Learning Single-Index Models in Gaussian Space

Rishabh Dudeja RD2714@COLUMBIA.EDU
Department of Statistics, Columbia University

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Computer Science Department, Columbia University

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We consider regression problems where the response is a smooth but non-linear function of a k-dimensional projection of p normally-distributed covariates, contaminated with additive Gaussian noise. The goal is to recover the range of the k-dimensional projection, i.e., the index space. This model is called the multi-index model, and the $k = 1$ case is called the single-index model. For the single-index model, we characterize the population landscape of a natural semi-parametric maximum likelihood objective in terms of the link function and prove that it has no spurious local minima. We also propose and analyze an efficient iterative procedure that recovers the index space up to error ϵ using a sample size $\tilde{O}(p^{O(R^2/\mu)} + p/\epsilon^2)$, where R and μ, respectively, parameterize the smoothness of the link function and the signal strength. When a multi-index model is incorrectly specified as a single-index model, we prove that essentially the same procedure, with sample size $\tilde{O}(p^{O(kR^2/\mu)} + p/\epsilon^2)$, returns a vector that is ϵ-close to being completely in the index space.

Keywords: Single-index models, multi-index models, non-convex optimization, semi-parametric models.

1. Introduction

Suppose we are given n data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ generated independently from the following regression model:

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p), \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right), \quad y_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \perp\!\!\!\perp \boldsymbol{x}_i. \tag{1}$$

Here, $\boldsymbol{x}_i \in \mathbb{R}^p$ are p-dimensional covariates or features and $y_i \in \mathbb{R}$ are the response variables. The function $f : \mathbb{R}^p \to \mathbb{R}$ is assumed to be smooth and unknown. In many applications of practical interest the function f is not an arbitrary p-variate function but depends on an *unknown* k dimensional projection of $\boldsymbol{x}$, that is,

$$f(\boldsymbol{x}) = g(\langle \boldsymbol{u}_1^\star, \boldsymbol{x}\rangle, \langle \boldsymbol{u}_2^\star, \boldsymbol{x}\rangle, \ldots, \langle \boldsymbol{u}_k^\star, \boldsymbol{x}\rangle), \tag{2}$$

where $g : \mathbb{R}^k \to \mathbb{R}$ and $\boldsymbol{u}_1^\star, \boldsymbol{u}_2^\star, \ldots, \boldsymbol{u}_k^\star$ are orthonormal vectors in $\mathbb{R}^p$. In the statistics community, this model is called the multi-index model. The special case $k = 1$ is called the single-index model; a simple example is the phase retrieval problem for real signals where $g(z) = z^2$. We note that in the multi-index model, the index vectors are themselves unidentifiable. However one can hope to identify the span of index vectors which we denote as $\mathcal{U}^\star \stackrel{\text{def}}{=} \text{Span}(\boldsymbol{u}_1^\star, \boldsymbol{u}_2^\star, \ldots, \boldsymbol{u}_k^\star)$.

We study the index model from a semiparametric point-of-view: the parameter of interest is the subspace $\mathcal{U}^\star$ and the link function g is treated as an unknown nuisance parameter. The advantages

of taking this view are two-fold. First, designing procedures which make weak assumptions on the link function g are robust to misspecification of the link function. Second, this point of view allows us to study the problem of *representation learning* or *feature engineering* in a simple setting. If one is successfully able to estimate the transformation $\boldsymbol{x} \mapsto (\langle \boldsymbol{u}_1^\star, \boldsymbol{x} \rangle, \ldots, \langle \boldsymbol{u}_k^\star, \boldsymbol{x} \rangle)$, there is hope of avoiding the curse of dimensionality in p dimensions and incurring a curse of dimensionality only in k dimensions.

There is a long line of work studying this model in statistics and machine learning. However existing works suffer from one or more of the following drawbacks:

1. Some procedures derive estimators that are solutions to highly non-convex optimization problems. It is unclear when these optimization problems are tractable.

2. Some existing procedures make *ad-hoc* assumptions on the unknown link function g. These assumptions seem to be required for specific procedures to work and don't seem to capture the inherent statistical difficulty of the problem.

3. The sample complexity analysis of some procedures suppresses the dependence of the ambient dimension p in the constants, which makes it unclear whether the sample complexity is polynomial or exponential in p.

In this paper we attempt to address some of these shortcomings. Our main contributions are summarized as follows:

1. For the single-index model, $k = 1$, we provide an explicit formula for the population loss of a natural semiparametric maximum likelihood estimate (SMLE). We show that the population loss has no spurious minima. However it may have degenerate critical points. In Theorem 5, we explicitly characterize the degeneracy of these critical points in terms of a single parameter corresponding to the unknown link function called the Order of Degeneracy of g.

2. Motivated by our analysis of population loss of the SMLE, we design an easy-to-analyze procedure to recover the index vector for the single-index model. In Theorem 14, we analyze the sample complexity of our procedure in terms of statistically motivated parameters R^2 and μ which quantify the smoothness and signal strength of the link function g; see Assumptions 3 and 4 for their definitions. Our procedure (Algorithm 2) requires $\tilde{\mathcal{O}}(p^{\mathcal{O}(R^2/\mu)} + p/\epsilon^2)$ samples where the $\tilde{\mathcal{O}}$ notation suppresses factors polynomial in $1/\delta, \ln(1/\epsilon)$ and $(R^2/\mu)^{R^2/\mu}$ to return an estimate $\hat{\boldsymbol{u}}$ that satisfies $\min(\|\hat{\boldsymbol{u}} - \boldsymbol{u}^\star\|_2, \|\hat{\boldsymbol{u}} + \boldsymbol{u}^\star\|_2) \leq \epsilon$ with probability $1 - O(\delta) - o_n(1) - o_p(1)$. Furthermore, our procedure runs in time $\mathcal{O}(np)$. Our procedure is a natural higher order generalization of ADE (Brillinger, 2012) and PHD (Li, 1992) estimators.

3. For the multi-index case, in Theorem 8, we show that the same procedure can be seen as optimizing an objective which has the desirable property that (nearly) every local extrema in the population objective is an element of the $\mathcal{U}^\star$. Finally, in Theorem 17, we show that with $\tilde{\mathcal{O}}(p^{\mathcal{O}(kR^2/\mu)}/\epsilon^2)$ samples, the procedure returns an estimate $\hat{\boldsymbol{u}}$ such that $\|\mathcal{P}_{\mathcal{U}^\star}(\hat{\boldsymbol{u}})\|_2 \geq 1 - \epsilon$. This means that our procedure for learning single-index model is robust even when a multi-index model is misspecified as a single-index model.

1.1. Related Work

There is a large literature dealing with estimation of index models. We briefly review the different approaches mentioning some representative papers for each approach.

Semiparametric Maximum Likelihood Estimators: A well known estimator for the index vector is the semiparemetric maximum likelihood estimator (SMLE). The basic idea behind SMLE is as follows: Suppose our best guess for the index vectors was $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k$. Given this guess, one could estimate the link function g using a non-parametric estimator such as a kernel smoothing estimator with bandwidth $h : \hat{g}_h(\langle \boldsymbol{u}_1, \boldsymbol{x} \rangle, \ldots, \langle \boldsymbol{u}_k, \boldsymbol{x} \rangle)$. One could also evaluate how good our guess for the index vectors was using a suitable goodness-of-fit statistic such as the Sum-of-Squared Errors $\text{SSE}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) = \sum_{i=1}^n (\hat{g}_h(\langle \boldsymbol{u}_1, \boldsymbol{x}_i \rangle, \ldots, \langle \boldsymbol{u}_k, \boldsymbol{x}_i \rangle) - y_i)^2$. One can then estimate the index vectors by minimizing the goodness of fit statistic. The SMLE is known to have excellent statistical properties in the asymptotic regime where the ambient dimension p is fixed and the number of samples $n \to \infty$ such as $\sqrt{n}$-consistency and asymptotic efficiency under *very weak* assumptions on the distribution of covariates. However, it is not clear whether the optimizing the SSE is tractable. Furthermore, the classical asymptotic analysis does not capture terms in the Mean Squared Error which might be more important in modern day scenarios, where p is often large and comparable to n. See the book by Horowitz (2009) for a nice review on results about the SMLE.

Gradient-Based Estimators: A second approach developed in a series of papers by Hristache et al. (2001b,a); Dalalyan et al. (2008) leverages the observation that the gradients $\nabla f(\boldsymbol{x})$ lies in the span of the index vectors and hence $\mathcal{U}^\star$ can be estimated by running PCA on non-parametric estimators of the gradients, for example the slope of a local-linear regression estimate:

$$(\hat{f}(\boldsymbol{x}), \widehat{\nabla f}(\boldsymbol{x})) = \arg \min_{c \in \mathbb{R}, \beta \in \mathbb{R}^p} \sum_{i=1}^n \left(y_i - c - \langle \boldsymbol{\beta}, \boldsymbol{x}_i - \boldsymbol{x} \rangle\right)^2 K_h\left(\|\boldsymbol{x} - \boldsymbol{x}_i\|\right)$$

for some kernel smoothing function K_h. The problem with this estimate is that it isn't clear if the estimate of the index vectors derived from this non-parametric gradient estimate would even be $\sqrt{n}$-consistent since the gradients are estimated at a slow rate. However Hristache et al. (2001b,a); Dalalyan et al. (2008) show that it is possible to iteratively improve this estimator to get a $\sqrt{n}$-consistent estimator of the span of the index vectors when $k < 4$ under very weak assumptions on the distribution of covariates. Furthermore, their procedure is also computationally tractable. However, their analysis suppresses the dependence of p in constant terms. More precisely, they show that estimating the span of the index vectors up to a tolerance ϵ requires $\mathcal{O}\left(1/\epsilon^2\right)$ samples (which is much better than $\mathcal{O}\left(1/\epsilon^p\right)$), but this $\mathcal{O}\left(\cdot\right)$ suppresses a 2^p factor in the sample size.

Moment-Based Estimators: Another line of work makes assumptions on the covariate distribution (e.g., Gaussian or elliptical). Such assumptions permit one to take advantage of Stein's Lemma and its generalizations to derive moment-based estimators for the index vectors. Specifically if $\boldsymbol{x}_i \sim \mathcal{N}\left(0, \boldsymbol{I}_p\right)$, then,

$$\mathbb{E}[y\boldsymbol{x}] = \mathbb{E}[\nabla f(\boldsymbol{x})], \quad \mathbb{E}[y\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[f(\boldsymbol{x})]\boldsymbol{I}_p = \mathbb{E}[\nabla^2 f(\boldsymbol{x})].$$

Since for multi-index models, $\mathbb{E}[\nabla f(\boldsymbol{x})] \in \mathcal{U}^\star$ and $\text{Range}(\mathbb{E}[\nabla^2 f(\boldsymbol{x})]) \subset \mathcal{U}^\star$, estimates of these moments derived from empirical averages can be used to estimate subspaces of the span of the index vectors. The estimator based on the first moment is called Average Derivative Estimator (ADE) and

was proposed by Brillinger (2012). The estimator based on the second moment is called Principal Hessian Directions (PHD) and was proposed by Li (1992). More recently, these estimators were revisited by Plan et al. (2017) and Neykov et al. (2016) in the context of single-index models. They analyze these estimators providing non-asymptotic bounds with explicit dependence on p. They also extend these estimators to the situation where the index vector has additional structure like sparsity. However, the key drawback of these estimators is that they are not guaranteed to estimate the entire span of the index vectors. For example consider the situation when $k = 1$ and $g(z) = H_3(z)$. Here H_3 denotes the Hermite Polynomial of degree 3. One can check for this example, $\mathbb{E}[\nabla f(\boldsymbol{x})] = \mathbf{0}$ and $\mathbb{E}[\nabla^2 f(\boldsymbol{x})] = \mathbf{0}$. Hence both ADE and PHD fail to recover the index vector. The underlying cause for this failure mode is that both ADE and PHD extract index vectors participating in the first two harmonics of the link function g and can miss out on index vectors involved in higher order harmonics of the link function g.

Slicing: A partial solution to the failure of moment-based estimators to the entire subspace $\mathcal{U}^\star$ is a technique called Slicing. This technique was introduced in Li (1991) and is based on the observation that almost surely with respect to y,

$$\mathbb{E}[\boldsymbol{x}|y] \in \mathcal{U}^\star, \quad \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T|y] - I) \subset \mathcal{U}^\star.$$

The advantage of slicing is that one can now estimate the sliced moments $\mathbb{E}[\boldsymbol{x}|y]$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T|y]$ for a number of different values of y and this hopefully reduces the chance of missing certain relevant directions. However, even Sliced Inverse Regression is guaranteed to consistently capture all the relevant directions under ad-hoc assumptions like:

$$\text{Rank}\left(\mathbb{E}_y\left[\mathbb{E}\left[\boldsymbol{x}|y\right]\mathbb{E}\left[\boldsymbol{x}|y\right]^T\right]\right) = k, \tag{3}$$

$$\text{Rank}\left(\mathbb{E}_y\left[\left(\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^T|y\right] - I\right)^2\right]\right) = k. \tag{4}$$

For Equation (3), it is easy to see that the phase retrieval problem violates this assumption since the link function $g(z) = z^2$ is even and hence $\mathbb{E}[\boldsymbol{x}|y] = 0$. We are not aware of any counterexamples to Equation (4) nor a proof that this assumption holds for an arbitrary link function. Due to assumptions like these, the analysis of slicing depends on parameters like the $\lambda_{\min}\left(\mathbb{E}_y\left[\mathbb{E}\left[\boldsymbol{x}|y\right]\mathbb{E}\left[\boldsymbol{x}|y\right]^T\right]\right)$ and the smoothness of the function $s(y) \stackrel{\text{def}}{=} \mathbb{E}[\boldsymbol{x}|y]$. It is not clear how these parameters relate to more natural notions of signal strength for this problem or to the smoothness of the underlying link function. We refer the reader to Babichev and Bach (2016) for a non-asymptotic analysis of Sliced Inverse Regression and a discussion of the failure modes of various slicing strategies.

Other related work: Recent concurrent work of Ge et al. (2017) uses techniques based on Hermite polynomials similar to ours to learn neural networks of the form $\sum_{i=1}^m a_i^\star \sigma(\langle \boldsymbol{b}_i^\star, \boldsymbol{x}\rangle)$ where $a_i^\star \geq 0$ and $\boldsymbol{b}_i$ are linearly independent. They assume that the link function σ is known (e.g., ReLU or $\tanh$ activation), and leverage this knowledge to design an objective depending on σ that has benign structure (e.g., no spurious local minima or degenerate critical points). Since we take a semi-parametric point-of-view, we are unable to do this. In particular we need to handle objectives with degenerate critical points.

1.2. Assumptions

We assume that we have n data points generated the regression model defined in Equation (1) with a multi-index regression function of order k (defined in Equation (2)). We make the following assumptions on the unknown link function (with $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$):

Assumption 1 (Normalization) $\mathbb{E}[g^2(\boldsymbol{z})] = 1$.

Assumption 2 (Bounded Link Function) $\|g\|_\infty < \infty$.

Remark 1 *Assumption 2 is not strictly required for our analysis. It can be relaxed to allow for link functions such as $g(z) = z^2$ in the case of phase retrieval. This is done in Appendix D.*

Assumption 3 (Smoothness) *g is twice differentiable, and $\mathbb{E}[(\frac{\partial^2 g(\boldsymbol{z})}{\partial z_i \partial z_j})^2] \leq R^2$ for all $i, j \in [k]$.*

Assumption 4 (Minimum Signal Strength) $\mathbb{E}[(\frac{\partial g(\boldsymbol{z})}{\partial z_i})^2] \geq \mu$ *for all* $i \in [k]$.

Remark 2 *We note that μ is a very natural notion of signal strength in this problem. If Assumption 4 is violated for some coordinate $i \in [k]$, we have*

$$\mathbb{E}\left[\left(\frac{\partial g(\boldsymbol{z})}{\partial z_i}\right)^2\right] = 0 \implies \frac{\partial g(\boldsymbol{z})}{\partial z_i} \overset{\text{a.s.}}{=} 0 \implies g(z_1, z_2, \ldots, z_k) = g'(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_k).$$

This means that the value of the response y is independent of the projection of the covariate along the direction $\boldsymbol{u}_i^\star$ and hence, we cannot possibly hope to recover it.

1.3. Notation

Notation for Hermite Polynomials: We will represent the unknown link function using orthogonal polynomials for the Gaussian measure called *Hermite polynomials*. Let $H_i(z)$ denote the (normalized) Hermite polynomial of degree i. These polynomials form an orthonormal basis for square-integrable univariate functions under the Gaussian measure. Hence, in the case of the single-index model ($k = 1$), the unknown link function g admits an expansion in the Hermite polynomial basis of the form:

$$g(z) = \sum_{l=0}^{\infty} a_l^\star H_l(z).$$

We define the following index sets: $\mathcal{I}_t \overset{\text{def}}{=} \{\boldsymbol{S} \in (\mathbb{N} \cup \{0\})^k : \sum_{i=1}^k S_i \leq t\}$ and $\mathcal{I}_\infty \overset{\text{def}}{=} \bigcup_{t=0}^\infty \mathcal{I}_t$; and we use the notation $H_{\boldsymbol{S}}(\boldsymbol{z})$ for $\boldsymbol{S} \in \mathcal{I}_\infty$ to denote the tensor-product Hermite polynomial basis: $H_{\boldsymbol{S}}(\boldsymbol{z}) \overset{\text{def}}{=} \prod_{i=1}^k H_{S_i}(z_i)$. Analogously, we use the notation $\boldsymbol{z}^{\boldsymbol{S}}$ for $\boldsymbol{S} \in \mathcal{I}_\infty$ to denote the monomial: $\boldsymbol{z}^{\boldsymbol{S}} \overset{\text{def}}{=} \prod_{i=1}^k z_i^{S_i}$. The tensor-product Hermite polynomials form an orthonormal basis for square integrable k-variate functions under the product Gaussian measure. Hence for the multi-index model, the unknown link function g admits an expansion of the form:

$$g(\boldsymbol{z}) = \sum_{\boldsymbol{S} \in \mathcal{I}_\infty} a_{\boldsymbol{S}}^\star H_{\boldsymbol{S}}(\boldsymbol{z}).$$

Notation for Linear Algebraic Aspects: For a vector $\boldsymbol{v} \in \mathbb{R}^p$, we use $\|\boldsymbol{v}\|_1$ to denote the L1 norm and $\|\boldsymbol{v}\|_2$ to denote the L2 norm. If the subscript is omitted, $\|\boldsymbol{v}\|$ refers to the L2 norm. For matrices, $\|\boldsymbol{A}\|$ represents the operator norm. The notation $\mathcal{P}_{\mathcal{U}^\star}$ and $\mathcal{P}_{\mathcal{U}^\star}^{\perp}$ refers to projection operators onto $\mathcal{U}^\star$ and the orthogonal complement $\mathcal{U}^{\star\perp}$, respectively. The unit sphere in $\mathbb{R}^p$ is denoted by $\mathbb{S}^{p-1}$. Finally, we define the matrix of orthonormal index vectors $\boldsymbol{U}^\star = [\boldsymbol{u}_1^\star, \boldsymbol{u}_2^\star, \ldots, \boldsymbol{u}_k^\star] \in \mathbb{R}^{p\times k}$.

Notation for Probabilistic Aspects: We use the notation $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ to represent the standard Gaussian distribution in p dimensions. In particular the statement $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ means the random variable $\boldsymbol{X}$ is drawn from a p-variate standard Gaussian distribution. Analogously $\boldsymbol{u}_0 \sim$ Uniform$(\mathbb{S}^{p-1})$ means that $\boldsymbol{u}_0$ is a uniformly random unit vector.

Outline: The remaining paper is organized as follows. In Section 2 we analyze the landscape of the semi-parametric MLE. We propose a simple objective for the estimating single-index models and analyze its landscape when a multi-index model is misspecified as a single-index one. In Section 3 we construct and analyze a procedure to estimate single-index models from finite samples. In Section 4 we analyze the behaviour of this procedure under a multi-index misspecification. We conclude with Section 5 and discuss some open problems. All omitted proofs are in the Appendix.

2. Landscape of Some Population Objectives

A commonly used estimator for single-index models is the semiparametric MLE which is defined as follows:

$$\hat{\boldsymbol{u}} = \arg\min_{u\in\mathbb{S}^{p-1}} \min_{h\in\mathcal{F}_L} \frac{1}{n}\sum_{i=1}^{n}(y_i - h(\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle))^2.$$

In the above display $\mathcal{F}_L$ is a suitable class of functions from $\mathbb{R}$ to $\mathbb{R}$. The parameter L controls the complexity of the function class and is tuned to achieve an optimal tradeoff between the bias and the variance of the resulting estimator. For example a simple choice for $\mathcal{F}_L$ would be the set of all degree L polynomials: $\mathcal{F}_L = \{g : \mathbb{R} \to \mathbb{R}, g(z) = \sum_{i=0}^{L} a_i H_i(z)\}$. For this function class, the SMLE becomes:

$$\hat{\boldsymbol{u}} = \arg\min_{u\in\mathbb{S}^{p-1}} \min_{\boldsymbol{a}\in\mathbb{R}^{L+1}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=0}^{L} a_l H_l(\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle)\right)^2 \tag{OPT 1}$$

It is not clear if Optimization Problem OPT 1 is tractable. The first step in understanding its complexity is to understand the landscape of the associated population loss:

$$R_L(\boldsymbol{u}) := \min_{\boldsymbol{a}\in\mathbb{R}^{L+1}} \mathbb{E}\left[\left(y - \sum_{l=0}^{L} a_l H_l(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)\right)^2\right].$$

It turns out that it is possible to give an explicit expression of the population loss due to a surprising algebraic property of Hermite Polynomials stated below.

Lemma 3 (O'Donnell, 2014) *Let $\boldsymbol{u}, \boldsymbol{v}$ be unit vectors in $\mathbb{R}^p$. For $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$,*

$$\mathbb{E}[H_l(\langle \boldsymbol{u}, \boldsymbol{x}\rangle) H_m(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)] = \begin{cases} 0 & l \neq m, \\ \langle \boldsymbol{u}, \boldsymbol{v}\rangle^l & l = m. \end{cases}$$

Our first result (Theorem 5) is that the objective $R_L(\boldsymbol{u})$ has precisely two local minima at $\boldsymbol{u} = \pm\boldsymbol{u}^\star$. However it may have degenerate critical points. The degeneracy of the critical points of $R_L(\boldsymbol{u})$ is determined by the degree of the smallest non-zero harmonic in the link function f. More precisely we define the following notion of the order of degeneracy of f:

Definition 4 (Order of Degenerarcy of f) *The order of degeneracy of f, denoted by OD(f), is defined as:*

$$\mathit{OD}(f) := \min\left\{l \in [L] : \mathbb{E}_{x\sim\mathcal{N}(0,\boldsymbol{I}_p)}\left[f(\boldsymbol{x})H_l(\langle\boldsymbol{x},\boldsymbol{u}^\star\rangle)\right] \neq 0\right\}.$$

Theorem 5 *The population loss admits the explicit form:*

$$R_L(\boldsymbol{u}) = \sigma^2 + \sum_{l=1}^{L} a_l^{\star 2}(1 - \langle\boldsymbol{u},\boldsymbol{u}^\star\rangle^{2l}).$$

The critical points of $R_L(\boldsymbol{u})$ are given by:

1. $\boldsymbol{u} = \pm\boldsymbol{u}^\star$*, which are global minima.*
2. $\boldsymbol{u} \in \{\boldsymbol{a} \in \mathbb{S}^{p-1} : \langle\boldsymbol{a},\boldsymbol{u}^\star\rangle = 0\}$*. All points in this subspace are global maxima. Furthermore, when* OD$(f) > 1$*, these local maxima are degenerate, that is, $\nabla^2 R_L(\boldsymbol{u}) = \boldsymbol{0}$.*

While the objective $R_L(\boldsymbol{u})$ has no spurious local minima, there are a few issues with it:

1. Since the objective squares the residuals, it increases the effective order of degeneracy of the function by a factor of two. This increases the number of samples required to guarantee that the landscape of the objective is well-behaved.
2. The coefficient vector $\boldsymbol{a}$ that minimizes the objective is dependent on the data. This dependence makes the analysis more complicated.

To avoid these difficulties, we instead optimize the following objective for some value of l and get an estimate $\hat{\boldsymbol{u}}_l$:

$$\hat{F}_l(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} y_i H_l(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle), \quad \hat{\boldsymbol{u}}_l = \arg\max_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \hat{F}_l(\boldsymbol{u}).$$

To understand why this objective is reasonable, we consider the population version of the objective:

$$\begin{aligned} F_l(\boldsymbol{u}) &= \mathbb{E}\left[f(\boldsymbol{x})H_l(\langle\boldsymbol{x},\boldsymbol{u}\rangle)\right] \\ &\overset{\text{Lemma 3}}{=} a_l^\star \langle\boldsymbol{u},\boldsymbol{u}^\star\rangle^l. \end{aligned}$$

Hence, $F_l(\boldsymbol{u})$ is extremized at $\hat{\boldsymbol{u}} = \pm\boldsymbol{u}^\star$ provided $a_l^\star \neq 0$. Intuitively, given a harmonic l, the objective $F_l(\boldsymbol{u})$ tries to orient the vector $\boldsymbol{u}$ in such a way that the energy in this harmonic is maximized.

Remark 6 (Degenerate Critical Points) *When $l > 2$, $F_l(\boldsymbol{u})$ has a degenerate critical points at all $\boldsymbol{u}^\perp$ that satisfy $\langle\boldsymbol{u}^\star,\boldsymbol{u}^\perp\rangle = 0$. In particular, this means that for $l > 2$, our objective does not satisfy the strict-saddle (Ge et al., 2015) or the ridable saddle properties (Sun et al., 2015). Hence, it is not immediately clear if these generic analysis methods can be applied here.*

Let us now consider the situation where a multi-index model of order k was misspecified as a single-index model. One might still hope that optimizing $F_l(\boldsymbol{u})$ does something reasonable. It turns out that the objective $F_l(\boldsymbol{u})$ indeed has this desirable property. One can write an explicit form for $F_l(\boldsymbol{u})$ when the data is generated by a multi-index model of order k. The key tool that allows us to do this is a multivariate analogue of Lemma 3 which is stated below.

Lemma 7 *Let $\boldsymbol{U} := [\boldsymbol{u}_1, \boldsymbol{u}_2 \ldots, \boldsymbol{u}_k]$ be a matrix in $\mathbb{R}^{p\times k}$ with orthonormal columns. Let $\boldsymbol{v}$ be a arbitrary unit vector. Then,*

$$1) \quad \mathbb{E}\left[H_{\boldsymbol{S}}(\boldsymbol{U}x)H_l(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)\right] = 0 \qquad \text{if } l \neq |\boldsymbol{S}|,$$

$$2) \quad \mathbb{E}\left[H_{\boldsymbol{S}}(\boldsymbol{U}x)H_l(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)\right] = \sqrt{\frac{l!}{S_1!S_2!\cdots S_k!}}\prod_{i=1}^{k}\langle \boldsymbol{u}_i, \boldsymbol{v}\rangle^{S_i} \qquad \text{if } l = |\boldsymbol{S}|.$$

Our next result (Theorem 8) characterizes the landscape of the objective $F_l(\boldsymbol{u})$ under a multi-index model and shows that it has no spurious local extrema.

Theorem 8 *Under the order k multi-index model, the population objective $F_l(\boldsymbol{u})$ has the following properties:*

1. *The population objective has the following explicit form:*

$$F_l(\boldsymbol{u}) = \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1=l} a^\star_{\boldsymbol{S}}\sqrt{\binom{l}{S_1, S_2, \ldots, S_k}}\prod_{i=1}^{k}\langle \boldsymbol{u}_i^\star, \boldsymbol{u}\rangle^{S_i}.$$

2. *Any local maximizer with $F_l(\boldsymbol{u}) > 0$ is contained in the subspace $\mathcal{U}^\star$.*

3. *Any local minimizer with $F_l(\boldsymbol{u}) < 0$ is contained in the subspace $\mathcal{U}^\star$.*

Remark 9 *We note that Theorem 8 falls short of showing that $F_l(\boldsymbol{u})$ has no spurious local minima or maxima. In particular existence of local maxima (or minima) not in $\mathcal{U}^\star$ such that $F_l(\boldsymbol{u}) = 0$ is not ruled out. However such local maxima are unlikely to be a cause of problems for procedures like gradient ascent (or descent). To see this consider the following procedure: Choose a random initialization point. With probability 1, we will have $F_l(\boldsymbol{u}_0) \neq 0$. If the objective is positive, run gradient ascent otherwise run gradient descent. Since gradient ascent with small enough step size is guaranteed to increase the objective, we will never get stuck in a local maxima with $F_l(\boldsymbol{u}) = 0$.*

3. Learning Single-index Models from Finite Samples

In Section 2 we showed that the population version of the objective $\hat{F}_l(\boldsymbol{v})$ is extremized at the true index vector $\boldsymbol{u}^\star$ provided that the energy of the link function in the harmonic l is not zero ($a_l^\star \neq 0$). In this section, we first design a procedure (Algorithm 1) that extracts an estimate of $\boldsymbol{u}^\star$ from harmonic l under the promise that $a_l^\star \neq 0$.

Algorithm 1 exhibits extremely fast convergence. In particular, it requires only two update steps to return an estimate. The underlying reason behind fast convergence for this algorithm is that the gradient of the population objective is *perfectly* aligned with the index vector $\boldsymbol{u}^\star$. The analysis of this procedure is presented in Theorem 10. At a high level, the analysis of this procedure involved the following steps:

Algorithm 1 Estimate-Index-Vector-from-Harmonic(S, l)

input Data $S = \{\boldsymbol{x}_i, y_i\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$; Degree of Harmonic $l \in \mathbb{N}$

output Index Estimate $\hat{\boldsymbol{u}}_l \in \mathbb{R}^p$.

1: Split S into two equal parts:
$S_1 := \{(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, \frac{n}{2}\}, S_2 := \{(\boldsymbol{x}_i, y_i) : i = \frac{n}{2} + 1, \ldots, n\}$

2: Define $\hat{F}_l(\boldsymbol{u}; S_1) := \frac{2}{n} \sum_{i=1}^{\frac{n}{2}} y_i H_l(\langle \boldsymbol{x}_i, \boldsymbol{u} \rangle)$ and $\hat{F}_l(\boldsymbol{u}; S_2) := \frac{2}{n} \sum_{i=\frac{n}{2}+1}^{n} y_i H_l(\langle \boldsymbol{x}_i, \boldsymbol{u} \rangle)$

3: Random Initialization: $\boldsymbol{u}_0 \sim \text{Uniform}(\mathbb{S}^{p-1})$

4: Compute two steps of iterative process: $\boldsymbol{u}_1 = \frac{\nabla \hat{F}_l(\boldsymbol{u}_0; S_1)}{\|\nabla \hat{F}_l(\boldsymbol{u}_0; S_1)\|_2}$, and then $\boldsymbol{u}_2 = \frac{\nabla \hat{F}_l(\boldsymbol{u}_1; S_2)}{\|\nabla \hat{F}_l(\boldsymbol{u}_1; S_2)\|_2}$.

5: **return** $\hat{\boldsymbol{u}}_l := \boldsymbol{u}_2$

1. Analysis of Random Initialization: We expect a uniformly random unit vector to have a projection of size $\Omega(1/\sqrt{p})$ on $\boldsymbol{u}^\star$. This means that the initialization is very close to a degenerate critical point and hence we see *very small gradients* of size $\mathcal{O}\left(1/p^l\right)$

2. Analysis of the stochastic fluctuations of the gradient: Because our gradients are heavy tailed, this is done via a standard truncation argument. We show that $\|\nabla F_l(\boldsymbol{u}) - \nabla \hat{F}_l(\boldsymbol{u})\|_2 \leq \tilde{\mathcal{O}}(\sqrt{p/n})$. Hence when $n \geq \mathcal{O}(p^l)$, the stochastic fluctuations don't kill off the small gradients we observe at the initialization. This initial sample size requirement of $\mathcal{O}\left(p^l\right)$ can be seen as the price to pay to escape the degenerate local critical points.

3. Analysis of Iterates: We show that because the gradient of the objective is perfectly aligned with the index vector, $\|\hat{\boldsymbol{u}}_l - \boldsymbol{u}^\star\|_2 \leq 1$ at the end of the first iteration and $\|\hat{\boldsymbol{u}}_l - \boldsymbol{u}^\star\|_2 \leq \tilde{\mathcal{O}}(\sqrt{p/n})$ at the end of the second iteration.

Theorem 10 *Given any $\epsilon, \delta \in (0, 1)$; with probability $1 - 2\exp(-p/32) - 5\delta - \frac{8}{n}$, the output $\hat{\boldsymbol{u}}_l$ of Algorithm 1 satisfies*

$$|\langle \hat{\boldsymbol{u}}_l, \boldsymbol{u}^\star \rangle| \geq 1 - \frac{100(\|f\|_\infty + 4\sigma) \cdot 2^{2l+1}}{l|a_l^\star|} \sqrt{\frac{2\max(p, \ln(1/\delta)) \ln^l(n)}{n}},$$

provided n is large enough so that the following holds:

$$\frac{n}{\ln^l(n)} \geq \frac{32 \cdot 10^4 (\|f\|_\infty + 4\sigma)^2}{l^2 {a_l^\star}^2} \frac{2^{2l}}{\delta^{2l-2}} \max(p, \ln(1/\delta)) p^{l-1}.$$

Remark 11 (Connections to 1-bit Compressed Sensing) *Consider the 1-bit compressed sensing problem where $g(z) = sign(z)$. One can check that for this link function, $a_1^\star = \sqrt{2/\pi} > 0$. Hence, when specialized to this case, Theorem 10 gives a sample complexity of $O(p \ln(p))$ which is optimal for unstructured signals up to log-factors.*

Remark 12 (Connections to Phase Retrieval) *Consider the phase retrieval problem where $g(z) = z^2 = \sqrt{2}H_2(z) + 1$. A common approach to get an estimate of $\mathbf{u}^\star$ is by computing the leading*

Algorithm 2 Learn-single-index-Model($S, R^2, \mu, \sigma^2, \|f\|_\infty, \delta$)

Input Data: $S = \{\boldsymbol{x}_i, y_i\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$;
smoothness parameter R^2; minimum signal strength parameter μ; noise variance σ^2; upper bound on link function $\|f\|_\infty$; confidence parameter δ.

Output Index Estimate $\hat{\boldsymbol{u}} \in \mathbb{R}^p$.

1: Split S into S_{train} and S_{test} such that: $m := |S_{\text{test}}| = 256 \cdot 2^{\frac{4R^2}{\mu}} R^4(\sigma^2 + \|f\|_\infty^2)/(\delta\mu^3)$
2: Let $L := \frac{2R^2}{\mu}$
3: Let $\hat{\boldsymbol{u}}_l :=$ Estimate-Index-Vector-From-Harmonic(S_{train}, l) for each $l \in \{1, 2, \ldots, L\}$.
4: Compute $T_l := \sum_{i \in S_{\text{test}}} y_i H_l(\langle \boldsymbol{x}_i, \hat{\boldsymbol{u}}_l \rangle)/m$ for each $l \in \{1, 2, \ldots, L\}$.
5: Let $l_{\text{best}} := \arg\max_{l \in [L]} |T_l|$
6: **return** $\hat{\boldsymbol{u}} := \boldsymbol{u}_{l_{\text{best}}}$

eigenvector of the matrix $\widehat{\boldsymbol{M}} = \sum_{i=1} y_i \boldsymbol{x}_i \boldsymbol{x}_i^T$. Using the variational characterization of the leading eigenvector, one can see that this is exactly the same as optimizing the objective $\hat{F}_2$. When specialized to this case, Theorem 10 gives a suboptimal sample complexity of $\mathcal{O}(p^2)$, but more specialized analyses of the same estimator using matrix perturbation tools do give the optimal sample complexity of $\mathcal{O}(p)$ (see, e.g., Candes et al., 2015).

We emphasize that Algorithm 1 succeeds only if we know a harmonic l such that $a_l^\star$ is not too small. In order to design an algorithm that learns single-index models with arbitrary link functions satisfying our assumptions we show that for any such link function, there exists a bounded $l_\sharp \in \mathbb{N}$ such that $a_{l_\sharp}^\star$ is not too small.

Lemma 13 (Existence of a good $l_\sharp$) *For a link function g that satisfies Assumptions 1, 3 and 4, there exists a $l_\sharp \leq \frac{2R^2}{\mu}$ such that, $l_\sharp |a_{l_\sharp}^\star|^2 \geq \frac{\mu^2}{4R^2}$.*

While Lemma 13 guarantees the *existence* of a $l_\sharp$ it does not tell us which value of $l_\sharp$ should be used for an unknown link function. A simple solution is to construct estimates $\hat{\boldsymbol{u}}_l$ for all values of $l \in \{1, 2, \ldots, 2R^2/\mu\}$ and choose the one with the best performance on a held-out data set via some goodness-of-fit statistic. This is implemented in Algorithm 2 for learning single-index models.

Theorem 14 *Given any $\epsilon, \delta \in (0, 1)$; with probability $1 - \frac{4R^2}{\mu} e^{-p/32} - \frac{12R^2}{\mu}\delta - \frac{16R^2}{n\mu}$, the estimate returned by Algorithm 2, $\hat{\boldsymbol{u}}$ satisfies*

$$|\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}} \rangle| \geq 1 - \frac{3200 \cdot 2^{\frac{4R^2}{\mu}} (\|f\|_\infty + 4\sigma) R^4}{\mu^2 \sqrt{\mu}} \sqrt{\frac{\max\left(p, \ln\left(\frac{1}{\delta}\right)\right) \ln^{\frac{2R^2}{\mu}}(n)}{n}},$$

provided that n satisfies

$$\frac{n}{\ln^{\frac{2R^2}{\mu}}(n)} \geq \frac{1024 \cdot 10^4 (\|f\|_\infty + 4\sigma)^2 R^4}{\mu^3} \cdot \frac{2^{\frac{4R^2}{\mu}}}{\delta^{\frac{4R^2}{\mu} - 2}} \max\left(p, \ln\left(\frac{1}{\delta}\right)\right) p^{\frac{2R^2}{\mu} - 1}.$$

Remark 15 *If we treat $\mu, \|f\|_\infty, \sigma, R$ and δ as fixed constants, Theorem 14 states that Algorithm 2 requires $\tilde{\mathcal{O}}(p^{2R^2/\mu} + p/\epsilon^2)$ samples to return an estimate $\hat{\boldsymbol{u}}$ such that $\min(\|\boldsymbol{u}^\star - \hat{\boldsymbol{u}}\|, \|\boldsymbol{u}^\star + \hat{\boldsymbol{u}}\|) \leq \epsilon$.*

Algorithm 3 Learn-single-index-Model-Robust$(S, \mu, R^2, K_{\text{max}})$

input Data $S = \{\boldsymbol{x}_i, y_i\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$; Smoothness Parameter R^2; Minimum Signal Strength μ; Upper Bound on true k, K_{max}

output Index Estimate $\hat{\boldsymbol{u}} \in \mathbb{R}^p$.

1: Set $L := \frac{2K_{\text{max}}R^2}{\mu} + K_{\text{max}} - 1$

2: Random Initialization: $\boldsymbol{u}_0 \sim \text{Uniform}(\mathbb{S}^{p-1})$

3: Compute: $l_{\text{best}} := \arg\max_{l \in \{1,2,\ldots,L\}} \|\nabla \hat{F}_l(\boldsymbol{u}_0)\|$.

4: **return** $\hat{\boldsymbol{u}} := \frac{\nabla \hat{F}_{l_{\text{best}}}(\boldsymbol{u})}{\|\nabla \hat{F}_{l_{\text{best}}}(\boldsymbol{u})\|_2}$

4. Learning Misspecified Single-index Models from Finite Samples

We recall that in Theorem 8 we showed that even when a multi-index model was misspecified as a single-index one, essentially all local extrema of the objective $F_l(\boldsymbol{u})$ belong to the index space $\mathcal{U}^\star$. In this section we show that with a finite sample size, with minor modifications (shown as Algorithm 3 in Appendix C), Algorithm 2 returns a vector approximately in the index space $\mathcal{U}^\star$ when the data is generated order k multi-index model.

Recall the form of the population loss F_l for a given harmonic l and given in Theorem 8. In particular, it is possible that for some l, the coefficients $a^\star_{\boldsymbol{S}} = 0$ for all $\boldsymbol{S}$ such that $\|\boldsymbol{S}\|_1 = l$. For such l, the estimate computed by Algorithm 3, $\hat{\boldsymbol{u}}_l$ is expected to be useless. Analogous to Lemma 13 in the single-index case, Lemma 16 shows that there exists a bounded $l_\sharp$ such that the associated coefficients $a^\star_{\boldsymbol{S}}$ are not too small.

Lemma 16 (Existence of a good $l_\sharp$) *Let g be a link function from $\mathbb{R}^k \to \mathbb{R}$ obeying Assumptions 1, 3 and 4. Then, there exists an $l_\sharp \in \mathbb{N}$ such that:*

$$l_\sharp \leq \frac{2kR^2}{\mu} + k - 1, \qquad \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1 = l_\sharp} a^2_{\boldsymbol{S}} \|\boldsymbol{S}\|_1 \geq \frac{\mu^2}{2(2R^2 + \mu)}.$$

Even if we knew a good $l_\sharp$ as guaranteed by Lemma 16, since we initialize using a uniformly random unit vector which will be close to orthogonal to the true index subspace $\mathcal{U}^\star$, the initial gradient we will observe will be very small. The key challenge here is to develop a lower bound on the norm of the observed gradient in terms of the minimum signal strenth parameter (μ) and the smoothness parameter (R). We address this challenge by exploiting the form of the population gradient and applying the Carbery-Wright anticoncentration inequality for Gaussian polynomials.

As in the single-index case, the gradient of the population objective lies in the index space $\mathcal{U}^\star$. Thus, Algorithm 3 returns an estimate with an arbitrarily small constant projection on $\mathcal{U}^{\star\perp}$ with a single update step. The analysis of the sampling error in the gradients and the first update step is identical to the single-index case. Sample complexity of this procedure is analyzed in Theorem 17.

Theorem 17 *There is an absolute constant C such that the following holds. Given any $\epsilon, \delta \in (0,1)$; with probability $1 - 2\delta K_{max}(2R^2/\mu + 1) - 4K_{max}\left(2R^2/\mu + 1\right) \cdot n^{-1} - 2\exp(-p/32)$, the*

estimate returned by Algorithm 3 satisfies $\|\mathcal{P}^{\perp}_{\mathcal{U}^\star}(\hat{\boldsymbol{u}})\|_2 \leq \epsilon$*, provided* n *satisfies*

$$\frac{n}{\ln^{K_{max}\left(\frac{2R^2}{\mu}+1\right)}(n)} \geq \frac{C(\|f\|_\infty+\sigma)^2(R^2+\mu)}{\epsilon^2\mu^2}\left(\frac{p\cdot K_{max}^4\left(\frac{R^2}{\mu}+1\right)^4 K_{CW}^2\ln(1/\delta)}{\delta^2}\right)^{K_{max}\left(\frac{2R^2}{\mu}+1\right)}.$$

In the above display, K_{max} *is an upper bound on the true* k *given to the algorithm and* $K_{CW}>1$ *is a universal constant appearing in the Carbery-Wright Theorem.*

Remark 18 *If we treat* $\|f\|_\infty, \sigma, R^2, \mu, \delta$ *as constants, Theorem 17 states that in order to return an estimate* $\hat{\boldsymbol{u}}$ *such that* $\|\mathcal{P}^{\perp}_{\mathcal{U}^\star}(\hat{\boldsymbol{u}})\|_2 \leq \epsilon$*, Algorithm 3 requires* $\tilde{O}(p^{O(K_{max}R^2/\mu)}/\epsilon^2)$*. Note that with* $K_{max}=1$ *this sample complexity is worse than the* $\tilde{O}(p^{O(R^2/\mu)}+p/\epsilon^2)$ *sample complexity of Algorithm 2 in the single-index case. Due to the more complex structure of the gradients for Algorithm 3, we are only able to analyze one update step in contrast to two update steps for Algorithm 2.*

5. Conclusion and Future Work

In this paper, we studied the problem of estimating the unknown index space $\mathcal{U}^\star$ for single and multi-index models under natural smoothness and minimum signal strength assumptions on the link function. In the case of single-index models, we characterized the population landscape of a natural semi-parametric MLE. We found that though the landscape has no spurious minima, but it may have degenerate critical points which cannot be distinguished from local minima using the first- and second-derivative information and can possibly create problems for gradient-based procedures. We analyzed a simple iterative procedure for estimating the index vector and showed that it returns an ϵ-close estimate of the true index vector with $\tilde{\mathcal{O}}(p^{\mathcal{O}(R^2/\mu)}+p/\epsilon^2)$ samples. The procedure had an appealing robustness property: if a multi-index model is misspecified as single-index, essentially the same procedure recovers a vector ϵ-close to the index space with $\tilde{\mathcal{O}}(p^{\mathcal{O}(K_{\max}R^2/\mu)}/\epsilon^2)$ samples.

A major open question that remains is whether the dependence of sample complexity on p can be made linear without sacrificing computational efficiency. The $p^{\mathcal{O}(R^2/\mu)}$ dependence in our sample complexities appears to be more of a price to pay to escape degenerate critical points than a information-theoretic requirement. A simple idea to explore would be to investigate if it is possible to transform the response using a transformation $\mathcal{T}$ such that the composed link function $\mathcal{T}\circ g$ has enough energy in harmonics of order $l=1,2$. The reason for choosing $l=1,2$ is that it is precisely for these values of l that the objective $F_l(\boldsymbol{u})$ has the strict saddle property. It seems likely that the optimal transformation $\mathcal{T}$ would be data-driven and have links to sliced inverse regression.

One drawback of our approach is that it is tailored to Gaussian covariates. Even when the covariates are i.i.d. and subgaussian, estimating the index vector is not possible without additional assumptions: Ai et al. (2014) exhibit a counter-example of two index vectors that cannot be distinguished using samples from a single index model with the sign-link function and i.i.d. Rademacher covariates. However, when the index vector is far from sparse (i.e., $\|\boldsymbol{u}\|_\infty \ll 1$), Ai et al. (2014) leverage high-dimensional central limit theorems to handle independent subgaussian designs. An interesting question for future work is to see if their techniques can be extended to our estimators.

Acknowledgments

DH was supported in part by NSF grant IIS-1563785 and a Sloan Fellowship.

References

Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

Dmitry Babichev and Francis Bach. Slice inverse regression with score functions. 2016.

David R Brillinger. A generalized linear model with "Gaussian" regressor variables. In *Selected Works of David Brillinger*, pages 589–606. Springer, 2012.

Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

A Carbery and J Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical research letters*, 8(3):233–248, 2001.

Arnak S Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9(Aug):1647–1678, 2008.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points: online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Joel L Horowitz. *Semiparametric and nonparametric methods in econometrics*, volume 12. Springer, 2009.

Marian Hristache, Anatoli Juditsky, Jörg Polzehl, and Vladimir Spokoiny. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001a.

Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001b.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Ker-Chau Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420): 1025–1039, 1992.

Matey Neykov, Zhaoran Wang, and Han Liu. Agnostic estimation for misspecified phase retrieval models. In *Advances in Neural Information Processing Systems*, pages 4089–4097, 2016.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Martin Wainwright. Basic tail and concentration bounds, 2015. URL https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf.

Appendix A. Missing Proofs from Section 2

A.1. Single-index model

Theorem 19 (Theorem 5 restated) *The population loss admits the explicit form:*

$$R_L(\boldsymbol{u}) = \sigma^2 + \sum_{l=1}^{L} a_l^{\star 2}(1 - \langle \boldsymbol{u}, \boldsymbol{u}^\star \rangle^{2l}).$$

The critical points of $R_L(\boldsymbol{u})$ are given by:

1. $\boldsymbol{u} = \pm \boldsymbol{u}^\star$*, these points are global minima.*
2. $\boldsymbol{u} \in \{\boldsymbol{a} \in \mathbb{S}^{p-1} : \langle \boldsymbol{a}, \boldsymbol{u}^\star \rangle = 0\}$*. All points in this subspace are global maxima. Furthermore, when OD$(f) > 1$, these local maxima are degenerate.*

Proof We first note that since $y = f(\boldsymbol{x}) + \epsilon$ and $\mathbb{E}[\epsilon] = 0, \mathbb{E}[\epsilon^2] = \sigma^2$. Using the Bias-Variance decomposition we have,

$$R_L(\boldsymbol{u}) = \sigma^2 + \min_{\boldsymbol{a} \in \mathbb{R}^{L+1}} \mathbb{E}\left[\left(f(\boldsymbol{x}) - \sum_{l=0}^{L} a_l H_l(\langle \boldsymbol{u}, \boldsymbol{x} \rangle)\right)^2\right].$$

Since the multivariate Hermite polynomials form an Orthonormal Basis for $L^2[\mathcal{N}(0, \boldsymbol{I}_p)]$, the value of $\boldsymbol{a}$ which minimizes the expected square loss is given by:

$$\begin{aligned} a_l(\boldsymbol{u}) &= \langle f, H_l(\langle \boldsymbol{u}, \boldsymbol{x} \rangle) \rangle \\ &\overset{\text{Lemma 3}}{=} a_l^\star \langle \boldsymbol{u}, \boldsymbol{u}^\star \rangle^l. \end{aligned}$$

Using the Pythagorean Theorem,

$$\begin{aligned} \mathbb{E}\left[\left(f(\boldsymbol{x}) - \sum_{l=0}^{L} a_l H_l(\langle \boldsymbol{u}, \boldsymbol{x} \rangle)\right)^2\right] &= \mathbb{E}[f^2(\boldsymbol{x})] - \mathbb{E}\left[\left(\sum_{l=0}^{L} a_l H_l(\langle \boldsymbol{u}, \boldsymbol{x} \rangle)\right)^2\right] \\ &= \sum_{l=1}^{L} a_l^{\star 2}(1 - \langle \boldsymbol{u}^\star, \boldsymbol{u} \rangle^{2l}). \end{aligned}$$

Hence we have,

$$R_L(\boldsymbol{u}) = \sigma^2 + \sum_{l=1}^{L} a_l^{\star 2}(1 - \langle \boldsymbol{u}, \boldsymbol{u}^\star \rangle^{2l}).$$

Differentiating the objective, we find that the (Riemannian) gradient is given by:

$$\nabla R_L(\boldsymbol{u}) = \left(-2 \sum_{i=1}^{L} l a_l^{\star 2} \langle \boldsymbol{u}, \boldsymbol{u}^\star \rangle^{2l-1} \right) (\boldsymbol{u}^\star - \langle \boldsymbol{u}^\star, \boldsymbol{u} \rangle \boldsymbol{u}) .$$

Solving for $\nabla R_L(\boldsymbol{u}) = 0$, we get that the only critical points are $\boldsymbol{u} = \pm \boldsymbol{u}^\star$ and $\boldsymbol{u} \in \{\boldsymbol{a} \in \mathbb{S}^{p-1} : \langle \boldsymbol{a}, \boldsymbol{u}^\star \rangle = 0\}$. Since $R_L(\boldsymbol{u}) \geq \sigma^2 \; \forall \boldsymbol{u} \in \mathbb{S}^{p-1}$, and $R_L(\pm \boldsymbol{u}^\star) = \sigma^2$, the points $\pm \boldsymbol{u}^\star$ are global minimizers. Analogously consider any $\boldsymbol{u}^\perp$ such that $\langle \boldsymbol{u}^\perp, \boldsymbol{u}^\star \rangle = 0$. Since $R_L(\boldsymbol{u}) \leq \sigma^2 + \sum_{l=1}^{L} a_l^{\star 2}$ $\forall \boldsymbol{u} \in \mathbb{S}^{p-1}$ and $R_L(\boldsymbol{u}^\perp) = \sigma^2 + \sum_{l=1}^{L} a_l^{\star 2}$, $\boldsymbol{u}^\perp$ is a global maximizer. To show that for some link functions g, these maximizers can be degenerate we consider a small perturbation at $\boldsymbol{u}^\perp$ in an arbitrary direction $\boldsymbol{u}$:

$$R_L \left(\sqrt{1 - \delta^2} \boldsymbol{u}^\perp + \delta \boldsymbol{u} \right) - R_L(\boldsymbol{u}^\perp) = - \sum_{l=1}^{L} a_l^{\star 2} \delta^{2l} \langle \boldsymbol{u}, \boldsymbol{u}^\star \rangle^{2l} .$$

One can see when $\mathrm{OD}(f) > 1$, $R_L \left(\sqrt{1 - \delta^2} \boldsymbol{u}^\perp + \delta \boldsymbol{u} \right) - R_L(\boldsymbol{u}^\perp) = o(\delta^2)$ demonstrating that the maximum is degenerate. ■

A.2. Multi-index model

Lemma 20 (Lemma 7 restated) *Let $\boldsymbol{U} := [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k]$ be a matrix in $\mathbb{R}^{p \times k}$ with orthonormal columns. Let $\boldsymbol{v}$ be a arbitrary unit vector. Then,*

$$\begin{aligned}
&1) \quad \mathbb{E}\left[H_{\boldsymbol{S}}(\boldsymbol{U}x) H_l(\langle \boldsymbol{v}, \boldsymbol{x} \rangle) \right] = 0 && \text{if } l < |\boldsymbol{S}|, \\
&2) \quad \mathbb{E}\left[H_{\boldsymbol{S}}(\boldsymbol{U}x) H_l(\langle \boldsymbol{v}, \boldsymbol{x} \rangle) \right] = \sqrt{\frac{l!}{S_1! S_2! \cdots S_k!}} \prod_{i=1}^{k} \langle \boldsymbol{u}_i, \boldsymbol{v} \rangle^{S_i} && \text{if } l = |\boldsymbol{S}|, \\
&3) \quad \mathbb{E}\left[H_{\boldsymbol{S}}(\boldsymbol{U}x) H_l(\langle \boldsymbol{v}, \boldsymbol{x} \rangle) \right] = 0 && \text{if } l > |\boldsymbol{S}|.
\end{aligned}$$

Proof We consider 2 cases:

Case 1: Consider $l \leq \|\boldsymbol{S}\|_1$. We can write $\boldsymbol{v} = \sum_{i=1}^{k} \alpha_i \boldsymbol{u}_i + \sqrt{1 - \|\boldsymbol{\alpha}\|^2} \boldsymbol{u}_\perp$. Here $\alpha_i = \langle \boldsymbol{u}_i, \boldsymbol{v} \rangle$ and $\boldsymbol{u}_\perp$ is a unit vector orthogonal to $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k$. We define $Z_i = \langle \boldsymbol{u}_i, \boldsymbol{x} \rangle$ and $Y = \langle \boldsymbol{u}_\perp, \boldsymbol{x} \rangle$. Then, $H_l(\langle \boldsymbol{v}, \boldsymbol{x} \rangle) = H_l(\sum_{i=1}^{k} \alpha_i Z_i + \sqrt{1 - \|\boldsymbol{\alpha}\|^2} Y)$ is a degree l polynomial in $k+1$ independent Gaussian random variables $Z_1, \ldots, Z_k, Y$ and admits an expansion of the form:

$$H_l \left(\sum_i \alpha_i Z_i + \sqrt{1 - \|\boldsymbol{\alpha}\|^2} Y \right) = \sum_{\boldsymbol{D}, d : \|\boldsymbol{D}\|_1 + d \leq l} c_{\boldsymbol{D}, d}(\boldsymbol{\alpha}, \boldsymbol{v}) \prod_{i=1}^{k} H_{D_i}(Z_i) H_d(Y). \tag{5}$$

Hence we have:

$$\mathbb{E}\left[H_{\boldsymbol{S}}(\boldsymbol{U}x) H_l(\langle \boldsymbol{v}, \boldsymbol{x} \rangle) \right] = c_{\boldsymbol{S}, 0}(\boldsymbol{\alpha}, \boldsymbol{v}).$$

First we note that $c_{\boldsymbol{S},0}(\boldsymbol{\alpha}, \boldsymbol{v}) = 0$ if $l < \|\boldsymbol{S}\|_1$. On the other hand, if $l = \|\boldsymbol{S}\|_1$, we can find the expression for $c_{\boldsymbol{S},0}(\boldsymbol{\alpha}, \boldsymbol{v})$ by comparing the coefficient for $\prod_{i=1}^k Z_i^{S_i}$ on both sides of equation Equation (5) and equating the two:

$$\left(\prod_{i=1}^k \alpha_i^{S_i}\right) \frac{\frac{l!}{S_1!S_2!\cdots S_k!}}{\sqrt{l!}} = \frac{c_{\boldsymbol{S}}(\boldsymbol{\alpha}, \boldsymbol{v})}{\sqrt{S_1!S_2!\cdots S_k!}},$$

which gives us the required result.

Case 2: Consider $l > \|\boldsymbol{S}\|_1$. Let $\{\boldsymbol{v}_i^\perp\}_{i=1}^k$ be unit vectors orthogonal to $\boldsymbol{v}$ and $\alpha_i = \langle \boldsymbol{u}_i, \boldsymbol{v}\rangle$. We can write:

$$\boldsymbol{u}_i = \alpha_i \boldsymbol{v} + \sqrt{1-\alpha_i^2}\boldsymbol{v}_i^\perp.$$

Hence,

$$\langle \boldsymbol{u}_i, \boldsymbol{x}\rangle = \alpha_i \langle \boldsymbol{v}, \boldsymbol{x}\rangle + \sqrt{1-\alpha_i^2}\left\langle \boldsymbol{v}_i^\perp, \boldsymbol{x}\right\rangle := \alpha_i W + \sqrt{1-\alpha_i^2}X_i.$$

Here $W, X_1, X_2, \ldots, X_k$ are marginally standard normal random variables. Furthermore, W is independent of $\{X_i\}_{i=1}^k$, however $\{X_i\}_{i=1}^k$ might be correlated. Next we observe:

$$H_{\boldsymbol{S}}(\boldsymbol{U}\boldsymbol{x}) = \prod_{i=1}^k H_{S_i}\left(\alpha_i W + \sqrt{1-\alpha_i^2}X_i\right).$$

is a degree $\|\boldsymbol{S}\|_1$ polynomial in the variables $X_1, \ldots, X_k, W$ and hence can be expanded in the basis:

$$H_{\boldsymbol{S}}(\boldsymbol{U}\boldsymbol{x}) = \sum_{\boldsymbol{D},d:\|D\|_1+d\le\|S\|_1} c'_{\boldsymbol{D},d}(\boldsymbol{\alpha}) H_d(W) H_{\boldsymbol{S}}(\boldsymbol{X}).$$

Hence:

$$\mathbb{E}\left[H_{\boldsymbol{S}}(\boldsymbol{U}x)H_l(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)\right] = \sum_{\boldsymbol{D},d:\|D\|_1+d\le\|S\|_1} c'_{\boldsymbol{D},d}(\boldsymbol{\alpha})\mathbb{E}[H_l(W)H_d(W)]\mathbb{E}[H_{\boldsymbol{S}}(\boldsymbol{X})].$$

Finally we note that $\mathbb{E}[H_l(W)H_d(W)] = 0$ for all $d \le \|\boldsymbol{S}\|_1 < l$. This gives us:

$$\mathbb{E}\left[H_{\boldsymbol{S}}(\boldsymbol{U}x)H_l(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)\right] = 0.$$

■

Theorem 21 (Theorem 8 restated) *Under the order k multi-index model, the population objective $F_l(\boldsymbol{u})$ has the following properties:*

1. *The population objective has the following explicit form:*

$$F_l(\boldsymbol{u}) = \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1=l} a^\star_{\boldsymbol{S}} \sqrt{\binom{l}{S_1, S_2, \ldots, S_k}} \prod_{i=1}^k \langle \boldsymbol{u}_i^\star, \boldsymbol{u}\rangle^{S_i}.$$

2. *Any local maximizer with $F_l(\boldsymbol{u}) > 0$ is contained in the subspace $\mathcal{U}^\star$.*
3. *Any local minimizer with $F_l(\boldsymbol{u}) < 0$ is contained in the subspace $\mathcal{U}^\star$.*

Proof We recall that the link function g has the following expansion in the Hermite Basis:

$$g(\boldsymbol{z}) = \sum_{\boldsymbol{S} \in \mathcal{I}_\infty} a^\star_{\boldsymbol{S}} \boldsymbol{z}^{\boldsymbol{S}}.$$

Hence,

$$\begin{aligned} F_l(\boldsymbol{u}) &= \mathbb{E}[y H_l(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)] \\ &= \mathbb{E}[g(\boldsymbol{U}^{\star T}\boldsymbol{x}) H_l(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)] \\ &\overset{\text{Lemma 7}}{=} \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1 = l} a^\star_{\boldsymbol{S}} \sqrt{\binom{l}{S_1, S_2, \ldots, S_k}} \prod_{i=1}^k \langle \boldsymbol{u}_i^\star, \boldsymbol{u}\rangle^{S_i}. \end{aligned}$$

We prove the second claim by contradiction. The proof for the third claim is analogous and is omitted. Consider a unit vector $\boldsymbol{u} \notin \mathcal{U}^\star$. Hence, we have,

$$\boldsymbol{u} = \sum_{i=1}^k \alpha_i \boldsymbol{u}_i^\star + \sqrt{1 - \|\boldsymbol{\alpha}_i\|^2}\boldsymbol{u}_\perp.$$

In the above display, $\|\boldsymbol{\alpha}\| < 1$ (otherwise we would have $\boldsymbol{u} \in \mathcal{U}^\star$) and the vector $\boldsymbol{u}_\perp \in \mathcal{U}^{\star\perp}$. We claim that the vector $\boldsymbol{u}$ cannot be a local maximizer. To show this we will construct a direction $\boldsymbol{\Delta}$ such that an arbitrarily small perturbation of size $\delta > 0$ in this direction is guaranteed to increase the objective. We construct this direction as follows:

$$\boldsymbol{\Delta} := \frac{\sum_{i=1}^k \alpha_i \boldsymbol{u}_i^\star}{\|\boldsymbol{\alpha}\|}.$$

To show this is a direction of increase we compute:

$$\begin{aligned} F_l\left(\sqrt{1-\delta^2}\boldsymbol{u} + \delta\boldsymbol{\Delta}\right) &= \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1 = l} a^\star_{\boldsymbol{S}} \sqrt{\binom{l}{S_1, S_2, \ldots, S_k}} \prod_{i=1}^k (\langle \boldsymbol{u}_i^\star, \boldsymbol{u}\rangle)^{S_i} \left(\sqrt{1-\delta^2} + \frac{\delta}{\|\boldsymbol{\alpha}\|}\right)^{S_i} \\ &= \left(\sqrt{1-\delta^2} + \frac{\delta}{\|\boldsymbol{\alpha}\|}\right)^l \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1 = l} a^\star_{\boldsymbol{S}} \sqrt{\binom{l}{S_1, S_2, \ldots, S_k}} \prod_{i=1}^k (\langle \boldsymbol{u}_i^\star, \boldsymbol{u}\rangle)^{S_i} \\ &= \left(\sqrt{1-\delta^2} + \frac{\delta}{\|\boldsymbol{\alpha}\|}\right)^l F_l(\boldsymbol{u}). \end{aligned}$$

We now analyze the leading multiplicative factor:

$$\begin{aligned} \sqrt{1-\delta^2} + \frac{\delta}{\|\boldsymbol{\alpha}\|} &\overset{\|\boldsymbol{\alpha}\| < 1}{>} \sqrt{1-\delta^2} + \delta \\ &\overset{\delta > 0}{>} 1. \end{aligned}$$

Hence we have $F_l\left(\sqrt{1-\delta^2}\boldsymbol{u} + \delta\boldsymbol{\Delta}\right) - F_l(\boldsymbol{u}) > 0$ and $\boldsymbol{u}$ is not a local maximum. ■

Appendix B. Missing Proofs from Section 3

B.1. Analysis of Algorithm 1

Theorem 22 (Theorem 10 restated) *With probability* $1 - 2\exp(-p/32) - 5\delta - \frac{8}{n}$*, the output* $\hat{\boldsymbol{u}}_l$ *of Algorithm 1 satisfies*

$$|\langle \hat{\boldsymbol{u}}_l, \boldsymbol{u}^\star \rangle| \geq 1 - \frac{100(\|f\|_\infty + 4\sigma) \cdot 2^{2l+1}}{l|a_l^\star|} \sqrt{\frac{2\max(p, \ln(1/\delta))\ln^l(n)}{n}},$$

provided n *is large enough so that the following holds:*

$$n \geq \frac{32 \cdot 10^4 (\|f\|_\infty + 4\sigma)^2}{l^2 a_l^{\star 2}} \frac{2^{2l}}{\delta^{2l-2}} \max(p, \ln(1/\delta)) p^{l-1} \ln^l(n).$$

Proof We begin by introducing some notation: For $t \in \{0, 1, 2\}$, we define:

$$\begin{aligned}
\alpha_t &:= |\langle \boldsymbol{u}^\star, \boldsymbol{u}_t \rangle|, \\
\Delta_t &:= \nabla \hat{F}_l(\boldsymbol{u}_{t-1}; S_t) - \mathbb{E}[\nabla \hat{F}_l(\boldsymbol{u}_{t-1}; S_t)].
\end{aligned}$$

Using Lemma 43, with probability $1 - 2\exp(-p/32) - \delta$,

$$\alpha_0 \geq \frac{\delta}{\sqrt{p}}.$$

We can further compute the expression for the gradients as each iteration:

$$\begin{aligned}
\nabla \hat{F}_l(\boldsymbol{u}_{t-1}, S_t) &= \frac{1}{|S_t|} \sum_{i \in S_t} y_i H_l'(\langle \boldsymbol{x}_i, \boldsymbol{u}_{t-1} \rangle) \boldsymbol{x}_i \\
&\overset{\text{Fact 2}}{=} \frac{\sqrt{l}}{|S_t|} \sum_{i \in S_t} y_i H_{l-1}(\langle \boldsymbol{x}_i, \boldsymbol{u}_{t-1} \rangle) \boldsymbol{x}_i.
\end{aligned}$$

Using Theorem 48 and a union bound, with probability $1 - 4\delta - \frac{8}{n}$,

$$\max(\|\Delta_1\|_2, \|\Delta_2\|_2) \leq 100(\|f\|_\infty + 4\sigma) \cdot 2^l \sqrt{\frac{2\max(p, \ln(1/\delta))\ln^l(n)}{n}}.$$

Next we derive a recursive lower bound on α_t:

$$\begin{aligned}
\alpha_t &= |\langle \boldsymbol{u}^\star, \boldsymbol{u}_t \rangle| \\
&= \left| \left\langle \boldsymbol{u}^\star, \frac{\nabla \hat{F}_l(\boldsymbol{u}^{t-1}, S_t)}{\|\nabla \hat{F}_l(\boldsymbol{u}^{t-1}, S_t)\|_2} \right\rangle \right|. \qquad (6)
\end{aligned}$$

Next we note that $\nabla \hat{F}_l(\boldsymbol{u}_{t-1}; S_t) = \mathbb{E}[\nabla \hat{F}_l(\boldsymbol{u}_{t-1}; S_t)] + \Delta_t$. Furthermore $\mathbb{E}[\nabla \hat{F}_l(\boldsymbol{u}_{t-1}; S_t)] = \nabla \mathbb{E}[\hat{F}_l(\boldsymbol{u}_{t-1}; S_t)] = \nabla a_l^\star \langle \boldsymbol{u}_{t-1}, \boldsymbol{u}^\star \rangle^l = l a_l^\star \langle \boldsymbol{u}_{t-1}, \boldsymbol{u}^\star \rangle^{l-1} \boldsymbol{u}^\star$. Substituting these into Equation (6), we get,

$$\begin{aligned}
\alpha_t = & \frac{|l a_l^\star \langle \boldsymbol{u}_{t-1}, \boldsymbol{u}^\star \rangle^{l-1} + \langle \boldsymbol{\Delta}_t, \boldsymbol{u}^\star \rangle|}{\| l a_l^\star \langle \boldsymbol{u}_{t-1}, \boldsymbol{u}^\star \rangle^{l-1} \boldsymbol{u}^\star + \boldsymbol{\Delta}_t \|_2} \\
\overset{\text{Triangle Ineq.}}{\geq} & \frac{l|a_l^\star \langle \boldsymbol{u}_{t-1}, \boldsymbol{u}^\star \rangle^{l-1}| - |\langle \boldsymbol{\Delta}_t, \boldsymbol{u}^\star \rangle|}{\| l a_l^\star \langle \boldsymbol{u}_{t-1}, \boldsymbol{u}^\star \rangle^{l-1} \boldsymbol{u}^\star + \boldsymbol{\Delta}_t \|_2} \\
\overset{\text{Cauchy Schwarz}}{\geq} & \frac{l|a_l^\star| \alpha_{t-1}^{l-1} - \|\boldsymbol{\Delta}_t\|_2}{\| l a_l^\star \langle \boldsymbol{u}_{t-1}, \boldsymbol{u}^\star \rangle^{l-1} \boldsymbol{u}^\star + \boldsymbol{\Delta}_t \|_2} \\
\overset{\text{Triangle Ineq.}}{\geq} & \frac{l a_l^\star \langle \boldsymbol{u}_{t-1}, \boldsymbol{u}^\star \rangle^{l-1} - \|\boldsymbol{\Delta}_t\|_2}{\| l a_l^\star \langle \boldsymbol{u}_{t-1}, \boldsymbol{u}^\star \rangle^{l-1} \boldsymbol{u}^\star \|_2 + \|\boldsymbol{\Delta}_t\|_2} \\
\geq & 1 - \frac{2\|\Delta_t\|_2}{|l a_l^\star| \alpha_{t-1}^{l-1}}
\end{aligned} \tag{7}$$

The condition on n assumed in the statement of theorem guarantees $\|\Delta_t\|_2 \leq \frac{l|a_l^\star|\alpha_0^{l-1}}{4}$. Applying Equation (7) with $t = 1$ yields:

$$\alpha_1 \geq \frac{1}{2}.$$

Applying Equation (7) with $t = 2$ yields,

$$\alpha_2 \geq 1 - \frac{2^{l+1}}{l|a_l^\star|}\|\boldsymbol{\Delta}_2\|_2 \geq 1 - \frac{100(\|f\|_\infty + 4\sigma) \cdot 2^{2l+1}}{l|a_l^\star|} \sqrt{\frac{2\max(p, \ln(1/\delta)) \ln^l(n)}{n}}.$$

■

B.2. Analysis of Algorithm 2

Lemma 23 (Lemma 13 restated) *For a link function g that satisfies Assumptions 1, 3 and 4, there exists a $l_\sharp \leq \frac{2R^2}{\mu}$ such that, $l_\sharp |a_{l_\sharp}^\star|^2 \geq \frac{\mu^2}{4R^2}$.*

Proof We first translate Assumptions 1, 3 and 4 into statements about the coefficients $\boldsymbol{a}^\star$:

$$\mathbb{E}_{\boldsymbol{z}}[g^2(\boldsymbol{z})] = 1 \implies \sum_{i=0}^{\infty} a_i^{\star 2} = 1.$$

Next we consider the minimum signal strength assumption (Assumption 4) and we note that:

$$\begin{aligned}
\frac{dg(z)}{dz} = & \sum_{i=1}^{\infty} a_i^\star H_i'(z) \\
\overset{\text{Fact 2}}{=} & \sum_{i=1}^{\infty} \sqrt{i} a_i^\star H_{i-1}(z).
\end{aligned}$$

Hence,

$$\mathbb{E}\left[\left(\frac{dg(z)}{dz}\right)^2\right] \geq \mu \implies \sum_{i=1}^{\infty} i a_i^{\star 2} \geq \mu. \tag{8}$$

Similarly the smoothness assumption can be written as:

$$\mathbb{E}\left[\left(\frac{d^2 g(z)}{dz^2}\right)^2\right] \leq R^2 \implies \sum_{i=2}^{\infty} i(i-1) a_i^{\star 2} \leq R^2. \tag{9}$$

We first note that, for any $L \in \mathbb{N}$,

$$\begin{aligned}\sum_{i=1}^{L} i a_i^2 = & \sum_{i=1}^{\infty} i a_i^2 - \sum_{i=L+1}^{\infty} i a_i^2 \\ \overset{\text{eq. (8)}}{\geq} & \left(\mu - \sum_{i=L+1}^{\infty} i a_i^2\right).\end{aligned} \tag{10}$$

Furthermore,

$$\begin{aligned}\sum_{i=L+1}^{\infty} i a_i^2 \leq & \frac{1}{L} \sum_{i=L+1}^{\infty} i(i-1) a_i^2 \\ \overset{\text{eq. (9)}}{\leq} & \frac{R^2}{L}.\end{aligned}$$

Substituting this in Equation (10), we get,

$$\frac{1}{L} \sum_{i=1}^{L} i a_i^2 \geq \frac{1}{L}\left(\mu - \frac{R^2}{L}\right).$$

Choosing $L = \frac{2R^2}{\mu}$, gives,

$$\max_{l \in [L]} l a_l^2 \geq \frac{1}{L} \sum_{i=1}^{L} i a_i^2 \geq \frac{\mu}{2L} = \frac{\mu^2}{4R^2}.$$

■

Theorem 24 (Theorem 14 restated) *With probability* $1 - \frac{4R^2}{\mu} e^{-p/32} - \frac{12R^2}{\mu}\delta - \frac{16R^2}{n\mu}$, *the estimate returned by Algorithm 2,* $\hat{\boldsymbol{u}}$ *satisfies*

$$|\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}} \rangle| \geq 1 - \frac{3200 \cdot 2^{\frac{4R^2}{\mu}} (\|f\|_\infty + 4\sigma) R^4}{\mu^2 \sqrt{\mu}} \sqrt{\frac{\max\left(p, \ln\left(\frac{1}{\delta}\right)\right) \ln^{\frac{2R^2}{\mu}}(n)}{n}},$$

provided that n *satisfies*

$$n \geq \frac{1024 \cdot 10^4 (\|f\|_\infty + 4\sigma)^2 R^4}{\mu^3} \cdot \frac{2^{\frac{4R^2}{\mu}}}{\delta^{\frac{4R^2}{\mu} - 2}} \max\left(p, \ln\left(\frac{1}{\delta}\right)\right) p^{\frac{2R^2}{\mu} - 1} \ln^{\frac{2R^2}{\mu}}(n).$$

Proof As in the description of Algorithm 2, we define:

$$L := \frac{2R^2}{\mu}, \quad m := |S_{\text{test}}|.$$

Theorem 13 guarantees us the existence of a $l_\sharp$ such that:

$$l_\sharp \leq L, \quad l_\sharp {a^\star_{l_\sharp}}^2 \geq \frac{\mu^2}{4R^2}.$$

We define the index set $\mathcal{L}_{\text{good}}$ as:

$$\mathcal{L}_{\text{good}} = \left\{ l \in [L] : |a^\star_l| \geq \frac{|a^\star_{l_\sharp}|}{2} \right\}.$$

The condition imposed n is such that, for each $l \in \mathcal{L}_{\text{good}}$, we can apply Theorem 10 and get with probability $1 - 2e^{-p/32} - 5\delta - \frac{8}{n}$, the estimates $\hat{\boldsymbol{u}}_l$ satisfy:

$$|\langle \hat{\boldsymbol{u}}_l, \boldsymbol{u}^\star \rangle| \geq 1 - \frac{100(\|f\|_\infty + 4\sigma) \cdot 2^{2l+1}}{|a^\star_l|} \sqrt{\frac{2\max(p, \ln(1/\delta)) \ln^l(n)}{n}}.$$

Next using the definition of the set $\mathcal{L}_{\text{good}}$ and a union bound, we have, with probability $1 - \frac{4R^2}{\mu} e^{-p/32} - \frac{10R^2}{\mu}\delta - \frac{16R^2}{\mu n}$, we have,

$$|\langle \hat{\boldsymbol{u}}_l, \boldsymbol{u}^\star \rangle| \geq 1 - \frac{1600 \cdot 2^{\frac{4R^2}{\mu}} (\|f\|_\infty + 4\sigma) R^2}{\mu\sqrt{\mu}} \sqrt{\frac{\max\left(p, \ln\left(\frac{1}{\delta}\right)\right) \ln^{\frac{2R^2}{\mu}}(n)}{n}} \geq \frac{1}{2} \quad \forall l \in \mathcal{L}_{\text{good}}. \tag{11}$$

Next we analyze the concentration of the goodness-fit-statistic. Using Lemma 49 and a union bound we have, with probability $1 - \frac{2R^2}{\mu}\delta$

$$|T_l - \mathbb{E}[T_l]| \leq \Delta \quad \forall l \in [L],$$

provided $m \geq \frac{2(\sigma^2 + \|f\|_\infty^2)}{\delta \Delta^2}$. We set:

$$\Delta = \frac{|a^\star_{l_\sharp}|}{4 \cdot 2^{l_\sharp}} \overset{\text{eq. (11), } l_\sharp \in \mathcal{L}_{\text{good}}}{\leq} \frac{\left| a^\star_{l_\sharp} \left\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}}_{l_\sharp} \right\rangle^{l_\sharp} \right|}{4}. \tag{12}$$

In order to ensure that this happens we need to make sure that

$$m \geq \frac{2(\sigma^2 + \|f\|_\infty^2)}{\delta \Delta^2} = \frac{32 \cdot 2^{2l_\sharp}(\sigma^2 + \|f\|_\infty^2)}{\delta |a_{l_\sharp}|^2}.$$

Since we set $m = \frac{256 \cdot 2^{\frac{4R^2}{\mu}} R^4 (\sigma^2 + \|f\|_\infty^2)}{\delta \mu^3}$ this requirement is indeed satisfied. Now we argue that the estimator returned by Algorithm 2 performs nearly as well as $\hat{\boldsymbol{u}}_{l_\sharp}$:

$$\begin{aligned} \left| a^\star_{l_{\text{best}}} \langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}}_{l_{\text{best}}} \rangle^{l_{\text{best}}} \right| &\overset{\text{Theorem 3}}{=} |\mathbb{E}[T_{l_{\text{best}}}]| \\ &\overset{\text{Triangle Ineq.}}{\geq} |T_{l_{\text{best}}}| - |T_{l_{\text{best}}} - \mathbb{E}[T_{l_{\text{best}}}]| \\ &\geq |T_{l_{\text{best}}}| - \Delta. \end{aligned}$$

Since l_{best} was the harmonic with the best goodness-of-fit statistic, $|T_{l_{\text{best}}}| \geq |T_{l_\sharp}|$. Substituting this in the previous display,

$$\begin{aligned}\left|a^\star_{l_{\text{best}}} \left\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}}_{l_{\text{best}}} \right\rangle^{l_{\text{best}}}\right| &\geq |T_{l_\sharp}| - \Delta \\ &\overset{\text{Triangle Ineq.}}{\geq} \mathbb{E}[T_{l^\sharp}] - 2\Delta \\ &\overset{\text{Theorem 3}}{=} \left|a^\star_{l_\sharp} \left\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}}_{l_\sharp} \right\rangle^{l_\sharp}\right| - 2\Delta \\ &\overset{\text{eq. (12)}}{\geq} \frac{\left|a^\star_{l_\sharp} \left\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}}_{l_\sharp} \right\rangle^{l_\sharp}\right|}{2}.\end{aligned}$$

Hence one of the two cases hold:

Case 1: $|a^\star_{l_{\text{best}}}| \geq |a^\star_{l_\sharp}|/2$
In this case, we know that $l_{\text{best}} \in \mathcal{L}_{\text{good}}$ and hence using Equation (11), the estimate returned by Algorithm 2 $\hat{\boldsymbol{u}}$ satisfies

$$|\left\langle \hat{\boldsymbol{u}}, \boldsymbol{u}^\star \right\rangle| \geq 1 - \frac{1600 \cdot 2^{\frac{4R^2}{\mu}} (\|f\|_\infty + 4\sigma) R^2}{\mu\sqrt{\mu}} \sqrt{\frac{\max\left(p, \ln\left(\frac{1}{\delta}\right)\right) \ln^{\frac{2R^2}{\mu}}(n)}{n}}.$$

Case 2: $\left|\left\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}}_{l_{\text{best}}} \right\rangle^{l_{\text{best}}}\right| \geq \left|\left\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}}_{l_\sharp} \right\rangle^{l_\sharp}\right|$.
This means:

$$\begin{aligned}|\left\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}}_{l_{\text{best}}} \right\rangle| &\geq \left|\left\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}}_{l_\sharp} \right\rangle^{\frac{l_\sharp}{l_{\text{best}}}}\right| \\ &\geq \left|\left\langle \boldsymbol{u}^\star, \hat{\boldsymbol{u}}_{l_\sharp} \right\rangle\right|^{\frac{2R^2}{\mu}} \\ &\geq \left(1 - \frac{1600 \cdot 2^{\frac{4R^2}{\mu}} (\|f\|_\infty + 4\sigma) R^2}{\mu\sqrt{\mu}} \sqrt{\frac{\max\left(p, \ln\left(\frac{1}{\delta}\right)\right) \ln^{\frac{2R^2}{\mu}}(n)}{n}}\right)^{\frac{2R^2}{\mu}} \\ &\geq 1 - \frac{3200 \cdot 2^{\frac{4R^2}{\mu}} (\|f\|_\infty + 4\sigma) R^4}{\mu^2\sqrt{\mu}} \sqrt{\frac{\max\left(p, \ln\left(\frac{1}{\delta}\right)\right) \ln^{\frac{2R^2}{\mu}}(n)}{n}}.\end{aligned}$$

Here in the last step we used the fact that $(1-x)^n \geq 1 - nx$ for every $n \in \mathbb{N}, x \in (0,1)$.

Combining the two cases and the probabilities of the various failures, we get the claim of the theorem. ■

Appendix C. Missing Proofs from Section 4

Lemma 25 (Lemma 16 restated) *Let g be a link function from $\mathbb{R}^k \to \mathbb{R}$ obeying Assumptions 1,3 and 4. Then, there exists an $l_\sharp \in \mathbb{N}$ such that:*

$$l_\sharp \leq \frac{2kR^2}{\mu} + k - 1, \qquad \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1 = l_\sharp} a^2_{\boldsymbol{S}} \|\boldsymbol{S}\|_1 \geq \frac{\mu^2}{2(2R^2 + \mu)}.$$

Proof The proof of this lemma is analogous to Theorem 13. We begin by translating the assumptions made on the link function into conditions on the coefficients $a_{\boldsymbol{S}}$. First we consider the minimum signal strength assumption (Assumption 4) and we note that,

$$\frac{\partial g}{\partial z_i}(\boldsymbol{z}) = \sum_{\boldsymbol{S}\in\mathcal{I}_\infty} \sqrt{S_i} a^\star_{\boldsymbol{S}} H_{\boldsymbol{S}^{(i)}}(\boldsymbol{z}).$$

In the above display, $\boldsymbol{S}^{(i)} := (S_1, S_2, \ldots, S_{i-1}, S_i - 1, S_{i+1}, \ldots, S_k)$. Hence we have,

$$\mathbb{E}\left[\left(\frac{\partial g}{\partial z_i}(\boldsymbol{z})\right)^2\right] \geq \mu \implies \sum_{\boldsymbol{S}\in\mathcal{I}_\infty} a^2_{\boldsymbol{S}} S_i \geq \mu \forall i \in [k],$$

$$\sum_{\boldsymbol{S}\in\mathcal{I}_\infty} a^2_{\boldsymbol{S}} S_i \geq \mu \quad \forall i \in [k] \implies \sum_{\boldsymbol{S}\in\mathcal{I}_\infty} a^2_{\boldsymbol{S}} \|\boldsymbol{S}\|_1 \geq \mu k. \tag{13}$$

Next we consider the smoothness assumption (Assumption 3). Analogously,

$$\mathbb{E}\left[\left(\frac{\partial^2 g}{\partial z_i^2}(\boldsymbol{z})\right)^2\right] \leq R^2 \implies \sum_{\boldsymbol{S}\in\mathcal{I}_\infty} a^2_{\boldsymbol{S}} S_i(S_i - 1) \leq R^2,$$

$$\sum_{\boldsymbol{S}\in\mathcal{I}_\infty} a^2_{\boldsymbol{S}} S_i(S_i - 1) \leq R^2 \quad \forall i \in [k] \implies \sum_{\boldsymbol{S}\in\mathcal{I}_\infty} a^2_{\boldsymbol{S}} (\|\boldsymbol{S}\|_2^2 - \|\boldsymbol{S}\|_1) \leq kR^2. \tag{14}$$

Consider any arbitrary $L \geq k - 1$. We have,

$$\sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1\leq L} a^2_{\boldsymbol{S}} \|\boldsymbol{S}\|_1 = \sum_{\boldsymbol{S}\in\mathcal{I}_\infty} a^2_{\boldsymbol{S}} \|\boldsymbol{S}\|_1 - \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1> L} a^2_{\boldsymbol{S}} \|\boldsymbol{S}\|_1. \tag{15}$$

Next we observe that, for any $\boldsymbol{S}$ such that $\|\boldsymbol{S}\|_1 > L \geq k - 1$,

$$\begin{aligned} \|\boldsymbol{S}\|_2^2 - \|\boldsymbol{S}\|_1 &\overset{\text{Cauchy Schwarz}}{\geq} \frac{\|\boldsymbol{S}\|_1^2}{k} - \|\boldsymbol{S}\|_1 \\ &= \|\boldsymbol{S}\|_1 \left(\frac{\|\boldsymbol{S}\|_1}{k} - 1\right) \\ &\geq \|\boldsymbol{S}\|_1 \left(\frac{L+1}{k} - 1\right). \end{aligned} \tag{16}$$

This allows us to upper bound:

$$\sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1> L} a^2_{\boldsymbol{S}} \|\boldsymbol{S}\|_1 \overset{\text{eq. (16)}}{\leq} \frac{k}{L+1-k} \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1> L} a^2_{\boldsymbol{S}} (\|\boldsymbol{S}\|_2^2 - \|\boldsymbol{S}\|_1) \tag{17}$$

$$\overset{\text{eq. (14)}}{\leq} \frac{k^2 R^2}{L+1-k}. \tag{18}$$

Substituting the bounds obtained in Equation (13) and Equation (18) into Equation (15) gives:

$$\begin{aligned} \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1\leq L} a^2_{\boldsymbol{S}} \|\boldsymbol{S}\|_1 &= \sum_{\boldsymbol{S}} a^2_{\boldsymbol{S}} \|\boldsymbol{S}\|_1 - \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1> L} a^2_{\boldsymbol{S}} \|\boldsymbol{S}\|_1 \\ &\geq \mu k - \frac{k^2 R^2}{L+1-k}. \end{aligned}$$

Setting $\frac{k^2R^2}{L+1-k} = \frac{\mu k}{2}$ gives, $L = \frac{2kR^2}{\mu} + k - 1$. Using this value of L guarantees,

$$\sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1 \le L} a_{\boldsymbol{S}}^2 \|\boldsymbol{S}\|_1 \ge \frac{\mu k}{2}.$$

In particular, this means there exists an $l_\sharp \le \frac{2kR^2}{\mu} + k - 1$, such that,

$$\sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1 = l_\sharp} a_{\boldsymbol{S}}^2 \|\boldsymbol{S}\|_1 \ge \frac{\mu k}{2L} > \frac{\mu^2}{2(2R^2+\mu)}.$$

■

Lemma 26 *Let $l_\sharp \in \mathbb{N}$ be as in Theorem 16. Let $\boldsymbol{u} \sim$ Uniform$(\mathbb{S}^{p-1})$. Then, with probability $1 - \delta - 2\exp(-p/32)$,*

$$\|\nabla F_{l_\sharp}(\boldsymbol{u})\|_2^2 \ge \left(\frac{\delta}{2kK_{CW}(l_\sharp - 1)\sqrt{p}}\right)^{2l_\sharp - 2} \frac{\mu^2}{2^{3l_\sharp + k + 1}(2R^2+\mu)}.$$

In the above display K_{CW} is a universal constant appearing in the Carbery-Wright Theorem (Theorem 44).

Proof Since $\boldsymbol{u}$ is a uniformly random unit vector, we can assume $\boldsymbol{u} := \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$ where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. By Theorem 7,

$$F_{l_\sharp}(\boldsymbol{u}) = \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1 = l_\sharp} a_{\boldsymbol{S}} \sqrt{\binom{l_\sharp}{S_1, S_2, \ldots, S_k}} \prod_{i=1}^k \langle \boldsymbol{u}, \boldsymbol{u}_i^\star \rangle^{S_i}.$$

Taking gradients,

$$\nabla F_{l_\sharp}(\boldsymbol{u}) = \frac{1}{\|\boldsymbol{g}\|_2^{l_\sharp - 1}} \sum_{i=1}^k D^{(i)}(\langle \boldsymbol{u}_1^\star, \boldsymbol{g}\rangle, \ldots, \langle \boldsymbol{u}_k^\star, \boldsymbol{g}\rangle) \boldsymbol{u}_i^\star, \tag{19}$$

where the functions $D^{(i)}$ are defined as follows:

$$D^{(i)}(z_1, z_2, \ldots, z_k) = \sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1 = l_\sharp} S_i a_{\boldsymbol{S}} \sqrt{\binom{l_\sharp}{S_1, S_2, \ldots, S_k}} \frac{\boldsymbol{z}^S}{z_i}.$$

We note that D_i is a degree $l_\sharp - 1$ polynomial in k independent Gaussian variables. Let $\boldsymbol{d}^{(i)} \in \mathbb{R}^{|\mathcal{I}_{l_\sharp}|}$ be its coefficient vector in the monomial basis. Let $\boldsymbol{B}_{l_\sharp} \in \mathbb{R}^{|\mathcal{I}_{l_\sharp}| \times |\mathcal{I}_{l_\sharp}|}$ be the linear transformation

that converts the monomial basis representation to the Hermite Basis representation. Then we have,

$$\begin{aligned}
\|D^{(i)}\|_2^2 &= \|\boldsymbol{B}_{l_\sharp}\boldsymbol{d}^{(i)}\|_2^2 \\
&\geq \lambda_{\min}\left(\boldsymbol{B}_{l_\sharp}^T\boldsymbol{B}_{l_\sharp}\right)\|\boldsymbol{d}^{(i)}\|_2^2 \\
&\overset{\text{Lemma 36}}{\geq} 2^{-3l_\sharp-k}\|\boldsymbol{d}^{(i)}\|_2^2 \\
&= 2^{-3l_\sharp-k}\sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1=l_\sharp} S_i^2 a_{\boldsymbol{S}}^2 \binom{l_\sharp}{S_1, S_2, \ldots, S_k} \\
&\geq 2^{-3l_\sharp-k}\sum_{\boldsymbol{S}:\|\boldsymbol{S}\|_1=l_\sharp} S_i a_{\boldsymbol{S}}^2.
\end{aligned}$$

Combining the above display with Lemma 16, we get,

$$\sum_{i=1}^{k}\|D^{(i)}\|_2^2 \overset{\text{Lemma 16}}{\geq} 2^{-3l_\sharp-k}\frac{\mu^2}{2(2R^2+\mu)}.$$

We know that $D^{(i)}$ is a polynomial in k independent Gaussian variables of degree $l_\sharp - 1$. Applying Carbery-Wright Theorem (Theorem 44),

$$\mathbb{P}\left[|D^{(i)}| \leq \|D^{(i)}\|_2\left(\frac{\delta}{kK_{CW}(l_\sharp-1)}\right)^{l_\sharp-1}\right] \leq \frac{\delta}{k}.$$

Furthermore, using Equation (19) and a union bound, we know that:

$$\mathbb{P}\left[\|\boldsymbol{g}\|_2^{2l_\sharp-2}\|\boldsymbol{\nabla}F_{l_\sharp}(\boldsymbol{u})\|_2^2 \leq \left(\frac{\delta}{kK_{CW}(l_\sharp-1)}\right)^{2l_\sharp-2}\frac{\mu^2}{2^{3l_\sharp+k+1}(2R^2+\mu)}\right] \leq \delta.$$

Using standard chi-square concentration (Fact 5),

$$\mathbb{P}[\|\boldsymbol{g}\|_2^2 > 1.5p] \leq 2\exp(-p/32).$$

Hence, using a union bound we have,

$$\mathbb{P}\left[\|\boldsymbol{\nabla}F_{l_\sharp}(\boldsymbol{u})\|_2^2 \leq \left(\frac{\delta}{2kK_{CW}(l_\sharp-1)\sqrt{p}}\right)^{2l_\sharp-2}\frac{\mu^2}{2^{3l_\sharp+k+1}(2R^2+\mu)}\right] \leq \delta + 2\exp(-p/32).$$

■

Theorem 27 (Theorem 17 restated) *Given any $\epsilon \in (0,1)$; with probability $1-2\delta K_{max}\left(\frac{2R^2}{\mu}+1\right)-\frac{4K_{max}\left(\frac{2R^2}{\mu}+1\right)}{n}-2\exp(-p/32)$, The estimate returned by Algorithm 3 satisfies*

$$\|\mathcal{P}_{\mathcal{U}^\star}^{\perp}(\hat{\boldsymbol{u}})\|_2 \leq \epsilon,$$

provided n satisfies

$$n \geq \frac{4 \cdot 10^4 (\|f\|_\infty + 4\sigma)^2 (2R^2 + \mu)}{\epsilon^2 \mu^2} \left(\frac{256 \ln(n) K_{max}^4 \left(\frac{2R^2}{\mu} + 1 \right)^4 K_{CW}^2 \ln(1/\delta)}{\delta^2} \cdot p \right)^{K_{max}\left(\frac{2R^2}{\mu}+1\right)}.$$

In the above display, K_{max} is an upper bound on the true k given to the algorithm and $K_{CW} > 1$ is a universal constant appearing in the Carbery-Wright Theorem.

Proof We begin by introducing some notation. We define:

$$\boldsymbol{\Delta}_l := \nabla \hat{F}(\boldsymbol{u}_0) - \nabla F_l(\boldsymbol{u}_0) \quad L := \frac{2K_{\max} R^2}{\mu} + K_{\max} - 1.$$

Applying Theorem 48 and a union bound, we know that with probability $1 - 2L\delta - \frac{4L}{n}$,

$$\max_{l \in [L]} \|\boldsymbol{\Delta}_l\|_2 \leq e(n),$$

where we define $e(n)$ as:

$$e(n) := 100(\|f\|_\infty + 4\sigma) \cdot 2^L \sqrt{\frac{\max\left(p, \ln(1/\delta) \ln^L(n)\right)}{n}}.$$

Theorem 13 guarantees the existence of $l_\sharp$ such that:

$$l_\sharp \leq \frac{2kR^2}{\mu} + k - 1 \leq L.$$

For this $l_\sharp$, Lemma 26 tells us with probability $1 - \delta - 2\exp(-p/32)$,

$$\|\nabla F_{l_\sharp}(\boldsymbol{u}_0)\|_2 \geq \omega,$$

where we define ω as:

$$\omega := \left(\frac{\delta}{2kK_{CW}(l_\sharp - 1)\sqrt{p}} \right)^{l_\sharp - 1} \frac{\mu}{\sqrt{2^{3l_\sharp + k + 1}(2R^2 + \mu)}}.$$

It is easy to check that once

$$n \geq \frac{4 \cdot 10^4 (\|f\|_\infty + 4\sigma)^2 (2R^2 + \mu)}{\epsilon^2 \mu^2} \left(\frac{256 \ln(n) K_{\max}^4 \left(\frac{2R^2}{\mu} + 1 \right)^4 K_{CW}^2 \ln(1/\delta)}{\delta^2} \cdot p \right)^{K_{\max}\left(\frac{2R^2}{\mu}+1\right)},$$

we have,

$$e(n) \leq \frac{\epsilon}{2\sqrt{p}} \omega.$$

We can now analyze the estimator returned by Algorithm 3. Let $\boldsymbol{u}_{k+1}^\star, \boldsymbol{u}_{k+2}^\star \ldots, \boldsymbol{u}_p^\star$ be an orthonormal basis for $\mathcal{U}^{\star\perp}$. Consider the projection of the estimator returned by the algorithm on any $\boldsymbol{u}_i$, $i \geq k+1$:

$$\begin{aligned} |\langle \boldsymbol{u}_i, \hat{\boldsymbol{u}} \rangle| &= \frac{\left|\left\langle \nabla \hat{F}_{l_{\text{best}}}(\boldsymbol{u}_0), \boldsymbol{u}_i^\star \right\rangle\right|}{\|\nabla \hat{F}_{l_{\text{best}}}(\boldsymbol{u}_0)\|} \\ &\leq \frac{|\langle \nabla F_{l_{\text{best}}}(\boldsymbol{u}_0), \boldsymbol{u}_i^\star \rangle| + |\langle \boldsymbol{\Delta}_{l_{\text{best}}}, \boldsymbol{u}_i^\star \rangle|}{\|\nabla \hat{F}_{l_{\text{best}}}(\boldsymbol{u}_0)\|}. \end{aligned} \tag{20}$$

Next we observe that since $\nabla F_{l_{\text{best}}}(\boldsymbol{u}_0) \in \mathcal{U}^\star$, we have,

$$\langle \nabla F_{l_{\text{best}}}(\boldsymbol{u}_0), \boldsymbol{u}_i^\star \rangle = 0. \tag{21}$$

By Cauchy-Schwarz inequality,

$$|\langle \boldsymbol{\Delta}_{l_{\text{best}}}, \boldsymbol{u}_i^\star \rangle| \leq \|\Delta_{l_{\text{best}}}\| \leq e(n) \leq \frac{\epsilon}{2\sqrt{p}}\omega. \tag{22}$$

Next using the definition of l_{best}, we know that,

$$\begin{aligned} \|\nabla \hat{F}_{l_{\text{best}}}(\boldsymbol{u}_0)\| &\geq \|\nabla \hat{F}_{l_\sharp}(\boldsymbol{u}_0)\| \\ &\overset{\text{Triangle Inequality}}{\geq} \|\nabla F_{l_\sharp}(\boldsymbol{u}_0)\| - e(n) \qquad (23) \\ &\geq \omega - e(n) \\ &\geq \omega\left(1 - \frac{\epsilon}{2\sqrt{p}}\right) \\ &\overset{\epsilon \in (0,1)}{\geq} \frac{\omega}{2}. \qquad (24) \end{aligned}$$

Substituting the bounds obtained in Equation (21), Equation (22) and Equation (24) into Equation (20) gives:

$$|\langle \boldsymbol{u}_i, \hat{\boldsymbol{u}} \rangle| \leq \frac{\epsilon}{\sqrt{p}}.$$

This implies,

$$\|\mathcal{P}_{\mathcal{U}^\star}^\perp(\hat{\boldsymbol{u}})\|_2^2 = \sum_{i=k+1}^{p} |\langle \boldsymbol{u}_i, \hat{\boldsymbol{u}} \rangle|^2 \leq \epsilon^2.$$

■

Appendix D. Handling Unbounded Link Functions

In this section, we relax the assumption that the link function g is bounded. We assume the link function g satisfies Assumptions 1, 3 and 4. But instead of the assumption that $\|g\|_\infty < \infty$, we assume that g and ∇g grow at most polynomially at infinity. More precisely, we assume, that the link function g is (d, T, C)-polynomial bounded (defined below).

Definition 28 *A link function $g : \mathbb{R}^k \to \mathbb{R}$ is (d, T, C)-polynomially bounded, if:*

$$\exists\, T > 0, r \in \mathbb{N} \text{ such that, } \forall\, t \geq T, \max(|g(\boldsymbol{z})|, \|\nabla g(\boldsymbol{z})\|_\infty) \leq C\|\boldsymbol{z}\|_\infty^d.$$

Remark 29 *A number of link functions of practical interest are not bounded but are polynomially bounded. For example, in the phase retrieval problem $g(z) = z^2$ which is $(2, 1, 2)$-polynomially bounded.*

The reduction from polynomially bounded link functions to bounded link functions involve the following steps:

1. We construct a auxiliary link function $\tilde{g}_t$ which is bounded. Here $t > 0$ is a parameter which we will choose appropriately at the end.

2. Next, we compute the ℓ_∞ norm bound, minimum signal strength parameter and smoothness parameter for the auxiliary link function. Hence Algorithms 1,2 and 3 will have the desired guarantees when the data is generated from a multi-index model with this link function.

3. Finally, we show that the total variation distance between the data distribution induced by the auxiliary link function and the true link function is small. Hence, if an algorithm succeeds with high probability with the auxiliary link function, it must succeed with high probability with the actual link function.

We first construct the auxiliary link function $\tilde{g}_t$. We first introduce some notation. Let $q(z) : \mathbb{R} \to \mathbb{R}$ be the function:

$$q(z) \stackrel{\text{def}}{=} \begin{cases} 0 & z \leq -2 \\ 2(z+2)^2 & -2 \leq z \leq -1.5 \\ 1 - 2(z+1)^2 & -1.5 \leq z \leq -1 \\ 1 & -1 \leq z \leq 1 \\ 1 - 2(z-1)^2 & 1 \leq z \leq 1.5 \\ 2(z-2)^2 & z \geq 2. \end{cases}$$

The above function is an approximation to the indicator function of the interval $[-1, 1]$ that is twice differentiable almost everywhere. In particular, we have,

$$q(z) = 1 \quad \forall \quad z \in [-1, 1], \quad q(z) = 0 \quad \forall \quad z \in (-\infty, -2] \cup [2, \infty), \quad 0 \leq q(z) \leq 1 \quad \forall \quad z \in \mathbb{R}.$$

Furthermore, we have that almost surely,

$$|q'(z)| \leq 2, \quad |q''(z)| \leq 4.$$

Finally we define the function $Q_t : \mathbb{R}^k \to \mathbb{R}$ as:

$$Q_t(\boldsymbol{z}) = \prod_{i=1}^{k} q\left(\frac{z_i}{t}\right).$$

We can now approximate the original link function with a bounded link function $\tilde{g}_t$ defined as:

$$\tilde{g}_t(\boldsymbol{z}) \stackrel{\text{def}}{=} \frac{g(\boldsymbol{z})Q_t(\boldsymbol{z})}{\sqrt{\mathbb{E}[g^2(\boldsymbol{z})Q_t^2(\boldsymbol{z})]}}.$$

The following lemma verifies that the auxiliary link function satisfies Assumptions 1,3 and 4. We recall that $\mathbb{1}\{\cdot\}$ denotes the indicator of an event.

Lemma 30 *The auxiliary link function $\tilde{g}_t$ satisfies*

1. $\|\tilde{g}_t\|_\infty \leq 2C(2t)^d$.
2. $\mathbb{E}[(\frac{\partial \tilde{g}_t(\boldsymbol{z})}{\partial z_i})^2] \geq \frac{\mu}{2}$ *for all* $i \in [k]$.
3. $\mathbb{E}[(\frac{\partial^2 \tilde{g}_t(\boldsymbol{z})}{\partial z_i \partial z_j})^2] \leq 10R^2$ *for all* $i, j \in [k]$

provided $t \geq \max(T, 8d, \sqrt{4\ln(C^2k)}, \sqrt{4\ln(C^2k(\mu+1)/\mu)}, \sqrt{4\ln(256kC^2(R^2+1)/R^2)})$.

Proof

1. We note that, since $\tilde{g}_t(\boldsymbol{z}) = 0 \; \forall \; \|\boldsymbol{z}\|_\infty > 2t$,

$$\|\tilde{g}_t\|_\infty \leq \frac{C(2t)^d}{\sqrt{\mathbb{E}[g^2(\boldsymbol{z})Q_t^2(\boldsymbol{z})]}}.$$

   Furthermore,

$$\begin{aligned}
\mathbb{E}[g^2(\boldsymbol{z})Q_t^2(\boldsymbol{z})] &= 1 - \mathbb{E}[g^2(\boldsymbol{z})(1 - Q_t^2(\boldsymbol{z})] \\
&\geq 1 - \mathbb{E}[g^2(\boldsymbol{z})\mathbb{1}\{\|\boldsymbol{z}\|_\infty > t\}] \\
&\geq 1 - C^2\mathbb{E}[\|\boldsymbol{z}\|_\infty^{2d}\mathbb{1}\{\|\boldsymbol{z}\|_\infty > t\}] \\
&\geq 1 - C^2k\mathbb{E}[|Z|^{2d}\mathbb{1}\{|Z| > t\}].
\end{aligned}$$

   From Lemma 42, we know that, $\mathbb{E}[|Z|^{2d}\mathbb{1}\{|Z| > t\}] \leq 0.25\exp(-t^2/4)$ provided $t > 8d$. Hence by choosing t such that $t > \max(\sqrt{4\ln(C^2k)}, 8d)$, we get $\|\tilde{g}_t\|_\infty \leq 2C(2t)^d$.

2. Consider the following sequence of inequalities:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial \tilde{g}_t(\boldsymbol{z})}{\partial z_i}\right)^2\right] &\geq \mathbb{E}\left[\left(\frac{\partial \tilde{g}_t(\boldsymbol{z})}{\partial z_i}\right)^2 \mathbb{1}\{\|\boldsymbol{z}\|_\infty \leq t\}\right] \\
&\geq \mathbb{E}\left[\left(\frac{\partial g(\boldsymbol{z})}{\partial z_i}\right)^2 \mathbb{1}\{\|\boldsymbol{z}\|_\infty \leq t\}\right] \\
&= \mu - \mathbb{E}\left[\left(\frac{\partial g(\boldsymbol{z})}{\partial z_i}\right)^2 \mathbb{1}\{\|\boldsymbol{z}\|_\infty > t\}\right] \\
&\geq \mu - kC^2\mathbb{E}[|Z|^{2d}\mathbb{1}\{|Z| > t\}].
\end{aligned}$$

   Choosing $t > \max(\sqrt{4\ln(C^2k(\mu+1)/\mu)}, 8d)$ gives us the required lower bound of $\frac{\mu}{2}$.

3. We note that,

$$\frac{\partial^2 \tilde{g}_t(\boldsymbol{z})}{\partial z_i \partial z_j} = \frac{1}{\sqrt{\mathbb{E}[g^2(\boldsymbol{z})Q_t^2(\boldsymbol{z})]}} \left(\frac{\partial^2 g(\boldsymbol{z})}{\partial z_i \partial z_j} + \frac{\partial^2 Q_t(\boldsymbol{z})}{\partial z_i \partial z_j} + \frac{\partial g}{\partial z_i}\frac{\partial Q_t}{\partial z_j} + \frac{\partial g}{\partial z_j}\frac{\partial Q_t}{\partial z_i} \right).$$

Furthermore,

$$\left| \frac{\partial Q_t}{\partial z_i} \right| \leq 2 \cdot \mathbb{1}\{t < \|\boldsymbol{z}\|_\infty < 2t\}$$
$$\left| \frac{\partial^2 Q_t}{\partial z_i \partial z_j} \right| \leq 4 \cdot \mathbb{1}\{t < \|\boldsymbol{z}\|_\infty < 2t\}.$$

Hence,

$$\mathbb{E}\left[\left(\frac{\partial \tilde{g}_t(\boldsymbol{z})}{\partial z_i} \right)^2 \right] \leq 8\mathbb{E}\left(\left(\frac{\partial^2 g(\boldsymbol{z})}{\partial z_i \partial z_j} \right)^2 + \left(\frac{\partial^2 Q_t(\boldsymbol{z})}{\partial z_i \partial z_j} \right)^2 + \left(\frac{\partial g}{\partial z_i}\frac{\partial Q_t}{\partial z_j} \right)^2 + \left(\frac{\partial g}{\partial z_j}\frac{\partial Q_t}{\partial z_i} \right)^2 \right)$$
$$= 8\left(R^2 + 16\mathbb{P}\left[t < \|\boldsymbol{z}\|_\infty < 2t \right] + 8\mathbb{E}\left[\left(\frac{\partial g}{\partial z_i} \right)^2 \mathbb{1}\{t < \|\boldsymbol{z}\|_\infty < 2t\} \right] \right).$$

Finally choosing $t > \max(\sqrt{4\ln(256kC^2(R^2+1)/R^2)}, 8d)$ gives us the required upper bound of $10R^2$.

■

Next, we bound the total variation distance between the measures induced on the data when the link function is g and the link function is $\tilde{g}_t$. We first introduce some notation:

$$\boldsymbol{X} \stackrel{\text{def}}{=} [\boldsymbol{x}_1 | \boldsymbol{x}_2 | \cdots | \boldsymbol{x}_n]^T$$
$$\boldsymbol{Y} \stackrel{\text{def}}{=} [y_1, y_2, \ldots, y_n]^T.$$

Let $\mathcal{P}$ be the measure induced on $(\boldsymbol{X}, \boldsymbol{Y})$ when the link function is g. Let $\tilde{\mathcal{P}}_t$ denote the measure induced on $(\boldsymbol{X}, \boldsymbol{Y})$ when the link function is $\tilde{g}_t$. We note that the under both these measures, the marginal density of $\boldsymbol{X}$ is the same. Let us denote the marginal density of $\boldsymbol{X}$ by $p(\cdot)$. Let the conditional density of $\boldsymbol{Y}|\boldsymbol{X} = X$ be denoted by $q(\cdot|X)$ and $\tilde{q}(\cdot|X)$ under $\mathcal{P}$ and $\tilde{\mathcal{P}}_t$ respectively. Finally, $\forall\, t > 0$, we define the event E_t as follows:

$$E_t \stackrel{\text{def}}{=} \{\|\boldsymbol{x}_i\|_\infty \leq t \;\forall\; i \in [n]\}.$$

We observe that if $X \in E_t$, then,

$$q(\cdot|X) = \tilde{q}(\cdot|X).$$

The following lemma bounds the total variation distance (denoted by $d_{\text{TV}}(\mathcal{P}, \tilde{\mathcal{P}}_t)$) between the measure $\mathcal{P}$ and $\tilde{\mathcal{P}}_t$.

Lemma 31 *In the setup introduced above, we have,*

$$d_{TV}(\mathcal{P}, \tilde{\mathcal{P}}_t) \leq 2np\exp(-t^2/2).$$

Proof From the definition of total variation distance we have,

$$\begin{aligned} d_{\mathrm{TV}}(\mathcal{P}, \tilde{\mathcal{P}}_t) &= \frac{1}{2} \int_{\mathbb{R}^{n\times p}} \int_{\mathbb{R}^n} |p(X)q(Y|X) - p(X)\tilde{q}(Y|X)| \, dY \, dX \\ &= \frac{1}{2} \int_{X \in E_t} \int_{\mathbb{R}^n} |p(X)q(Y|X) - p(X)\tilde{q}(Y|X)| \, dY \, dX \\ &+ \frac{1}{2} \int_{X \notin E_t} \int_{\mathbb{R}^n} |p(X)q(Y|X) - p(X)\tilde{q}(Y|X)| \, dY \, dX \, . \end{aligned}$$

Using the fact that $q(\cdot|X) = \tilde{q}(\cdot|X)$, when $X \in E_t$, we find the first term in the above display is 0. Turning our attention to the second term, we note,

$$\begin{aligned} \frac{1}{2} \int_{\mathbb{R}^n} |q(Y|X) - \tilde{q}(Y|X)| \, dY &= d_{\mathrm{TV}}(q(\cdot|X), \tilde{q}(\cdot|X)) \\ &\leq 1. \end{aligned}$$

Hence,

$$\begin{aligned} d_{\mathrm{TV}}(\mathcal{P}, \tilde{\mathcal{P}}_t) &= \frac{1}{2} \int_{X \notin E_t} \int_{\mathbb{R}^n} |p(X)q(Y|X) - p(X)\tilde{q}(Y|X)| \, dY \, dX \\ &\leq \int_{X \notin E_t} p(X) \, dX \\ &= \mathbb{P}[\boldsymbol{X} \notin E_t]. \end{aligned}$$

Applying standard concentration bounds for a gaussian random variable and a union bound gives us the required result. ■

The following theorem gives a general reduction showing that any algorithm which is able to estimate multi-index models with bounded link functions is also able to estimate multi-index models with polynomially bounded link functions.

Theorem 32 *Let $\mathcal{A}$ be any algorithm which returns an estimate $\hat{\mathbf{u}}$ that satisfies $\|\mathcal{P}_{\mathcal{U}^\star}^{\perp} \hat{\mathbf{u}}\|_2 \leq \epsilon$ with probability atleast $1 - \delta$ given data generated from a multi-index model with normalized link function g with $\|g\|_\infty \leq B$, minimum signal strength parameter μ and smoothness parameter R^2 provided the number of samples n satisfies*

$$n \geq N(p, k, \sigma^2, \mu, R^2, B, \epsilon, \delta).$$

Then the same algorithm returns an estimate $\hat{\mathbf{u}}$ that satisfies $\|\mathcal{P}_{\mathcal{U}^\star}^{\perp} \hat{\mathbf{u}}\|_2 \leq \epsilon$ with probability atleast $1 - 2\delta$ given data generated from a multi-index model with (d, T, C)-polynomially bounded normalized link function with minimum signal strength parameter μ and smoothness parameter R^2 provided the number of samples n satisfies,

$$n \geq N(p, k, \sigma^2, \mu/2, 10R^2, B', \epsilon, \delta)$$

where

$$B' = 2C \max\left(T^d, 8^d d^d, 4^{d/2} \ln^{d/2}\left(\frac{256C^2(R^2+1)(\mu+1)}{\mu R^2 \delta} \cdot knp\right)\right).$$

Proof To construct $\tilde{g}_t$ choose t as:

$$t = \max\left(T, 8d, \sqrt{4\ln\left(\frac{256C^2(R^2+1)(\mu+1)}{\mu R^2\delta}\cdot knp\right)}\right).$$

By Lemma 30, we have, that $\tilde{g}_t$ satisfies

$$\begin{aligned} \|\tilde{g}_t\|_\infty &\le B' \\ \mathbb{E}\left[\left(\frac{\partial \tilde{g}_t(\boldsymbol{z})}{\partial z_i}\right)^2\right] &\ge \frac{\mu}{2} && \forall\, i \in [k] \\ \mathbb{E}\left[\left(\frac{\partial^2 \tilde{g}_t(\boldsymbol{z})}{\partial z_i \partial z_j}\right)^2\right] &\le 10R^2 && \forall\, i,j \in [k]. \end{aligned}$$

Hence when given data generated using link function $\tilde{g}_t$, the algorithm succeeds with probability $1-\delta$ provided,

$$n \ge N(p, k, \sigma^2, \mu/2, 10R^2, B', \epsilon, \delta).$$

On the other hand, by Lemma 31,

$$d_{\mathrm{TV}}(\mathcal{P}, \tilde{\mathcal{P}}_t) \le \delta.$$

Hence, by the definition of Total Variation distance, given data generated from g, the same algorithm succeeds with probability atleast $1-2\delta$. ■

Remark 33 *We note that in Theorems 14 and 17, the sample complexity depends only polynomially on $\|g\|_\infty$. Hence, if the link function is unbounded but polynomially bounded, the same guarantees hold provided the number of samples $n \ge \tilde{\mathcal{O}}(\mathrm{poly}(p)/\epsilon^2)$ where the $\tilde{O}$ notation suppresses factors that are logarithmic in p but can possibly be exponential in the link function parameters like $R, \frac{1}{\mu}, d$.*

Appendix E. Properties of Hermite Polynomials

Fact 1 (Explicit Form of Hermite Polynomials) *The (normalized) Hermite Polynomial of degree i is given by:*

$$H_i(x) = \sqrt{i!}\sum_{m=0}^{\lfloor i/2\rfloor}\frac{(-1)^m}{m!}\frac{x^{i-2m}}{(i-2m)!2^m}.$$

Fact 2 (Differentiating Hermite Polynomials) *The derivative of the Hermite Polynomial of degree i is given by:*

$$H_i'(x) = \sqrt{i}H_{i-1}(x).$$

The following lemma gives an upper bound on the value of a Hermite polynomial on a compact interval. This will be helpful in analyzing the concentration properties of Hermite polynomials via a truncation argument.

Lemma 34 *For all $\lambda \geq i$,*

$$\sup_{|x| \leq \lambda} |H_i(x)| \leq \frac{i\lambda^i}{\sqrt{i!}}.$$

Proof We make the following crude approximations to get an upper bound:

$$\begin{aligned} |H_i(x)| &\overset{\text{Fact 1}}{=} \left| \sqrt{i!} \sum_{m=0}^{\lfloor i/2 \rfloor} \frac{(-1)^m}{m!} \frac{x^{i-2m}}{(i-2m)!2^m} \right| \\ &\leq \sqrt{i!} \sum_{m=0}^{\lfloor i/2 \rfloor} \frac{1}{m!} \frac{|x|^{i-2m}}{(i-2m)!2^m} \\ &\leq \sqrt{i!} \sum_{m=0}^{\lfloor i/2 \rfloor} \frac{1}{m!} \frac{\lambda^{i-2m}}{(i-2m)!2^m}. \end{aligned}$$

Next we note that for $\lambda \geq i$, $\lambda^i/i!$ is the dominant term in the above summation. This is because:

$$\begin{aligned} \frac{\frac{1}{m!}\frac{\lambda^{i-2m}}{(i-2m)!2^m}}{\frac{\lambda^i}{i!}} &= \frac{(i-2m+1)\cdot(i-2m+2)\cdots(i-1)\cdot i}{\lambda^{2m}2^m m!} \\ &\leq \left(\frac{i}{\lambda}\right)^{2m} \\ &\leq 1. \end{aligned}$$

Hence we have,

$$\begin{aligned} \sup_{|x| \leq \lambda} |H_i(x)| &\leq \frac{i}{2} \cdot \sqrt{i!} \frac{\lambda^i}{i!} \\ &\leq \frac{i\lambda^i}{\sqrt{i!}}. \end{aligned}$$

■

We will also need bounds on the maximum coefficient in the monomial representation of the Hermite Polynomial of degree i.

Lemma 35 *Let B_i be the maximum absolute coefficient in the monomial expansion of the degree i Hermite Polynomial. Then,*

$$B_i \leq 2^i.$$

Proof Using the explicit formula for Hermite Polynomials (Fact 1), we have,

$$B_i = \max_{0 \le m \le \lfloor \frac{i}{2} \rfloor} \frac{\sqrt{i!}}{m!(i-2m)!2^m}.$$

Hence, it suffices to bound the right hand side for a fixed value of $m \le \lfloor \frac{i}{2} \rfloor$. We observe that,

$$\begin{aligned}
\frac{\sqrt{i!}}{m!(i-2m)!2^m} &= \frac{(2m)!}{2^m m!\sqrt{i!}} \binom{i}{2m} \\
&\overset{\text{Lemma 50}}{\le} \frac{(2m)!}{2^m m!\sqrt{i!}} \cdot 2^i \\
&\overset{2m \le i}{\le} \frac{\sqrt{2m!}}{2^m m!} \cdot 2^i \\
&= \frac{2^i}{2^m} \sqrt{\binom{2m}{m}} \\
&\overset{\text{Lemma 50}}{\le} 2^i.
\end{aligned}$$

■

Appendix F. Condition Number of Monomial Basis

We define the index set $\mathcal{I}_t$ as: $\mathcal{I}_t \overset{\text{def}}{=} \{\boldsymbol{S} \in (\mathbb{N} \cup \{0\})^k : \|\boldsymbol{S}\|_1 \le t\}$. Any arbitrary polynomial with degree at most t in k variables is of the form:

$$V(\boldsymbol{z}) = \sum_{\boldsymbol{S} \in \mathcal{I}_l} v_{\boldsymbol{S}} \boldsymbol{z}^{\boldsymbol{S}}.$$

We have used the notation $\boldsymbol{z}^{\boldsymbol{S}} := \prod_{i=1}^k z_i^{S_i}$. We can associate every degree l polynomial $V(\boldsymbol{z})$ with a coefficient vector $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{I}_l|}$. We can also write the polynomial $V(\boldsymbol{z})$ in terms of the Hermite basis:

$$V(\boldsymbol{z}) = \sum_{\boldsymbol{S} \in \mathcal{I}_t} v'_{\boldsymbol{S}} H_{\boldsymbol{S}}(\boldsymbol{z}).$$

Let $\boldsymbol{B}_t \in \mathbb{R}^{|\mathcal{I}_t| \times |\mathcal{I}_t|}$ denote the invertible linear map that converts the monomial representation $\boldsymbol{v}$ to the Hermite representation $\boldsymbol{v}'$. That is, $\boldsymbol{v}' = \boldsymbol{B}_t \boldsymbol{v}$. The main goal of this section is to obtain lower bounds on $\lambda_{\min}\left(\boldsymbol{B}_t^T \boldsymbol{B}_t\right)$. We note that since the $\boldsymbol{B}_t^T \boldsymbol{B}_t$ is positive definite, $\lambda_{\min}\left(\boldsymbol{B}_t^T \boldsymbol{B}_t\right) = \frac{1}{\lambda_{\max}\left(\boldsymbol{B}_t^{-1} \boldsymbol{B}_t^{-T}\right)}$. Let $\boldsymbol{b}_{\boldsymbol{S}}$ denote the coefficient representation of the Hermite Polynomial $H_{\boldsymbol{S}}(\boldsymbol{z})$ in the monomial basis. One can see that the columns of $\boldsymbol{B}_t^{-1}$ are precisely $\boldsymbol{b}_{\boldsymbol{S}}$ for $\boldsymbol{S} \in \mathcal{I}_t$.

Lemma 36 (Condition Number of Monomial Basis)

$$\lambda_{min}\left(\boldsymbol{B}_t^T \boldsymbol{B}_t\right) \ge 2^{-3t-k}.$$

Proof Instead of trying to lower bound $\lambda_{\min}\left(\boldsymbol{B}_t^T\boldsymbol{B}_t\right)$, we upper bound $\lambda_{\max}\left(\boldsymbol{B}_t^{-1}\boldsymbol{B}_t^{-T}\right)$.

$$\begin{aligned}\lambda_{\max}\left(\boldsymbol{B}_t^{-1}\boldsymbol{B}_t^{-T}\right) &= \left\|\sum_{\boldsymbol{S}\in\mathcal{I}_l}\boldsymbol{b}_{\boldsymbol{S}}\boldsymbol{b}_{\boldsymbol{S}}^T\right\| \\ &\le \sum_{\boldsymbol{S}\in\mathcal{I}_l}\|\boldsymbol{b}_{\boldsymbol{S}}\|^2 \\ &\le |\mathcal{I}_t|\sum_{\boldsymbol{S}\in\mathcal{I}_t}\|\boldsymbol{b}_{\boldsymbol{S}}\|_\infty^2 \\ &\le |\mathcal{I}_t|^2\max_{\boldsymbol{S}\in\mathcal{I}_t}\|\boldsymbol{b}_{\boldsymbol{S}}\|_\infty^2. \end{aligned} \tag{25}$$

Next we give an upper bound for $\|\boldsymbol{b}_{\boldsymbol{S}}\|_\infty$. We recall that,

$$H_{\boldsymbol{S}}(\boldsymbol{z}) = \prod_{i=1}^k H_{S_i}(z_i).$$

Using the explicit formula for Hermite Polynomials (Fact 1) and the fact that the largest coefficient in the monomial representation of the degree l (univariate) Hermite Polynomial is 2^l (Lemma 35), we see that,

$$\|\boldsymbol{b}_{\boldsymbol{S}}\|_\infty \le 2^{S_1+S_2+\cdots+S_k} \le 2^t.$$

Furthermore using standard combinatorial arguments,

$$|\mathcal{I}_t| = \binom{t+k-1}{k-1} \overset{\text{Lemma 50}}{\le} 2^{t+k-1}.$$

Substituting these bounds in Equation (25) gives:

$$\lambda_{\max}\left(\boldsymbol{B}_t^{-1}\boldsymbol{B}_t^{-T}\right) \le 2^{3t+k-1} \implies \lambda_{\min}\left(\boldsymbol{B}_t^T\boldsymbol{B}_t\right) \ge 2^{-3t-k+1}.$$

■

Appendix G. Concentration Results

In this section we collect some basic concentration results which will be useful in our analysis.

We first recall the definitions of subgaussian and subexponential random variables from Wainwright (2015).

Definition 37 (Subgaussian Random Variables) *A random variable X with $\mathbb{E}[X] = \mu$ is called σ-subgaussian if:*

$$\forall\lambda\in\mathbb{R}, \mathbb{E}[\exp(\lambda(X-\mu))] \le \exp(\lambda^2\sigma^2/2).$$

Definition 38 (Subexponential Random Variables) *A random variable X with $\mathbb{E}[X] = \mu$ is called (ν, b)-subexponential if:*

$$\forall |\lambda| < \frac{1}{b}, \mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp(\lambda^2\nu^2/2).$$

Next we recall the standard concentration bounds for sum of independent subgaussian and subexponential random variables.

Fact 3 (Hoeffding Bound) *Let X_i be independent subgaussian random variables with mean μ_i variance proxies σ_i. Then,*

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) > t\right] \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).$$

Fact 4 (Subexponential Concentration) *Let X_i be independent subexponential random variables with parameters (ν_i, b_i). Define:*

$$\nu_\star := \sqrt{\sum_{i=1}^{n}\nu_i^2}$$
$$b_\star := \max_{i\in[n]} b_i.$$

Then,

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) > t\right] \leq \max\left(\exp\left(-\frac{t^2}{2\nu_\star^2}\right), \exp\left(-\frac{t}{2b_\star}\right)\right).$$

Fact 5 (Chi Square Concentration) *Let $Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then, $\forall t \in (0,1)$, we have,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} Z_k^2 - 1\right| > t\right] \leq 2\exp(-nt^2/8).$$

Lemma 39 (Product of a Gaussian RV and a bounded RV) *Let X_1 be a 1-subgaussian random variable and let X_2 be a bounded random variable. Then X_1X_2 is $8\|X_2\|_\infty$-subgaussian.*

Proof Omitted. ■

Lemma 40 (Conditioning preserves subgaussianity) *Let Z_1, Z_2 be jointly gaussian with $\mathbb{E}[Z_1] = \mathbb{E}[Z_2] = 0$, $\mathbb{E}[Z_1^2] = \mathbb{E}[Z_2^2] = 1$ and $\mathbb{E}[Z_1Z_2] = \rho$. Let $\mathcal{Z} = \{|Z_2| \leq \lambda\}$. Then conditioned on $\mathcal{Z}$, the distribution of Z_1 is 1-subgaussian.*

Proof Omitted. ■

Lemma 41 *Let $Z \sim \mathcal{N}(0,1)$ and let W be a ω-subgaussian random variable independent from Z. The WZ is $(4\omega, 4\omega)$-subexponential.*

Proof

$$\begin{aligned}\mathbb{E}[\exp(\lambda WZ)] &= \mathbb{E}_W[\mathbb{E}[\exp(\lambda ZW|W]] \\ &= \mathbb{E}_W[\exp(\lambda^2 W^2/2] \\ &= \mathbb{E}_W\left[\sum_{q=0}^{\infty} \frac{\lambda^{2q} W^{2q}}{q!}\right].\end{aligned}$$

Since W is ω-subgaussian, $\mathbb{E}[W^{2q}] \leq q! 4^q \omega^{2q}$. Substituting this bound we get, if $4\lambda^2\omega^2 < 1$,

$$\begin{aligned}\mathbb{E}[\exp(\lambda WZ)] &\leq \frac{1}{1-4\lambda^2\omega^2} \\ &= 1 + \frac{4\lambda^2\omega^2}{1-4\lambda^2\omega^2}.\end{aligned}$$

Furthermore, if $\lambda^2 \leq \frac{1}{8\omega^2}$

$$\mathbb{E}[\exp(\lambda WZ)] \leq \exp(8\lambda^2\omega^2).$$

Hence we conclude WZ is $(4\omega, 4\omega)$-subexponential. ■

Below, $\mathbb{1}\{P\}$ is the zero-one indicator function for a predicate P.

Lemma 42 *Let $Z \sim \mathcal{N}(0,1)$. Let $l \in \mathbb{N}$. Then, for all $\lambda > 4l$, we have,*

$$\mathbb{E}[|Z|^l \mathbb{1}\{|Z| > \lambda\}] \leq \frac{1.6}{\lambda} \exp(-\lambda^2/4).$$

Proof

$$\mathbb{E}[|Z|^l \mathbb{1}\{|Z| > \lambda\}] = \sqrt{\frac{2}{\pi}} \int_{\lambda}^{\infty} z^l \exp(-z^2/2) dz.$$

Next by comparing the taylor series of x^l and $\exp(x^2/4)$ we note that $\exp(x^2/4) \geq x^l$ for all $x > 4l$. Since $\lambda > 4l$, we have,

$$\begin{aligned}\mathbb{E}[|Z|^l \mathbb{1}\{|Z| > \lambda\}] &\leq \sqrt{\frac{2}{\pi}} \int_{\lambda}^{\infty} \exp(-z^2/4) dz \\ &\leq \frac{2}{\lambda}\sqrt{\frac{2}{\pi}} \int_{\lambda}^{\infty} \frac{1}{2} z \exp(-z^2/4) dz \\ &= \frac{2}{\lambda}\sqrt{\frac{2}{\pi}} \exp(-\lambda^2/4) \\ &< \frac{1.6}{\lambda} \exp(-\lambda^2/4).\end{aligned}$$

■

Lemma 43 (Anticoncentration of a Uniformly Random Unit Vector) *Let $\boldsymbol{v} \sim$ Unif$(\mathbb{S}^{p-1})$ and let $\boldsymbol{u}$ be a fixed unit vector. Then,*

$$\mathbb{P}\left[|\langle \boldsymbol{u}, \boldsymbol{v} \rangle| \leq \frac{\delta}{\sqrt{p}} \right] \leq 2e^{-p/32} + \delta.$$

Proof Let $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. We know that,

$$\frac{g_1}{\|\boldsymbol{g}\|_2} \overset{\mathrm{d}}{=} \langle \boldsymbol{u}, \boldsymbol{v} \rangle .$$

Therefore we have,

$$\begin{aligned}
\mathbb{P}\left[|\langle \boldsymbol{u}, \boldsymbol{v} \rangle| \leq \frac{\delta}{\sqrt{p}} \right] &= \mathbb{P}\left[\frac{|g_1|}{\|\boldsymbol{g}\|_2} \leq \frac{\delta}{\sqrt{p}} \right] \\
&\leq \mathbb{P}\left[|g_1| \leq \frac{\delta}{\sqrt{p}} \|\boldsymbol{g}\|_2, \|\boldsymbol{g}\|_2^2 \leq 1.5p \right] + \mathbb{P}\left[\|g\|_2^2 > 1.5p \right] \\
&\overset{\text{Fact 5}}{\leq} \mathbb{P}\left[|g_1| \leq \frac{\delta}{\sqrt{p}} \|\boldsymbol{g}\|_2, \|\boldsymbol{g}\|_2^2 \leq 1.5p \right] + 2e^{-p/32} \\
&\leq \mathbb{P}[|g_1| \leq \delta\sqrt{1.5}] + 2\exp(-p/32). \qquad (26)
\end{aligned}$$

Next we note that, if $Z \sim \mathcal{N}(0, 1)$,

$$\begin{aligned}
\mathbb{P}[|Z| \leq c] &= \int_{-c}^{c} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&\leq \frac{2c}{\sqrt{2\pi}}.
\end{aligned}$$

Substituting the above display in Equation (26),

$$\begin{aligned}
\mathbb{P}\left[|\langle \boldsymbol{u}, \boldsymbol{v} \rangle| \leq \frac{\delta}{\sqrt{p}} \right] &\leq \frac{2\delta\sqrt{1.5}}{\sqrt{2\pi}} + 2e^{-p/32} \\
&< \delta + 2e^{-p/32}.
\end{aligned}$$

■

Theorem 44 (Carbery and Wright, 2001; O'Donnell, 2014) *Let $P : \mathbb{R}^p \to \mathbb{R}$ be a polynomial of degree atmost L. Then,*

$$\mathbb{P}_{\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)} \left[|P(\boldsymbol{z})| \leq \frac{\|P(\boldsymbol{z})\|_2 \delta^L}{K_{CW}^L L^L} \right] \leq \delta.$$

Here K_{CW} is a universal constant.

G.1. Concentration of Gradients

Let $h : \mathbb{R}^p \to \mathbb{R}$ be a arbitrary bounded function. Let $(\boldsymbol{x}_i, y_i)$ be independent and identically distributed observations from the following model:

$$\begin{aligned}\boldsymbol{x}_i &\sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_p\right), \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}\left(0, \sigma^2\right), \\ y_i &= h(\boldsymbol{x}_i) + \epsilon_i.\end{aligned}$$

The gradient of the objectives we consider are of the form:

$$\frac{1}{n}\sum_{i=1}^{n} y_i H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i.$$

Hence, we will often be interested in analyzing the deviation of the following empirical average from its expectation:

$$\frac{1}{n}\sum_{i=1}^{n} y_i H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i - \mathbb{E}[y H_l(\langle \boldsymbol{x}, \boldsymbol{u}\rangle)\boldsymbol{x}].$$

In the above display, H_l denotes the Hermite polynomial of degree l and $\boldsymbol{u}$ is a fixed unit vector.

Since the aforementioned vector involves higher moments of gaussians, we proceed via a standard truncation argument. We define the events:

$$\begin{aligned}\mathcal{E}_i(\lambda) &= \{|\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle| \le \lambda\}, \\ \mathcal{E}_{\text{all}}(\lambda) &= \bigcap_{i=1}^{n} \mathcal{E}_i.\end{aligned}$$

We will represent the conditional distribution of $\boldsymbol{x}_i$ on the event $\mathcal{E}_i$ with the random variable $\tilde{x}_i$. The basic idea behind introducing this event is that conditioned on this event, all the random variables involved are either subgaussian or subexponential and standard concentration inequalities apply. We analyze the concentration of this random vector via a sequence of intermediate lemmas:

1. In Lemma 45 we analyze the difference between expectations conditioned on the event $\mathcal{E}_{\text{all}}(\lambda)$ and the unconditional expectations. This is important since conditioned on $\mathcal{E}_{\text{all}}(\lambda)$ all quantities concentrate to their respective expectations conditioned on $\mathcal{E}_{\text{all}}(\lambda)$ while we want to show they concentrate near their unconditional expectations.
2. In Lemma 46 we analyze the concentration of the signal part, $\frac{1}{n}\sum_{i=1}^{n} y_i H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i$.
3. In Lemma 47 we analyze the concentration of the noise part, $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i$.
4. In Lemma 48, we combine the above mentioned intermediate results into a ready-to-use theorem.

Lemma 45 *Consider the setup outlined above. Let $\boldsymbol{a}$ be any fixed unit vector. Then we have,*

$$\mathbb{E}[\epsilon_i H_l(\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle \boldsymbol{x}_i] - \mathbb{E}[\epsilon_i H_l(\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle)\boldsymbol{x}_i|\mathcal{E}_i(\lambda)] = 0,$$

$$|\mathbb{E}\left[h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\left\langle \boldsymbol{x}_i, \boldsymbol{a}\right\rangle\right] - \mathbb{E}\left[h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\left\langle \boldsymbol{x}_i, \boldsymbol{a}\right\rangle|\mathcal{E}_i(\lambda)\right]| \leq \frac{\|h\|_\infty}{2}\exp(-\lambda^2/8),$$

provided $\lambda \geq 8l$.

Proof For the first equality, we note that ϵ_i is independent of $\boldsymbol{x}_i$ and hence,

$$\mathbb{E}[\epsilon_i H_l(\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle)\boldsymbol{x}_i|\mathcal{E}_i(\lambda)] = \mathbb{E}[\epsilon_i H_l(\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle \boldsymbol{x}_i] = 0.$$

For the second claim we note that,

$$\mathbb{E}\left[h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i\right] = \mathbb{E}\left[h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i|\,\mathcal{E}_i\right]\mathbb{P}[\mathcal{E}_i] + \mathbb{E}\left[h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i|\mathcal{E}_i^c\right]\mathbb{P}[\mathcal{E}_i^c].$$

Hence,

$$|\mathbb{E}\left[h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\left\langle \boldsymbol{x}_i, \boldsymbol{a}\right\rangle\right] - \mathbb{E}\left[h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\left\langle \boldsymbol{x}_i, \boldsymbol{a}\right\rangle|\mathcal{E}_i(\lambda)\right]| = |\mathbb{E}\left[h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\left\langle \boldsymbol{x}_i, \boldsymbol{a}\right\rangle \mathbb{1}\{\mathcal{E}_i^c\}\right]|.$$

Furthermore,

$$\begin{aligned}
|\mathbb{E}\left[h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\left\langle \boldsymbol{x}_i, \boldsymbol{a}\right\rangle \mathbb{1}\{\mathcal{E}_i^c\}\right]| \leq\ & \mathbb{E}\left[|h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\left\langle \boldsymbol{x}_i, \boldsymbol{a}\right\rangle|\mathbb{1}\{\mathcal{E}_i^c(\lambda)\}\right] \\
\leq\ & \|h\|_\infty \mathbb{E}\left[|H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\left\langle \boldsymbol{x}_i, \boldsymbol{a}\right\rangle|\mathbb{1}\{\mathcal{E}_i^c(\lambda)\}\right] \\
\overset{\text{Lemma 34}}{\leq}\ & \frac{l\|h\|_\infty}{\sqrt{l!}}\mathbb{E}[|\left\langle \boldsymbol{a}, \boldsymbol{x}_i\right\rangle\left\langle \boldsymbol{u}, \boldsymbol{x}_i\right\rangle^l|\mathbb{1}\{\mathcal{E}_i(\lambda)^c\}] \\
\overset{\text{Cauchy-Schwarz}}{\leq}\ & \frac{l\|h\|_\infty}{\sqrt{l!}}\sqrt{\mathbb{E}\left[(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)^{2l}\mathbb{1}\{\mathcal{E}_i^c(\lambda)\}\right]} \\
\overset{\text{Lemma 42}}{\leq}\ & \frac{1.3l\|h\|_\infty}{\sqrt{l!}\sqrt{\lambda}}\exp(-\lambda^2/8) \\
\overset{\lambda \geq 8l}{\leq}\ & \frac{\|h\|_\infty}{2}\exp(-\lambda^2/8).
\end{aligned}$$

■

Lemma 46 *In the setup introduced above, with probability, $1 - \delta - \frac{2}{n}$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle)\boldsymbol{x}_i - \mathbb{E}[h(\boldsymbol{x})H_l(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)\boldsymbol{x}]\right\|_2 \leq 100\|h\|_\infty \cdot 2^l\sqrt{\frac{\max(p, \ln(1/\delta))\ln^l(n)}{n}}.$$

Proof Let $\boldsymbol{E}_n := \frac{1}{n}\sum_{i=1}^n h(\boldsymbol{x}_i)\boldsymbol{x}_i - \mathbb{E}[h(\boldsymbol{x})\boldsymbol{x}]$. Let $\mathcal{N}$ denote a $\frac{1}{2}$-packing of the unit sphere in $\mathbb{R}^p$ with $|\mathcal{N}| \leq 5^p$. Using standard arguments, we know that,

$$\|E_n\|_2 = \sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2 \leq 1}\langle a, E_n\rangle \leq 2\max_{\boldsymbol{a}\in\mathcal{N}}\left\langle \boldsymbol{a}, \boldsymbol{E}_n\right\rangle.$$

Hence we have,

$$\mathbb{P}\left[\|\boldsymbol{E}_n\|_2 > 2t\right] \leq \mathbb{P}[\|\boldsymbol{E}_n\|_2 > 2t|\mathcal{E}_{\text{all}}(\lambda)] + \mathbb{P}[\mathcal{E}_{\text{all}}(\lambda)^c] \tag{27}$$

$$\leq \mathbb{P}\left[\max_{\boldsymbol{a}\in\mathcal{N}} \langle \boldsymbol{a}, \boldsymbol{E}_n\rangle > t|\mathcal{E}_{\text{all}}(\lambda)\right] + \mathbb{P}[\mathcal{E}_{\text{all}}(\lambda)^c]$$

$$\leq 5^p \mathbb{P}\left[\langle \boldsymbol{a}, \boldsymbol{E}_n\rangle > t|\mathcal{E}_{\text{all}}(\lambda)\right] + \mathbb{P}[\mathcal{E}_{\text{all}}(\lambda)^c]. \tag{28}$$

For a fixed unit vector $\boldsymbol{a}$,

$$\langle \boldsymbol{a}, \boldsymbol{E}_n\rangle = \frac{1}{n}\sum_{i=1}^n h(\boldsymbol{x}_i)\langle \boldsymbol{x}_i, \boldsymbol{a}\rangle H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle) - \mathbb{E}[h(\boldsymbol{x})\langle \boldsymbol{x}, \boldsymbol{a}\rangle H_l(\langle \boldsymbol{x}, \boldsymbol{u}\rangle)]$$

$$\overset{\text{Lemma 45}}{\leq} \underbrace{\frac{1}{n}\sum_{i=1}^n h(\boldsymbol{x}_i)\langle \boldsymbol{x}_i, \boldsymbol{a}\rangle H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle) - \mathbb{E}[h(\boldsymbol{x})\langle \boldsymbol{x}, \boldsymbol{a}\rangle H_l(\langle \boldsymbol{x}, \boldsymbol{u}\rangle)|\mathcal{E}]}_{\tilde{E}_n} + \frac{\|h\|_\infty}{2}\exp(-\lambda^2/8).$$

The final task is to analyze the concentration of $\tilde{E}_n$ conditioned on the event $\mathcal{E}_{\text{all}}(\lambda)$. We note that on the event $\mathcal{E}_{\text{all}}(\lambda)$, $\langle \boldsymbol{x}_i, \boldsymbol{a}\rangle$ is 1-subgaussian. By Lemma 34, $|h(\boldsymbol{x}_i)H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)| \leq 2\|h\|_\infty\lambda^{l+1}$. By Lemma 39, $h(x_i)\langle \boldsymbol{x}_i, \boldsymbol{a}\rangle H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)$ is $16\|h\|_\infty\lambda^l$-subgaussian. Applying the Hoeffding bound (Fact 3), we get,

$$\mathbb{P}\left[\tilde{E}_n > 16\|h\|_\infty\lambda^l\gamma\right] \leq \exp\left(-\frac{nt^2}{2}\right).$$

Substituting this bound in Equation (28) gives us:

$$\mathbb{P}\left[\|\boldsymbol{E}_n\|_2 > 32\|h\|_\infty\lambda^l\gamma t + \|h\|_\infty\exp(-\lambda^2/8)\right] \leq 5^p\exp\left(-\frac{nt^2}{2}\right) + \mathbb{P}[\mathcal{E}_{\text{all}}(\lambda)^c].$$

By standard results on gaussian concentration, $\mathbb{P}[\mathcal{E}_{\text{all}}(\lambda)^c] \leq 2n\exp(-\lambda^2/2)$. We set:

$$\lambda = \sqrt{4\ln(n)},$$

$$\gamma = 3\sqrt{\frac{\max p, \ln(1/\delta)}{n}},$$

and conclude,

$$\mathbb{P}\left[\|\boldsymbol{E}_n\|_2 > 100\|h\|_\infty \cdot 2^l\sqrt{\frac{\max(p, \ln(1/\delta))\ln^l(n)}{n}}\right] \leq \delta + \frac{2}{n}.$$

■

Lemma 47 *Consider the setting introduced above. With probability* $1 - \delta - \frac{2}{n}$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i\right\|_2 \leq 400\sigma \cdot 2^l\sqrt{\frac{\max(p, \ln(1/\delta))\ln^l(n)}{n}},$$

provided,

$$n > 9 \max\left(p, \ln\frac{1}{\delta}\right).$$

Proof Let $\boldsymbol{F}_n := \frac{1}{n}\sum_{i=1}^n \epsilon_i H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i$. Via arguments analogous to the proof of Lemma 46, we get,

$$\mathbb{P}[\|\boldsymbol{F}_n\|_2 > 2t] \le 5^p \mathbb{P}[\langle \boldsymbol{F}_n, \boldsymbol{a}\rangle > t | \mathcal{E}_{\text{all}}(\lambda)] + \mathbb{P}[\mathcal{E}_{\text{all}}(\lambda)^c].$$

Here $\boldsymbol{a}$ is a fixed unit vector. Furthermore,

$$\langle \boldsymbol{F}_n, \boldsymbol{a}\rangle = \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{x}_i, \boldsymbol{a}\rangle\, H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\epsilon_i.$$

On the event $\mathcal{E}_{\text{all}}(\lambda)$, $H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle) \le 2\lambda^l$ and $\langle \boldsymbol{x}_i, \boldsymbol{a}\rangle$ is 1-subgaussian. By Lemma 39, $\langle \boldsymbol{x}_i, \boldsymbol{a}\rangle\, H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)$ is $16\lambda^l$-subgaussian. By Lemma 41, $\frac{\langle \boldsymbol{x}_i, \boldsymbol{a}\rangle H_l(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\epsilon_i}{\sigma}$ is $(64\lambda^l, 64\lambda^l)$ subexponential. By concentration of sum of independent subexponential random variables (Fact 4), we have,

$$\mathbb{P}[\|\boldsymbol{F}_n\|_2 > 64t\lambda^l\sigma] \le 5^p \max\left(\exp(-nt^2/2), \exp(-nt/2)\right) + 2n\exp(-\lambda^2/2).$$

We set

$$t = 3\sqrt{\frac{\max\left(p, \ln\frac{1}{\delta}\right)}{n}},$$
$$\lambda = \sqrt{4\ln(n)}.$$

When $n > 9\max\left(p, \ln\frac{1}{\delta}\right)$, we have $t < 1$ and we obtain,

$$\mathbb{P}\left[\|\boldsymbol{F}_n\|_2 > 400\sigma \cdot 2^l \sqrt{\frac{\max(p, \ln(1/\delta))\ln^l(n)}{n}}\right] \le \delta + \frac{2}{n}.$$

■

We are now ready to state our main concentration result.

Theorem 48 *In the setting introduced above, we have, with probability* $1 - 2\delta - \frac{4}{n}$,

$$\left\|\frac{1}{n}\sum_{i=1}^n y_i H_l(\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle)\boldsymbol{x}_i - \mathbb{E}[yH_l(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)\boldsymbol{x}]\right\|_2 \le 100(\|h\|_\infty + 4\sigma)\cdot 2^l\sqrt{\frac{\max(p, \ln(1/\delta)\ln^l(n)}{n}}.$$

Proof This follows directly from Lemma 46 and Lemma 47 and a union bound. ■

G.2. Concentration of Goodness-of-Fit Statistic

The goodness of fit statistic we consider is:

$$T_l(\boldsymbol{u}) = \frac{1}{|S_{\text{test}}|} \sum_{i \in S_{\text{test}}} y_i H_l(\langle \boldsymbol{u}, \boldsymbol{x}_i \rangle).$$

We define $m := |S_{\text{test}}|$. We recall that by Lemma 3, $\mathbb{E}[T(\boldsymbol{u})] = a_l^\star \langle \boldsymbol{u}^\star, \boldsymbol{u} \rangle^l$. The following lemma analyzes the concentration of $T_l(\boldsymbol{u})$ about its expectation.

Lemma 49 *With probability* $1 - \delta$*,*

$$|T_l(\boldsymbol{u}) - \mathbb{E}[T_l(\boldsymbol{u})]| \leq \Delta,$$

provided $m \geq \frac{2(\sigma^2 + \|h\|_\infty^2)}{\delta \Delta^2}$.

Proof Since we don't need exponential tail bounds, we use Chebychev's Inequality for simplicity. We first bound the variance,

$$\text{Var}\left(y H_l(\langle \boldsymbol{x}, \boldsymbol{u} \rangle)\right) \leq \mathbb{E}\left[y^2 H_l^2(\langle \boldsymbol{x}, \boldsymbol{u} \rangle\right].$$

Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ we have $y^2 \leq 2(\epsilon^2 + h(\boldsymbol{x})^2)$. Hence,

$$\begin{aligned} \text{Var}\left(y H_l(\langle \boldsymbol{x}, \boldsymbol{u} \rangle)\right) \leq & \quad 2\left(\mathbb{E}\left[\epsilon^2 H_l^2(\langle \boldsymbol{u}, \boldsymbol{x} \rangle)\right] + \mathbb{E}\left[h^2(\boldsymbol{x}) H_l^2(\langle \boldsymbol{x}, \boldsymbol{u} \rangle)\right]\right) \\ \leq & \quad 2\left(\sigma^2 + \|h\|_\infty^2\right). \end{aligned}$$

Applying Chebychev Inequality with this variance upper bound, we get,

$$\mathbb{P}\left[|T_l(\boldsymbol{u}) - \mathbb{E}[T_l(\boldsymbol{u})]| > \Delta\right] \leq \frac{2(\sigma^2 + \|h\|_\infty^2)}{m \Delta^2}.$$

Hence if $m \geq \frac{2(\sigma^2 + \|h\|_\infty^2)}{\delta \Delta^2}$, the above probability is bounded by δ. ■

Appendix H. Miscellaneous Results

Fact 6 (Stirling's Approximation)

$$\sqrt{2\pi} n^{n+1/2} e^{-n} \leq n! \leq e n^{n+1/2} e^{-n}.$$

Lemma 50 (Upper Bound on Multinomial Coefficient)

$$\binom{kt}{t, t, \ldots, t} \leq e\sqrt{tk} \left(\frac{k^{2t}}{2\pi t}\right)^{k/2}.$$

Proof This follows from Fact 6. ■

Lemma 51 *Define* $\mu_{2k} = \mathbb{E}[Z^{2k}]$ *where* $Z \sim \mathcal{N}(0,1)$. *Then,*

$$\mu_{2k} = \frac{2k!}{2^k k!} \leq \frac{e}{\sqrt{\pi}} \left(\frac{2k}{e}\right)^k.$$

Proof This follows from Fact 6 ■