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| = k}} \prod_{i=1}^{n} \left(\mathbb{E}[h_{\alpha_i}(x)] \right)^2 \leq 1 + 2 \sum_{k=3}^{D} \left(\frac{k^{10.5} n}{d^{3/2} p} \right)^{k/3}.$$

This is true if $d^{1.5} p > n D^{c_2}$ for a sufficiently large constant $c_2 > 0$ such that $\frac{k^{10.5} n}{d^{3/2} p} < 1/4$. □

Appendix E. Additional Proofs

Lemma 16. *Choosing the ortho-normal basis* $\mathbf{E} = (\mathbf{u}^*, \mathbf{e}_2, ..., \mathbf{e}_d) \in \mathbb{R}^{d\times d}$ *such that* $\mathbf{u} = \langle \mathbf{u}, \mathbf{u}^*\rangle \mathbf{u}^* + \sqrt{1 - \langle \mathbf{u}, \mathbf{u}^*\rangle^2}\mathbf{e}_2$.

$$\left\| \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2 = \sqrt{\left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle^2 \left(1 + \frac{\langle \mathbf{u}, \mathbf{u}^*\rangle^2}{1 - \langle \mathbf{u}, \mathbf{u}^*\rangle^2}\right) + \left\| \frac{\phi'(\mathbf{X}\mathbf{u})^\top \mathbf{X}(I_d - \mathbf{u}^*\mathbf{u}^{*\top} - \mathbf{e}_2\mathbf{e}_2^\top)}{n} \right\|_2^2}$$

Proof.

$$\left\| \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2 = \sqrt{\left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle^2 + \left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{e}_2 \right\rangle^2 + \sum_{i=3}^d \left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{e}_i \right\rangle^2}$$

By $\langle \mathbf{u}, \mathbf{u}^*\rangle \left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle + \sqrt{1 - \langle \mathbf{u}, \mathbf{u}^*\rangle^2} \left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{e}_2 \right\rangle = 0$ we can expand to

$$\left\| \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2 = \sqrt{\left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle^2 \left(1 + \frac{\langle \mathbf{u}, \mathbf{u}^*\rangle^2}{1 - \langle \mathbf{u}, \mathbf{u}^*\rangle^2}\right) + \sum_{i=3}^d \left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, e_i \right\rangle^2}$$

□

Lemma 17. *Choose a ortho-normal basis* $\mathbf{E} = (\mathbf{u}^*, \mathbf{e}_2, ..., \mathbf{e}_d) \in \mathbb{R}^{d\times d}$ *such that* $\mathbf{u} = \langle \mathbf{u}, \mathbf{u}^*\rangle \mathbf{u}^* + \sqrt{1 - \langle \mathbf{u}, \mathbf{u}^*\rangle^2}\mathbf{e}_2$. *Thus*

$$\left\| \frac{\phi'(\mathbf{X}\mathbf{u})^\top \mathbf{X}(I_d - \mathbf{u}^*\mathbf{u}^{*\top} - \mathbf{e}_2\mathbf{e}_2^\top)}{n} \right\|_2 = O\left(\sqrt{d} \left\| \frac{\phi'(\mathbf{X}\mathbf{u})}{n} \right\|_2 \right)$$

with probability at least $1 - O\left(\frac{1}{s}\right)$.

Proof. First notice that $\mathbf{X}(I_d - \mathbf{u}^*\mathbf{u}^{*\top} - \mathbf{e}_2\mathbf{e}_2^\top) \sim \mathcal{N}(0, (I_d - \mathbf{u}^*\mathbf{u}^{*\top} - \mathbf{e}_2\mathbf{e}_2^\top))^n$ and thus that for some vector $\mathbf{v}$

$$\mathbb{P}\left[\left\| \sum_{i=1}^n \frac{\mathbf{v}_i}{\|\mathbf{v}\|_2} \mathbf{X}_i(I_d - \mathbf{u}^*\mathbf{u}^{*\top} - \mathbf{e}_2\mathbf{e}_2^\top) \right\|_2 \geq \sqrt{d} + t \right] \leq 2\exp\left(-\Theta\left(t^2\right)\right)$$

The lemma follows by using $t = \sqrt{\log(s)}$ and choosing $\mathbf{v} = \frac{\phi'(\mathbf{X}\mathbf{u})}{n}$ □

Lemma 18. *For any* $c_3 \in (0, 1)$ *such that if* $\left\| \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2 \leq (1 - c_3) \frac{\left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle}{\langle \mathbf{u}, \mathbf{u}^*\rangle}$ *and* $\frac{\eta \left\langle \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle}{\langle \mathbf{u}, \mathbf{u}^*\rangle} \geq 1$ *we have*

$$\frac{\langle \mathbf{u}^*, \mathbf{u} + \eta \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle}{\left\| \mathbf{u} + \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle \right\|_2} \geq \langle \mathbf{u}, \mathbf{u}^*\rangle \left(1 + \frac{c_6}{2}\right)$$

Proof.

$$\begin{aligned}
\left\langle \mathbf{u}^*, \frac{\mathbf{u} + \eta \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}}{\left\| \mathbf{u} + \eta \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2} \right\rangle &\geq \langle \mathbf{u}, \mathbf{u}^*\rangle \frac{1 + \eta \langle \mathbf{u}, \mathbf{u}^*\rangle^{-1} \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle}{\sqrt{1 + \eta^2 \left\| \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2^2}} \\
&\geq \langle \mathbf{u}, \mathbf{u}^*\rangle \frac{1 + \eta \langle \mathbf{u}, \mathbf{u}^*\rangle^{-1} \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle}{1 + (1 - c_3)\eta \langle \mathbf{u}, \mathbf{u}^*\rangle^{-1} \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle} \geq \langle \mathbf{u}, \mathbf{u}^*\rangle \left(1 + \frac{c_3}{2}\right)
\end{aligned}$$

□

Lemma 19. *If* $\eta\langle \mathbf{u}, \mathbf{u}^*\rangle^{-1} \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle \geq \left\| \eta \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2^2$

$$\frac{\langle \mathbf{u}^*, \mathbf{u} + \eta \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle}{\left\| \mathbf{u} + \eta \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2} \geq \langle \mathbf{u}, \mathbf{u}^* \rangle \min \left\{ 1 + \eta \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle, 2 \right\}$$

Proof.

$$\begin{aligned}
\left\langle \mathbf{u}^*, \frac{\mathbf{u} + \eta \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}}{\left\| \mathbf{u} + \eta \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2} \right\rangle &\geq \langle \mathbf{u}, \mathbf{u}^* \rangle \frac{1 + \eta \langle \mathbf{u}, \mathbf{u}^* \rangle^{-1} \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle}{\sqrt{1 + \left\| \eta \frac{\sum_{i=1}^n g_{\mathbf{u}}(\mathbf{X}_i)}{n} \right\|_2^2}} \\
&\geq \langle \mathbf{u}, \mathbf{u}^* \rangle \sqrt{1 + \eta \langle \mathbf{u}, \mathbf{u}^* \rangle^{-1} \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle} \geq \langle \mathbf{u}, \mathbf{u}^* \rangle \min \left\{ 1 + \eta \left\langle \sum_{i=1}^n \frac{g_{\mathbf{u}}(\mathbf{X}_i)}{n}, \mathbf{u}^* \right\rangle, 2 \right\}
\end{aligned}$$

□

Lemma 20. *For any* $\phi(\cdot)$ *and* $\mathbf{x}, \mathbf{u}, \mathbf{u}^* \in \mathbb{R}^d$ *have*

$$\langle g_{\mathbf{u}}(\mathbf{x}), \mathbf{u}^* \rangle = \frac{\partial \phi(\langle \mathbf{x}, \mathbf{u} \rangle)}{\partial \langle \mathbf{x}, \mathbf{u} \rangle} (\langle \mathbf{x}, \mathbf{u}^* \rangle - \langle \mathbf{x}, \mathbf{u} \rangle \langle \mathbf{u}, \mathbf{u}^* \rangle)$$

Proof. First recall the definition $g_{\mathbf{u}}(\mathbf{x}) = (I_d - \mathbf{u}\mathbf{u}^T) \frac{\partial \phi(\langle \mathbf{u}, \mathbf{x} \rangle}{\partial u}$. Let us choose a new ortho-normal basis $\mathbf{E} = (\mathbf{u}^*, \mathbf{e}_2, ..., \mathbf{e}_d) \in \mathbb{R}^{d \times d}$. By $\langle \mathbf{x}, \mathbf{u} \rangle = \langle \mathbf{x}, \mathbf{u}^* \rangle \langle \mathbf{u}, \mathbf{u}^* \rangle + \left(\sum_{i=2}^d \langle \mathbf{x}, e_i \rangle \langle \mathbf{u}, e_i \rangle \right)$

$$\begin{aligned}
\langle g_{\mathbf{u}}(\mathbf{x}), \mathbf{u}^* \rangle &= \left\langle (I - \mathbf{u}\mathbf{u}^T) \frac{\partial \phi(\langle \mathbf{x}, \mathbf{u} \rangle)}{\partial \mathbf{u}}, \mathbf{u}^* \right\rangle \\
&= \frac{\partial \phi(\langle \mathbf{x}, \mathbf{u} \rangle)}{\partial \langle \mathbf{x}, \mathbf{u} \rangle} \left(\left(1 - \langle \mathbf{u}, \mathbf{u}^* \rangle^2 \right) \langle \mathbf{x}, \mathbf{u}^* \rangle - \langle \mathbf{u}, \mathbf{u}^* \rangle \left(\sum_{i=2}^d \langle \mathbf{u}, \mathbf{e}_i \rangle \langle \mathbf{x}, \mathbf{e}_i \rangle \right) \right) \\
&= \frac{\partial \phi(\langle \mathbf{x}, \mathbf{u} \rangle)}{\partial \langle \mathbf{x}, \mathbf{u} \rangle} \left((1 - \langle \mathbf{u}, \mathbf{u}^* \rangle^2) \langle \mathbf{x}, \mathbf{u}^* \rangle - \langle \mathbf{u}, \mathbf{u}^* \rangle (\langle \mathbf{x}, \mathbf{u} \rangle - \langle \mathbf{u}, \mathbf{u}^* \rangle \langle \mathbf{x}, \mathbf{u}^* \rangle) \right) \\
&= \frac{\partial \phi(\langle \mathbf{x}, \mathbf{u} \rangle)}{\partial \langle \mathbf{x}, \mathbf{u} \rangle} (\langle \mathbf{x}, \mathbf{u}^* \rangle - \langle \mathbf{x}, \mathbf{u} \rangle \langle \mathbf{u}, \mathbf{u}^* \rangle)
\end{aligned}$$

□