Computational-Statistical Gaps in Gaussian Single-Index Models

Alex Damian[1], Loucas Pillaud-Vivien[2], Jason D. Lee[3], and Joan Bruna[4,5]

[1]PACM, Princeton University
[2]Ecole de Ponts ParisTech, CERMICS
[3]Electrical Engineering Department, Princeton University
[4]Courant Institute of Mathematical Sciences, New York University
[5]Center for Data Science, New York University

March 14, 2024

Abstract

Single-Index Models are high-dimensional regression problems with planted structure, whereby labels depend on an unknown one-dimensional projection of the input via a generic, non-linear, and potentially non-deterministic transformation. As such, they encompass a broad class of statistical inference tasks, and provide a rich template to study statistical and computational trade-offs in the high-dimensional regime.

While the information-theoretic sample complexity to recover the hidden direction is linear in the dimension d, we show that computationally efficient algorithms, both within the Statistical Query (SQ) and the Low-Degree Polynomial (LDP) framework, necessarily require $\Omega(d^{k^\star/2})$ samples, where $k^\star$ is a "generative" exponent associated with the model that we explicitly characterize. Moreover, we show that this sample complexity is also sufficient, by establishing matching upper bounds using a partial-trace algorithm. Therefore, our results provide evidence of a sharp computational-to-statistical gap (under both the SQ and LDP class) whenever $k^\star > 2$. To complete the study, we construct smooth and Lipschitz deterministic target functions with arbitrarily large generative exponents $k^\star$.

Contents

1 Introduction

1.1 Problem Setup

The focus of this paper is to study a family of high-dimensional inference tasks characterized by *planted* low-dimensional structure. In the context of supervised learning, where a learner observes a dataset $\{(x_i, y_i)\}_{i=1}^n$ with input features $x \in \mathbb{R}^d$ and labels $y \in \mathbb{R}$, the natural starting point is to consider data with hidden one-dimensional structure, and where features are drawn from the standard Gaussian measure γ_d in $\mathbb{R}^d$:

Definition 1.1 (Gaussian Single-Index Model)**.** We say that a joint distribution $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R})$ follows a Gaussian single index model if there exists a probability measure $\mathsf{P} \in \mathcal{G} \subset \mathcal{P}(\mathbb{R}^2)$ and $w^\star \in S^{d-1}$ such that $\mathbb{P} = [R_{w^\star} \otimes \mathrm{Id}]_\# [\gamma_{d-1} \otimes \mathsf{P}]$, where $R_{w^\star} \in \mathcal{O}_d$ is any orthogonal matrix whose last column is $w^\star$, i.e. of the form $R_{w^\star} = [R_\perp \; w^\star]$ and

$$\mathcal{G} = \left\{\nu_{(z,y)} \in \mathcal{P}(\mathbb{R} \times \mathbb{R});\; \nu_z = \gamma_1; \mathbb{E}_\nu[Y^2] < \infty;\; \mathbb{D}_{\chi^2}(\nu || \gamma_1 \otimes \nu_y) > 0\right\} \tag{1}$$

is the class of non-separable bivariate probability measures, whose first marginal is the standard one-dimensional Gaussian, and whose second-order moment w.r.t. its second argument is finite.

In words, a single index model is a joint distribution in $\mathbb{R}^d \times \mathbb{R}$ that admits a product structure $\mathrm{d}\mathbb{P}(x, y) = \mathrm{d}\gamma_{d-1}(R_\perp^\top x)\, \mathrm{d}\mathsf{P}(w^\star \cdot x, y)$ into a non-informative component $R_\perp^\top x$ of dimension $d-1$, and an informative component, determined precisely by P and the direction $w^\star$. The Gaussian setting further specifies the marginal distribution of the features. We will denote the planted model by $\mathbb{P}_{w^\star, \mathsf{P}}$ (or simply $\mathbb{P}_{w^\star}$ when the context is clear). We have used the conventions that, for any random variable X, P_x stands for the law of X under P, e.g. $\mathsf{P}_z = \gamma_1$, and P_y the marginal of

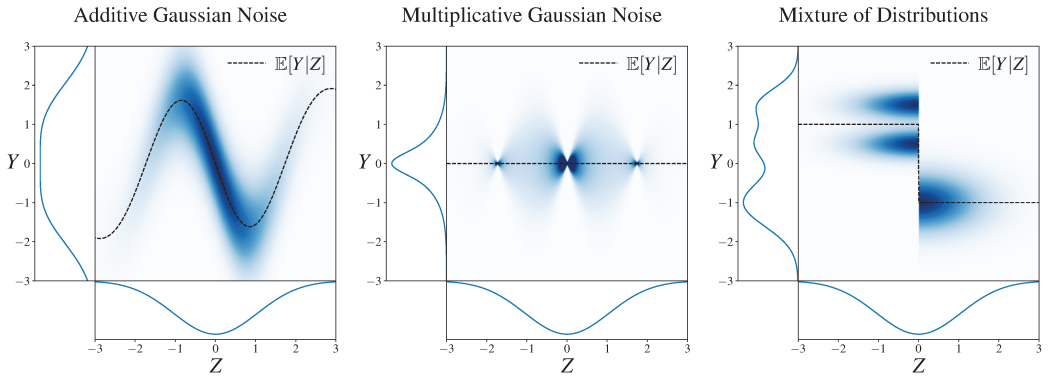


Figure 1: Visualization of the joint density P of (Z, Y) for the additive noise model, multiplicative noise model, and mixture of distributions model. The heatmap shows the density of P and the plots to the left of and below the heatmap show the densities of the marginals P_y and P_z respectively.

Y. Similarly, we will make use of notations $\mathsf{P}_{z|y}, \mathsf{P}_{y|z}$, that stand respectively for the conditional probability laws of Z given Y and Y given Z.

Note that in this language, Gaussian single-index models are closely related to Non-Gaussian Component Analysis (NGCA) [DKS17]. In NGCA, a d-dimensional distribution admits a product structure in terms of a univariate non-Gaussian marginal and a $d-1$ Gaussian distribution. In our case, the planted non-Gaussian component is two-dimensional, but the statistician is given one direction of the non-Gaussian subspace (the label Y).

Throughout the paper, we use the letter Z to denote the one-dimensional (Gaussian) random variable $w^\star \cdot X$. When there exists $\sigma : \mathbb{R} \to \mathbb{R}$ such that $\mathsf{P}_{y|z}(\cdot, z) = \delta_{\sigma(z)}$, we say that (X, Y) follows a deterministic single-index model as $Y = \sigma(Z) = \sigma(w^\star \cdot X)$, where σ is said to be *link function*. However, Definition 1.1 allows for additional randomness in the label, as long as its distribution only depends on x through $z = w^\star \cdot x$. Examples of non-deterministic single-index models include additive noise, where $Y = \sigma(Z) + \xi$ and ξ is an independent random variable, e.g. $\xi \sim N(0, 1)$; multiplicative noise, where $Y = \xi\sigma(Z)$, Mixture of distributions, where $Y \sim \mu_1$ if $Z \geq 0$ and $Y \sim \mu_2$ if $Z < 0$, or Massart-type noise, where $Y = \xi\sigma(Z)$ and $\mathbb{P}(\xi = 1) = 1 - \eta(Z)$ and $\mathbb{P}(\xi = -1) = \eta(Z)$; see Figure 1.

In the remainder of this paper, and unless stated otherwise, we assume that P is known, so that the inference task reduces to estimating the hidden direction $w^\star$ drawn from the uniform prior over S^{d-1}, after obserivng n iid samples from $\mathbb{P}_{w^\star,\mathsf{P}}$. By instantiating specific choices for P, one recovers several well-known statistical inference problems, such as linear recovery, phase-recovery, one-bit compressed sensing, generalized linear models or Non-Gaussian Component Analysis [DKS17], as well as close variants of Tensor PCA [MR14]. An important common theme across these different statistical models over recent years has been to understand *computational-to-statistical gaps*, namely comparing the required amount of samples needed to estimate $w^\star$, using either computationally efficient methods or brute-force search. Computational efficiency may be measured either by restricting estimation algorithms to belong to certain computational models, or by establishing reductions to problems believed to be computationally hard. In this paper, we focus our attention on *Statistical-Query* algorithms [Kea98], which capture a broad class of learning algorithms including robust gradient-descent methods, as well as the *Low-Degree Polynomial* method [Hop18, KWB19],

a flexible framework to assess the average-case complexity of statistical inference tasks.

The primary focus of this work is to derive the optimal sample complexity for recovering $w^\star$ given i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ from a single-index model (Definition 1.1). Perhaps unsuprinsingly, one can recover $w^\star$ up to error ϵ by brute-force search when $n = \Theta(d/\epsilon^2)$ (Theorem 6.1), in accordance with related statistical inference tasks. However, when restricted to SQ and LDP algorithms, our main results will establish a tight sample complexity of order $n = \Theta(d^{k^\star/2})$, with matching upper and lower bounds, where $k^\star = k^\star(\mathsf{P})$ is an exponent associated with P that we explicitly characterize (Definition 2.4). We thus obtain evidence of a computational-statistical gap of polynomial scale for a broad class of inference problems.

1.2 Background and Related Work

Single-index models (also called generalized linear models) have a long history in the statistics literature [MN83, Ich93, HJS01, HMSW04, DJS08, DH18]. When the link function σ is monotonic, there are perceptron-like algorithms (e.g. Isotron/GLM-tron [KS09, KKSK11]) that recover the ground truth signal with $n = \Theta(d)$ samples where d is the dimension of the data.

Perhaps the simplest example of a generalized linear model with a non-monotonic link function is phase retrieval in which $\sigma(u) = |u|$. In contrast to the monotonic case, phase retrieval exhibits a conjectured *statistical-computational gap* in the noisy setting [MM18, BKM+19, MLKZ20] In the presence of label noise, there are constants $c_{\mathrm{IT}} < c_{\mathrm{ALG}}$ such that recovery is information theoretically possible with $n/d \geq c_{\mathrm{IT}}$ but is conjectured to be computationally hard unless $n/d \geq c_{\mathrm{ALG}}$. Note, however, that this is not true in the noise-free setting as there are computationally efficient algorithms based on lattice reduction that achieve the information-theoretic threshold [AHSS17, SZB21].

For both monotonic and quadratic link functions, the optimal sample complexity scales linearly in the input dimension d. In addition, this rate can even be recovered by directly optimizing the maximum likelihood objective using simple first order methods such as online stochastic gradient descent (SGD). However, the problem is significantly harder when the link function is more complicated. [BAGJ21] demonstrated that for general single index models, online SGD on the maximum likelihood objective succeeds if and only if $n = \tilde{\Theta}(d^{\max(1,l^\star-1)})$ where $l^\star$ is the *information exponent* of the single index model, which is defined to be the degree of the first nonzero Hermite coefficient of σ (Definition 2.2). However, the significance of the information exponent extends far past optimizing the maximum likelihood objective. It has appeared as the fundamental object in determining the sample complexity in many follow up works [BBSS22, DLS22, DNGL23, DKL+23, ABAM23].

[DLS22] formalized this by proving a correlational statistical query (CSQ) lower bound which shows that either $n \gtrsim d^{\max(1,l^\star/2)}$ samples or exponentially many queries are necessary for learning a single index model with information exponent $l^\star$. In addition, this lower bound is tight as it is possible to learn a single index model with $\tilde{\Theta}(d^{\max(1,l^\star/2)})$ samples in polynomial time by training wide three layer neural networks [CBL+20] or by smoothing the landscape of the maximum likelihood objective [DNGL23].

The information exponent arose as the fundamental object governing sample complexities given *correlational queries* of the form $\mathbb{E}[Yh(X)]$. This is rich enough to capture gradient methods with mean squared error loss as they only interact with the data through correlational queries:

$$L(\theta) = \mathbb{E}[(f_\theta(X) - Y)^2] \quad \Longrightarrow \quad \nabla_\theta L(\theta) = \mathbb{E}[(f_\theta(X) - \underbrace{Y)\nabla_\theta f_\theta(X)}_{\text{correlational query}}] \,. \tag{2}$$

However, methods outside of the correlational statistical query (CSQ) framework can break the $n \gtrsim d^{\max(1,l^\star/2)}$ sample complexity barrier [CM20, MM18, BKM+19, MLKZ20, DTA+24]. The general technique for these methods is to first apply a pre-processing function $\mathcal{T}$ to the label Y to lower the information exponent to 2 before running a CSQ algorithm. This is possible because the information exponent defined by [BAGJ21] is not composition invariant, i.e. it is possible that the information exponent of $(X, \mathcal{T}(Y))$ is strictly smaller than the information exponent of (X, Y). In fact, [MM18, BKM+19, MLKZ20] give a necessary and sufficient condition on P that enables such $\mathcal{T}$ to lower the information exponent to 2. Using the notation described in Definition 1.1, the condition on P writes:

$$\mathbb{E}\left[\mathbb{E}\left[Z^2 - 1|Y\right]^2\right] \neq 0 \ , \tag{3}$$

where expectations are taken with respect to $(Z, Y) \sim \mathsf{P}$. Such 'pre-processing' methods that go beyond the CSQ lower bound are in fact instances of SQ-algorithms, which interact with data via general queries of the form $\mathbb{E}[\phi(X, Y)]$ for a broad class of test functions ϕ. In particular, these works imply that the SQ complexity for learning single-index models satisfying (3) is $n = \Theta(d)$. The natural follow-up question –and the main focus of this work– is to quantify the statistical-computational gap for *arbitrary* P, going beyond the criterion given by Eq.(3) and identifying necessary and sufficient conditions for efficient recovery.

1.3 Summary of Main Results

Our first main result establishes sample complexity lower bounds required by any SQ-algorithm to solve the single-index problem determined by P:

Theorem 1.2 (SQ lower bound, informal version of Theorem 3.2)**.** *Given $\mathsf{P} \in \mathcal{G}$ and n i.i.d. samples from $\mathbb{P}_{w^\star,\mathsf{P}}$, there exists an explicit exponent $k^\star = k^\star(\mathsf{P}) < \infty$ such that no polynomial time SQ-algorithm can succeed in recovering $w^\star$ unless $n \gtrsim d^{k^\star/2}$.*

Additionally, this SQ complexity coincides with the rate where the low-degree method succeeds:

Theorem 1.3 (Low-degree method detection lower bound, informal version of Theorem 3.5)**.** *Under the low-degree conjecture (Conjecture 3.4), if $k^\star = k^\star(\mathsf{P})$ is the exponent in Theorem 1.2, then given n i.i.d. samples from either $\mathbb{P}_{w^\star,\mathsf{P}}$ where $w^\star \sim \mathrm{Unif}(S^{d-1})$ or a null distribution $\mathbb{P}_0 := \gamma_d \otimes \mathsf{P}_y$, no polynomial time algorithm can distinguish $\mathbb{P}_{w^\star,\mathsf{P}}$ from $\mathbb{P}_0$ unless $n \gtrsim d^{k^\star/2}$.*

We denote the associated exponent $k^\star(\mathsf{P})$ as the *generative exponent* of the model, in contrast with the information exponent, and analyze its main properties in Section 2. In the meantime, the attentive reader might already anticipate that indeed one has $k^\star(\mathsf{P}) \leq l^\star(\mathsf{P})$.

Our second main result shows that these computational lower bounds are tight, by exhibiting a polynomial-time algorithm, based on the partial-trace estimator, that succeeds as soon as $n \gtrsim d^{k^\star/2}$:

Theorem 1.4 (informal version of Theorem 4.1 and Corollary 4.4)**.** *Given $\mathsf{P} \in \mathcal{G}$ and n i.i.d. samples from $\mathbb{P}_{w^\star,\mathsf{P}}$, there exists an efficient algorithm that succeeds in estimating $w^\star$ when $n \gtrsim d^{max(1,k^\star/2)}$, even when P is misspecified.*

Combined with an information-theoretic sample complexity upper bound $n = O(d)$ (Theorem 6.1) —which follows from relatively standard arguments, Theorem 1.4 thus establishes a sharp computational-to-statistical gap (under both the SQ and the LDP frameworks) as soon as $k^\star(\mathsf{P}) > 2$. Our last main contribution shows that for any k, there exists smooth distributions such that $k^\star(\mathsf{P}) = k$:

Theorem 1.5 (informal version of Theorem 5.1 and Theorem 5.2)**.** *For any $k \in \mathbb{N}$ and $\tau \geq 0$, there exists $\sigma \in C_b^\infty(\mathbb{R})$ such that $Z \sim \gamma$, $Y = \sigma(Z) + W$, with $W \sim N(0, \tau^2)$ defines a joint distribution $(Y, Z) \sim \mathsf{P}$ satisfying $k^\star(\mathsf{P}) = k$.*

Notations Given a probability measure μ defined over $\mathbb{R}^m$, we denote by $L^2(\mathbb{R}^m, \mu)$ the space of μ-measurable, square-integrable functions. For $f, g \in L^2(\mathbb{R}^m, \mu)$, we write $\langle f, g \rangle_\mu = \mathbb{E}_{X\sim\mu}[f(X)g(X)]$ and $\|f\|_\mu^2 = \langle f, f\rangle_\mu$. We use $\gamma = \gamma_1$ to denote the standard Gaussian measure $N(0, 1)$ and for $d \geq 1$ we use γ_d to denote the standard isotropic Gaussian measure in $\mathbb{R}^d$, $N(0, I_d)$.

Hermite Polynomials We define the normalized Hermite polynomials $\{h_k\}_{k\geq 0} \in L^2(\mathbb{R}, \gamma_1)$ by

$$h_k(z) := \frac{(-1)^k}{\sqrt{k!}} \frac{1}{\gamma_1(z)} \frac{\partial^k \gamma_1(z)}{\partial z^k} \ , \ z \in \mathbb{R}, \ k \in \mathbb{N} \ . \tag{4}$$

These polynomials satisfy the orthogonality relations $\mathbb{E}_{Z\sim\gamma_1}[h_j(Z)h_k(Z)] = \mathbf{1}_{j=k}$.

Acknowledgements: We thank Guy Bresler, Yatin Dandi, Ilias Diakonikolas, Daniel Hsu, Florent Krzakala, Theodor Misiakievicz, Tselil Schramm, Denny Wu, Ilias Zadik and Lenka Zdeborova for useful feedback during the completion of this work, which was partially developed during 2023's Summer School "Statistical Physics and ML back together again" in Cargese. AD acknowledges support from a NSF Graduate Research Fellowship. AD and JDL acknowledge support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, NSF CIF 2212262, ONR Young Investigator Award, and NSF CAREER Award 2144994. JB was partially supported by the Alfred P. Sloan Foundation and awards NSF RI-1816753, NSF CAREER CIF 1845360, NSF CHS-1901091 and NSF DMS-MoDL 2134216.

2 The Generative Exponent

Let us start by defining the information exponent of [BAGJ21, DH18] in our framework. We begin with Parseval's identity for Hermite polynomials. We define σ such that $\sigma(Z) := \mathbb{E}_\mathsf{P}[Y|Z]$ is the conditional expectation of Y given Z, thus $\sigma \in L^2(\mathbb{R}, \gamma)$ and we have

Fact 2.1 (Spectral Variance Decomposition)**.** *The variance of $\sigma(Z)$ verifies the expansion*

$$\mathrm{Var}_\mathsf{P}[\sigma(Z)] = \sum_{l\geq 1} \beta_l^2 \quad \textit{where} \quad \beta_l := \mathbb{E}_\mathsf{P}[Y h_l(Z)] \ . \tag{5}$$

Therefore if σ is not constant, there exists $l \geq 1$ such that $\beta_l \neq 0$. We define the information exponent $l^\star$ as the first such l:

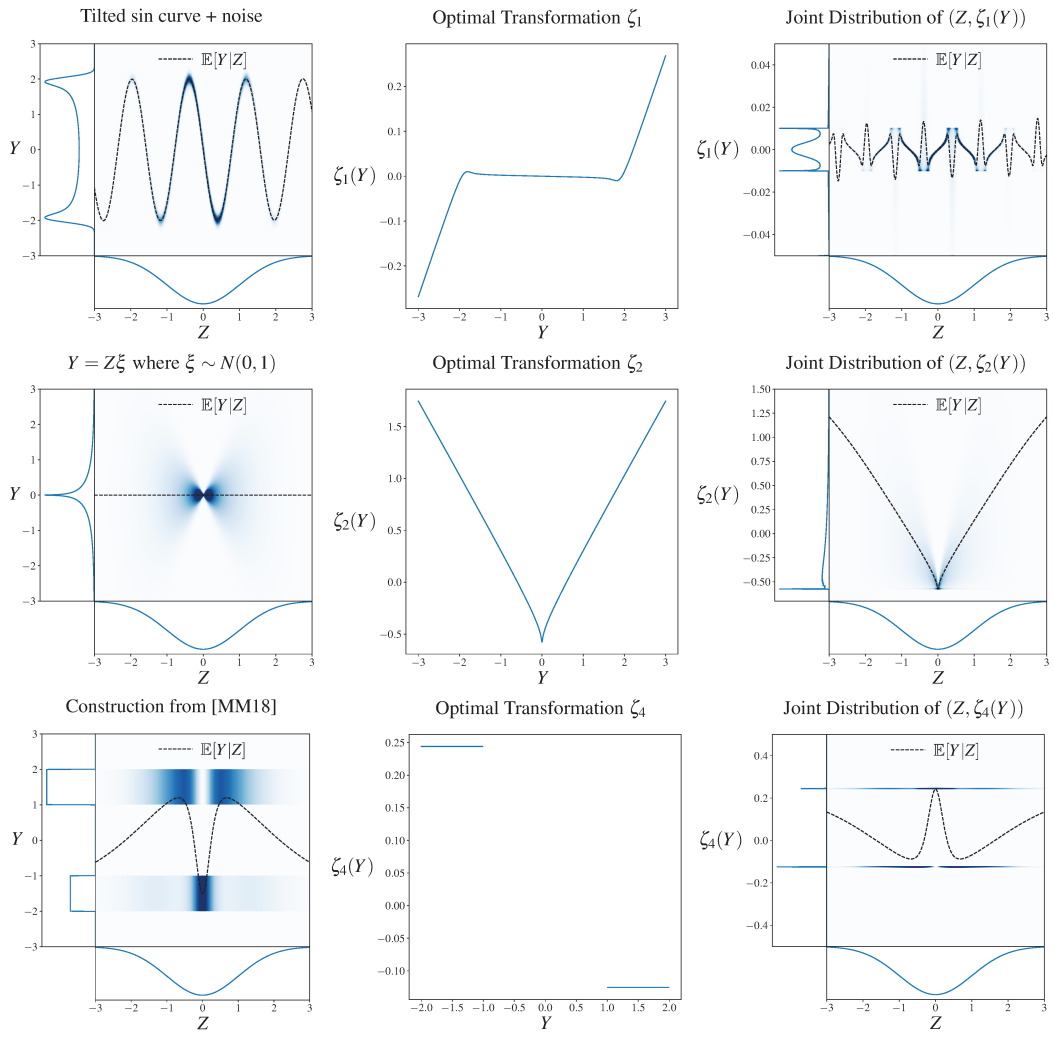


Figure 2: We plot three examples of a joint distribution P of (Z, Y), the *witness* function $\xi_{k^\star}(y)$, and the joint distribution of $(Z, \xi_k(Y))$. In the first example, $Y = c_1 x + \sin(c_2 Z)$ for constants c_1, c_2 such that $\beta_1 = 0$. The transformation ξ_1 zeros out the bulk and amplifies the caps of the curve in order to lower the information exponent from $l^\star(\mathsf{P}) = 2$ to $l^\star((\mathrm{Id} \otimes \zeta_1)_\# \mathsf{P}) = 1$. As a result, $k^\star(\mathsf{P}) = 1$. In the second example, the model $Y = Z\xi$ has multiplicative Gaussian noise and $E[Y|Z] = 0$ so this model has $l^\star(\mathsf{P}) = \infty$. The transformation ζ_2 interpolates between $\sqrt{|Y|}$ for $|Y| \approx 0$ and $|Y|$ for $|Y|$ farther from 0. The transformed distribution (right column) now has $l^\star((\mathrm{Id} \otimes \xi_k)_\# \mathsf{P}) = 2$ so $k^\star(\mathsf{P}) = 2$. The third example is the distribution used in [MM18] as an example where $k^\star > 2$. In this case, we verify that $k^\star = l^\star = 4$ so the tight sample complexity for the single index model corresponding to this choice of P is $n \gtrsim d^2$.

Definition 2.2 (Information Exponent revisited)**.** The *information exponent* of $\mathsf{P} \in \mathcal{G}$ is defined by:

$$l^\star(\mathsf{P}) := \min\{l \geq 1; \beta_l \neq 0\} \quad \text{where} \quad \beta_l := \mathbb{E}_{\mathsf{P}}[Y h_l(Z)]. \tag{6}$$

Note that Definition 2.2 only depends on P through the conditional expectation σ, and is therefore agnostic to any form of label noise. Below, we define another index, the *generative exponent*. As for the information exponent, its definition follows from an expansion, but for this exponent, we do it through the expansion of the χ^2-mutual information of $(Z, Y) \sim \mathsf{P}$, i.e. the χ^2-divergence between P and the product of its marginals $I_{\chi^2}[\mathsf{P}] := \mathbb{D}_{\chi^2}[\mathsf{P}||\mathsf{P}_z \otimes \mathsf{P}_y]$:

Lemma 2.3 (Mutual Information Decomposition)**.** *We have the following expansion*

$$I_{\chi^2}[\mathsf{P}] = \sum_{k \geq 1} \lambda_k^2 \quad \textit{where} \quad \lambda_k := \|\zeta_k\|_{\mathsf{P}_y} \quad \textit{and} \quad \zeta_k := \mathbb{E}[h_k(Z)|Y]. \tag{7}$$

The proof of this Lemma is postponed to Section D of the Appendix. We note that because $\mathbb{E}_{\mathsf{P}}[Y^2] < \infty$, the conditional expectation $\zeta_k := \mathbb{E}[h_k(Z)|Y]$ is well defined. In addition, $\zeta_k \in L^2(\mathbb{R}, \mathsf{P}_y)$ because for each k, $\|\zeta_k\|^2_{\mathsf{P}_y} \leq \mathbb{E}_Y \mathbb{E}_{Z|Y}[h_k(Z)^2] = \mathbb{E}_Z[h_k(Z)^2] = 1$. Therefore when P is not a product measure, as is required by Definition 1.1, we have that $I_{\chi^2}[\mathsf{P}] > 0$ so there exists $k \geq 1$ such that $\lambda_k \neq 0$. We define the generative exponent $k^\star$ as the first such k:

Definition 2.4 (Generative Exponent)**.** The generative exponent of $\mathsf{P} \in \mathcal{G}$ is defined by:

$$k^\star(\mathsf{P}) := \min\{k \geq 1; \lambda_k \neq 0\} \quad \text{where} \quad \lambda_k := \|\zeta_k\|_{\mathsf{P}_y} \quad \text{and} \quad \zeta_k := \mathbb{E}_{\mathsf{P}}[h_k(Z)|Y]. \tag{8}$$

Observe that β_k and ζ_k are related by $\beta_k = \langle y, \zeta_k \rangle_{\mathsf{P}_y}$. We can thus reinterpret the exponent $l^\star(\mathsf{P})$ as the smallest k such that ζ_k has non-zero correlation with linear functions in $L^2(\mathbb{R}, \mathsf{P}_y)$, capturing the correlational structure of CSQ queries. This also implies that the generative exponent is at most the Information exponent, i.e. $k^\star(\mathsf{P}) \leq l^\star(\mathsf{P})$ because $\lambda_k = 0 \Rightarrow \beta_k = 0$.

The function $\zeta_k(y)$ measures the k-th Hermite moment of the conditional 'generative' process $Z|Y = y$; as such, the fact that $\lambda_k > 0$ 'witnesses' χ^2-mutual information carried by order-k moments, and reciprocally $\lambda_k = 0$ indicates the absence of order-k exploitable information, even after conditioning on the observed labels. This is illustrated in Figure 2 and formalized by our SQ and low-degree lower bounds; see next section.

Remark 2.5. *Observe that it is possible to build distributions* P *such that* $k^\star(\mathsf{P}) < \infty$*, but* $l^\star(\mathsf{P}) = \infty$*, i.e. such that* $\mathbb{E}[Y|Z]$ *is constant. Consider for example* $\mathsf{P} = \varphi_\# \gamma_2$*, with* $\varphi(z, \xi) = (z, z\xi)$*. We have* $\mathbb{E}[Y|Z] = 0$*, hence* $l^\star(\mathsf{P}) = \infty$*. However,* $k^\star(\mathsf{P}) = 2$ *(see Figure 2). This is reflected by the fact that squaring the labels reduces the problem to noisy phase retrieval as* $\mathbb{E}[Y^2|Z = z] = z^2$*.*

Finally, we show that the generative exponent can be expressed as the smallest information exponent over all possible transformations of the label by squared-integrable functions:

Proposition 2.6 (A Variational Representation)**.** *The generative exponent* $k^\star(\mathsf{P})$ *can be written as:*

$$k^\star(\mathsf{P}) = \inf_{\mathcal{T} \in L^2(\mathsf{P}_y)} l^\star((\mathrm{Id} \otimes \mathcal{T})_\# \mathsf{P}) \ . \tag{9}$$

This provides a 'user-friendly' characterization of the generative exponent: for any polynomial Q of degree $k < k^\star(\mathsf{P})$ and any measurable test function $\mathcal{T} \in L^2(\mathbb{R}, \mathsf{P}_y)$, Proposition 2.6 tells that we must have $\mathbb{E}_{\mathsf{P}}[\mathcal{T}(Y)Q(Z)] = 0$. In addition, because $l^\star$ only depends on the conditional expectation $\mathbb{E}[\mathcal{T}(Y)|Z]$, this variational representation extends to non-deterministic channels, i.e. if $Y \to Y'$ is a Markov chain with $\mathbb{E}[(Y')^2] < \infty$ then $k^\star(Z, Y') \geq k^\star(Z, Y)$. In particular, no post-processing of Y, either random or deterministic, can reduce the generative exponent.

We conclude this section with some representatitve examples for deterministic models, illustrated in Figures 3 and 2.

Example 2.7 (Explicit Examples of generative exponent)**.** *We give the following explicit examples:*

(i) For σ a polynomial, we have $k^\star(\sigma) \leq 2$ and $k^\star(\sigma) = 2$ iff σ is even. In particular, $k^\star(h_j) = 1$ if j is odd and $k^\star(h_j) = 2$ if j is even.

(ii) For $\sigma(z) = z^2 e^{-z^2}$, we have $k^\star(\sigma) = 4$.

(iii) From [MM18, Remark 3], if $y \in \{-1, 1\}$ is boolean with $P(Y = 1|Z = z) = \mathbb{E}_{W\sim\gamma}[\tanh(c_1 z)^2 - \tanh(c_2 z)^2]$ for carefully chosen c_1, c_2 then $k^\star(\sigma) = 4$.

Finally, an immediate consequence of Proposition 2.6 is the following:

Corollary 2.8 (invariance to bijections)**.** *If $\varphi : \mathbb{R} \to \mathbb{R}$ is a bijection such that $\varphi, \varphi^{-1} \in L^2(\mathsf{P}_y)$, then $k^\star(\sigma) = k^\star(\varphi \circ \sigma)$.*

3 Computational Lower Bounds

3.1 From CSQ to SQ lower bounds

Recall that the CSQ complexity of Gaussian single-index models is established by considering the correlation $\mathbb{E}_X[\mathbb{E}[Y|X] \cdot \mathbb{E}[Y'|X]]$, where $(X, Y) \sim \mathbb{P}_w$ and $(X, Y') \sim \mathbb{P}_{w'}$ are two hypothesis in the class. This correlation admits a closed-form representation via the Ornstein-Ulhenbeck semigroup in $L^2(\mathbb{R}, \gamma_1)$ and the Hermite decomposition of $\sigma(z) = \mathbb{E}_{\mathsf{P}}[Y|Z = z] = \sum_k \beta_k h_k(z)$:

$$\mathbb{E}_X[\mathbb{E}[Y|X] \cdot \mathbb{E}[Y'|X]] = \sum_{k \geq l^\star} m^k \beta_k^2 \,, \quad \text{where } m = w \cdot w' \,. \tag{10}$$

For randomly chosen $w, w' \sim \mathrm{Unif}(S^{d-1})$, the correlation $m = w \cdot w'$ is of order $d^{-1/2}$ so the kth term in this expansion is of order $d^{-k/2}$. Therefore, the first nonzero term of this expansion, which is of order $d^{-l^\star/2}$, dominates the search problem. This is used in the proof of the CSQ lower bound that CSQ algorithms require $n \gtrsim d^{l^\star/2}$ samples to learn $w^\star$ [DLS22, ABAM22].

However, unlike the CSQ complexity, which is determined by average pairwise correlations, the SQ-complexity is determined by the χ^2 symmetrized divergence:

$$\chi_0^2(\mathbb{P}_w, \mathbb{P}_{w'}) := \mathbb{E}_{\mathbb{P}_0}\left[\frac{\mathrm{d}\mathbb{P}_w}{\mathrm{d}\mathbb{P}_0}\frac{\mathrm{d}\mathbb{P}_{w'}}{\mathrm{d}\mathbb{P}_0}\right] - 1 \quad \text{where} \quad \mathbb{P}_0 = \gamma_d \otimes \mathsf{P}_y \,. \tag{11}$$

Remarkably, one can exhibit an analog of (10), where the coefficients β_k are replaced by λ_k:

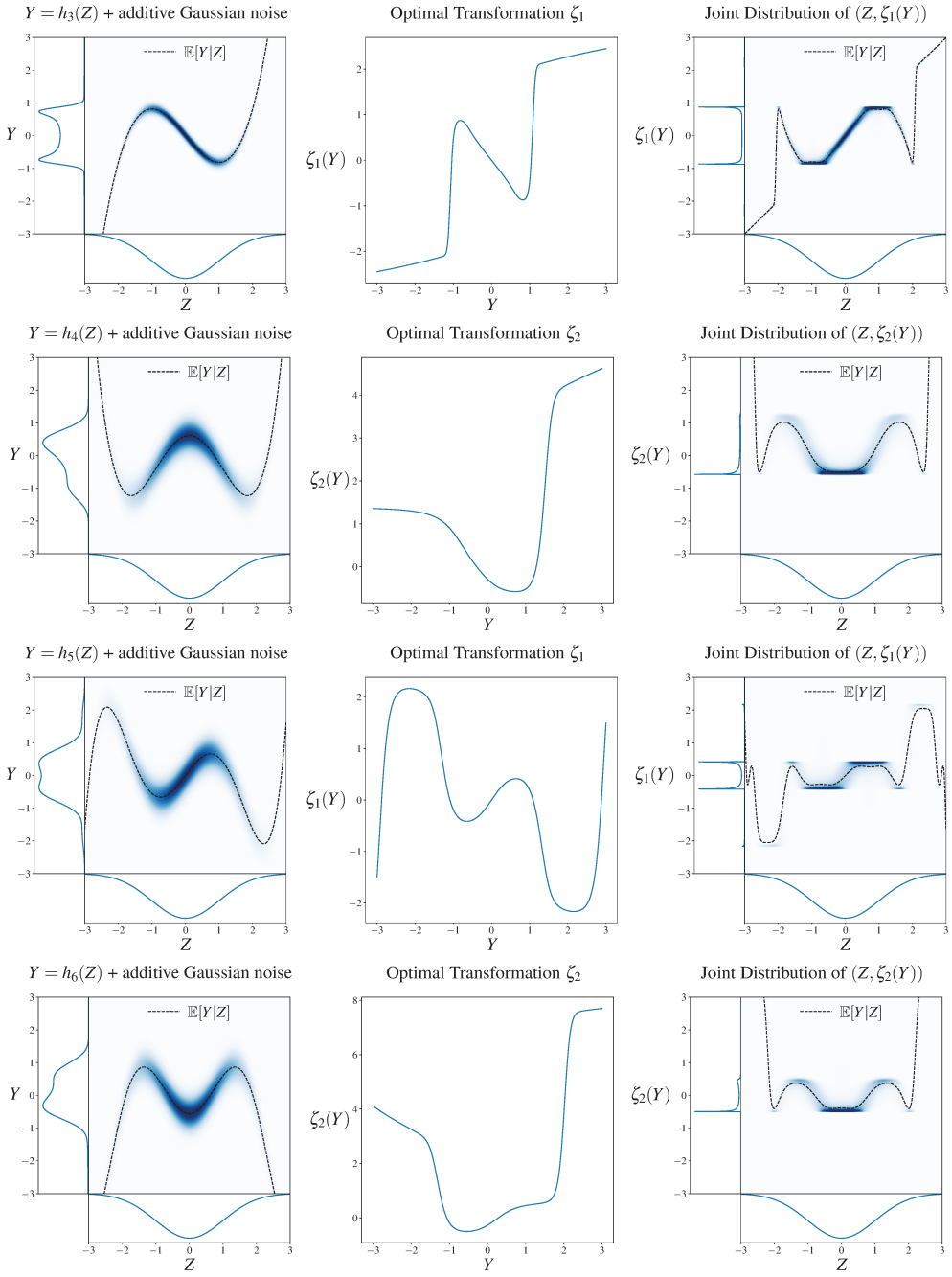


Figure 3: For every k, $Y = h_k(Z)$ has generative exponent 1 if k is odd and 2 if k is even. In particular, the difficulty of learning the single index model defined by $\mathsf{P} = (\mathrm{Id} \otimes h_k)_{\#}\gamma_1$ does not grow with k.

Lemma 3.1 (χ^2 Representation)**.** *Let $w, w' \in S^{d-1}$, and denote $m = w \cdot w'$, then we have*

$$\chi_0^2(\mathbb{P}_w, \mathbb{P}_{w'}) = \sum_{k \geq k^\star} m^k \lambda_k^2 \ . \tag{12}$$

The proof of this Lemma is postponed to Section E of the Appendix. As above, m is of order $d^{-1/2}$ so this sum is dominated by the first nonzero term in the sequence, which is given precisely by the generative exponent $k^\star = k^\star(\mathsf{P})$.

3.2 SQ Framework for Single-Index Models

Appendix B reviews the basic framework to establish SQ query complexity for search problems, which has been used for a variety of statistical inference tasks, e.g. [DKS17, DH21], and instantiates it for the particular setting of the single-index model; see Section B.2. We consider the statistical query oracle VSTAT(n) which responds to query $h : \mathbb{R}^d \times \mathbb{R} \to [0, 1]$ with $\hat{h}$ which satisfies: $|\hat{h} - p| \leq \sqrt{p(1-p)/n} + 1/n$, where $p = \mathbb{E}_{X,Y}[h(X, Y)]$.

The primary challenge in proving an SQ lower bound is computing the *SQ-dimension*; see Definition B.5. This is enabled in our setting thanks to the key Lemma 3.1. This result leads to a control of the relative χ^2-divergence of the form $\chi_0^2(\mathbb{P}_w, \mathbb{P}_{w'}) \leq \lambda_{k^\star}^2 m^{k^\star} + \frac{m^{k^\star}}{1-m}$, with $m = w \cdot w'$ (Lemma E.2), which yields the following SQ lower bound (proved in Appendix E.2):

Theorem 3.2 (Statistical Query Lower Bound)**.** *Assume that (X, Y) follow a Gaussian single-index model (Definition 1.1) with generative exponent $k^\star$. Let $q = d^r$ for any $r \leq d^{1/4}$ and assume that $\lambda_{k^\star}^2 \geq r d^{-1/2}$. Then there exists a constant $c_{k^\star}$ depending only on $k^\star$ such that to return $\hat{w}$ with $|\hat{w} \cdot w^\star| \geq \tilde{\omega}(d^{-1/2})$, any algorithm using q queries to* VSTAT(n) *requires:*

$$n \geq \frac{c_{k^\star}}{\lambda_{k^\star}^2} \left(\frac{d}{r^2} \right)^{\frac{k^\star}{2}} \ . \tag{13}$$

Therefore under the standard heuristic that VSTAT(n) measures the algorithmic complexity of an algorithm using n samples, any algorithm that uses poly(d) queries must use $n \gtrsim d^{k^\star/2}$ samples. A remarkable feature of eq. (13) is the absence of polylog factors; this is achieved thanks to an improved construction of the discrete hypothesis set $\mathcal{D}_D$ (Definition B.5) using spherical codes; see Lemma E.3.

Remark 3.3 (Choice of Null Model)**.** *In the proof of Theorem 3.2, we compare $\mathbb{P}_w$ with the null model $\mathbb{P}_0 := \gamma_d \otimes \mathsf{P}_y$. As a result, Theorem 3.2 primarily measures the detection threshold. Indeed, [DH21] observed a detection/estimation gap for the related problem of tensor PCA and proved that SQ algorithms require $n \gtrsim d^{\frac{k^\star}{2}+1}$ for estimation. We leave open the possibility that such a gap exists under the SQ framework for our setting as well. However, because Algorithm 1 succeeds in estimating $w^\star$ with $n \asymp d^{k^\star/2}$ (see Theorem 4.1) this would be an artifact of the SQ framework.*

3.3 The Low Degree Polynomial Method

In this section we prove a lower bound for the class of *low-degree polynomials*, which has been used to argue for the existence of statistical-computational gaps in a wide variety of problems [KWB19, BEAH+22, Wei23]. This has been formalized through various versions of the low degree

conjecture [Hop18, Hypothesis 2.1.5, Conjecture 2.2.4], for which we state an informal version for distinguishing between two distributions, $\mathbb{P}$ (signal) and $\mathbb{P}_0$ (null). We define the likelihood ratio $\mathcal{R} := \frac{d\mathbb{P}}{d\mathbb{P}_0}$ and the low-degree likelihood ratio $\mathcal{R}_{\le D}$ by the orthogonal projection in $L^2(\mathbb{P}_0)$ of $\mathcal{R}$ onto polynomials of degree at most D. The low-degree conjecture states that low-degree polynomials (i.e. $D \approx \log(n)$) act as a proxy for polynomial time algorithms:

Conjecture 3.4 (Low-Degree Conjecture, informal)**.** [1] *Let $D = \omega(\log(d))$. Then if $\|\mathcal{R}_{\le D}\|_{\mathbb{P}_0} = 1 + o_d(1)$, no polynomial time algorithm can distinguish $\mathbb{P}$ and $\mathbb{P}_0$. If $\|\mathcal{R}_{\le D}\|_{\mathbb{P}_0} = O_d(1)$, they can do so with only constant probability.*

By definition, we have that $\|\mathcal{R}\|^2_{L^2(\mathbb{P}_0)} = 1 + \mathbb{D}_{\chi^2}[\mathbb{P}||\mathbb{P}_0] = 1 + I_{\chi^2}[\mathsf{P}]$, which hints at the fact that the decomposition lemma 3.1 and its associated generative exponent will also be driving the behavior of the Low-Degree estimator. This is indeed the case: our main result in this section computes the low degree likelihood ratio $\|\mathcal{R}_{\le D}\|_{\mathbb{P}_0}$ and shows that it remains near 1 unless $n \gtrsim d^{k^\star/2}$:

Theorem 3.5 (Low-Degree Method Lower Bound)**.** *Let P be a single index model, let $\mathbb{P}$ be the distribution of $\mathbb{P}_w$ when w is chosen randomly from $\mathrm{Unif}(S^{d-1})$. Let $X \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^n$ denote a sequence of n i.i.d. inputs and targets drawn from P. Let $\mathbb{P}_0 := \gamma_d \otimes \mathsf{P}_y$ denote the null distribution. Let $\mathcal{R}(x, y)$ denote the likelihood ratio $\frac{d\mathbb{P}}{d\mathbb{P}_0}(x, y)$ and let $\mathcal{R}_{\le D}(x, y)$ denote the orthogonal projection in $L^2(\mathbb{P}_0)$ of $\mathcal{R}(x, y)$ onto polynomials of degree at most D in x. Then if $\frac{D}{\lambda_k^2} \ll \sqrt{d}$ and $\delta := \frac{n}{d^{k^\star/2}}$,*

$$\|\mathcal{R}_{\le D}\|^2_{\mathbb{P}_0} = (1 + o_d(1)) \sum_{j=0}^{\lfloor \frac{D}{k^\star} \rfloor} \mathbf{1}_{2|k^\star j}(k^\star j - 1)!! \frac{(\lambda_{k^\star}^2 \delta)^j}{j!}.$$

The proof is in Appendix E.3. This provides an immediate corollary which proves impossibility results for weak and strong recovery for all polynomial time algorithms under the low-degree conjecture (Conjecture 3.4):

Corollary 3.6. *Let $\mathbb{P}, \mathbb{P}_0, \mathcal{R}$ be as in Theorem 3.5 and assume $k^\star > 1$. Then,*

- ***Weak Detection:*** *If $D = \log(d)^2$ and $n \le d^\gamma$ for $\gamma < \frac{k^\star}{2}$, then $\|\mathcal{R}_{\le D}\|_{\mathbb{P}_0} = 1 + o_d(1)$.*
- ***Strong Detection:*** *For any $D \ll \sqrt{d}$, if $n \le \frac{d^{\frac{k^\star}{2}}}{D^{\frac{k^\star}{2}-1}}$, then $\|\mathcal{R}_{\le D}\|_{\mathbb{P}_0} = O_d(1)$.*

This shows that under Conjecture 3.4, $n \gtrsim d^{\frac{k^\star}{2}}$ samples are necessary for polynomials of degree $D = \log(d)^2$ to distinguish between $\mathbb{P}$ and $\mathbb{P}_0$. Note that because recovery is strictly harder than detection, this also implies that recovering $w^\star$ from $\mathbb{P}$ also requires $n \gtrsim d^{\frac{k^\star}{2}}$ samples, which is matched by our upper bound (Theorem 4.1). We also remark that the $\frac{d^{k^\star/2}}{D^{k^\star/2-1}}$ threshold in Corollary 3.6 matches the optimal known computational-statistical trade-off for tensor PCA [BGL17, WAM19].

3.4 Discussion

Relationship between SQ and LD lower bounds Our statistical query lower bound (Theorem 3.2) and our low-degree polynomial lower bound (Theorem 3.5 and Corollary 3.6) have similar

[1] The conjecture is stated for inference problems with some level of robustness to noise; this excludes known failures of LD methods to capture computational hardness in problems with algebraic structure; see [ZSWB22, DK22].

forms. The statistical query lower bound says that given d^r statistical queries or a polynomial of degree D (which can be computed in d^D time) we need:

$$n \gtrsim \frac{d^{k^\star/2}}{r^{k^\star}} \quad \text{or} \quad n \gtrsim \frac{d^{k^\star/2}}{D^{k^\star/2-1}}$$

samples respectively. A common theme in these lower bounds is the possibility of trading off statistics with computation. In particular, for $k^\star > 2$, given $n = \delta d^{k^\star/2}$ samples, $w^\star$ is always learnable given sufficiently many queries (p sufficiently large) or a sufficiently high degree polynomial (D sufficiently large). These bounds differ slightly in their dependence on r, D: the SQ bound depends on $r^{k^\star}$ while the low-degree bound depends on $D^{k^\star/2-1}$. We believe the low-degree result is tight, as it exactly matches existing statistical-computational tradeoffs that are known for tensor PCA [BGL17, WAM19]. This gap is due to the fact that our low-degree lower bound relies on *average* correlations over a prior, while the SQ lower bound relies on *worst case* correlations over a discrete set of points, which we constructed using spherical codes (see Lemma E.3).

Indeed, the proof of the low-degree lower bound (Theorem 3.5) suggests an algorithm for achieving this continuous statistical-computational tradeoff by computing higher order tensors than in Algorithm 1. Specifically, if p is a sufficiently large even integer, then we consider

$$T := \frac{1}{\binom{n}{p}} \sum_{S \subset [n], |S|=p} \mathrm{Sym}\left(\bigotimes_{i \in S} y_i \boldsymbol{h}_k(x_i) \right) \in \mathbb{R}^{d^{k^\star p}}.$$

We can now apply partial trace to T to reduce it to a matrix M and then return the top eigenvector of M. Note that by construction, $\mathbb{E}[T] = (w^\star)^{\otimes k^\star p}$ so $\mathbb{E}[M] = w^\star w^{\star T}$. We leave the analysis of M and the question of whether M obeys Gaussian universality (as in Section 4) to future work.

Finally, we mention the work [BBH+21], which provides generic 'translations' between SQ and low-degree lower bounds under mild conditions, at the expense of a loss in the parameters (number of queries to degree of polynomial). It would be interesting to quantify this loss in our setting.

NGCA *aka* 'Gaussian Pancakes' The SQ lower bounds that we present here for single index models are similar in essence to the SQ lower bounds for Gaussian hypothesis testing from [DKS17]. In short, given a non-gaussian univariate distribution $\mu \in \mathcal{P}(\mathbb{R})$, this problem considers distributions $\mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$ of the form $\mathbb{Q} = \mathbb{Q}_{\mu,w^\star} = R_\#(\gamma_{d-1} \otimes \mu)$, where $R \in \mathcal{O}_d$ as in Definition 1.1. The relevant quantity that governs the SQ-complexity of recovering the planted direction $w^\star$ (where again $w^\star = Re_d$) is the smallest degree k such that $\mathbb{E}_\mu[h_k(z)] \neq 0$. As such, there is a direct reduction from the single index model setting to NGCA:

Proposition 3.7 (Single-Index to NGCA Reduction)**.** *An SQ algorithm that solves NGCA with planted distribution μ of exponent k and direction $w^* \in S^{d-1}$ yields an efficient SQ algorithm to solve the single-index model with generative exponent k and direction w^*.*

As a result, our SQ-hardness results imply the hardness results in [DKS17]. An intriguing question is whether the reduction could go the other way, namely whether the SQ-hardness of Gaussian pancakes implies the SQ-hardness of learning certain single-index models. While we do not answer this question in the present paper, we show in Appendix E.5 a reduction from NGCA to a slight variant of the single-index model.

Tensor PCA Both our upper bound (Theorem 4.1) and lower bound (Theorem 3.2) were heavily inspired by related work in the tensor PCA literature. Specifically, the partial trace estimator in Algorithm 1 is related to the partial trace estimator for tensor PCA [HSSS16, DH24] which is the simplest method that achieves the conjectured optimal rate of $n = \Theta(d^{k/2})$ for tensor PCA. The derivation of our lower bounds were also inspired by similar results for tensor PCA, including the SQ lower bound for tensor PCA proved by [DH21], and the Low-Degree method lower bound of [KWB19, Theorem 3.3].

CLWE and periodic structures [SZB21] proved a lower bound for learning single-index models. They showed that under a cryptographic hardness assumption, there is a noisy single-index problem such that regardless of the samples n, no polynomial time algorithm can recover the ground truth $w^\star$. Our results imply a similar lower bound for the special case of SQ algorithms without the cryptographic hardness assumption (Lemma E.4). We also note that in the noise-free setting, the situation is different, and in fact there are (non-SQ) polynomial-time algorithms, such as LLL, that break the SQ-lower bound for certain single-index models and related NGCA [SZB21, ZSWB22, DK22] by exploiting periodic structures in the link function.

4 Partial Trace Estimator

Algorithm 1: Tensor Power Iteration with Partial Trace Warm Start

Input: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, moment k

Draw $n/2$ fresh samples $\mathcal{D}_0$ from $\mathcal{D}$

if *k is even* **then**

  Construct the partial trace matrix: $M_n \leftarrow \frac{1}{|\mathcal{D}_0|}\sum_{(x,y)\in\mathcal{D}_0} y\boldsymbol{h}_k(x)[I^{\otimes\frac{k-2}{2}}]$;

  $v \leftarrow v_1(M_n)$, the eigenvector corresponding to the top eigenvalue of M_n (in absolute value)

else if *k is odd* **then**

  Construct the partial trace vector: $v_n \leftarrow \frac{1}{|\mathcal{D}_0|}\sum_{(x,y)\in\mathcal{D}_1} y\boldsymbol{h}_k(x)[I^{\otimes\frac{k-1}{2}}]$;

  Normalize $v \leftarrow v_n/\|v_n\|$;

  if $k \geq 3$ **then**

    Run $\log(d)$ steps of tensor power iteration:

    for $i = 1, ..., \log(d)$ **do**

      Draw $n/2^{i+2}$ fresh samples $\mathcal{D}_i$ from $\mathcal{D}$

      $v \leftarrow \frac{1}{\mathcal{D}_i}\sum_{(x,y)\in\mathcal{D}_i} y\boldsymbol{h}_k(x)[v^{\otimes(k-1)}]$;

      $v \leftarrow v/\|v\|$;

    end

Run one final step of tensor power iteration:

  Draw $n/4$ fresh samples $\mathcal{D}_{\text{fin}}$ from $\mathcal{D}$

  $v \leftarrow \frac{1}{\mathcal{D}_{\text{fin}}}\sum_{(x,y)\in\mathcal{D}_{\text{fin}}} y\boldsymbol{h}_k(x)[v^{\otimes(k-1)}]$;

  $v \leftarrow v/\|v\|$;

Output: v

We have shown so far that Theorems 3.2 and 3.5 extend the 'necessary' criterion presented in Eq. (3) from [MM18, BKM+19, MLKZ20] to arbitrary exponents $k^\star > 2$, as indeed a simple calculation shows that $\mathbb{E}[\mathbb{E}[Z^2 - 1|Y]^2] = 2\lambda_2^2$. Concerning the 'sufficient' direction of Eq.(3), the

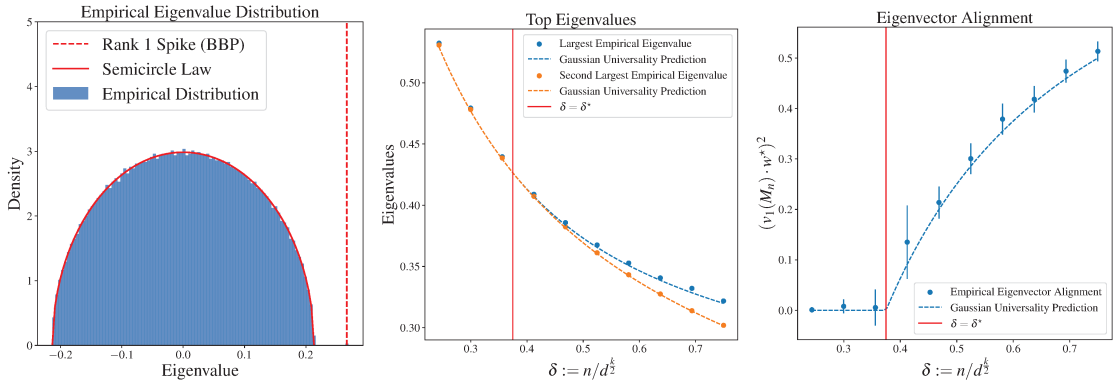


Figure 4: We empirically verify our observation in Lemma F.9 that M_n satisfies Gaussian universality for $k^\star > 2$. We take $k^\star = 4$ and compute M_n for $d = 4096$ and varying $\delta = \frac{n}{d^2}$. The target function is $Y = \zeta_4(Z^2 e^{-Z^2})$ (see Figure 2). Here $\delta^\star = \frac{\mathbb{E}[Y^2]}{\beta_k^2 k(k-1)!!}$ is the predicted BBP threshold [BBAP05]. The dots represent the medians over 10 random seeds of each quantity (eigenvalues and eigenvector correlation) and the error bars represent the standard deviation over these 10 trials.

algorithms that match the previous lower bound for $k^\star \le 2$ suggest a meta-strategy for the general case $k^\star > 2$: first apply a transformation $\mathcal{T}$ of the labels to reduce to correlational queries, and then apply an optimal CSQ algorithm. Moreover, Theorem 3.5 further illuminates the matter, since it identifies the optimal degree-D test.

The definition of the generative exponent reveals that the desired pre-processing is precisely $\mathcal{T}(y) = \zeta_{k^\star}(y)$. We can then combine this pre-processing with an order-$k^\star$ optimal CSQ algorithm, based on the partial-trace algorithm [HSSS16, FD23, DNGL23], described in Algorithm 1. As shown in Appendix F, this algorithm achieves the promised optimal sample complexity:

Theorem 4.1 (Algorithm 1 Statistical Guarantee)**.** *Let $\{(x_i, y_i)\}_{i=1}^n$ be i.i.d. samples from $\mathbb{P}_{w^\star, \mathtt{P}}$. For any $\epsilon > 0$ and $k \ge 1$, there exists a constant C_k depending only on k and a denoiser $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ such that if $\tilde{\lambda}_k = \lambda_k / \log^k(3/\lambda_k)$ and $n \ge C_k[d^{k/2}/\tilde{\lambda}_k^2 + d/(\tilde{\lambda}_k^2 \epsilon^2)]$, Algorithm 1 applied to samples $\{(x_i, \mathcal{T}(y_i))\}_{i=1}^n$ returns a vector $\hat{w}$ with $(\hat{w} \cdot w^\star)^2 \ge 1 - \epsilon^2$ with probability greater than $1 - 2e^{-d^c}$ for a constant $c = c(k)$ depending only on k.*

The partial trace algorithm has been analysed in the context of Tensor PCA [HSSS16, FD23], and provides efficient recovery of the planted direction up to the computational threshold [DH21]. At the technical level, the primary challenge we face is that the errors in Algorithm 1 are heavy tailed and non-Gaussian. Explicitly, while it is possible to reduce learning a single index model to a form of tensor PCA, the resulting noise matrix has highly correlated entries with heavy tails. For example, while the individual entries of the resulting noise tensor are of order $n^{-1/2}$, as for tensor PCA, the operator norm of this tensor is of order $d^{k/2}$ rather than $d^{1/2}$ for tensor PCA.

Characterizing the precise weak recovery threshold A remarkable consequence of our analysis is that, when $k^\star > 2$ is even, after applying the partial trace operation to reduce this tensor to a matrix, this matrix obeys Gaussian universality [BvH23], and in particular, the spectrum of the partial trace matrix M_n converges to a semicircle law. Explicitly, this means that the spectrum of M_n

is close to the spectrum of the Gaussian matrix G with $\mathbb{E}[G] = \mathbb{E}[M_n]$ and $\mathbb{E}[G \otimes G] = \mathbb{E}[M_n \otimes M_n]$. Therefore to understand the spectral properties of M_n, it suffices to compute the the covariance of M_n, which we do in Lemma F.10. From Lemma F.10, we see that when $k^\star \geq 4$, the Gaussian equivalent matrix G can be approximated by another Gaussian matrix $G^\star$ (in the space probability space) which satisfies:

$$G^\star \stackrel{d}{=} \beta_{k^\star} w^\star w^{\star T} + \sqrt{\frac{\mathbb{E}[\mathcal{T}(Y)^2] d^{k^\star/2}}{n k^\star (k^\star - 1)!!}} \cdot W, \quad \|G - G^\star\|_2 \lesssim \sqrt{\frac{d^{\frac{k^\star - 1}{2}}}{n}}, \tag{14}$$

where $\beta_{k^\star} = \mathbb{E}[\mathcal{T}(Y) h_{k^\star}(Z)]$ is the signal strength after applying the appropriate label transformation, W is a standard Wigner matrix, and the bound on $\|G - G^\star\|_2$ holds with high probability. We can therefore leverage existing results for spiked Wigner matrices to get exact constants for weak recovery when considering this partial trace estimator. With high probability, the spectrum of $G^\star - \mathbb{E}G^\star$ is contained in the interval $[-2R(1 + o_d(1)), 2R(1 + o_d(1))]$ where

$$R^2 := \frac{\mathbb{E}[\mathcal{T}(Y)^2]}{k^\star (k^\star - 1)!!} \cdot \frac{d^{k^\star/2}}{n}. \tag{15}$$

In addition, the spectrum of G exhibits a BBP transition [BBAP05], i.e. there is an outlier eigenvalue at $(\beta_{k^\star} + R^2/\beta_{k^\star})(1 + o_d(1))$. It is also known that for $\beta_{k^\star} \geq R$, $(v_1(G^\star) \cdot w^\star)^2 \to 1 - (R/\beta_{k^\star})^2$ where $v_1(G^\star)$ is the eigenvector of $G^\star$ corresponding to the largest eigenvalue (in absolute value). However, as we only prove universality of the spectrum of M_n (and not of its eigenvectors), this does not directly imply that $v_1(M_n) \cdot w^\star$ has the same behavior. Nevertheless, we empirically verify this property for M_n in Figure 4. The combination of Gaussian universality with the BBP transition implies there is a phase transition for weak recovery of $w^\star$ at the cutoff:

$$\delta := \frac{n}{d^{k^\star/2}} = \frac{\mathbb{E}[Y^2]}{\beta_{k^\star}^2 k^\star (k^\star - 1)!!} =: \delta^\star, \tag{16}$$

However, unlike for phase retrieval, in which the spectral estimator gives a tight threshold for estimation [MM18], this constant can be improved by increasing the order of the tensor before applying partial trace (see Section 3.4). This is consistent with the known statistical-computational tradeoffs for tensor PCA ([BGL17, WAM19]). We note also that the Gaussian universality does *not* hold for $k^\star = 2$ in which R and σ^2 are both order d/n, corresponding to the setting studied in [MM18].

Remark 4.2 (Memory and Runtime of Algorithm 1)**.** *It is possible to implement Algorithm 1 with $O(d)$ memory (not counting the memory to store the dataset), and runtime $\tilde{O}(d^{\frac{k}{2}+1})$. This requires using Lemma F.3 and Lemma F.8 to avoid explicitly computing the Hermite tensor $\boldsymbol{h}_k(x)$.*

Extension to unknown P The only instance where we used the knowledge of P in the previous algorithm is in Lemma F.2, where a suitable thresholding $\mathcal{T}$ is applied to the labels. This guarantees that $\eta = \mathbb{E}[\mathcal{T}(Y) h_k(Z)] \neq 0$, leading to the recovery of $w^\star$ in the high-dimensional regime. However, it is sufficient to consider a label transformation $\tilde{\mathcal{T}}$ that has *non-negligible correlation* with ζ_k. For instance, this can be guaranteed in a setting where P is only known to belong to a certain non-parametric class of distributions, as we now illustrate. Given $\mathsf{P} \in \mathcal{G}$, let $\{\phi_k\}_k$ denote the orthogonal polynomial basis of $L^2(\mathbb{R}, \mathsf{P}_y)$. We now decompose $\zeta_{k^\star}$ in this basis: $\zeta_{k^\star} = \sum_l v_l \phi_l$.

Assumption 4.3 (Source Condition). *For $m \geq 1$, define $\varepsilon_m := \lambda_{k^\star}^{-2} \sum_{l \geq m} v_l^2 \leq 1$, then there exists N such that $\varepsilon_m < 1$ for $m \geq N$.*

This assumption is mild as it only prevents extreme cases for which the first harmonics $(v_l)_{l \leq N}$ have infinitesimally small mass. Note that we choose a polynomial basis only for convenience[2]. We can now build $\tilde{\mathcal{T}}$ as a truncated random polynomial spanned by the first M terms; Lemma F.14 shows that such a random choice already provides a label transformation $\tilde{\mathcal{T}}$ with non-negligible correlation $\tilde{\eta} = \mathbb{E}_{\mathsf{P}}[\tilde{\mathcal{T}}(Y) h_{k^\star}(Z)]$, with high probability over the draw of the random polynomial.

Plugging this lemma directly in the proof of Theorem 4.1 and via a union-bound, we obtain an analogous guarantee for the recovery of $w^\star$, where the price of not knowing P only manifests itself in the worsening of the constant, from η to $\tilde{\eta}$. We can now go one step further, and ignore $k^\star$ altogether. This may be achieved using the technique from [DH18, Algorithm 2], that tries all possible $k \in \{1, K\}$ and outputs the direction yielding the best goodness-of-fit on a held-out dataset of size L. The overall procedure, described in Algorithm 2 in Appendix F.8, enjoys the following guarantees:

Corollary 4.4 (Partial Trace for unknown P and $k^\star$). *Let $\{(x_i, y_i)\}_{i=1}^n$ be i.i.d. samples from $\mathbb{P}_{w^\star, \mathsf{P}}$ with $k^\star(\mathsf{P}) = k^\star$. Then, under Assumption 4.3, if $n \geq \Omega(\lambda_{k^\star}^2 d^{k^\star/2} + d/\epsilon^2)$, $L \gtrsim \delta^{-4} \log(1/\tilde{\delta})$, the procedure described in Algorithm 2 with $K = k^\star$ returns $\hat{w}$ satisfying $(\hat{w} \cdot w^\star)^2 \geq 1 - \epsilon^2$ with probability greater than $1 - e^{-d^\kappa} - \delta - 2\tilde{\delta} k^\star$.*

5 Existence of Smooth Distributions for any generative exponent $k^\star$

Now that we have identified the precise sample and query complexity of the single index model in terms of the exponent $k^\star(\mathsf{P})$, we turn to the question of characterizing the class $\{\mathsf{P}; k^\star(\mathsf{P}) = k\}$ for any k. We focus our attention on the additive gaussian noisy setting, namely $(Z, Y) \sim \mathsf{P}$ satisfy $Y = \sigma(Z) + \tau\xi$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is a link function, and $\xi \sim \gamma_1$ is independent of Z. When $\tau = 0$ we recover the deterministic case.

In this context, the Information exponent provides a transparent description in terms of the Hermite decomposition of σ; in particular for each k one can easily construct analytic functions such that $l^\star(\mathsf{P}) = l^\star(\sigma) = k$ (e.g. the degree-k Hermite polynomial). Such structure is absent in the generative exponent setting: observe that in virtue of Lemma 2.6, the exponent only depends on the set of level sets $\{\{u \in \mathbb{R}; \sigma(u) = t\}\}_t$, which does not easily lend itself to harmonic analysis.

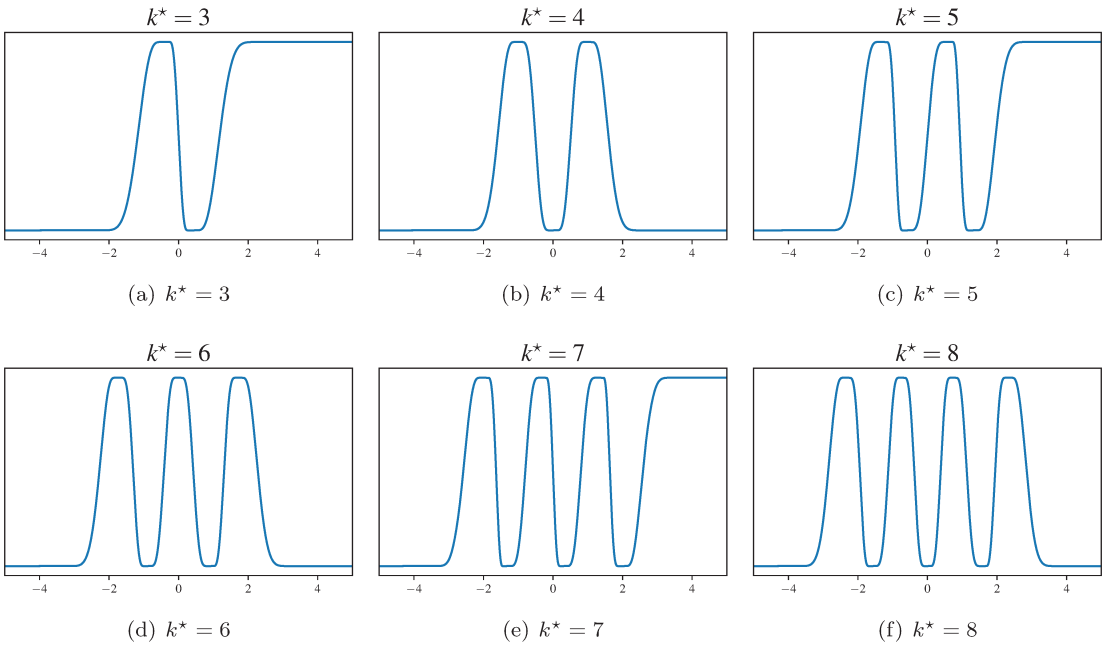
A simple example of link function with $k^\star > 2$ is $\sigma(x) = x^2 e^{-x^2}$. For this σ, $k^\star = 4$ which implies the single index model determined by σ is unlearnable in polynomial time without $n \gtrsim d^2$ samples.[3] However, this construction does not easily generalize to higher $k^\star$. Nonetheless, we are able to establish the existence of smooth link functions with prescribed generative exponent:

Theorem 5.1 (Smooth Single-Index models with prescribed $k^\star$). *For each k, there exists $\sigma \in C_b^\infty(\mathbb{R})$ such that the deterministic single index model $\mathsf{P} = (\mathrm{Id} \otimes \sigma)_\# \gamma_1$ satisfies $k^\star(\mathsf{P}) = k$.*

[2] We note that from observed data $\{(x_i, y_i)\}_{i \leq n}$, one could in particular estimate the marginal P_y using a (scalar) non-parametric kernel density estimator, and therefore estimate the first terms of the orthogonal polynomial basis. For simplicity, and w.l.o.g., we will assume that such basis elements are available.

[3] Examples of σ with $k^\star > 2$ were discovered in prior works as well, e.g. [MM18, Remark 3]. See Figure 2 for a figure of this construction along with the corresponding ζ_4.

The main idea of the proof, presented in Appendix G.1, is to build the link function via the coarea formula, by evolving a one-parameter family of level sets $\{S_t := \sigma^{-1}(t)\}_t$. Each level set $S_t = \{z_{1,t}, \dots, z_{n_t,t}\}$ needs to be such that $\mathbb{E}[h_k(z)|z \in \sigma^{-1}(t)] = 0$, which becomes a polynomial equation in the points $z_{j,t}$. We first determine a suitable form for S_0, based on Hermite-Gauss quadrature, and then obtain S_t, $t \in [0, T)$ as the solution of an ODE that enforces that this condition is preserved through 'time' using a Vandermonde-type kernel. See Figure 5 for examples of link functions σ with generative exponents $k^\star = 3, ..., 8$.



(a) $k^\star = 3$ (b) $k^\star = 4$ (c) $k^\star = 5$

(d) $k^\star = 6$ (e) $k^\star = 7$ (f) $k^\star = 8$

Figure 5: Explicit constructions of σ with different prescribed generative exponents. These were generated by numerically integrating the ODE in eq. (181).

Finally, we establish that adding Gaussian noise to the labels preserves the generative exponent:

Theorem 5.2 (Additive Gaussian noise preserves generative exponent)**.** *For $\tau \geq 0$ and $\sigma : \mathbb{R} \to \mathbb{R}$, we denote $\Phi_{\tau,\sigma}(u, v) = (u, \sigma(u) + \tau v) \in \mathbb{R}^2$. Then the additive noisy model $\tilde{\mathsf{P}} = (\Phi_{\tau,\sigma})_\# \gamma_2$ satisfies $k^\star(\tilde{\mathsf{P}}) = k^\star(\mathsf{P})$, where $\mathsf{P} = (\mathrm{Id} \otimes \sigma)_\# \gamma_1$.*

The proof of Theorem 5.2 reveals that non-Gaussian noise distributions μ also yield the same conclusion, provided the characteristic function $\varphi(\xi) = \mathbb{E}_{Z\sim\mu}[e^{-i\xi Z}]$ satisfies $\varphi(\xi) \neq 0$ for all ξ. Under these conditions, like the information exponent, the generative exponent is 'oblivious' to additive noise (albeit with potentially smaller signal strength $\tilde{\lambda}_{k^\star}$).

6 Information-Theoretic Sample-Complexity

We conclude the analysis of the single-index model by obtaining an upper bound for the sample complexity, irrespective of the estimation procedure. This question has been addressed in several planted problems of similar structure [MM18, DH21, MR14], providing the groundwork needed to establish the following result:

Theorem 6.1 (Information-Theoretic Upper Bound)**.** *For all $k \geq 1$, there exists a procedure which returns $\hat{w}$ with $(w \cdot w^\star)^2 \geq 1 - \epsilon^2$ with probability at least $1 - 2e^{-d}$ if $n = \tilde{\Theta}\left(\frac{d}{\lambda_k^2 \epsilon^2}\right)$.*

In essence, the (inefficient) estimator optimizes a certain correlation over the sphere, and uses a 'corrector' step from [DH21, Theorem 5] to yield a tight dependence on ϵ. Combined with the SQ and LDP-lower bounds, these results thus establish a sharp computation-to-statistical gap (under both the SQ and LDP frameworks) for the single-index problem as soon as $k^\star(\mathsf{P}) > 2$.

7 Conclusions

In this work, we have explored the question of efficiently estimating the planted structure from data drawn from a single-index model. We identified the generative exponent $k^\star(\mathsf{P})$ as the fundamental quantity driving the sample complexity required for recovering the hidden direction, and proved tight matching upper and lower bounds, using the partial trace algorithm and the SQ and LDP frameworks respectively. Taken together, our results give a unified perspective on a variety of planted high-dimensional problems, and provide evidence of a tight computational-to-statistical gap as soon as $k^\star(\mathsf{P}) > 2$.

That said, there are nuances that future work should aim to address: On one hand, our partial trace algorithm includes power iterations, which are not *bona-fide* SQ, making the fine-grained comparison between our upper and SQ-lower bounds somewhat murky [DH21]. On the other hand, our Low-Degree lower bound concerns the detection problem (that is, testing $\mathbb{P}_w$ vs $\mathbb{P}_0$ for some w), while the natural setting for us is the recovery problem. Although this mismatch does not impact the tightness of the rate $d^{k^\star/2}$, in other high-dimensional inference problems it reveals more fine-grained information, e.g. [PWBM18], such as sharp recovery thresholds.

One natural extension of our work is to the *multi-index* setting, in which the labels depend on a projection of the input onto a subspace of dimension $r \geq 1$. A multi-index model can be described by an orthogonal matrix $W^\star \in \mathbb{R}^{r \times d}$ and a joint distribution $\mathsf{P} \in \mathcal{P}(\mathbb{R}^r \times \mathbb{R})$ over (Z, Y) where $z = W^\star x$. The goal in the multi-index setting is to recover the subspace defined by $W^\star$. This already gives rise to rich structure not present in the single-index setting. For example, the natural generalization of the information exponent [AGJ21] is the *leap complexity* [ABAM23, DTA+24, BBPV23]. However, it remains unclear what the natural generalization of the generative exponent is, even for $k^\star = 2$.

References

[ABAM22] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. *arXiv preprint arXiv:2202.08658*, 2022. 9

[ABAM23] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics, 2023. 4, 19

[AGJ21] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *J. Mach. Learn. Res.*, 22:106–1, 2021. 19

[AHSS17] Alexandr Andoni, Daniel Hsu, Kevin Shi, and Xiaorui Sun. Correspondence retrieval. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 105–126. PMLR, 07–10 Jul 2017. 4

[BAGJ21] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research (JMLR)*, 22:106–1, 2021. 4, 5, 6

[BBAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, pages 1643–1697, 2005. 15, 16

[BBH+21] M Brennan, G Bresler, S Hopkins, J Li, and T Schramm. Statistical query algorithms and low-degree tests are almost equivalent. In *Conference on Learning Theory*, 2021. 13

[BBPV23] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023. 19

[BBSS22] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *arXiv preprint arXiv:2210.15651*, 2022. 4, 50

[BEAH+22] Afonso S Bandeira, Ahmed El Alaoui, Samuel Hopkins, Tselil Schramm, Alexander S Wein, and Ilias Zadik. The franz-parisi criterion and computational trade-offs in high dimensional statistics. *Advances in Neural Information Processing Systems*, 35:33831–33844, 2022. 11

[BGL17] Vijay Bhattiprolu, Venkatesan Guruswami, and Euiwoong Lee. Sum-of-squares certificates for maxima of random tensors on the sphere, 2017. 12, 13, 16

[BKM+19] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, March 2019. 4, 5, 14

[BvH23] Tatiana Brailovskaya and Ramon van Handel. Universality and sharp matrix concentration inequalities, 2023. 15, 61

[CBL+20] Minshuo Chen, Yu Bai, Jason D Lee, Tuo Zhao, Huan Wang, Caiming Xiong, and Richard Socher. Towards understanding hierarchical learning: Benefits of neural representations. *Advances in Neural Information Processing Systems*, 2020. 4

[CM20] Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In *Conference on Learning Theory*, pages 1161–1227. PMLR, 2020. 5

[DH18] Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1887–1930. PMLR, 06–09 Jul 2018. 4, 6, 17, 51

[DH21] Rishabh Dudeja and Daniel Hsu. Statistical query lower bounds for tensor pca, 2021. 11, 14, 15, 19, 59

[DH24] Rishabh Dudeja and Daniel Hsu. Statistical-computational trade-offs in tensor pca and related problems via communication complexity, 2024. 14

[DJS08] Arnak S Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *The Journal of Machine Learning Research*, 9:1647–1678, 2008. 4

[DK22] Ilias Diakonikolas and Daniel Kane. Non-gaussian component analysis via lattice basis reduction. In *Conference on Learning Theory*, pages 4535–4547. PMLR, 2022. 12, 14

[DKL+23] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time, 2023. 4

[DKS17] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017. 3, 11, 13

[DLS22] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, 2022. 4, 9

[DNGL23] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *arXiv preprint arXiv:2305.10633*, 2023. 4, 15, 35, 60, 61

[DR07] P.J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Dover Books on Mathematics Series. Dover Publications, 2007. 55

[DTA+24] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborov'a, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *arXiv preprint arXiv:2402.03220*, 2024. 5, 19

[FD23] Michael Jacob Feldman and David Donoho. Sharp recovery thresholds of tensor pca spectral algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 15

[FGR+17] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S. Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *J. ACM*, 64(2), 2017. 24, 25

[HJS01] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics*, 29(3):595–623, 2001. 4

[HMSW04] Wolfgang Härdle, Marlene Müller, Stefan Sperlich, and Axel Werwatz. *Nonparametric and semiparametric models*, volume 1. Springer, 2004. 4

[Hop18] Samuel Hopkins. *Statistical inference and the sum of squares method.* PhD thesis, Cornell University, 2018. 3, 12

[HSSS16] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191, 2016. 14, 15

[Ich93] Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of econometrics*, 58(1-2):71–120, 1993. 4

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998. 3

[KKSK11] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011. 4

[KS09] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009. 4

[KWB19] Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019. 3, 11, 14

[MLKZ20] Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions, 2020. 4, 5, 14

[MM18] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval, 2018. 4, 5, 7, 9, 14, 16, 17, 19, 28, 47

[MN83] P. McCullagh and J.A. Nelder. *Generalized Linear Models.* Monographs on Statistics and Applied Probability. Springer US, 1983. 4

[MR14] Andrea Montanari and Emile Richard. A statistical model for tensor pca, 2014. 3, 19

[NNW12] Jelani Nelson, Huy Nguyen, and David P. Woodruff. On deterministic sketching and streaming for sparse recovery and norm estimation, 2012. 31

[PWBM18] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018. 19

[SZB21] Min Jae Song, Ilias Zadik, and Joan Bruna. On the cryptographic hardness of learning single periodic neurons. *Advances in Neural Processing Systems (NeurIPS)*, 2021. 4, 14

[WAM19] Alexander S. Wein, Ahmed El Alaoui, and Cristopher Moore. The kikuchi hierarchy and tensor pca, 2019. 12, 13, 16

[Wei23] Alexander S Wein. Average-case complexity of tensor decomposition for low-degree polynomials. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1685–1698, 2023. 11

[ZSWB22] Ilias Zadik, Min Jae Song, Alexander S Wein, and Joan Bruna. Lattice-based methods surpass sum-of-squares in clustering. In *Conference on Learning Theory*, pages 1247–1248. PMLR, 2022. 12, 14

A Additional Notation

A.1 Tensor Notation

Throughout this section let $T \in (\mathbb{R}^d)^{\otimes k}$ be a k-tensor.

Definition A.1 (Tensor Action)**.** For a j tensor $A \in (\mathbb{R}^d)^{\otimes j}$ with $j \leq k$, we define the action $T[A]$ of T on A by

$$(T[A])_{i_1,\ldots,i_{k-j}} := \sum_{i_{k-j+1},\ldots,i_k=1}^{d} T_{i_1,\ldots,i_k} A^{i_{k-j+1},\ldots,i_k} \in (\mathbb{R}^d)^{\otimes(k-j)}. \tag{17}$$

We will also use $\langle T, A\rangle$ to denote $T[A] = A[T]$ when A, T are both k tensors. Note that this corresponds to the standard dot product after flattening A, T.

Definition A.2 (Permutation/Transposition)**.** Given a k-tensor T and a permutation $\pi \in S_k$, we use $\pi(T)$ to denote the result of permuting the axes of T by the permutation π, i.e.

$$\pi(T)_{i_1,\ldots,i_k} := T_{i_{\pi(1)},\ldots,i_{\pi(k)}}. \tag{18}$$

Definition A.3 (Symmetrization)**.** We define $\mathrm{Sym}_k \in (\mathbb{R}^d)^{\otimes 2k}$ by

$$(\mathrm{Sym}_k)_{i_1,\ldots,i_k,j_1,\ldots,j_k} = \frac{1}{k!} \sum_{\pi \in S_k} \delta_{i_{\pi(1)},j_1} \cdots \delta_{i_{\pi(k)},j_k} \tag{19}$$

where S_k is the symmetric group on $1,\ldots,k$. Note that Sym_k acts on k tensors T by

$$(\mathrm{Sym}_k[T])_{i_1,\ldots,i_k} = \frac{1}{k!} \sum_{\pi \in S_k} \pi(T). \tag{20}$$

i.e. $\mathrm{Sym}_k[T]$ is the symmetrized version of T.

We will also overload notation and use Sym to denote the symmetrization operator, i.e. if T is a k-tensor, $\mathrm{Sym}(T) := \mathrm{Sym}_k[T]$.

Lemma A.4. *For any tensor T,*

$$\|\mathrm{Sym}(T)\|_F \leq \|T\|_F. \tag{21}$$

Proof.

$$\|\mathrm{Sym}(T)\|_F = \left\| \frac{1}{k!} \sum_{\pi \in S_k} \pi(T) \right\|_F \leq \frac{1}{k!} \sum_{\pi \in S_k} \|\pi(T)\|_F = \|T\|_F \tag{22}$$

because permuting the indices of T does not change the Frobenius norm. ∎

B SQ Framework for Search Problems

The relevant framework is developed in [FGR+17]. We provide a quick recap of the relevant definitions.

B.1 Main Ingredients

Definition B.1 (Search Problem over Distributions)**.** Let X be a domain, $\mathcal{D}$ be a set of distributions over X, $\mathcal{F}$ be a set of solutions, and $\mathcal{Z} : \mathcal{D} \to 2^{\mathcal{F}}$ be a map to the set of valid solutions. The distributional search problem is to find a valid solution $f \in \mathcal{Z}(D)$ given oracle access to samples from an unknown $D \in \mathcal{D}$. We will also use $\mathcal{Z}_f$ to denote the set of distributions $\mathcal{D}$ for which f is a valid solution.

Definition B.2 (STAT oracle)**.** Let $D \in \mathcal{D}$ be the unknown distribution. Given a tolerance τ and a query $h : X \to [-1, 1]$, the STAT(τ) oracle returns a value within τ of $\mathbb{E}_{x\sim D}[h(x)]$.

Definition B.3 (Relative Pairwise Correlation)**.** Given two distributions D_1, D_2 and a reference distribution D,

$$\chi_D(D_1, D_2) := \int \frac{D_1(x)D_2(x)}{D(x)}dx - 1. \tag{23}$$

Definition B.4 ((γ, β)-correlation)**.** We say that a set of m distributions $\mathcal{D} = \{D_1, \ldots, D_m\}$ is (γ, β) correlated relative to a distribution D_0 over X if $|\chi_D(D_i, D_j)| \le \gamma$ for $i \neq j$ and $|\chi_{D_0}(D_i, D_i)| \le \beta$ for all $i \in [m]$.

Definition B.5 (SQ Dimension)**.** Given a search problem $\mathcal{Z}$ and parameters γ, β, we define the statistical query dimension $\mathcal{SD}(\mathcal{Z}, \gamma, \beta)$ to be the largest integer m such that there exists a distribution D_0 over X and a finite set of distributions $\mathcal{D}_D \subset \mathcal{D}$ with $|\mathcal{D}_D| \ge m$ such that for any $f \in \mathcal{F}$, $\mathcal{D}_f := \mathcal{D}_D \setminus \mathcal{Z}_f$ is (γ, β)-correlated relative to D_0.

The following lemma is from [FGR+17, Corollary 3.12]:

Lemma B.6 (General SQ Lower Bound)**.** *For any $\gamma' > 0$, any SQ algorithm requires at least $\mathcal{SD}(\mathcal{Z}, \gamma, \beta) \cdot \frac{\gamma'}{\beta-\gamma}$ queries to* STAT$(\sqrt{\gamma + \gamma'})$ *or* VSTAT$\left(\frac{1}{3(\gamma+\gamma')}\right)$ *to solve $\mathcal{Z}$.*

We will use the following corollary which is equivalent to Lemma B.6:

Corollary B.7. *For any $\gamma, \beta, \tau \ge 0$, any algorithm requires at least $\mathcal{SD}(\mathcal{Z}, \gamma, \beta) \cdot \frac{\frac{3}{n}-\gamma}{\beta-\gamma}$ queries to* VSTAT(n) *to solve $\mathcal{Z}$.*

B.2 Instantation for the Single-Index Problem

We now instantiate this framework for the single-index model from Definition 1.1:

- **Domain:** $X = \mathbb{R}^d \times \mathbb{R}$ (represents the (X, Y) pair).
- **Distributions:** $\mathcal{D} = \{\mathbb{P}_w : w \in S^{d-1}\}$.
- **Solution Set:** $\mathcal{F} = S^{d-1}$.
- **Valid Solutions:** $\mathcal{Z}(\mathbb{P}_{w^\star}) = \{w \in \mathcal{F} \ : \ |w \cdot w^\star| \ge \tilde{\Theta}(d^{-1/2})\}$.
- **Inverse:** $\mathcal{Z}_w = \{\mathbb{P}_{w^\star} \ : \ w^\star \in S^{d-1} \text{ and } |w \cdot w^\star| \ge \tilde{\Theta}(d^{-1/2})\}$,
- **Reference Distribution:** $D = \gamma_d \otimes \mathsf{P}_y$.

C Hermite Polynomials and Hermite Tensors

We provide a brief review of the properties of Hermite polynomials and Hermite tensors.

Definition C.1. Let $u \in \mathbb{R}^d$. We define the kth normalized Hermite tensor $\boldsymbol{h}_k \in (\mathbb{R}^d)^{\otimes k}$ by

$$\boldsymbol{h}_k(u) := \frac{(-1)^k}{\sqrt{k!}} \frac{\nabla^k \gamma_d(u)}{\gamma_d(u)} \tag{24}$$

where $\gamma_d(u) := \frac{e^{-\|u\|^2/2}}{(2\pi)^{d/2}}$ is the PDF of a standard Gaussian in d dimensions.

Note that when $d = 1$, this definition reduces to the standard univariate Hermite polynomials $\{h_k\}$, which are orthonormal with respect to γ_1:

$$\mathbb{E}_{u\sim\gamma_1}[h_j(u)h_k(u)] = \delta_{jk} \ . \tag{25}$$

Furthermore, if u, v are correlated Gaussians with correlation α, this inner product scales with α^k. Explicitly,

$$\mathbb{E}_{u,v\sim\gamma_2^{(\alpha)}}[h_j(u)h_k(v)] = \delta_{jk}\alpha^k \quad \text{where} \quad \gamma_2^{(\alpha)} = N\left(0, \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}\right). \tag{26}$$

The orthogonality property also has a tensor analogue:

$$\mathbb{E}_{u\sim\gamma_d}[\boldsymbol{h}_j(u) \otimes \boldsymbol{h}_k(u)] = \delta_{jk}\mathrm{Sym}_k. \tag{27}$$

Equivalently, for any j tensor A and k tensor B:

$$\mathbb{E}_{u\sim\gamma_d}\left[\langle \boldsymbol{h}_j(x), A\rangle \langle \boldsymbol{h}_k(x), B\rangle\right] = \delta_{jk}\langle \mathrm{Sym}(A), \mathrm{Sym}(B)\rangle. \tag{28}$$

The Hermite tensors in $\mathbb{R}^d$ are related to the univariate Hermite polynomials by the identity:

$$h_k(u \cdot v) = \langle \boldsymbol{h}_k(u), v^{\otimes k}\rangle \quad \text{for all } u \in \mathbb{R}^d, v \in S^{d-1}. \tag{29}$$

D Proofs of Section 2

Fact 2.1 (Spectral Variance Decomposition)**.** *The variance of $\sigma(Z)$ verifies the expansion*

$$\mathrm{Var}_{\mathsf{P}}[\sigma(Z)] = \sum_{l\geq 1} \beta_l^2 \quad \textit{where} \quad \beta_l := \mathbb{E}_{\mathsf{P}}[Y h_l(Z)] \ . \tag{5}$$

Proof. We have $\mathrm{Var}_{\mathsf{P}}[\sigma(Z)] = \mathbb{E}[\sigma(Z)^2] - \mathbb{E}[\sigma(Z)]^2$ and by decomposition of σ into $\{h_k\}_{k\geq 0}$, the orthogonal basis of $L^2(\mathbb{R}, \gamma_1)$, we have

$$\mathbb{E}[\sigma(Z)^2] = \sum_{l\geq 0} \mathbb{E}[\sigma(Z)h_l(Z)]^2 = \sum_{l\geq 0} \mathbb{E}\left[\mathbb{E}[Y|Z]h_l(Z)\right]^2 = \sum_{l\geq 0} \mathbb{E}[Y h_l(Z)]^2 \ , \tag{30}$$

where the last equality stems from the property of conditional expectation. Finally, substracting $\mathbb{E}[\sigma(Z)]^2 = \beta_0^2$ ends the proof of the equality. ■

We begin by computing the Hermite expansion of $\frac{d\mathsf{P}}{d\mathsf{P}_0}$ where $\mathsf{P}_0 = \mathsf{P}_z \otimes \mathsf{P}_y$ is the null distribution:

Lemma D.1. *We have the following expansion in $L^2(\mathsf{P}_0)$:*

$$\frac{d\mathsf{P}}{d\mathsf{P}_0}(z,y) = \sum_{k\geq 0} \zeta_k(y) h_k(z). \tag{31}$$

Proof. We can directly compute the kth Hermite coefficient of the likelihood ratio as a function of Y:

$$\mathbb{E}_{\mathsf{P}_0}\left[\frac{d\mathsf{P}}{d\mathsf{P}_0}(Z,Y)\boldsymbol{h}_k(Z)|Y\right] = \mathbb{E}_{\mathsf{P}}[\boldsymbol{h}_k(Z)|Y] = \zeta_k. \tag{32}$$

Therefore the Hermite expansion of $\frac{d\mathsf{P}}{d\mathsf{P}_0}$ is,

$$\frac{d\mathsf{P}}{d\mathsf{P}_0}(z,y) \overset{L^2(\mathsf{P}_0)}{=} \sum_{k\geq 0} \zeta_k(y) h_k(z). \tag{33}$$

■

We can now restate and prove Lemma 2.3.

Lemma 2.3 (Mutual Information Decomposition)**.** *We have the following expansion*

$$I_{\chi^2}[\mathsf{P}] = \sum_{k\geq 1} \lambda_k^2 \quad \textit{where} \quad \lambda_k := \|\zeta_k\|_{\mathsf{P}_y} \quad \textit{and} \quad \zeta_k := \mathbb{E}[h_k(Z)|Y]. \tag{7}$$

Proof of Lemma 2.3. Recall that the mutual information is given by

$$I_{\chi^2}[\mathsf{P}] = \mathbb{E}_{\mathsf{P}_0}\left[\left(\frac{d\mathsf{P}}{d\mathsf{P}_0}\right)^2\right] - 1. \tag{34}$$

Therefore by Lemma D.1, this is equal to

$$I_{\chi^2}[\mathsf{P}] = \sum_{k\geq 0} \mathbb{E}_{\mathsf{P}_y}[\zeta_k(Y)^2] - 1 = \sum_{k\geq 0} \lambda_k^2 - 1 = \sum_{k\geq 1} \lambda_k^2. \tag{35}$$

■

Proposition 2.6 (A Variational Representation)**.** *The generative exponent $k^\star(\mathsf{P})$ can be written as:*

$$k^\star(\mathsf{P}) = \inf_{\mathcal{T}\in L^2(\mathsf{P}_y)} l^\star((\mathrm{Id}\otimes\mathcal{T})_\#\mathsf{P}) \ . \tag{9}$$

Proof. Let $k^\star = k^\star(\mathsf{P})$. For any $k < k^\star$ and $\mathcal{T} \in L^2(\mathbb{R}, \mathsf{P}_y)$, By properties of the conditional expectation:

$$\mathbb{E}\left[\mathcal{T}(Y)h_k(Z)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{T}(Y)h_k(Z)|Y\right]\right] = \mathbb{E}\left[\mathcal{T}(Y)\mathbb{E}\left[h_k(Z)|Y\right]\right] = \mathbb{E}\left[\mathcal{T}(Y)\zeta_k(Y)\right] = 0. \tag{36}$$

Therefore for all $\mathcal{T} \in L^2(\mathbb{R}, \mathsf{P}_y)$, we have $k^\star(\mathsf{P}) \leq l^\star((\mathrm{Id} \otimes \mathcal{T})_\# \mathsf{P})$ and hence taking the infimum over such $\mathcal{T}$, it yields that

$$k^\star \leq \inf_{\mathcal{T} \in L^2(\mathsf{P}_y)} l^\star((\mathrm{Id} \otimes \mathcal{T})_\# \mathsf{P}) \ . \tag{37}$$

Next, let define for all $y \in \mathbb{R}$, $\mathcal{T}^*(y) := \zeta_{k^\star}(y) = \mathbb{E}[h_{k^\star}(Z)|Y = y]$. Note that $\mathcal{T} \in L^2(\mathsf{P}_y)$ and by the same calculation as previously, we have

$$\mathbb{E}\left[\mathcal{T}^*(Y) h_{k^\star}(Z)\right] = \mathbb{E}\left[\mathcal{T}^*(Y) \zeta_{k^\star}(Y)\right] = \mathbb{E}\left[\zeta^2_{k^\star}(Y)\right] = \|\zeta_{k^\star}\|^2_{\mathsf{P}_y} > 0 \ , \tag{38}$$

which concludes the proof of the theorem.

■

Example 2.7 (Explicit Examples of generative exponent)**.** *We give the following explicit examples:*

(i) For σ a polynomial, we have $k^\star(\sigma) \leq 2$ and $k^\star(\sigma) = 2$ iff σ is even. In particular, $k^\star(h_j) = 1$ if j is odd and $k^\star(h_j) = 2$ if j is even.

(ii) For $\sigma(z) = z^2 e^{-z^2}$, we have $k^\star(\sigma) = 4$.

(iii) From [MM18, Remark 3], if $y \in \{-1, 1\}$ is boolean with $P(Y = 1|Z = z) = \mathbb{E}_{W \sim \gamma}[\tanh(c_1 z)^2 - \tanh(c_2 z)^2]$ for carefully chosen c_1, c_2 then $k^\star(\sigma) = 4$.

Proof. When σ is a polynomial, we need to show that there exists $g \in \mathsf{P}_y$ such that

$$\mathbb{E}[g(\sigma(z)) h_l(z)] \neq 0 \tag{39}$$

for $l = 1$ or $l = 2$. Since both σ and h_l are monotonic after their largest root, picking $g(t) = \mathbf{1}_{t \in [R-\delta, R+\delta]}$ for sufficiently large R and $\delta > 0$ yields

$$\mathbb{E}[g(\sigma(z)) h_l(z)] = \mathbb{E}[h_l(z) \mathbf{1}_{z \in A}] \ , \tag{40}$$

where either $A = I$ is a single interval (if σ has odd degree) or $A = I_- \cup I_+$ two intervals if σ has even degree. We have $\mathbb{E}[h_1(z)\mathbf{1}_{z \in A}] \neq 0$ whenever $A = I$ or $I_- \neq -I_+$, and $\mathbb{E}[h_2(z)\mathbf{1}_{z \in A}] \neq 0$ otherwise. To conclude, observe that we can find R and δ such that $I_- \neq -I_+$ iff σ is not even.

Lemma D.2. *Let $\sigma(Z) := Z^2 \exp\left(-Z^2\right)$. Then the single index model defined by $\mathsf{P} := (\mathrm{Id} \otimes \sigma)_\# \gamma_1$ satisfies $k^\star(\mathsf{P}) = 4$.*

Proof of Lemma D.2. As σ is even it suffices to prove that $\lambda_2 = 0$ and $\lambda_4 > 0$. By Lemma G.2, to prove that $\lambda_4 = 0$ it suffices to check that

$$\sum_{z \in \sigma^{-1}(y)} \mathrm{sign}(\sigma'(z)) z \gamma(z) \tag{41}$$

is constant in y P_y-almost everywhere. Therefore it suffices to check this for $0 < y < 1$, as $\max_z \sigma(z) = 1$. On this interval, $\sigma^{-1}(y) = \{-z_2(y), -z_1(y), z_1(y), z_2(y)\}$ with $\sigma'(z_1(y)) > 0$ and

$\sigma'(z_2(y)) < 0$. Therefore for $y \in (0,1)$,

$$\begin{aligned}
&\sum_{z \in \sigma^{-1}(y)} \operatorname{sign}(\sigma'(z)) z\gamma(z) \\
&= 2[z_1(y)\gamma(z_1(y)) - z_2(y)\gamma(z_2(y))] \\
&= 2[\sqrt{y} - \sqrt{y}] \\
&= 0.
\end{aligned} \tag{42}$$

Next, we need to verify that $\zeta_4 \neq 0$. In fact, we claim that $\beta_4 \neq 0$, i.e. $l^\star(\mathsf{P}) = 4$. We have that:

$$\begin{aligned}
\mathbb{E}[\sigma(Z)h_4(Z)] &= \int_{-\infty}^{\infty} Z^2 e^{-Z^2} \cdot He_4(Z) \cdot \frac{e^{-Z^2/2}}{\sqrt{2\pi}} dZ \\
&= \int_{-\infty}^{\infty} (Z^6 - 6Z^4 + 3Z^2) \cdot \frac{e^{-3Z^2/2}}{\sqrt{2\pi}} dZ \\
&= -\frac{4\sqrt{3}}{27} \\
&\neq 0
\end{aligned} \tag{43}$$

by routine Gaussian integration. ■

■

E Proofs of Section 3

The main goal of this section is to prove Theorem 3.2. We begin this section by proving some necessary intermediate results, and then conclude the section proving the aforementioned theorem. Finally we conclude providing results on reduction from the NGCA to the single index model and back.

E.1 Proof of the preliminary results

We begin by generalizing Lemma D.1 to the full distribution $\mathbb{P}_w$ over (X, Y) where the null distribution is now $\mathbb{P}_0 := \gamma_d \otimes \mathsf{P}_y$.

Lemma E.1. *For any $w \in S^{d-1}$ we have the following expansion in $L^2(\mathbb{P}_0)$:*

$$\frac{d\mathbb{P}_w}{d\mathbb{P}_0}(X,Y) \overset{L^2(\mathbb{P}_0)}{=} \sum_{k \geq 0} \zeta_k(Y) h_k(X \cdot w) \tag{44}$$

Proof. Note that by Definition 1.1, $d\mathbb{P}_w(X,Y) = \gamma_{d-1}(X^\perp)\mathsf{P}(X \cdot w, Y)$ and $d\mathbb{P}_0(X,Y) = \gamma_{d-1}(X^\perp)\mathsf{P}_0(X \cdot w, Y)$. Therefore by Lemma D.1,

$$\frac{d\mathbb{P}_w}{d\mathbb{P}_0}(X,Y) = \frac{d\mathsf{P}}{d\mathsf{P}_0}(X \cdot w, Y) = \sum_{k \geq 0} \zeta_k(Y) h_k(X \cdot w). \tag{45}$$

■

We can now prove the expansion of $\chi_0^2(\mathbb{P}_w, \mathbb{P}_{w'})$ presented in Lemma 3.1.

Lemma 3.1 (χ^2 Representation)**.** *Let $w, w' \in S^{d-1}$, and denote $m = w \cdot w'$, then we have*

$$\chi_0^2(\mathbb{P}_w, \mathbb{P}_{w'}) = \sum_{k \geq k^\star} m^k \lambda_k^2 \ . \tag{12}$$

Proof of Lemma 3.1. We have:

$$\chi_0^2(\mathbb{P}_w, \mathbb{P}_{w'}) := \mathbb{E}_{\mathbb{P}_0}\left[\frac{d\mathbb{P}_w}{d\mathbb{P}_0} \cdot \frac{d\mathbb{P}_{w'}}{d\mathbb{P}_0}\right] - 1. \tag{46}$$

Therefore by Lemma E.1 and the orthogonality property of Hermite polynomials, this is equal to

$$\chi^2_{\mathbb{P}_0}(\mathbb{P}_w, \mathbb{P}_{w'}) = \sum_{k \geq 0} \mathbb{E}[\zeta_k(Y)^2](w \cdot w')^k - 1 = \sum_{k \geq 1} \lambda_k^2 m^k.$$

■

E.2 Proof of Theorem 3.2

We now move towards proving Theorem 3.2. For this we introduce two intermediate results in Lemma E.2 and E.3.

Lemma E.2. *Let $m = w \cdot w'$. We have*

$$\chi_0^2(\mathbb{P}_w, \mathbb{P}_{w'}) \leq \lambda_{k^\star}^2 m^{k^\star} + \frac{m^{k^\star+1}}{1-m}. \tag{47}$$

Proof. From the previous lemma, we have that

$$\begin{aligned}
\chi_0^2(\mathbb{P}_w, \mathbb{P}_{w'}) &= \sum_{k \geq k^\star} \lambda_k^2 m^k \\
&= \lambda_{k^\star}^2 m_{k^\star} + \sum_{k > k^\star} \lambda_k m^k \\
&\leq \lambda_{k^\star}^2 m_{k^\star} + \sum_{k > k^\star} m^k \\
&= \lambda_{k^\star}^2 m_{k^\star} + \frac{m^{k^\star+1}}{1-m}.
\end{aligned} \tag{48}$$

■

The following lemma shows that there are a large number of nearly orthogonal vectors:

Lemma E.3. *There exists an absolute constant C such that for any $m \leq d^{d^{1/4}}$, there exist m vectors $w_1, \ldots, w_m$ with $\max_{i \neq j} |w_i \cdot w_j| \leq \epsilon$ for $\epsilon = \sqrt{\frac{C \log_d(m)^2}{d}}$.*

Proof. From [NNW12], we have that there exists a constant c such that such points exist whenever

$$d \le c\epsilon^{-2}\left(\frac{\log m}{\log\log m + \log(1/\epsilon)}\right)^2. \tag{49}$$

We will take

$$\epsilon = \sqrt{\frac{C\log_d(m)^2}{d}}. \tag{50}$$

for a sufficiently large constant C. Then we have that

$$\log m \le 1/\epsilon \iff \log^2 m \le \frac{d\log^2 d}{C^2\log^2 m} \iff \log m \le \frac{d^{1/4}\log^{1/2} d}{C^{1/2}} \tag{51}$$

which is true by the assumption on m. In addition, note that $\log(1/\epsilon) \lesssim \log(d)$. Therefore,

$$\epsilon^{-2}\left(\frac{\log m}{\log\log m + \log(1/\epsilon)}\right)^2 \gtrsim \frac{d}{C\log_d(m)^2}\cdot\left(\frac{\log m}{\log d}\right)^2 = \frac{d}{C} \tag{52}$$

and taking C sufficiently large completes the proof. ∎

Now we are in place to prove the main result of the section, Theorem 3.2.

Proof of Theorem 3.2. Let $m < d^{d^{1/4}}$ be a positive integer to be chosen later. From Lemma E.3, there exist m vectors $w_1, \dots, w_m$ such that $|w_i \cdot w_j| \le \epsilon$ for all $i \ne j$ where $\epsilon = \sqrt{\frac{C\log_d(m)^2}{d}}$ for an absolute constant C. Let $\mathcal{D} = \{\mathbb{P}_{w_i} \ : \ i \in [m]\}$. For any $\mathbb{P}_{w_i}, \mathbb{P}_{w_j}$ with $i \ne j$,

$$\chi_0^2(D_{w_1}, D_{w_2}) \le \lambda_{k^\star}^2 (w_1\cdot w_2)^{k^\star} + \frac{(w_1\cdot w_2)^{k^\star+1}}{1-(w_2\cdot w_2)} \le \lambda_{k^\star}^2\epsilon^{k^\star} + 2\epsilon^{k^\star+1}. \tag{53}$$

Therefore for $\lambda_{k^\star}^2 > 2\epsilon$, we can bound this by $2\lambda_{k^\star}^2\epsilon^{k^\star}$. Therefore $\mathcal{SD}(\mathcal{Z}, 2\lambda_{k^\star}^2\epsilon^{k^\star}, 1) \ge m$. Then by Corollary B.7,

$$q \ge m\cdot\frac{\frac{3}{n} - 2\lambda_{k^\star}^2\epsilon^{k^\star}}{1-2\lambda_{k^\star}^2\epsilon^{k^\star}} \ge \frac{1}{2}m\cdot\left(\frac{3}{n} - 2\lambda_{k^\star}^2\epsilon^{k^\star}\right) \tag{54}$$

which implies

$$\frac{3}{n} \le 2\lambda_{k^\star}^2\epsilon^{k^\star} + \frac{2q}{m}. \tag{55}$$

Now setting $m = 2qn$ gives that

$$n \ge \frac{1}{\lambda_{k^\star}^2\epsilon^{k^\star}} \gtrsim \frac{1}{\lambda_{k^\star}^2}\cdot\left(\frac{d}{\log_d^2(2qn)}\right)^{k^\star/2}. \tag{56}$$

Now let $c_{k^\star}$ be a sufficiently small constant which depends only on $k^\star$ and assume for the sake of contradiction that

$$n \le \frac{c_{k^\star}}{\lambda_{k^\star}^2}\cdot\left(\frac{d}{\log_d^2(q)}\right)^{k^\star/2}. \tag{57}$$

Then we must have that

$$\log_d(2n) \le \frac{k^\star+1}{2} + \log_d\left(2c_{k^\star}\right) \le k^\star. \tag{58}$$

Therefore for the algorithm to succeed we must have

$$n \gtrsim \frac{1}{\lambda_{k^\star}^2}\left(\frac{d}{\log_d^2(2qn)}\right)^{k^\star/2} \ge \frac{1}{\lambda_{k^\star}^2}\left(\frac{d}{(\log_d(q)+k^\star)^2}\right)^{k^\star/2} \gtrsim_{k^\star} \frac{1}{\lambda_{k^\star}^2}\left(\frac{d}{\log_d(q)}\right)^{k^\star/2} = \frac{n}{c_{k^\star}}. \tag{59}$$

Therefore for sufficiently small $c_{k^\star}$ we have derived a contradiction so we must have that

$$n \ge \frac{c_{k^\star}}{\lambda_{k^\star}^2} \cdot \left(\frac{d}{\log_d^2(q)}\right)^{k^\star/2} \tag{60}$$

which completes the proof. ■

Lemma E.4 (Lower Bound for Highly Periodic Neuron)**.** *Let $\sigma(Z) = \cos(2\pi\gamma Z)$. Then for w, w' with $m := |w \cdot w'| < \frac{1}{8\pi\gamma}$,*

$$\chi_0^2(\mathbb{P}_w, \mathbb{P}_{w'}) \lesssim \exp\left(-2\pi^2\gamma^2\right) + \frac{m^\gamma}{1-m}. \tag{61}$$

In addition, when $\gamma = cd^{1/4}$ and $m = d^{-1/4}$ for a sufficiently small constant c, any algorithm requires $m \gtrsim d^{d^{1/4}}$ queries to VSTAT(n) *to recover $w^\star$ when $\mathsf{P} = \gamma(z)\delta_{\sigma(z)}(y)$ unless $n \gtrsim d^{d^{1/4}}$.*

Proof. We will begin by computing $\mathsf{P}_{z|y}$. Note that the level sets are given by:

$$\{z \ : \ \sigma(z) = y\} = \bigcup_{j\in\mathbb{Z}} \{Z_j^+(y), Z_j^-(y)\} \quad \text{where} \quad Z_j^\pm(y) := \frac{\pm\arccos(y) + 2\pi j}{2\pi\gamma}. \tag{62}$$

Therefore by the coarea formula,

$$\mathsf{P}_{z|y} = \frac{\sum_{j\in\mathbb{Z}} \gamma(Z_j^+(y))\delta_{Z_j^+(y)}(z) + \gamma(Z_j^-(y))\delta_{Z_j^+(y)}(z)}{\sum_{j\in\mathbb{Z}} \gamma(Z_j^+(y)) + \gamma(Z_j^-(y))}. \tag{63}$$

We will now use the Poisson summation formula. For any f,

$$\sum_{j\in\mathbb{Z}} f(Z_j^+(y)) = \gamma \sum_{j\in\mathbb{Z}} \cos(j\arccos(y))\hat{f}(\gamma j). \tag{64}$$

Plugging in $f = \gamma$ gives

$$\sum_{j\in\mathbb{Z}} \gamma(Z_j^+(y)) = \gamma \sum_{j\in\mathbb{Z}} \cos(j\arccos(y)) \exp\left(-2\pi^2 j^2\gamma^2\right) \tag{65}$$

$$= \gamma + \gamma \sum_{j\neq 0} \cos(j\arccos(y)) \exp\left(-2\pi^2 j^2\gamma^2\right). \tag{66}$$

This second term can be bounded by

$$2\sum_{j>1}\exp\bigl(-2\pi^2 j^2\gamma^2\bigr) \le 2\sum_{j>1}\exp\bigl(-2\pi^2 j\gamma^2\bigr) = \frac{2}{\exp(2\pi^2\gamma^2)-1} \le 4\exp\bigl(-2\pi^2\gamma^2\bigr). \tag{67}$$

We can conduct the same analysis with Z_j^- so the normalization factor is $2\gamma(1+O(e^{-c\gamma^2}))$. We now apply the same technique to compute $\{\zeta_k\}$. We have that for $k \le \gamma$,

$$\begin{aligned}
&\left|\sum_{j\in\mathbb{Z}}\gamma(Z_j^+(y))h_k(Z_j^+(y))\right| \\
&= \gamma\left|\sum_{j\in\mathbb{Z}}\cos(j\arccos(y))(2\pi j\gamma)^k\exp\bigl(-2\pi^2 j^2\gamma^2\bigr)\exp(-ji\pi/2)\right| \\
&= \left|2\gamma\sum_{j>1}\cos(j\arccos(y))(2\pi j\gamma)^k\exp\bigl(-2\pi^2 j^2\gamma^2\bigr)\exp(-ji\pi/2)\right| \\
&\le 2\gamma\sum_{j>1}(2\pi j\gamma)^k\exp\bigl(-2\pi^2 j^2\gamma^2\bigr) \\
&\le 2\gamma(2\pi\gamma)^k\sum_{j>1}j^k\exp\bigl(-2\pi^2 j\gamma^2\bigr) \\
&\le 2\gamma(2\pi\gamma)^k\frac{\exp\bigl(2k\pi^2\gamma^2\bigr)+k!\exp\bigl(2(k-1)\pi^2\gamma^2\bigr)}{(\exp(2\pi^2\gamma^2)-1)^{k+1}} \\
&\lesssim \gamma(4\pi\gamma)^k\exp\bigl(-2\pi^2\gamma^2\bigr).
\end{aligned} \tag{68}$$

Therefore for $m < \frac{1}{8\pi\gamma}$,

$$\chi_0^2(\mathbb{P}_w,\mathbb{P}_{w'}) \lesssim \exp\bigl(-2\pi^2\gamma^2\bigr)\sum_{k\le\gamma}(4\pi\gamma m_j)^k + \frac{m^\gamma}{1-m} \tag{69}$$

$$\le 2\exp\bigl(-2\pi^2\gamma^2\bigr) + \frac{m^\gamma}{1-m}. \tag{70}$$

The SQ lower bound now directly follows from Corollary B.7. ■

E.3 Low Degree Polynomial Lower Bound

Using Lemma E.1, we can directly prove Theorem 3.5:

Theorem 3.5 (Low-Degree Method Lower Bound)**.** *Let* P *be a single index model, let* $\mathbb{P}$ *be the distribution of* $\mathbb{P}_w$ *when* w *is chosen randomly from* $\mathrm{Unif}(S^{d-1})$*. Let* $X \in \mathbb{R}^{n\times d}$ *and* $Y \in \mathbb{R}^n$ *denote a sequence of* n *i.i.d. inputs and targets drawn from* P*. Let* $\mathbb{P}_0 := \gamma_d \otimes \mathsf{P}_y$ *denote the null distribution. Let* $\mathcal{R}(x,y)$ *denote the likelihood ratio* $\frac{d\mathbb{P}}{d\mathbb{P}_0}(x,y)$ *and let* $\mathcal{R}_{\le D}(x,y)$ *denote the orthogonal projection in* $L^2(\mathbb{P}_0)$ *of* $\mathcal{R}(x,y)$ *onto polynomials of degree at most* D *in* x*. Then if* $\frac{D}{\lambda_k^2} \ll \sqrt{d}$ *and* $\delta := \frac{n}{d^{k^\star/2}}$*,*

$$\|\mathcal{R}_{\le D}\|_{\mathbb{P}_0}^2 = (1+o_d(1))\sum_{j=0}^{\lfloor\frac{D}{k^\star}\rfloor}\mathbf{1}_{2|k^\star j}(k^\star j-1)!!\frac{(\lambda_{k^\star}^2\delta)^j}{j!}.$$

Proof of Theorem 3.5. Let $\mathcal{R}_w := \frac{d\mathbb{P}_w}{d\mathbb{P}_0}$ denote the likelihood ratio conditioned on w. We begin by computing the full likelihood ratio:

$$\mathcal{R}((x_1,y_1),\dots,(x_n,y_n)) = \frac{\mathbb{E}_w[\prod_{i=1}^n \mathbb{P}_w[x_i,y_i]]}{\prod_{i=1}^n \mathbb{P}[x_i]\mathbb{P}[y_i]} = \mathbb{E}_w\left[\prod_{i=1}^n \mathcal{R}_w(x_i,y_i)\right].$$

Then by Lemma E.1, we can expand this as

$$\mathcal{R} = \mathbb{E}_w\left[\prod_{i=1}^n \left(\sum_{k\geq 0} \zeta_k(y_i) h_k(x_i\cdot w)\right)\right].$$

We will isolate the low degree part with respect to $\{x_1,\dots,x_n\}$, which we denote by $\mathcal{R}_{\leq D}$. To compute this, we need to switch the product and the summation:

$$\mathcal{R} = \mathbb{E}_w\left[\sum_{p=0}^{\infty} \sum_{k_1+\dots+k_n=p} \left(\prod_{i=1}^n \zeta_{k_i}(y_i) h_{k_i}(x_i\cdot w)\right)\right].$$

We note that each term on the right hand side is a polynomial in $x_1,\dots,x_n$ of degree p which is orthogonal to all polynomials of degree less than p. Therefore $\mathcal{R}_{\leq D}$ is given by:

$$\mathcal{R}_{\leq D} = \mathbb{E}_w\left[\sum_{p=0}^{D} \sum_{k_1+\dots+k_n=p} \left(\prod_{i=1}^n \zeta_{k_i}(y_i) h_{k_i}(x_i\cdot w)\right)\right].$$

We can now use the orthogonality property of Hermite polynomials to compute the norms with respect to the null distribution $\mathbb{P}_0$. If if w, w' are independent draws from the prior on w then:

$$\begin{aligned}\|\mathcal{R}_{\leq D}\|^2_{L^2(\mathbb{P}_0)} &= \mathbb{E}_{w,w'}\left[\sum_{p=0}^{D} \sum_{k_1+\dots+k_n=p} \left(\prod_{i=1}^n \lambda_{k_i}^2 (w\cdot w')^{k_i}\right)\right] \\ &= \sum_{p=0}^{D} \mathbb{E}[(w\cdot w')^p] \sum_{k_1+\dots+k_n=p} \left(\prod_{i=1}^n \lambda_{k_i}^2\right).\end{aligned}$$

Let z be a random variable with distribution $w\cdot w'$ where w, w' are drawn independently from the prior on w, and let $\mathcal{P}_{\leq D}$ be the projection operator onto polynomials of degree at most D in z. Then we can rewrite the above expression as:

$$\|\mathcal{R}_{\leq D}\|^2_{L^2(\mathbb{P}_0)} = \mathbb{E}_z\left[\mathcal{P}_{\leq D}\left[\left(\sum_{k\geq 0} \lambda_k^2 z^k\right)^n\right]\right].$$

By linearity of expectation and of the projection operator $\mathcal{P}_{\leq D}$, we can expand this using the binomial theorem:

$$\|\mathcal{R}_{\leq D}\|^2_{L^2(\mathbb{P}_0)} = \sum_{j\geq 0} \binom{n}{j} \mathbb{E}\left[\mathcal{P}_{\leq D}\left[\left(\sum_{k\geq k^\star} \lambda_k^2 z^k\right)^j\right]\right].$$

We can now lower bound $\|\mathcal{R}_{\le D}\|^2_{L^2(\mathbb{P}_0)}$ by isolating the terms where $k = k^\star$:

$$\begin{aligned}\|\mathcal{R}_{\le D}\|^2_{L^2(\mathbb{P}_0)} &\ge \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \mathbb{E}\Big[\mathcal{P}_{\le D}\Big[\lambda_{k^\star}^{2j} z^{k^\star j}\Big]\Big] \\ &= \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \lambda_k^{2j} \mathbf{1}_{2|k^\star j} \frac{(k^\star j - 1)!!}{\prod_{i=0}^{k^\star j/2-1}(d+2i)} =: \mathcal{R}^\star.\end{aligned}$$

We can similarly upper bound this expression by using that $\lambda_k^2 \le 1$. Plugging this in for $k > k^\star$ gives:

$$\begin{aligned}\|\mathcal{R}_{\le D}\|^2_{L^2(\mathbb{P}_0)} &\le \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \mathbb{E}\left[\mathcal{P}_{\le D}\left[\left(\lambda_{k^\star}^2 z^{k^\star} + \frac{z^{k^\star+1}}{1-z}\right)^j\right]\right] \\ &= \mathcal{R}^\star + \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \mathbb{E}\left[\mathcal{P}_{\le D}\left[\sum_{i=1}^{j} \binom{j}{i} (\lambda_{k^\star}^2 z^{k^\star})^{j-i} \left(\frac{z^{k^\star+1}}{1-z}\right)^i\right]\right] \\ &= \mathcal{R}^\star + \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \mathbb{E}\left[\sum_{i=1}^{j} \binom{j}{i} (\lambda_{k^\star}^2 z^{k^\star})^{j-i} \mathcal{P}_{\le D-k^\star(j-i)}\left[\left(\frac{z^{k^\star+1}}{1-z}\right)^i\right]\right] \\ &\le \mathcal{R}^\star + \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \mathbb{E}\left[\sum_{i=1}^{j} \binom{j}{i} (\lambda_{k^\star}^2 z^{k^\star})^{j-i} \left[\left(\frac{z^{k^\star+1}}{1-z}\right)^i\right]\right] \\ &= \mathcal{R}^\star + \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \sum_{i=1}^{j} \binom{j}{i} \lambda_{k^\star}^{2(j-i)} \mathbb{E}\left[\frac{z^{k^\star j+i}}{(1-z)^i}\right] \\ &\le \mathcal{R}^\star + C \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \sum_{i=1}^{j} \binom{j}{i} \lambda_{k^\star}^{2(j-i)} \mathbb{E}\left[z^{k^\star j+i}\right]\end{aligned}$$

where the last line follows from [DNGL23, Lemma 26]. We can now relate the $k^\star j + i$ and the $k^\star j$-th moments:

$$\begin{aligned}\|\mathcal{R}_{\le D}\|^2_{L^2(\mathbb{P}_0)} &\le \mathcal{R}^\star + C \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \sum_{i=1}^{j} \binom{j}{i} \lambda_{k^\star}^{2(j-i)} \|z\|_{jk^\star}^{jk^\star} \left(\frac{j(k^\star+1)}{d}\right)^{i/2} \\ &= \mathcal{R}^\star + C \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \lambda_{k^\star}^{2j} \|z\|_{jk^\star}^{jk^\star} \left(\left(1 + \sqrt{\frac{j(k^\star+1)}{\lambda_{k^\star}^4 d}}\right)^j - 1\right) \\ &\le \mathcal{R}^\star + \sqrt{\frac{CD}{\lambda_{k^\star}^4 d}} \sum_{j=0}^{\lfloor D/k^\star \rfloor} \binom{n}{j} \lambda_{k^\star}^{2j} \|z\|_{jk^\star}^{jk^\star} \\ &\le \mathcal{R}^\star(1 + o_d(1)),\end{aligned}$$

which completes the proof. ■

This implies the following corollary:

Corollary 3.6. *Let $\mathbb{P}, \mathbb{P}_0, \mathcal{R}$ be as in Theorem 3.5 and assume $k^\star > 1$. Then,*

- ***Weak Detection:*** *If $D = \log(d)^2$ and $n \le d^\gamma$ for $\gamma < \frac{k^\star}{2}$, then $\|\mathcal{R}_{\le D}\|_{\mathbb{P}_0} = 1 + o_d(1)$.*
- ***Strong Detection:*** *For any $D \ll \sqrt{d}$, if $n \le \frac{d^{\frac{k^\star}{2}}}{D^{\frac{k^\star}{2}-1}}$, then $\|\mathcal{R}_{\le D}\|_{\mathbb{P}_0} = O_d(1)$.*

Proof. The weak recovery threshold follows directly from Theorem 3.5 by setting $\delta = d^{\gamma - \frac{k^\star}{2}}$. For the strong recovery threshold, we have by Theorem 3.5,

$$\begin{aligned}
\|\mathcal{R}_{\le D}\|_{\mathbb{P}_0}^2 &\lesssim \sum_{j=0}^{\lfloor \frac{D}{k^\star} \rfloor} \mathbf{1}_{2|k^\star j} (k^\star j - 1)!! \frac{(\lambda_{k^\star}^2 \delta)^j}{j!} \\
&\lesssim \sum_{j=0}^{\lfloor \frac{D}{k^\star} \rfloor} \left(\frac{e}{j}\right)^j \left(\frac{k^\star j}{e}\right)^{k^\star j/2} (\lambda_{k^\star}^2 \delta)^j \\
&= \sum_{j=0}^{\lfloor \frac{D}{k^\star} \rfloor} \left(\frac{\delta k^{\star k^\star/2} j^{k^\star/2-1}}{e^{k^\star/2-1}}\right)^j \\
&\le \sum_{j=0}^{\infty} \left(\frac{\delta k^\star D^{k^\star/2-1}}{e^{k^\star/2-1}}\right)^j \\
&\le \sum_{j=0}^{\infty} (2/e)^j \\
&= \frac{1}{1-2/e}
\end{aligned}$$

which completes the proof. ∎

E.4 Reduction from Single-Index Model to NGCA

We restate and prove Proposition 3.7.

Proposition 3.7 (Single-Index to NGCA Reduction)**.** *An SQ algorithm that solves NGCA with planted distribution μ of exponent k and direction $w^* \in S^{d-1}$ yields an efficient SQ algorithm to solve the single-index model with generative exponent k and direction w^*.*

Proof. Let us fix a single index problem, $\mathbb{P}_{w^*,\mathsf{P}}$ with generative exponent $k = k^\star(\mathsf{P})$, and direction $w^* \in S^{d-1}$. By definition of the generative exponent, we have that $\|\zeta_k\|_{\mathsf{P}_y} > 0$. Define $A_+ = \{y; \zeta_k(y) > 0\}$, $A_- = \{y; \zeta_k(y) < 0\}$ and $\alpha := \max\{\mathsf{P}_y(A_+), \mathsf{P}_y(A_-)\}$. Then $\alpha > 0$ as otherwise ζ_k would be 0 P_y-a.e. In the following, we consider without loss of generality that $\alpha = \mathsf{P}_y(A_+)$.

Let us define $\mu(dz) = \alpha^{-1} \int_{A_+} \mathsf{P}_{z|y}(dz, y)\, \mathsf{P}_y(dy) \in \mathcal{P}(\mathbb{R})$, where we recall that $\mathsf{P}_{z|y}$ is the conditional distribution of Z given Y. Let us assume that we have access to $\{(x_i, y_i)\}_{i \le n}$, i.i.d. samples distributed according to $\mathbb{P}_{w^*,\mathsf{P}}$. Now, let us perform *rejection sampling*, that is, for all $i \le n$, we keep x_i if and only if $y_i \in A_+$. This builds a data set of $\tilde{n}(n)$ samples $\{\tilde{x}_i\}_{i \le \tilde{n}(n)}$ that are i.i.d. from a NGCA-model with distribution μ and direction $w^\star$. From $\alpha > 0$, we have that

$\mathbb{E}_\mu[h_k(Z)] = \alpha^{-1}\int \zeta_k(y)\mathsf{P}_y(dy) > 0$. Moreover, the number of samples obtained after rejection sampling will concentrate (with binomial tails) at $\tilde{n} = \alpha n$, with α independent of dimension. Hence any SQ-algorithm that solves the built NGCA model, i.e. that is able to find the direction w^*, will solve the later single-index one efficiently. ■

E.5 Reduction from NGCA to Single-Index Model Variant

Given $X' \sim \mathbb{Q}_{\mu,w^\star}$, we consider $X \sim \gamma_d$ drawn independently from X', and $Y = h_k(X \cdot X')$. Assume wlog that $w^\star$ is the first canonical vector, and let us write $x' = (x'_1, \bar{x}')$ and $x = (x_1, \bar{x})$. The conditional distribution of Y given X satisfies

$$\begin{aligned} p(y|x) &= \int p(y|x,x')\mathbb{Q}(dx') = \int_{\mathbb{R}} \left(\int_{\mathbb{R}^{d-1}} \delta_{y-h_k(x_1x'_1+\bar{x}\cdot\bar{x}')}\gamma_{d-1}(d\bar{x}') \right) \mu(dx'_1) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \delta_{y-h_k(x_1x'_1+\|\bar{x}\|z)}\gamma_1(dz) \right) \mu(dx'_1) \\ &= \mathbb{E}_{x'_1\sim\mu}\left[\int_{\{z;h_k(x_1x'_1+\|\bar{x}\|z)=y\}} \frac{\gamma_1(z)}{\|\bar{x}\||h'_k(x_1x'_1+\|\bar{x}\|z)|} d\mathcal{H}_0(z) \right] \\ &= \tilde{\psi}_k\left(y, x_1, \|x\|\right) \ , \end{aligned} \tag{71}$$

where we used the coarea formula and the rotational symmetry of the Gaussian measure. In other words, the conditional distribution $p(y|x)$ now depends on two scalar summary statistics: the projection along one hidden direction *and* the norm of x. The limitation of this construction, however, is the fact that the signal strength x_1 is of order $O(1)$, while the fluctuations of the uninformative norm $\|x\|$ are also of order $\Theta(1)$, leading to a presumably harder estimation task than that of Definition 1.1.

F Proofs of Section 4

We begin by defining the un-normalized Hermite polynomials and Hermite tensors:

Definition F.1.

$$\mathrm{He}_k(x) := \sqrt{k!}h_k(x) \quad \text{and} \quad \mathbf{He}_k(x) := \sqrt{k!}\boldsymbol{h}_k(x). \tag{72}$$

Unlike h_k, He_k naturally tensorizes, i.e.

$$\mathbf{He}_k(x)_{i_1,\dots,i_k} = \prod_{i=1}^{d} \mathrm{He}_{|\{j\,:\,i_j=i\}|}(x_i). \tag{73}$$

For example,

$$\mathbf{He}_3(x)_{1,1,2} = \mathrm{He}_2(x_1)\mathrm{He}_1(x_2). \tag{74}$$

F.1 Truncating $\zeta_k(Y)$

Throughout this section, let

$$\tilde{\lambda}_k^2 := \frac{\lambda_k^2}{\log(3/\lambda_k)^{k/2}}. \tag{75}$$

Lemma F.2. *Let $\mathsf{P} \in \mathcal{G}$. Then for any k, there exists a bounded function $\mathcal{T} : \mathbb{R} \to [-1, 1]$ such that $\mathbb{E}_{\mathsf{P}}[\mathcal{T}(Y)h_k(Z)] \gtrsim \tilde{\lambda}_k^2$ and $\mathbb{E}_{\mathsf{P}}[\mathcal{T}(Y)^2] \lesssim \tilde{\lambda}_k^2$.*

Proof. Let $k^\star = k^\star(\mathsf{P})$ and recall $\zeta_{k^\star}(y) := \mathbb{E}_{\mathsf{P}}[h_{k^\star}(Z)|Y = y]$ where $(Z, Y) \sim \mathsf{P}$. We will fix a truncation radius R and define $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ by $\mathcal{T}(y) := \frac{1}{R}\zeta_{k^\star}(y)\mathbf{1}_{|\zeta_{k^\star}(y)|\le R}$. Note that $|\mathcal{T}(y)| \le 1$ by definition. Then,

$$\begin{aligned} R \cdot \mathbb{E}_{\mathsf{P}}[\mathcal{T}(Y)h_{k^\star}(Z)] &= \mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)^2] - \mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)^2\mathbf{1}_{|\zeta_{k^\star}(Y)|\ge R}] \\ &= \lambda_{k^\star}^2 - \mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)^2\mathbf{1}_{|\zeta_{k^\star}(Y)|\ge R}]. \end{aligned} \tag{76}$$

The first term is nonzero because by the definition of $k^\star$. Therefore it suffices to prove that the second term vanishes as $R \to \infty$. Using Lemma I.4 and Markov's inequality, we can bound this second term by:

$$\begin{aligned} \left|\mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)^2\mathbf{1}_{|\zeta_{k^\star}(Y)|\ge R}]\right| &\le \sqrt{\mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)^2 h_{k^\star}(Z)^2]\mathbb{P}[|\zeta_{k^\star}(Y)| \ge R]} \\ &\lesssim \sqrt{\mathbb{E}[\zeta_k(Y)^2]\log(3/\mathbb{E}[\zeta_k(Y)^2])^{k^\star}} \cdot \frac{\mathbb{E}[\zeta_k(Y)^2]}{R} \\ &= \frac{\log(3/\lambda_{k^\star})^{k^\star/2}\lambda_{k^\star}^2}{R}. \end{aligned} \tag{77}$$

Therefore taking $R = C\log(3/\lambda_{k^\star})^{k^\star/2}$ for a sufficiently large constant C gives:

$$\mathbb{E}_{\mathsf{P}}[\mathcal{T}(Y)h_{k^\star}(Z)] \gtrsim \frac{\lambda_{k^\star}^2}{\log(3/\lambda_{k^\star})^{k^\star/2}}. \tag{78}$$

In addition, $\mathcal{T}$ also satisfies:

$$\mathbb{E}_{\mathsf{P}}[\mathcal{T}(Y)^2] \le \frac{1}{R}\mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)^2] \lesssim \frac{\lambda_{k^\star}^2}{\log(3/\lambda_{k^\star})^{k^\star/2}}. \tag{79}$$

■

F.2 Tensor Power Iteration

The following lemma shows that Tensor Power Iteration can be computed with $O(d)$ memory as it does not require storing the tensor $\boldsymbol{h}_k(x)$.

Lemma F.3 (Efficient Tensor Power Iteration)**.** *For any $v \in S^{d-1}$,*

$$\mathbf{He}_k(x)[v^{\otimes(k-1)}] = x\mathrm{He}_{k-1}(x \cdot v) - (k-1)v\mathrm{He}_{k-2}(x \cdot v). \tag{80}$$

Proof. Because $\mathbf{He}_k(x)$ tensorizes, we have that for $w \perp v$,

$$\mathbf{He}_k(x)[v^{\otimes(k-1)}] \cdot v = \mathrm{He}_k(x \cdot v) \quad \text{and} \quad \mathbf{He}_k(x)[v^{\otimes(k-1)}] \cdot w = (x \cdot w)\mathrm{He}_{k-1}(x \cdot v). \tag{81}$$

Therefore,

$$\begin{aligned} \mathbf{He}_k(x)[v^{\otimes(k-1)}] &= x\mathrm{He}_{k-1}(x \cdot v) + v[\mathrm{He}_k(x \cdot v) - (x \cdot v)\mathrm{He}_{k-1}(x \cdot v)] \\ &= x\mathrm{He}_{k-1}(x \cdot v) - (k-1)v\mathrm{He}_{k-2}(x \cdot v) \end{aligned} \tag{82}$$

by the two term recurrence for He_k. ■

Lemma F.4 (Tensor Power Iteration Convergence)**.** *Let $c, C = c(k), C(k)$ be constants depending only on k. Let $v \in S^{d-1}$ with $\alpha := v \cdot w^\star$. Let $\mathcal{D} = \{(x_i, y_i)\}_{i \in [n]}$ be a dataset of size n with $|y_i| \le 1$. Let*

$$\hat{v} = \frac{1}{n} \sum_{(x,y) \in \mathcal{D}} y \boldsymbol{h}_k(x)[v^{\otimes(k-1)}], \quad \text{and} \quad v' = \frac{\hat{v}}{\|\hat{v}\|} \ . \tag{83}$$

We will denote the effective signal strength by $\tilde{\beta}_k$:

$$\tilde{\beta}_k := \frac{\mathbb{E}[Y h_k(Z)]}{\sqrt{\mathbb{E}[Y^2] \log^k(3/\mathbb{E}[Y^2])}} \tag{84}$$

and we will assume without loss of generality that $\tilde{\beta}_k > 0$. Then for $\alpha \in [d^{-1/4}, 1/2]$ we have that if

$$n \ge \frac{Cd}{\tilde{\beta}_k^2 \alpha^{2k-4}}, \tag{85}$$

then with probability at least $1 - 2e^{-d^c}$, $v' \cdot w \ge 2(v \cdot w)$. In addition, for any $\epsilon > 0$ if

$$n \ge \frac{Cd}{\epsilon \tilde{\beta}_k^2 \alpha^{2k-2}}, \tag{86}$$

then with probability at least $1 - 2d^{-c}$, $v' \cdot w^\star \ge 1 - \epsilon$.

Proof. Let $\tilde{\lambda}^2 := \mathbb{E}[Y^2] \log^k(3/\mathbb{E}[Y^2])$. Note that $\mathbb{E}[\hat{v}] = w^\star \beta_k \alpha^{k-1}$. In addition for any w, we have by Lemma I.4 that

$$\mathbb{E}[(\hat{v} \cdot w)^2] - (\mathbb{E}[\hat{v}] \cdot w)^2 \le \frac{1}{n} \mathbb{E}[Y^2 \boldsymbol{h}_k(x)[v^{\otimes(k-1)}, w]^2] \lesssim \frac{\tilde{\lambda}^2}{n}. \tag{87}$$

Therefore by Lemma I.3, for any $w \in S^{d-1}$, with probability at least $1 - \delta$,

$$|(\hat{v} - \mathbb{E}[\hat{v}]) \cdot w| \lesssim \sqrt{\frac{\tilde{\lambda}^2 \log(1/\delta)}{n}} + \frac{\log(n/\delta)^{k/2}}{n} \lesssim \sqrt{\frac{\tilde{\lambda}^2 \log(1/\delta)}{n}}, \tag{88}$$

because $n \ge d$. Next, for any x let $x^\perp := P_{w,v}^\perp x$ and decompose $\hat{v}^\perp$ as:

$$\hat{v}^\perp = \frac{1}{\sqrt{k}} \cdot \frac{1}{n} \sum_{(x,y) \in \mathcal{D}} y x^\perp h_{k-1}(v \cdot x). \tag{89}$$

Then $x^\perp$ is independent of $y, h_{k-1}(v \cdot x)$ so this is equal in distribution to

$$N\left(0, \frac{1}{k} \cdot \frac{1}{n^2} \sum_{(x,y)\in\mathcal{D}} y^2 h_{k-1}(v\cdot x)^2 I_d\right). \tag{90}$$

Therefore by the standard χ^2 tail bound, with probability at least $1 - 2e^{-d}$,

$$\|v^\perp\|^2 \lesssim \frac{d}{n} \cdot \frac{1}{n} \sum_{(x,y)\in\mathcal{D}} y^2 h_{k-1}(v\cdot x)^2. \tag{91}$$

Again by Lemma I.3 we have that with probability at least $1-\delta$,

$$\|v^\perp\|^2 \lesssim \frac{d}{n} \cdot \left[\tilde{\lambda}^2 + \sqrt{\frac{\tilde{\lambda}\log(1/\delta)}{n}} + \frac{\log(n/\delta)^k}{n}\right]. \tag{92}$$

Therefore for $n \geq d$, we have with probability $1 - 2e^{-d^c}$ that $\|v^\perp\|^2 \lesssim \frac{\tilde{\lambda}^2 d}{n}$. In combination with the above bound on $|(\hat{v} - \mathbb{E}[\hat{v}]) \cdot w|$, this gives that with probability at least $1 - 2e^{-d^c}$,

$$\|\hat{v} - \mathbb{E}[\hat{v}]\|^2 \lesssim \frac{\tilde{\lambda}^2 d}{n}.$$

In the first case when $\alpha \in [d^{-1/4}, 1/4]$ and $n \gtrsim \frac{d}{\tilde{\beta}_k^2 \alpha^{2k-4}}$, this gives that with probability at least $1 - 2e^{-d^c}$,

$$\begin{aligned}
v' \cdot w^\star &= \frac{\hat{v}\cdot w^\star}{\|\hat{v}\|} \\
&\geq \frac{\beta_k \alpha^{k-1} + (\hat{v} - \mathbb{E}[\hat{v}]) \cdot w^\star}{\beta_k \alpha^{k-1} + \|\hat{v} - \mathbb{E}[\hat{v}]\|} \\
&\geq \frac{\frac{3}{4}\beta_k \alpha^{k-1}}{\beta_k \alpha^{k-1} + \|\hat{v} - \mathbb{E}[\hat{v}]\|} \\
&\geq \frac{\frac{3}{4}\beta_k \alpha^{k-1}}{\beta_k \alpha^{k-1} + \beta_k \alpha^{k-2}/8} \\
&= \frac{6\alpha}{8\alpha + 1} \\
&\geq 2\alpha
\end{aligned} \tag{93}$$

because $\alpha \leq 1/4$. In the second case when $n \gtrsim \frac{d}{\epsilon \tilde{\beta}_k^2 \alpha^{2k-2}}$, we have with probability at least $1 - 2e^{-d^c}$,

$$v' \cdot w^\star = \frac{\hat{v}\cdot w^\star}{\|\hat{v}\|} \geq \frac{\beta_k \alpha^{k-1} + (\hat{v} - \mathbb{E}[\hat{v}]) \cdot w^\star}{\beta_k \alpha^{k-1} + \|\hat{v} - \mathbb{E}[\hat{v}]\|} \geq \frac{\beta_k \alpha^{k-1}(1 - \epsilon/2)}{\beta_k \alpha^{k-1}(1 + \epsilon/2)} \geq 1 - \epsilon \tag{94}$$

which completes the proof. ■

F.3 Partial Trace

We begin by defining the un-normalized orthogonal polynomials for χ^2 random variables:

Definition F.5. Let $p_k^{(d)}(r)$ be defined by

$$p_k^{(d)}(r) = \sum_{j=0}^{k} \binom{k}{j} r^j (-1)^{k-j} \prod_{i=j}^{k-1} (d+2i). \tag{95}$$

These satisfy the following orthogonality relation:

Lemma F.6. *Let $r \sim \chi^2(d)$. Then,*

$$\mathbb{E}[p_j^{(d)}(r) p_k^{(d)}(r)] = \delta_{jk} k! 2^k \prod_{i=0}^{k-1} (d+2i). \tag{96}$$

Proof. For any $j \leq k$,

$$\begin{aligned}
\mathbb{E}[r^j p_k^{(d)}(r)] &= \sum_{l=0}^{k} \binom{k}{l} \mathbb{E}[r^{l+j}](-1)^{k-l} \prod_{i=l}^{k-1} (d+2i) \\
&= \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} \prod_{i=0}^{l+j-1} (d+2i) \prod_{i=l}^{k-1} (d+2i) \\
&= \prod_{i=0}^{k-1} (d+2i) \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} \prod_{i=l}^{l+j-1} (d+2i) \\
&= \prod_{i=0}^{k-1} (d+2i) \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} \binom{\frac{d}{2}+l+j-1}{j} 2^j j! \\
&= \mathbf{1}_{j=k} 2^k k! \prod_{i=0}^{k-1} (d+2i)
\end{aligned} \tag{97}$$

where the last line followed from the fact that the kth finite difference of the polynomial $g(l) = \binom{\frac{d}{2}+l+j-1}{j}$ is 0 unless $j = k$, in which case it is 1. ∎

These are related to the Hermite polynomials by the following lemma:

Lemma F.7. *For any $x \in \mathbb{R}^d$,*

$$p_k^{(d)}(\|x\|^2) = \mathbf{He}_{2k}(x)[I^{\otimes k}] \tag{98}$$

Proof. Note that both sides are monic polynomials in $\|x\|^2$. In addition, the right hand side is an orthogonal family of polynomials in $\|x\|^2$ because for $j \neq k$,

$$\mathbb{E}\big[\boldsymbol{h}_{2j}(x)[I^{\otimes j}] \boldsymbol{h}_{2k}(x)[I^{\otimes k}]\big] = 0 \tag{99}$$

because $\mathbb{E}[\boldsymbol{h}_{2j} \otimes \boldsymbol{h}_{2k}] = 0$. Because the set of monic orthogonal polynomials is unique, we must have that $p_k^{(d)}(\|x\|^2) = (2k!)^{1/2} \boldsymbol{h}_{2k}(x)[I^{\otimes k}]$ for all x. ∎

Lemma F.8 (Efficient Computation of Partial Trace)**.** *If $k = 2j + 1$,*

$$\mathbf{He}_k(x)[I^{\otimes j}] = p_j^{(d+2)}(\|x\|^2)x \tag{100}$$

and if $k = 2j + 2$. Then,

$$\mathbf{He}_k(x)[I^{\otimes j}] = p_j^{(d+4)}(\|x\|^2)xx^T - p_j^{(d+2)}(\|x\|^2)I_d. \tag{101}$$

Proof. We start with Lemma F.7 for $j + 1$:

$$\mathbf{He}_{2j+2}(x)[I^{\otimes(j+1)}] = p_{j+1}^{(d)}(\|x\|^2). \tag{102}$$

Differentiating both sides with respect to x in the v direction gives:

$$(2j + 2)\,\mathrm{Sym}(v \otimes \mathbf{He}_{2j+1}(x))[I^{\otimes j+1}] = (2j + 2)(x \cdot v)p_j^{(d+2)}(\|x\|^2). \tag{103}$$

Note that the left hand side is simply equal to $\mathbf{He}_{2j+1}(x)[I^{\otimes j}] \cdot v$, so and rearranging gives the result for k odd. Next, we will differentiate again in the v direction. Then we have that:

$$(2j + 1)\,\mathrm{Sym}(v \otimes v \otimes \mathbf{He}_{2j}(x))[I^{\otimes j+1}] \tag{104}$$

$$= (2j)(x \cdot v)^2 p_{j-1}^{(d+4)}(\|x\|^2) + p_j^{(d+2)}(\|x\|^2). \tag{105}$$

Of the $(2j+1)!!$ pairings on the left hand side, $(2j-1)!!$ pair v with itself and then pair all indices of $\mathbf{He}_{2j}(x)$ while $2j(2j-1)!!$ pair $2j-2$ indices of $\mathbf{He}_{2j}(x)$ and the remaining two with v. Therefore the left hand side is equal to:

$$p_j^{(d)}(\|x\|^2) + (2j)v^T\mathbf{He}_{2j}(x)[I^{\otimes(j-1)}]v. \tag{106}$$

Therefore we must have

$$v^T\mathbf{He}_{2j}(x)[I^{\otimes(j-1)}]v = (x \cdot v)^2 p_{j-1}^{(d+4)}(\|x\|^2) + \frac{p_j^{(d+2)}(\|x\|^2) - p_j^d(\|x\|^2)}{2j}. \tag{107}$$

Because $p_j^{(d+2)} - p_j^{(d)} = -(2j)p_{j-1}^{(d+2)}$, this reduces to

$$v^T\mathbf{He}_{2j}(x)[I^{\otimes(j-1)}]v = (x \cdot v)^2 p_{j-1}^{(d+4)}(\|x\|^2) - p_{j-1}^{(d+2)}. \tag{108}$$

As this is true for any $v \in S^{d-1}$, this completes the proof. ∎

F.4 The Even Case

First we will start with the even case. We will show that $v := v_1(M_n)$ has good alignment with $w^\star$.

Lemma F.9. *Let $k \geq 4$ and let M_n be the partial trace matrix in Algorithm 1 and assume that $|y_i| \leq 1$ for all i. Define the effective signal strength*

$$\tilde{\beta}_k := \frac{\mathbb{E}[Yh_k(Z)]}{\sqrt{\mathbb{E}[Y^2]\log^k(3/\mathbb{E}[Y^2])}}. \tag{109}$$

Then for $k \geq 4$, with probability at least $1 - 2e^{-d^c}$,

$$\|M_n - \mathbb{E}[M_n]\|_2 \lesssim \sqrt{\frac{\beta_k^2 d^{k/2}}{\tilde{\beta}_k^2 n}}. \tag{110}$$

In addition, for

$$n \gtrsim \frac{d^{k/2}}{\epsilon \tilde{\beta}_k^2}, \tag{111}$$

we have that $v = v_1(M_n)$ satisfies $(v \cdot w^\star)^2 \geq 1 - \epsilon$.

Proof. Let $\eta = \frac{\beta_k^2}{\tilde{\beta}_k^2} = \mathbb{E}[Y^2] \log^k(3/\mathbb{E}[Y^2])$. We will use Lemma I.5. First we compute $\sigma^2 := \|\mathbb{E}[(M_n - M)^2]\|_2$.

$$\begin{aligned}
\sigma^2 &:= \|\mathbb{E}[(M_n - M)^2]\|_2 \\
&= \frac{1}{n}\|\mathbb{E}[(M_1 - M)^2]\|_2 \\
&\lesssim \frac{1}{n} \cdot \|\mathbb{E}[M_1^2]\|_2 \\
&= \frac{1}{n} \cdot \left\| \mathbb{E}_{X,Y}\left[Y^2 \boldsymbol{h}_{k^\star}(X) \left[I^{\otimes \frac{k^\star - 2}{2}} \right]^2 \right] \right\|_2.
\end{aligned} \tag{112}$$

Now by Lemma I.4, for any $v \in S^{d-1}$,

$$\begin{aligned}
&\mathbb{E}_{X,Y}\left[Y^2 \left\| \boldsymbol{h}_{k^\star}(X) \left[I^{\otimes \frac{k^\star - 2}{2}}, v \right] \right\|^2 \right] \\
&\lesssim \eta \cdot \mathbb{E}\left[\left\| \boldsymbol{h}_{k^\star}(X) \left[I^{\otimes \frac{k^\star - 2}{2}}, v \right] \right\|^2 \right] \\
&\leq \eta d \left\| \mathrm{Sym}\left(I^{\otimes \frac{k^\star - 2}{2}} \otimes v \right) \right\|^2 \\
&\leq \eta d^{k/2}.
\end{aligned} \tag{113}$$

Therefore,

$$\sigma^2 \lesssim \frac{\eta d^{k/2}}{n}. \tag{114}$$

We bound σ_* similarly:

$$\begin{aligned}
\sigma_*^2 &:= \sup_{v,w\in S^{d-1}} \mathbb{E}[(v^T(M_n - M)w)^2] \\
&= \frac{1}{n}\mathbb{E}[(u^T(M_1 - M)v)^2] \\
&\lesssim \frac{1}{n}\mathbb{E}[(u^T M_1 v)^2] \\
&= \frac{1}{n}\mathbb{E}_{X,Y}[Y\boldsymbol{h}_{k^\star}(X)[I^{\otimes\frac{k^\star-2}{2}}, u, v]^2] \\
&\le \frac{1}{n}\mathbb{E}_X[\boldsymbol{h}_{k^\star}(X)[I^{\otimes\frac{k^\star-2}{2}}, u, v]^2] \\
&= \frac{1}{n}\cdot\left\|\mathrm{Sym}\left(I^{\otimes\frac{k^\star-2}{2}}\otimes u\otimes v\right)\right\|_F^2 \\
&\le \frac{1}{n}\cdot\|I\|_F^{k^\star-2} \\
&= \frac{d^{\frac{k^\star}{2}-1}}{n}.
\end{aligned} \tag{115}$$

Next, by Lemma F.12 and Lemma I.6 we have that

$$\overline{R} \lesssim \frac{d^{\frac{k^\star+2}{4}} + \log(n)^{\frac{k^\star}{2}} d^{\frac{k^\star}{4}}}{n} \tag{116}$$

and for any $\delta' \ge 0$, with probability at least $1-\delta'$ we have that

$$\max_{i\in[n]} \frac{\|M_i\|_2}{n} \lesssim \frac{d^{\frac{k^\star+2}{4}} + \log(n/\delta)^{\frac{k^\star}{2}} d^{\frac{k^\star}{4}}}{n}. \tag{117}$$

Therefore with probability at least $1 - n\exp\left(-d^{1/k^\star}\right)$, we have

$$\max_{i\in[n]} \frac{\|M_i\|_2}{n} \lesssim \frac{d^{\frac{k^\star+2}{4}}}{n}. \tag{118}$$

Now we can set

$$R = C\sqrt{\frac{d^{\frac{k^\star+2}{4}}}{n}\cdot\frac{d^{\frac{k^\star}{4}}}{n^{1/2}}} = C\cdot\frac{d^{\frac{k^\star+1}{2}}}{n^{3/4}} \tag{119}$$

for a sufficiently large constant C and apply Lemma I.5 to get that with probability at least $1 - de^{-t} - ne^{-d^{1/k}}$,

$$\|M_n - \mathbb{E}M_n\|_2 \le 2\sigma + O\left(\frac{d^{\frac{k^\star-2}{4}}}{n^{1/2}}t^{1/2} + \frac{d^{\frac{3k^\star+1}{12}}}{n^{7/12}}t^{2/3} + \frac{d^{\frac{k^\star+2}{4}}}{n}t\right). \tag{120}$$

Now for $k^\star \ge 4$ we can set $t = d^c$ for $c < 1/8$ to get that with probability at least $1 - de^{-d^c} - ne^{-d^{1/k}}$, for any $n \ge d^{\frac{k^\star}{2}}$ we have

$$\|M_n - \mathbb{E}M_n\|_2 \lesssim \sigma = \sqrt{\frac{\eta d^{k/2}}{n}}. \tag{121}$$

The conclusion for $v = v_1(M_n)$ now directly follows from the Davis-Kahan inequality. ■

Lemma F.10. *Assume that $(Z, Y) \sim \mathsf{P}$ and $k^\star(\mathsf{P}) > 2$. Then for any $k \geq 4$, if we define Σ by*

$$\Sigma := \frac{d^{(k-2)/2}}{k(k-1)!!} \cdot \left[\mathbb{E}[Y^2 \mathrm{He}_4(Z)] w^{\star \otimes 4} + 4\mathbb{E}[Y^2 \mathrm{He}_2(Z)] T + 2\mathbb{E}[Y^2] \operatorname{Sym}_2\right] \tag{122}$$

where $T = 6 \operatorname{Sym}(w^{\star \otimes 2} \otimes I) - w^{\star \otimes 2} \otimes I - I \otimes w^{\star \otimes 2}$, then

$$\|\mathbb{E}[M_1 \otimes M_1] - \Sigma\|_{op} \lesssim d^{\frac{k-4}{2}}. \tag{123}$$

Proof. We will temporarily switch to the un-normalized Hermite polynomials He_k. If $k = 2j + 2$, this is equal to

$$\begin{aligned}
&k!\mathbb{E}[Y^2 \boldsymbol{h}_k(X)[I^{\otimes \frac{k-2}{2}}] \otimes \boldsymbol{h}_k(X)[I^{\otimes \frac{k-2}{2}}]] \\
&\mathbb{E}[Y^2 \mathbf{He}_k(X)[I^{\otimes \frac{k-2}{2}}] \otimes \mathbf{He}_k(X)[I^{\otimes \frac{k-2}{2}}]] \\
&= \mathbb{E}\Big[Y^2 \Big[p_j^{(d+4)}(\|x\|) x x^T - p_j^{(d+2)}(\|x\|^2) I\Big] \otimes \Big[p_j^{(d+4)}(\|x\|) x x^T - p_j^{(d+2)}(\|x\|^2) I\Big]\Big] \\
&= \mathbb{E}\Big[Y^2 p_j^{(d+4)}(\|x\|)^2 x^{\otimes 4}\Big] \\
&\quad - \mathbb{E}\Big[Y^2 p_j^{(d+4)}(\|x\|) p_j^{(d+2)}(\|x\|) x^{\otimes 2} \otimes I\Big] \\
&\quad - \mathbb{E}\Big[Y^2 p_j^{(d+4)}(\|x\|) p_j^{(d+2)}(\|x\|) I \otimes x^{\otimes 2}\Big] \\
&\quad + \mathbb{E}\Big[Y^2 p_j^{(d+2)}(\|x\|)^2 I \otimes I\Big].
\end{aligned} \tag{124}$$

We will now compute this term by term. We will use $O(d^{j-1})$ to refer to error terms whose tensor operator norms are bounded by $O(d^{j-1})$. For the first term, we have that

$$\begin{aligned}
&\mathbb{E}\Big[Y^2 p_j^{(d+4)}(\|x\|)^2 x^{\otimes 4}\Big] \\
&= 2^j j! d^j \mathbb{E}[Y^2 x^{\otimes 4}] + O(d^{j-1}) \\
&= 2^j j! d^j [\mathbb{E}[Y^2 \mathrm{He}_4(Z)] w^{\star \otimes 4} \\
&\qquad + 6\mathbb{E}[Y^2 \mathrm{He}_2(Z)] \operatorname{Sym}(w^{\star \otimes 2} \otimes I) \\
&\qquad + 3\mathbb{E}[Y^2] \operatorname{Sym}(I \otimes I)] + O(d^{j-1}).
\end{aligned} \tag{125}$$

For the second and third terms,

$$\begin{aligned}
&\mathbb{E}\Big[Y^2 p_j^{(d+4)}(\|x\|) p_j^{(d+2)}(\|x\|) x^{\otimes 2} \otimes I\Big] \\
&= 2^j j! d^j \mathbb{E}[Y^2 x^{\otimes 2} \otimes I] + O(d^{j-1}) \\
&= 2^j j! d^j [\mathbb{E}[Y^2 \mathrm{He}_2(Z)] w^{\star \otimes 2} \otimes I + \mathbb{E}[Y^2] I \otimes I] + O(d^{j-1}).
\end{aligned} \tag{126}$$

Finally for the last term we have:

$$\begin{aligned}
&\mathbb{E}\Big[Y^2 p_j^{(d+2)}(\|x\|)^2 I \otimes I\Big] \\
&= 2^j j! d^j \mathbb{E}[Y^2] I \otimes I + O(d^{j-1}).
\end{aligned} \tag{127}$$

Renormalizing by $k!$ to reduce back to the normalized Hermite polynomials h_k gives:

$$\begin{aligned}
&\mathbb{E}[Y^2 \boldsymbol{h}_k(X)[I^{\otimes \frac{k-2}{2}}] \otimes \boldsymbol{h}_k(X)[I^{\otimes \frac{k-2}{2}}]] \\
&= \frac{d^j}{k(k-1)!!} \cdot \Big[\mathbb{E}[Y^2 \mathrm{He}_4(Z)] w^{\star \otimes 4} \\
&\quad + \mathbb{E}[Y^2 \mathrm{He}_2(Z)][6\,\mathrm{Sym}(w^{\star \otimes 2} \otimes I) - w^{\star \otimes 2} \otimes I - I \otimes w^{\star \otimes 2}] \\
&\quad + \mathbb{E}[Y^2][3\mathrm{Sym}(I \otimes I) - I \otimes I]\Big] + O(d^{j-1}) \\
&= \frac{d^j}{k(k-1)!!} \cdot \Big[\mathbb{E}[Y^2 \mathrm{He}_4(Z)] w^{\star \otimes 4} + 4\mathbb{E}[Y^2 \mathrm{He}_2(Z)] T + 2\mathbb{E}[Y^2]\,\mathrm{Sym}_2\Big] + O(d^{j-1}) \qquad (128)
\end{aligned}$$

where $T = 6\,\mathrm{Sym}(w^{\star \otimes 2} \otimes I) - w^{\star \otimes 2} \otimes I - I \otimes w^{\star \otimes 2}$. ■

F.5 The Odd Case

Next, we will study the odd case. This is much simpler than the even case as it doesn't require matrix concentration. However, it is not possible to directly reach ϵ error with this step. We therefore analyze the sample complexity for reaching $v \cdot w^\star \geq d^{-1/4}$:

Lemma F.11. *Let k be odd and let v_n be the vector from stage 1 of Algorithm 1. Assume that $|y_i| \leq 1$. Denote the effective signal strength $\tilde{\beta}_k$ by*

$$\tilde{\beta}_k := \frac{\mathbb{E}[Y h_k(Z)]}{\sqrt{\mathbb{E}[Y^2] \log^k(3/\mathbb{E}[Y^2])}}. \qquad (129)$$

Then for a sufficiently large constant $C = C(k)$, if

$$n = \frac{C d^{k/2}}{\tilde{\beta}_k^2} \qquad (130)$$

with probability at least $1 - 2e^{-d^c}$, $\frac{v_n}{\|v_n\|} \cdot w^\star \geq d^{-1/4}$. Furthermore, if $n \geq C\frac{d^{\frac{k+1}{2}}}{\epsilon \tilde{\beta}_k^2}$, we have that with probability at least $1 - 2e^{-d^c}$, $v \cdot w^\star \geq 1 - \epsilon$.

Proof. As in Lemma F.9, let $\eta := \lambda_k^2 \log^k(3/\lambda_k)$. We will begin by computing the variance. Note that for any $v \in S^{d-1}$, by Lemma I.4 we have:

$$\begin{aligned}
&\mathbb{E}[Y^2 \boldsymbol{h}_k(X)[I^{\otimes \frac{k-1}{2}}, v]^2] \\
&\lesssim \eta \mathbb{E}[\boldsymbol{h}_k(X)[I^{\otimes \frac{k-1}{2}}, v]^2] \\
&\leq \eta d^{\frac{k-1}{2}}. \qquad (131)
\end{aligned}$$

Therefore by Lemma I.3, with probability at least $1 - \delta$ we have

$$(v_n - \mathbb{E}[v_n]) \cdot w \lesssim \sqrt{\frac{\eta \log(1/\delta) d^{\frac{k-1}{2}}}{n}} + \frac{\log(n/\delta)^{k/2+1} d^{\frac{k-1}{2}}}{n}. \qquad (132)$$

For the norm, recall that if $k = 2j + 1$,

$$\sqrt{k!}\boldsymbol{h}_k(x)[I^{\otimes \frac{k-1}{2}}] = xp_j^{(d)}(\|x\|^2). \tag{133}$$

Let $\bar{x}_i := \frac{x_i}{\|x_i\|}$ and let $\bar{x}_i^\perp = \bar{x}_i - w^\star(\bar{x}_i \cdot w^\star)$. Then $\bar{x}_i$ is independent of $\|x_i\|$ and $x_i \cdot w^\star$. Therefore viewed as a function of $\{\bar{x}_i^\perp\}$, $v_n^\perp$ is a sub-Gaussian vector with constant $\sigma^2 = \frac{1}{n}\sum_{i=1}^n y_i^2 p_j^{(d)}(\|x\|)^2$. Therefore with probability at least $1 - 2e^{-d}$, $\|v_n^\perp\| \lesssim \sigma\sqrt{d}$ so it suffices to bound σ. We have by Lemma I.3,

$$\sigma^2 \lesssim \eta d^j + \tilde{O}(d^j/\sqrt{n}). \tag{134}$$

Therefore, with probability at least $1 - 2e^{-d^c}$, $\sigma^2 \lesssim \lambda_k^2 \log^k(3/\lambda_k) d^j$ and

$$\|v_n^\perp\| \lesssim \sqrt{\frac{\eta d^{\frac{k+1}{2}}}{n}}. \tag{135}$$

Combining this with the bound of $(v_n - \mathbb{E}[v_n]) \cdot w^\star$ this gives with probability at least $1 - e^{-d^c}$,

$$\|v_n - \mathbb{E}[v_n]\| \lesssim \sqrt{\frac{\eta d^{\frac{k+1}{2}}}{n}}. \tag{136}$$

Combining everything gives that when $n = Cd^{k/2}/\tilde{\beta}_k^2$,

$$\frac{v_n}{\|v_n\|} \cdot w^\star \geq \frac{\beta_k + (v_n - \mathbb{E}[v_n]) \cdot w^\star}{\beta_k + \|v_n - \mathbb{E}[v_n]\|} \gtrsim \sqrt{\frac{\beta_k^2 n}{\eta d^{\frac{k+1}{2}}}} = \sqrt{\frac{\tilde{\beta}_k^2 n}{d^{\frac{k+1}{2}}}} = Cd^{-1/4}. \tag{137}$$

In addition, when $n = Cd^{\frac{k+1}{2}}/(\epsilon\tilde{\beta}_k^2)$,

$$\frac{v_n}{\|v_n\|} \cdot w^\star \geq \frac{\beta_k + (v_n - \mathbb{E}[v_n]) \cdot w^\star}{\beta_k + \|v_n - \mathbb{E}[v_n]\|} \geq \frac{\beta_k(1 - \epsilon/2)}{\beta_k(1 + \epsilon/2)} \geq 1 - \epsilon. \tag{138}$$

■

F.6 Proof of Theorem 4.1

We are now ready to prove Theorem 4.1:

Proof of Theorem 4.1. By Lemma F.2 there exists a truncation of ζ_k, $\mathcal{T} : \mathbb{R} \to [-1, 1]$ such that the effective signal strength $\tilde{\lambda}_k$ satisfies:

$$\tilde{\lambda}_k^2 = \frac{\mathbb{E}_{\mathsf{P}}[\mathcal{T}(Y)h_{k^\star}(Z)]^2}{\mathbb{E}_{\mathsf{P}}[\mathcal{T}(Y)^2]\log^k(3/\mathbb{E}_{\mathsf{P}}[\mathcal{T}(Y)^2])} \gtrsim \frac{\lambda_{k^\star}^2}{\log(3/\lambda_k)^{2k^\star}}. \tag{139}$$

We will first show that the output of the first stage satisfies $v \cdot w^\star = \Theta(1)$. For $k = 1$, this follows directly from Lemma F.11. For $k = 2$, the result follows from [MM18, Theorem 2]. For $k \geq 2$ with k even, the result follows from Lemma F.9. Finally, when $k \geq 3$ with k odd we have

that after the partial trace warm start, by Lemma F.11 $v \cdot w^\star \geq d^{-1/4}$. Then until $v \cdot w^\star = 1/4$, each step of tensor power iteration will double $v \cdot w^\star$ with $n \gtrsim \frac{d}{(v \cdot w^\star)^{2k-4} \tilde{\beta}_k^2}$ by Lemma F.4. This will converge to $v \cdot w^\star = 1/4$ in $\log\left(d^{1/4}\right) \leq \log(d)$ steps. Finally, by Lemma F.4 one more step of tensor power iteration with $n \geq \frac{d}{\epsilon \tilde{\beta}_k^2}$ gives $(v \cdot w^\star)^2 \geq 1 - \epsilon$. As every step happens with probability at least $1 - 2e^{-d^c}$, a union bound gives that the final success probability is also $1 - 2e^{-d^c}$ for constant c depending only on k. ■

F.7 Concentration

Lemma F.12. *Let $X \sim \gamma_d$, let k be an even number, and define $M := \boldsymbol{h}_k(x) \left[I^{\otimes \frac{k-2}{2}}\right]$. Then,*

$$\mathbb{E}[\|M_2\|_2^2]^{1/2} \leq d^{\frac{k+2}{4}} \quad \text{and} \quad \mathbb{E}\left[|\|M\|_2 - \mathbb{E}\|M\|_2|^p\right]^{1/p} \lesssim p^{k/2} d^{\frac{k}{4}}. \tag{140}$$

Proof. For the first inequality we have:

$$\begin{aligned}
\mathbb{E}[\|M_2\|_2^2] &\leq \mathbb{E}[\|M_2\|_F^2] \\
&= \sum_{i,j} \mathbb{E}[\boldsymbol{h}_k(X)[I^{\otimes \frac{k-2}{2}}, e_i, e_j]^2] \\
&= \sum_{i,j} \left\|\mathrm{Sym}\left(I^{\otimes \frac{k-2}{2}} \otimes e_i \otimes e_j\right)\right\|_F^2 \\
&\leq d^2 \cdot d^{\frac{k-2}{2}} = d^{\frac{k+2}{2}}.
\end{aligned} \tag{141}$$

For the moment bound, first note that

$$\mathbb{E}[\mathrm{Tr}(M)^2] = \mathbb{E}[\boldsymbol{h}_k(X)[I^{\otimes k/2}]^2] \leq \|I\|_F^k = d^{k/2}. \tag{142}$$

Next, note that by symmetry, there exist polynomials p, q of degree at most $\frac{k^\star}{2} - 1$ such that $M = p(\|x\|^2)xx^T + q(\|x\|)^2 I$. We can expand $\mathrm{Tr}(M)$ as:

$$\mathrm{Tr}(M) = p(\|x\|^2)\|x\|^2 + q(\|x\|^2)d. \tag{143}$$

Therefore both $p(\|x\|^2)\|x\|^2$ and $q(\|x\|^2)d$ must have variance bounded by $d^{k/2}$. By Gaussian hypercontractivity, they also have p norms bounded by $(p-1)^{k/2} d^{k/4}$. Then,

$$\begin{aligned}
\|M\|_2 &= \max\left(|q(\|x\|^2)|, |p(\|x\|^2)\|x\|^2 + q(\|x\|^2)|\right) \\
&\leq |p(\|x\|^2)|\|x\|^2 + |q(\|x\|^2)|
\end{aligned} \tag{144}$$

so letting C denote the mean of the right hand side, if we subtract C from both sides we get that

$$\mathbb{E}[|\|M\|_2 - C|^p]^{1/p} \lesssim p^{k/2} d^{k/4}. \tag{145}$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[|\|M\|_2 - \mathbb{E}\|M\|_2|^p\right]^{1/p} &\leq \mathbb{E}[|\|M\|_2 - C|^p]^{1/p} + |\mathbb{E}\|M\|_2 - C| \\
&\leq 2\mathbb{E}[|\|M\|_2 - C|^p]^{1/p} \\
&\lesssim p^{k/2} d^{k/4}
\end{aligned} \tag{146}$$

where the second to last line follows from Jensen's inequality. ■

Lemma F.13. *Let $X \sim \gamma_d$, let k be an even number, let $v \in S^{d-1}$ and define $M := \boldsymbol{h}_k(x)\left[I^{\otimes \frac{k-3}{2}}, v\right]$. Then,*

$$\mathbb{E}[\|M_2\|_2^2]^{1/2} \le d^{\frac{k+1}{4}} \quad \text{and} \quad \mathbb{E}\left[|\|M\|_2 - \mathbb{E}\|M\|_2|^p\right]^{1/p} \lesssim p^{k/2} d^{\frac{k-1}{4}}. \tag{147}$$

Proof. As above we have

$$\begin{aligned}
\mathbb{E}[\|M_2\|_2^2] &\le \mathbb{E}[\|M_2\|_F^2] \\
&= \sum_{i,j} \mathbb{E}[\boldsymbol{h}_k(X)[I^{\otimes \frac{k-3}{2}}, v, e_i, e_j]^2] \\
&= \sum_{i,j} \left\| \mathrm{Sym}\left(I^{\otimes \frac{k-3}{2}} \otimes v \otimes e_i \otimes e_j\right)\right\|_F^2 \\
&\le d^2 \cdot d^{\frac{k-3}{2}} = d^{\frac{k+1}{2}}.
\end{aligned} \tag{148}$$

In addition, As above we have

$$\mathbb{E}[\mathrm{Tr}(M)^2] = \mathbb{E}[\boldsymbol{h}_k[I^{\otimes \frac{k-1}{2}}, v]^2] \le \|I\|_F^{k-1} = d^{\frac{k-1}{2}}. \tag{149}$$

The remainder of the proof is identical to the proof of Lemma F.12. ∎

F.8 Proofs for Unknown P Learning

Lemma F.14 (Unknown P label transformation)**.** *Assume $M \ge N$ and Assumption 4.3. Let $\theta \sim \mathrm{Unif}(\mathcal{S}^{M-1})$ and consider $\Psi = \sum_{l=1}^M \theta_l \phi_l$. Let $R > 0$ and define $\tilde{\mathcal{T}}(y) := \frac{1}{R}\Psi(y)\mathbf{1}_{|\Psi(y)|\le R}$. Then with probability greater than $1-\delta$ over the draw of θ, for $R = \frac{3^{k^\star} 4\sqrt{M}}{\lambda_{k^\star}\delta\sqrt{1-\varepsilon_M}}$ we have*

$$\tilde{\eta} := \left|\mathbb{E}_{\mathsf{P}}[\tilde{\mathcal{T}}(Y) h_{k^\star}(Z)]\right| \gtrsim \delta^2 \lambda_{k^\star}^2 > 0\ , \tag{150}$$

where $\gtrsim$ hides constants in $k^\star$ and appearing in Assumption 4.3.

Proof. Let $A_M \zeta_{k^\star} = \sum_{l \le M} \upsilon_l \phi_l$ the $L^2(\mathbb{R}, \mathsf{P}_y)$-orthogonal projection of $\zeta_{k^\star}$ onto the space spanned by degree-M polynomials. We have

$$\begin{aligned}
R\,\mathbb{E}_{\mathsf{P}}[\tilde{\mathcal{T}}(Y) h_{k^\star}(Z)] &= \langle \Psi, \zeta_{k^\star}\rangle_{\mathsf{P}_y} - \mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)\Psi(Y)\mathbf{1}_{|\Psi(Y)|\ge R}] \\
&= \langle \Psi, A_M\zeta_{k^\star}\rangle_{\mathsf{P}_y} - \mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)\Psi(Y)\mathbf{1}_{|\Psi(Y)|\ge R}] \\
&= \|A_M\zeta_k\| \left\langle \Psi, \frac{A_M\zeta_{k^\star}}{\|A_M\zeta_k\|}\right\rangle_{\mathsf{P}_y} - \mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)\Psi(Y)\mathbf{1}_{|\Psi(Y)|\ge R}]\ .
\end{aligned} \tag{151}$$

Now, following the proof of Lemma F.2 we bound the second term in the RHS:

$$\begin{aligned}
\left|\mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)\Psi(Y)\mathbf{1}_{|\Psi(Y)|\ge R}]\right| &\le \sqrt{\|\Psi(Y)\|_{\mathsf{P}_y}^2 \mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)^2\mathbf{1}_{|\Psi(Y)|\ge R}]} \\
&= \sqrt{\mathbb{E}_{\mathsf{P}}[\zeta_{k^\star}(Y)^2\mathbf{1}_{|\Psi(Y)|\ge R}]} \\
&\le \sqrt{\mathbb{E}_{\mathsf{P}}[h_{k^\star}(Z)^4]\mathbb{P}[|\Psi(Y)|\ge R]} \\
&\le \frac{3^{k^\star}\sqrt{\mathbb{E}[\Psi(Y)^2]}}{R} = \frac{3^{k^\star}}{R}\ ,
\end{aligned} \tag{152}$$

where we have used the fact that $\|\Psi\| = 1$.

Now, thanks to Assumption 4.3, and $M \geq N$, there exists $\kappa > 0$ such that $\|A_M \zeta_{k^\star}\|^2 = \lambda_{k^\star}^2(1-\varepsilon_M) \geq \kappa \lambda_{k^\star}^2$. We thus obtain

$$\begin{aligned}\left|\mathbb{E}_{\mathsf{P}}[\tilde{\mathcal{T}}(Y)h_{k^\star}(Z)]\right| &\geq \frac{1}{R}\lambda_{k^\star}\sqrt{\kappa}\left|\left\langle \Psi, \frac{A_M\zeta_{k^\star}}{\|A_M\zeta_k\|}\right\rangle_{\mathsf{P}_y}\right| - \frac{3^{k^\star}}{R^2} \\ &= \frac{\lambda_{k^\star}\sqrt{\kappa}}{R}|\theta\cdot v| - \frac{3^{k^\star}}{R^2}\ . \end{aligned} \tag{153}$$

Finally, we conclude with a basic anti-concentration property of the uniform measure on $\mathcal{S}_{M-1}$:

Lemma F.15 (Anti-Concentration of the Uniform measure, [BBSS22, Lemma A.7])**.** *Let $\theta \sim \mathrm{Unif}(\mathcal{S}_{M-1})$ and $\theta_0 \in \mathcal{S}_{M-1}$ arbitrary. For any $\epsilon > 0$, we have $\mathbb{P}[|\theta\cdot\theta_0| \leq \epsilon] \leq 4\epsilon\sqrt{M}$.*

Putting everything together, and picking $\epsilon = \delta/(4\sqrt{M})$ in Lemma F.15, we obtain that with probability greater than $1-\delta$, for $R \geq \frac{3^{k^\star}4\sqrt{M}}{\lambda_{k^\star}\delta\sqrt{\kappa}}$,

$$\begin{aligned}\left|\mathbb{E}_{\mathsf{P}}[\tilde{\mathcal{T}}(Y)h_{k^\star}(Z)]\right| &\geq \frac{\lambda_{k^\star}\delta\sqrt{\kappa}}{4\sqrt{M}R} - \frac{3^{k^\star}}{R^2} \\ &\geq \frac{\lambda_{k^\star}\delta\sqrt{\kappa}}{8\sqrt{M}R}\ . \end{aligned} \tag{154}$$

■

Algorithm 2: Partial Trace Algorithm, unknown P and $k^\star$

Input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, largest degree M, largest exponent $k^\star$, signal strength $\lambda_{k^\star}$, basis $(\phi_l)_{l\leq M}$

Set $R = C/\lambda_{k^\star}^2$ and $\tilde{R} = \tilde{C}$.

Split $\mathcal{D}$ into train $\mathcal{D}'$ and validation $\tilde{D}$ such that $|\tilde{D}| = L$.

Draw random $\theta \in \mathrm{Unif}(\mathcal{S}_{M-1})$ and form $\Psi = \sum_{l\leq M}\theta_l\phi_l$.

for $k \leq k^\star$ **do**

  Run Algorithm 1 on $\mathcal{D}'$ with $\mathcal{T} = R^{-1}\Psi\mathbf{1}_{|\Psi|\leq R}$ to obtain $\hat{w}_k$.

  Compute $F_k = \frac{1}{L}\sum_{l=1}^L \Psi(y_l)h_k(x_l\cdot\hat{w}_k)\mathbf{1}_{|\Psi(y_l)|\leq\tilde{R}}$.

end

Define $\hat{k} = \arg\max_k |F_k|$.

Output: $\hat{w}_{\hat{k}}$

Corollary 4.4 (Partial Trace for unknown P and $k^\star$)**.** *Let $\{(x_i, y_i)\}_{i=1}^n$ be i.i.d. samples from $\mathbb{P}_{w^\star,\mathsf{P}}$ with $k^\star(\mathsf{P}) = k^\star$. Then, under Assumption 4.3, if $n \geq \Omega(\lambda_{k^\star}^2 d^{k^\star/2} + d/\epsilon^2)$, $L \gtrsim \delta^{-4}\log(1/\tilde{\delta})$, the procedure described in Algorithm 2 with $K = k^\star$ returns $\hat{w}$ satisfying $(\hat{w}\cdot w^\star)^2 \geq 1-\epsilon^2$ with probability greater than $1 - e^{-d^\kappa} - \delta - 2\tilde{\delta}k^\star$.*

Proof. The proof is adapted from [DH18, Theorem 14].

Theorem 4.1 together with Lemma F.14 ensures that, provided $n \gg \frac{d^{k^\star/2}}{\delta^2 \lambda_{k^\star}^2}$, with probability greater than $1 - \delta - e^{-d^\kappa}$, $\|\hat{w}_{k^\star} - w^\star\| \lesssim \sqrt{d^{k^\star/2}/n}$, as well as $|\langle \Psi, \zeta_{k^\star}\rangle| \geq C\lambda_{k^\star}\delta$.

We now study the accuracy of our goodness-of-fit statistic:

Lemma F.16 (Concentration of Goodness-of-Fit Statistic)**.** *For any $k \in \{1, k^\star\}$, $\tilde{R} > 0$ and $\delta > 0$, we have*

$$\mathbb{P}_{\mathcal{D}'}\left\{ \left| F_k - (\hat{w}_k \cdot w^\star)^k \langle \Psi, \zeta_k \rangle_{\mathsf{P}_y} \right| \leq \frac{3^k}{\tilde{R}} + C_K \frac{\tilde{R}\sqrt{\log(1/\delta)}}{\sqrt{L}} + \frac{\tilde{R}}{L} \log(1/\delta) \log(L/\delta)^{k/2} \right\} \geq 1 - 2\delta \ . \tag{155}$$

For $L \gg \log \delta^{-1} \log\big(L\delta^{-1}\big)^k$, a union bound over $\{1, k^\star\}$ thus yields, with probability greater than $1 - 2\tilde{\delta}k^\star$,

$$\begin{aligned} \left| F_k - (\hat{w}_k \cdot w^\star)^k \langle \Psi, \zeta_k \rangle_{\mathsf{P}_y} \right| &\leq \inf_{\tilde{R}} \frac{3^{k^\star}}{\tilde{R}} + \tilde{C}_K \frac{\tilde{R}\sqrt{\log\Big(1/\tilde{\delta}\Big)}}{\sqrt{L}} \\ &= O\left(\frac{\log \tilde{\delta}^{-1}}{L} \right)^{1/4} \\ &:= \Delta(\tilde{\delta}, L) \quad , \ \forall \ k \in \{1, k^\star\} \ . \end{aligned} \tag{156}$$

Let us now relate the performance of our estimator $\hat{w}_{\hat{k}}$ in terms of the 'good' estimator $\hat{w}_{k^\star}$. Following [DH18, Theorem 14], we have

$$\begin{aligned} |(\hat{w}_{\hat{k}} \cdot w^\star)^{\hat{k}}| \cdot |\langle \Psi, \zeta_{\hat{k}} \rangle| &\geq F_{\hat{k}} - \Delta \\ &\geq F_{k^\star} - \Delta \\ &\geq |(\hat{w}_{k^\star} \cdot w^\star)^{k^\star}| \cdot |\langle \Psi, \zeta_{k^\star} \rangle| - 2\Delta \\ &> 0 \end{aligned} \tag{157}$$

whenever $\Delta(\tilde{\delta}, L) < C\lambda_{k^\star}\delta \leq \frac{1}{2}|\langle \Psi, \zeta_{k^\star}\rangle|$. But this implies that $|\langle \Psi, \zeta_{\hat{k}}\rangle|$, which means that $\hat{k} = k^\star$. ∎

Proof of Lemma F.16. We have $\mathbb{E}[F_k^2] \leq \tilde{R}^2 \mathbb{E}[h_k(X \cdot \hat{w}_k)^2] \leq \tilde{R}^2$, and by Gaussian hypercontractivity,

$$\mathbb{E}[F_k^l] \leq \tilde{R}^l \mathbb{E}[h_k(Z)^l] \leq \tilde{R}^l (l-1)^{kl/2} \ . \tag{158}$$

We can then apply Lemma lemma I.3, to $F_k - \mathbb{E}[F_k]$ to deduce that for any $\delta > 0$,

$$\mathbb{P}\left[|F_k - \mathbb{E}[F_k]| \gtrsim_k \frac{\tilde{R}\sqrt{\log(1/\delta)}}{\sqrt{L}} + \frac{\tilde{R}}{L} \log(1/\delta) \log(L/\delta)^{k/2} \right] \leq 2\delta \ . \frac{\tilde{R}^2}{Lt^2} \ . \tag{159}$$

Next we bound the effect of the truncation:

$$
\begin{aligned}
|\mathbb{E}[F_k] - \mathbb{E}[\Psi(Y)h_k(X\cdot\hat{w}_k)]| &= \left|\mathbb{E}_{\mathbb{P}}\left[\Psi(Y)h_{k^\star}(X\cdot\hat{w}_k)\cdot\mathbf{1}_{|\Psi(Y)|>\tilde{R}}\right]\right| \\
&\le \sqrt{\|\Psi(Y)\|^2_{\mathsf{P}_y}\mathbb{E}[h_k(X\cdot\tilde{w}_k)^2\mathbf{1}_{|\Psi(Y)|\ge\tilde{R}}]} \\
&= \sqrt{\mathbb{E}[h_k(X\cdot\tilde{w}_k)^2\mathbf{1}_{|\Psi(Y)|\ge R}]} \\
&\le \sqrt{\mathbb{E}[h_k(Z)^4]\mathbb{P}[|\Psi(Y)|\ge\tilde{R}]} \\
&\le \frac{3^k\sqrt{\mathbb{E}[\Psi(Y)^2]}}{\tilde{R}} = \frac{3^k}{\tilde{R}}\ ,
\end{aligned}
\tag{160}
$$

and finally let us compute

$$
\begin{aligned}
\mathbb{E}[\Psi(Y)h_k(X\cdot\hat{w}_k)] &= \mathbb{E}_Y[\Psi(Y)\mathbb{E}_{Z|Y}\mathbb{E}_{W\sim\gamma}h_k((\hat{w}_k\cdot w^\star)Z + \sqrt{1-(\hat{w}_k\cdot w^\star)^2}W)] \\
&= (\hat{w}_k\cdot w^\star)^k\mathbb{E}_Y[\Psi(Y)\mathbb{E}_{Z|Y}h_k(Z)] \\
&= (\hat{w}_k\cdot w^\star)^k\langle\Psi,\zeta_k\rangle_{\mathsf{P}_y}\ ,
\end{aligned}
\tag{161}
$$

where we used the fact that Hermite polynomials are eigenfunctions of the Ornstein-Ulhenbeck semigroup. ■

G Proofs of Section 5

G.1 Proof of Theorem 5.1

Throughout this section, we will occasionally use the un-normalized Hermite polynomials (Definition F.1) which will simplify the notation.

Lemma G.1. *There exists a function $f:[0,1]\to[0,1]$ such that $f(0)=0$, $f(1)=1$, f is strictly monotonic and for all $k\in\mathbb{N}$, $f^{(k)}(0)=f^{(k)}(1)=0$.*

Proof. Let

$$
f(x) = \frac{1}{Z}\int_0^x \exp\left\{-\frac{1}{s(1-s)}\right\}ds \quad \text{where} \quad Z = \int_0^1 \exp\left\{-\frac{1}{s(1-s)}\right\}ds \tag{162}
$$

Then it is clear that $f(0)=0$, $f(1)=1$, and if $a<b$,

$$
f(b)-f(a) = \frac{1}{Z}\int_a^b \exp\left\{-\frac{1}{s(1-s)}\right\}ds > 0 \tag{163}
$$

so f is monotonic. Finally, we have that

$$
\begin{aligned}
f^{(k)}(0) &= \frac{1}{Z}\left\{\frac{d^{k-1}}{dx^{k-1}}\exp\left\{-\frac{1}{x(1-x)}\right\}\right\} \\
&= \lim_{x\to 0}\left\{\frac{p_k(x)}{q_k(x)}\exp\left\{-\frac{1}{x(1-x)}\right\}\right\} \\
&= 0.
\end{aligned}
\tag{164}
$$

where $p_k(x), q_k(x)$ are polynomials in x. The proof for $f^{(k)}(1)=0$ is identical. ■

In the deterministic setting $\mathsf{P} = (\mathrm{Id} \otimes \sigma)_{\#}\gamma$ where $\sigma \in C^1(\mathbb{R})$, the condition that $\zeta_k = 0$ in $L^2(\mathbb{R}, \mathsf{P}_y)$ reduces to $\mathbb{E}[g(Y)h_k(Z)] = 0$ for any $g \in L^2_{\mathsf{P}_y}$. From Sard's theorem, the set of critical values of σ, ie, the set $S_\sigma := \{y \in \mathbb{R};\, \exists z \in \sigma^{-1}(y)\, s.t.\, \sigma'(z) = 0\}$ has Lebesgue measure zero. In particular, we have that a.e. $\sigma^{-1}(y)$ is a discrete set with $\sigma'(z) \neq 0$ for any $z \in \sigma^{-1}(y)$. Therefore, applying the coarea formula leads to

$$\begin{aligned} 0 &= \mathbb{E}[g(Y)h_k(Z)] \\ &= \int g(\sigma(z))h_k(z)\gamma(z)dz \\ &= \int g(\sigma(z))h_k(z)\gamma(z)\frac{|\sigma'(z)|}{|\sigma'(z)|}dz \\ &= \int g(y)\left(\int_{\sigma^{-1}(y)} \frac{h_k(z)\gamma(z)}{|\sigma'(z)|}d\mathcal{H}_0(z)\right)dy\ , \end{aligned} \tag{165}$$

and therefore

$$\begin{aligned} 0 &= \int_{\sigma^{-1}(y)} \frac{h_k(z)\gamma(z)}{|\sigma'(z)|}d\mathcal{H}_0(z) \\ &= \sum_i \frac{h_k(\sigma^{-1}(y)_i)\gamma(\sigma^{-1}(y)_i)}{|\sigma'(\sigma^{-1}(y)_i)|} \quad , \text{ for } y \notin S_\sigma\ . \end{aligned} \tag{166}$$

We first verify an equivalent integral condition:

Lemma G.2 (Integral Form)**.** *The condition* $\int_{\sigma^{-1}(y)} \frac{h_k(z)\gamma(z)}{|\sigma'(z)|}dz = 0$ *for* P_y*-a.e.* y *is equivalent to the following quantity being constant in* y*:*

$$\int_{\sigma^{-1}(y)} \gamma(z)h_{k-1}(z)\mathrm{sign}(\sigma'(z))d\mathcal{H}_0(z) = C\ . \tag{167}$$

Proof. Follows directly from integrating with respect to y, and using that $[h_k\gamma]' = -h_{k-1}\gamma$. ■

Definition G.3. Given $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$, we define $Q(u)$ by:

$$Q(u) := \mathbb{E}_{z\sim\gamma}\prod_{i=1}^n (u_i + z) = \sum_{i=0}^{\lfloor n/2 \rfloor} (2i-1)!!e_{n-2i}(u)\ , \tag{168}$$

where $e_k(u)$ is the kth elementary symmetric polynomial on $u_1, \ldots, u_n$.

Definition G.4. Given n distinct points $u_1, \ldots, u_n$ we define $v(u) \in \mathbb{R}^n$, $u = (u_1, \ldots, u_n)$ by

$$v(u)_i = \frac{Q(u_1, \ldots, \hat{u}_i, \ldots, u_n)}{\prod_{i\neq j}(u_j - u_i)}, \tag{169}$$

where $(u_1, \ldots, \hat{u}_i, \ldots, u_n)$ is the $(n-1)$-dimensional vector in which the i-th coordinate has been removed.

Lemma G.5. *For any distinct points $u_1, \dots, u_n$ and for $0 \le k \le n$ we have*

$$\sum_{i=1}^{n} \mathrm{He}_k(u_i) v(u)_i = \begin{cases} 1 & k = 0 \\ 0 & 0 < k < n \\ (-1)^{n+1} Q(u) & k = n \end{cases}. \tag{170}$$

Proof. We will first prove the result for $k < n$. Let $A \in \mathbb{R}^{n \times n}$ be defined by $A_{ij} := \mathrm{He}_{i-1}(x_j)$. Then we can decompose $A = CV$ where $C \in R^{n \times n}$ contains the coefficients of the Hermite polynomials, i.e. C_{ij} is the coefficient of x^j in $\mathrm{He}_i(x)$, and $V \in R^{n \times n}$ is a Vandermonde matrix defined by $V_{ij} = x_j^{i-1}$. Note that C is invertible (since it is triangular) and as the x_i are distinct, V is invertible as well so A is invertible. Then the unique solution to $Av^\star = e_1$ is given by $v^\star = V^{-1}C^{-1}e_1$. By the formula for converting Hermite polynomials to monomials, we have that $C^{-1} = |C|$ so $(C^{-1}e_1)_j = (j-2)!!\mathbf{1}_{2|j-1}$. Therefore from the formula for the inverse of a Vandermonde matrix,

$$\begin{aligned} v_i^\star &= \sum_{2|j-1} (V^{-1})_{ij} (j-2)!! \\ &= \sum_{2|j-1} \frac{e_{n-j}(x_1, \dots, \hat{x}_i, \dots, x_n)}{\prod_{k \ne i}(x_k - x_i)} (j-2)!! \\ &= \frac{\sum_{2|j} e_{n-j-1}(x_1, \dots, \hat{x}_i, \dots, x_n)(j-1)!!}{\prod_{j \ne i}(x_j - x_i)} \\ &= \frac{Q(x_1, \dots, \hat{x}_i, \dots, x_n)}{\prod_{j \ne i}(x_j - x_i)} \\ &= v(x)_i. \end{aligned} \tag{171}$$

Therefore $v^\star = v(x)$ so $Av(x) = e_1$. Next, pick some x_{n+1} which is distinct from $x_1, \dots, x_n$. By the above computation we know that

$$\sum_{i=1}^{n+1} \mathrm{He}_n(x_i) v(x_1, \dots, x_{n+1})_i = 0 \tag{172}$$

which implies that

$$\begin{aligned} \sum_{i=1}^{n} \mathrm{He}_n(x_i) v(x_1, \dots, x_{n+1})_i &= -\mathrm{He}_n(x_{n+1}) v(x_1, \dots, x_{n+1})_i \\ &= -\mathrm{He}_n(x_{n+1}) \frac{Q(x_1, \dots, x_n)}{\prod_{i=1}^{n}(x_i - x_{n+1})}. \end{aligned} \tag{173}$$

We can now take $x_{n+1} \to \infty$ on both sides. For the left hand side we have for $i \le n$:

$$\begin{aligned} \lim_{x_{n+1} \to \infty} v(x_1, \dots, x_{n+1})_i &= \lim_{x_{n+1} \to \infty} \frac{Q(x_1, \dots, \hat{x}_i, \dots, x_n, x_{n+1})}{(x_{n+1} - x_i) \prod_{j \ne i, j \le n}(x_j - x_i)} \\ &= \lim_{x_{n+1} \to \infty} \frac{\mathbb{E}_{z \sim N(0,1)}[\prod_{j \ne i}(x_j + z)]}{\prod_{j \ne i}(x_j - x_i)} \\ &= v(x_1, \dots, x_n)_i. \end{aligned} \tag{174}$$

On the right hand side we have:

$$\lim_{x_{n+1}\to\infty} \frac{\mathrm{He}_n(x_{n+1})}{\prod_{i=1}^n (x_i - x_{n+1})} = (-1)^n. \tag{175}$$

Putting it together gives that

$$\sum_{i=1}^n \mathrm{He}_n(x_i) v(x_1, \ldots, x_n)_i = (-1)^{n+1} Q(x_1, \ldots, x_n) \tag{176}$$

which completes the proof. ■

We will use the following well-know result for the Gauss-Hermite quadrature ([DR07]):

Lemma G.6 (Gauss-Hermite Quadrature)**.** *Let $r_1, \ldots, r_n$ be the roots of He_n. Then*

$$v(r_1, \ldots, r_n)_i = \frac{1}{n\mathrm{He}_{n-1}(r_i)^2} > 0. \tag{177}$$

Lemma G.7. *For any $k^\star \geq 0$ and any $\epsilon > 0$ there exist points $x_1, \ldots, x_{k^\star}$ such that if $r_1, \ldots, r_{k^\star}$ are the roots of $\mathrm{He}_{k^\star}(x)$, we have*

$$|x_i - r_i| \leq \epsilon \ \forall i \quad \textit{and} \quad Q(x_1, \ldots, x_n) \neq 0. \tag{178}$$

Proof. Note that Q is a polynomial in n variables of degree n which can have only finitely many roots. Therefore it is not possible for all points $x \in \bigtimes_{i=1}^n [r_i - \epsilon, r_i + \epsilon]$ to be roots of Q. ■

We are now ready to prove Theorem 5.1:

Theorem 5.1 (Smooth Single-Index models with prescribed $k^\star$)**.** *For each k, there exists $\sigma \in C_b^\infty(\mathbb{R})$ such that the deterministic single index model $\mathsf{P} = (\mathrm{Id} \otimes \sigma)_\# \gamma_1$ satisfies $k^\star(\mathsf{P}) = k$.*

Proof. Let $r_1, \ldots, r_{k^\star}$ be the roots of $\mathrm{He}_{k^\star}$. From Lemma G.6, we know that $v(r_1, \ldots, r_{k^\star}) > 0$. Therefore by continuity there exists δ such that

$$|x_i - r_i| \leq \delta \ \forall i \implies v(x_1, \ldots, x_n) > 0. \tag{179}$$

Then by Lemma G.7 there exist $x_1(0), \ldots, x_n(0)$ with $|x_i(0) - r_i| \leq \delta/2$ such that $Q(x_1(0), \ldots, x_n(0)) \neq 0$. Again by continuity there exists ϵ such that for all $x_1, \ldots, x_n$ with $|x_i - x_i(0)| \leq \epsilon$ for all i, we have both

$$v(x_1, \ldots, x_n) > 0 \quad \text{and} \quad \mathrm{sign}(Q(x_1, \ldots, x_n)) = \mathrm{sign}(Q(x(0))). \tag{180}$$

Now let $\gamma(x) := \frac{e^{-x^2/2}}{\sqrt{2\pi}}$ be the PDF of a standard Gaussian and let x evolve according to the ODE:

$$x_i'(t) = \gamma(x_i)^{-1} v(x_1, \ldots, x_n)_i \quad \text{for} \quad i = 1, \ldots, n. \tag{181}$$

We will run this ODE for $t \in [-\tau, \tau]$ for τ sufficiently small so that $\|x(t) - x(0)\|_1 < \epsilon$ for all $t \in [-\tau, \tau]$. We will also define the quantity:

$$Z_k(t) := \sum_{i=1}^n \mathrm{He}_{k-1}(x_i(t)) \gamma(x_i(t)). \tag{182}$$

Then using that $\frac{d}{dx}\mathrm{He}_{k-1}(x)\gamma(x) = -\mathrm{He}_k(x)\gamma(x)$, we have that for any $1 \le k \le k^\star$,

$$\begin{aligned} Z_k'(t) &= -\sum_{i=1}^n \mathrm{He}_k(x_i)\gamma(x_i)x_i'(t) \\ &= -\sum_{i=1}^n \mathrm{He}_k(x_i)v(x(t))_i \\ &= \begin{cases} 0 & k < k^\star \\ (-1)^{k^\star+1}Q(x(t)) & k = k^\star \end{cases}. \end{aligned} \tag{183}$$

Therefore we have that $Z_k(t) = Z_k(0)$ for all $k < k^\star$ and $|Z_{k^\star}(t) - Z_{k^\star}(-t)| > 0$.

Note that by construction, $x_i'(t) > 0$ for all $t \in [-\tau, \tau]$. Therefore $x_i : [-\tau, \tau] : \mathbb{R}$ is injective and for ϵ sufficiently small, the images $\{x_i([-\tau, \tau])\}_i$ don't intersect. We can now define $\hat{\sigma}$ by

$$\hat{\sigma}(x) := \begin{cases} \left|x_i^{-1}(x)\right| & x \in x_i([-\tau, \tau]) \\ \tau & \text{otherwise.} \end{cases} \tag{184}$$

Note that $\hat{\sigma}$ is smooth except at $\hat{\sigma}^{-1}(0) \cup \hat{\sigma}^{-1}(\tau)$. Now let $f : [0,1] \to [0,1]$ be the mollifier constructed in Lemma G.1. Then if $\hat{\sigma}$ has generative exponent $k^\star$, $\sigma(x) := f(\hat{\sigma}(x)/\tau)$ also has generative exponent $k^\star$ and is smooth. Therefore it suffices to show that $\hat{\sigma}$ has generative exponent $k^\star$.

We will compute $\mathbb{E}[\mathrm{He}_k(x)|y]$ for $y \in (0, \tau]$. First, let us consider $y = \tau$. Using $\mathbb{E}[\mathrm{He}_k(x)] = 0$ we can write:

$$\begin{aligned} \mathbb{E}[\mathrm{He}_k(x)|y = \tau] &= -\mathbb{E}[\mathrm{He}_k(x)|y < \tau] \\ &= -\sum_{i=1}^n \int_{x_i(-\tau)}^{x_i(\tau)} \mathrm{He}_k(x)\gamma(x)dx \\ &= \sum_{i=1}^n \mathrm{He}_{k-1}(x_i(\tau))\gamma(x_i(\tau)) - \mathrm{He}_{k-1}(x_i(-\tau))\gamma(x_i(-\tau)) \\ &= Z_k(\tau) - Z_k(-\tau). \end{aligned} \tag{185}$$

By the above calculation this is 0 for $k < k^\star$ because $Z_k(\tau) = Z_k(-\tau) = Z(0)$ and it is nonzero for $k = k^\star$. Next, let us consider $y \in (0, \tau)$. Then $\hat{\sigma}^{-1}(y)$ is a set of discrete points at which σ is smooth so y has a continuous density and we can apply the co-area formula (166):

$$\begin{aligned} \mathbb{E}[\mathrm{He}_k(x)|y] &\propto \sum_{x \in \hat{\sigma}^{-1}(y)} \frac{\mathrm{He}_k(x)\gamma(x)}{|\hat{\sigma}'(x)|} \\ &= \sum_{i=1}^n \frac{\mathrm{He}_k(x_i(y))\gamma(x_i(y))}{|\hat{\sigma}'(x_i(y))|} + \frac{\mathrm{He}_k(x_i(-y))\gamma(x_i(-y))}{|\hat{\sigma}'(x_i(-y))|}. \end{aligned} \tag{186}$$

From the definition of $\hat{\sigma}$ we have that if $x \in x_i([-\tau, \tau])$:

$$\hat{\sigma}'(x) = \frac{\mathrm{sign}(x_i^{-1}(x))}{x_i'(x_i^{-1}(x))}. \tag{187}$$

Therefore we can simplify the above expression as:

$$\mathbb{E}[\mathrm{He}_k(x)|y] \propto \sum_{i=1}^{n} \mathrm{He}_k(x_i(y))\gamma(x_i(y))\,|x_i'(y)| + \mathrm{He}_k(x_i(-y))\gamma(x_i(-y))\,|x_i'(-y)|\,. \tag{188}$$

Now because $x_i'(t) > 0$ for all $t \in [-\tau, \tau]$, we can remove the absolute values to get:

$$\begin{aligned}\mathbb{E}[\mathrm{He}_k(x)|y] &\propto \sum_{i=1}^{n} \mathrm{He}_k(x_i(y))\gamma(x_i(y))x_i'(y) + \mathrm{He}_k(x_i(-y))\gamma(x_i(-y))x_i'(-y) \\ &= -[Z_k'(y) - Z_k'(-y)] \end{aligned} \tag{189}$$

which we computed above. In particular, this is 0 for $k < k^\star$ and nonzero for $k = k^\star$ for any $y > 0$. Therefore, $\mathbb{E}_y\left[\mathbb{E}_x[\mathrm{He}_k(x)|y]^2\right] = 0$ for $k < k^\star$ and is nonzero for $k = k^\star$ which completes the proof. ■

G.2 Proof of Theorem 5.2

Theorem 5.2 (Additive Gaussian noise preserves generative exponent)**.** *For $\tau \geq 0$ and $\sigma : \mathbb{R} \to \mathbb{R}$, we denote $\Phi_{\tau,\sigma}(u, v) = (u, \sigma(u) + \tau v) \in \mathbb{R}^2$. Then the additive noisy model $\tilde{\mathsf{P}} = (\Phi_{\tau,\sigma})_{\#}\gamma_2$ satisfies $k^\star(\tilde{\mathsf{P}}) = k^\star(\mathsf{P})$, where $\mathsf{P} = (\mathrm{Id} \otimes \sigma)_{\#}\gamma_1$.*

Proof. Since $k^\star(\mathsf{P}) = k^\star$, by Lemma F.2 there exists $g : \mathbb{R} \to [-1, 1]$ such that $\mathbb{E}[g(Y)h_{k^\star}(Z)] \neq 0$. We consider $g_R(y) := g(y)\mathbf{1}_{|y|\leq R}$. For R sufficiently large, we claim that $\mathbb{E}[g_R(Y)h_{k^\star}(Z)] \neq 0$. We have that

$$\begin{aligned}|\mathbb{E}[g_R(Y)h_{k^\star}(Z)] - \mathbb{E}[g(Y)h_{k^\star}(Z)]| &= \left|\mathbb{E}[g(Y)h_{k^\star}(Z)\mathbf{1}_{|y|\geq R}]\right| \\ &\leq \sqrt{\mathbb{E}[g(Y)^2 h_{k^\star}(Z)^2]\mathbb{P}[|Y| \geq R]} \\ &\leq \sqrt{\mathbb{E}[Y^2]/R^2} \end{aligned} \tag{190}$$

which vanishes as $R \to \infty$. Therefore for sufficiently large R we have $\mathbb{E}[g_R(Y)h_{k^\star}(Z)] \neq 0$. Now $g_R \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Let us consider its Fourier representation $\mathcal{T}(y) = \int \hat{g_R}(\xi)e^{i\xi y}d\xi$. Then

$$\mathbb{E}[g_R(Y)h_{k^\star}(Z)] = \int \hat{g}_R(\xi)\mathbb{E}[e^{i\xi Y}h_{k^\star}(Z)]d\xi\ , \tag{191}$$

which shows that there must exist ξ such that $\phi_\xi(y, z) = e^{i\xi y}h_{k^\star}(z)$ satisfies $\mathbb{E}_{\mathsf{P}}[\phi_\xi(Y, Z)] \neq 0$. Now, let $G_\tau(y, z) = \tau\gamma(\tau y)\delta_z$, where γ is the standard Gaussian density. By definition we have $\tilde{\mathsf{P}} = \mathsf{P} * G_\tau := \mathcal{G}_\tau\mathsf{P}$. Recall that $\mathcal{G}_\tau$ is self-adjoint in $L^2(\mathbb{R})$. Thus

$$\begin{aligned}\mathbb{E}_{\tilde{P}}[\phi_\xi(Y, Z)] &= \int \phi_\xi(y, z)d\tilde{P}(y, z) \\ &= \int \phi_\zeta(y, z)d(\mathcal{G}_\tau\mathsf{P})(y, z) \\ &= \int \mathcal{G}_\tau^*(\phi)(y, z)d\mathsf{P}(y, z) \\ &= \exp\left(-\xi^2\tau^2\right)\int \phi(y, z)d\mathsf{P}(y, z) \\ &= \exp\left(-\xi^2\tau^2\right)\mathbb{E}_{\mathsf{P}}[\phi_\xi(Y, Z)] \neq 0\ . \end{aligned} \tag{192}$$

This shows that $k^\star(\tilde{\mathsf{P}}) \le k^\star(\mathsf{P})$. But note that by Proposition 2.6, we also have $k^\star(\tilde{\mathsf{P}}) \ge k^\star(\mathsf{P})$. Therefore $k^\star(\tilde{\mathsf{P}}) = k^\star(\mathsf{P})$.

■

H Proofs of Section 6

Theorem 6.1 (Information-Theoretic Upper Bound)**.** *For all $k \ge 1$, there exists a procedure which returns $\hat{w}$ with $(w \cdot w^\star)^2 \ge 1 - \epsilon^2$ with probability at least $1 - 2e^{-d}$ if $n = \tilde{\Theta}\left(\frac{d}{\lambda_k^2 \epsilon^2}\right)$.*

Proof. We will show that

$$n \ge C_k \frac{d \log\left(\frac{3}{\lambda_k}\right)^{k/2} \log\left(\frac{3}{\lambda_k \epsilon}\right)}{\lambda_k^2 \epsilon^2} = \tilde{\Theta}\left(\frac{d}{\lambda_k^2 \epsilon^2}\right) \ , \tag{193}$$

where C_k is an constant depending only on k, is sufficient for recovery whp.

Throughout the proof we will use $\lesssim$ to hide constants that depend only on k. Let $\mathcal{N}_\delta$ be an δ net of S^{d-1} with $|\mathcal{N}_\delta| \le (\frac{3}{\delta})^d$. We define $g(z, y) := \zeta_k(y) h_k(z)$ and $\lambda_k = \|\zeta_k\|_{L^2(\mathsf{P}_y)}$. Fix a truncation radius R and define $L_n(w)$ by

$$L_n(w) := \frac{1}{n} \sum_{i=1}^n g(w \cdot x_i, y_i) \mathbf{1}_{|g(w \cdot x_i, y_i)| \le R} \tag{194}$$

We will consider the estimator:

$$\hat{w} \in \arg\min_{\hat{w} \in S^{d-1}} \max_{w \in \mathcal{N}_\delta} \left| L_n(w) - \lambda_k^2 (w \cdot \hat{w})^k \right| . \tag{195}$$

We will begin by concentrating L_n. First note that by Lemma I.4,

$$\mathbb{E}[g(w \cdot X, Y)^2] = \mathbb{E}[\xi_k(Y)^2 h_k(w \cdot X)^2] \lesssim \lambda_k^2 \log\left(3/\lambda_k^2\right)^{k/2}. \tag{196}$$

Therefore by Bernstein's inequality we have that for any $w \in S^{d-1}$, with probability at least $1 - 2e^{-\iota}$,

$$|L_n(w) - \mathbb{E}[L_n(w)]| \lesssim \sqrt{\frac{\lambda_k^2 \log\left(\frac{3}{\lambda_k^2}\right)^{\frac{k}{2}} \iota}{n}} + \frac{R\iota}{n}. \tag{197}$$

Therefore by a union bound we have that with probability at least $1 - 2e^{-d}$,

$$\sup_{w \in \mathcal{N}_\delta} |L_n(\hat{w}) - \mathbb{E}[L_n(\hat{w})]| \lesssim \sqrt{\frac{\lambda_k^2 \log\left(\frac{3}{\lambda_k^2}\right)^{\frac{k}{2}} d \log(3/\delta)}{n}} + R\frac{d \log(3/\delta)}{n}. \tag{198}$$

Next we bound the effect of truncation on $\mathbb{E}[L_n(w)]$:

$$\begin{aligned}
&|\mathbb{E}[L_n(w)] - \mathbb{E}[g(w \cdot X, Y)]| \\
&= \left|\mathbb{E}[g(w \cdot X, Y) \mathbf{1}_{|g(w \cdot X, Y)| > R}]\right| \\
&\le \sqrt{\mathbb{E}[g(w \cdot X, Y)^2] \mathbb{P}[|g(w \cdot X, Y)| > R]} \\
&\le 3^k \sqrt{\mathbb{P}[|g(w \cdot X, Y)| > R]}.
\end{aligned} \tag{199}$$

To control this probability, we use the moment method. By Jensen's inequality and Gaussian hypercontractivity we have for all $p \geq 1$,

$$\mathbb{E}[|g(w \cdot X, Y)|^p] \leq \sqrt{\mathbb{E}[\xi_k(Y)^{2p}]\mathbb{E}[h_k(Z)^{2p}]} \leq \mathbb{E}[h_k(Z)^{2p}] \leq (2p)^{kp}. \tag{200}$$

Therefore for $R \geq (2e)^k$, we can take $p = R^{1/k}/(2e)$ to get

$$\mathbb{P}[|g(w \cdot X, Y)| > R] \leq \frac{(2p)^{kp}}{R^p} = \exp\left(-\frac{k}{2e}R^{1/k}\right). \tag{201}$$

Plugging this in gives

$$|\mathbb{E}[L_n(w)] - \mathbb{E}[g(w \cdot X, Y)]| \leq 3^k \exp\left(-\frac{k}{4e}R^{1/k}\right). \tag{202}$$

Finally, because the k-th Hermite coefficient of ν is λ_k^2, we have that

$$\mathbb{E}[g(w \cdot X, Y)] = \lambda_k^2 (w \cdot w^\star)^k. \tag{203}$$

Combining everything gives that with probability at least $1 - 2e^{-d}$:

$$\begin{aligned}
&\max_{w \in \mathcal{N}_\delta} \left|\lambda_k^2 (w \cdot \hat{w})^k - \lambda_k^2 (w \cdot w^\star)^k\right| \\
&\leq \max_{w \in \mathcal{N}_\delta} \left|L_n(w) - \lambda_k^2 (w \cdot \hat{w})^k\right| + \max_{w \in \mathcal{N}_\delta} \left|L_n(w) - \lambda_k^2 (w \cdot w^\star)^k\right| \\
&\leq 2 \max_{w \in \mathcal{N}_\delta} \left|L_n(w) - \lambda_k^2 (w \cdot w^\star)^k\right| \\
&= 2 \max_{w \in \mathcal{N}_\delta} |L_n(w) - \mathbb{E}[g(w \cdot X, Y)]| \\
&\lesssim \sup_{w \in \mathcal{N}_\delta} |L_n(\hat{w}) - \mathbb{E}[L_n(\hat{w})]| + 3^k \exp\left(-\frac{k}{4e}R^{1/k}\right) \\
&\lesssim \sqrt{\frac{\lambda_k^2 \log\left(\frac{3}{\lambda_k^2}\right)^{\frac{k}{2}} d \log(3/\delta)}{n}} + R\frac{d\log(3/\delta)}{n} + 3^k \exp\left(-\frac{k}{4e}R^{1/k}\right).
\end{aligned} \tag{204}$$

Therefore by [DH21, Lemma 25], we have that

$$\begin{aligned}
&\lambda_k^2 \min(\|\hat{w} - w^\star\|, \|\hat{w} + w^\star\|) \\
&\lesssim \delta + \sqrt{\frac{\lambda_k^2 \log\left(\frac{3}{\lambda_k^2}\right)^{\frac{k}{2}} d \log(3/\delta)}{n}} + R\frac{d\log(3/\delta)}{n} + \exp\left(-\frac{k}{4e}R^{1/k}\right).
\end{aligned} \tag{205}$$

Now take $R = (4e \log(3/\delta))^k$. Then,

$$\lambda_k^2 \min(\|\hat{w} - w^\star\|, \|\hat{w} + w^\star\|) \lesssim \delta + \sqrt{\frac{\lambda_k^2 \log\left(\frac{3}{\lambda_k^2}\right)^{\frac{k}{2}} d \log(3/\delta)}{n}} + \frac{d\log(3/\delta)^{k+1}}{n} \tag{206}$$

and taking $\delta = O(\epsilon \lambda_k^2)$ and using that $3/\delta \geq \log(3/\delta)^k$ completes the proof. ■

I Concentration Lemmas

Lemma I.1 (Gaussian Hypercontractivity)**.** *Let f be a polynomial of degree k. Then for $p \geq 2$,*

$$\mathbb{E}_{X\sim\gamma}[|f(X)|^p]^{2/p} \leq (p-1)^k \ \mathbb{E}_{x\sim\gamma}[f(X)^2]. \tag{207}$$

Such moments imply the following tail bound (e.g. see [DNGL23, Lemma 24]):

Lemma I.2. *Let $\delta \geq 0$ and let X be a mean zero random variable satisfying*

$$\mathbb{E}[|X|^p]^{1/p} \leq Bp^{k/2} \quad \textit{for} \quad p = \frac{2\log(1/\delta)}{k} \tag{208}$$

for some k. Then with probability at least $1-\delta$, $|X| \leq B(ep)^{k/2}$.

We will combine this with the following tail bound which can be easily proved with a routine truncation argument:

Lemma I.3. *Let $X_1, \ldots, X_n$ be independent mean zero random variables such that for all $p \geq 2$, $\|X_i\|_p \leq Bp^{k/2}$ for some k and let $\sigma = \sum_{i=1}^n \mathbb{E}[\|X_i\|^2]$. Let $Y = \sum_{i=1}^n X_i$. Then with probability at least $1-2\delta$,*

$$\|Y\| \lesssim_k \sigma\sqrt{\log(1/\delta)} + B\log(1/\delta)\log(n/\delta)^{k/2}. \tag{209}$$

Proof. Let R be a truncation radius to be chosen later and define $\tilde{X}_i = X_i \mathbf{1}_{\|X_i\|\leq R}$. We have that with probability at least $1-\delta$, $\|X_1\| \leq C_k B\log^{k/2}(1/\delta)$. Therefore by a union bound we have that with probability at least $1-\delta$, $\max_i \|X_i\| \leq C_k B\log^{k/2}(n/\delta) =: R$. Let $\tilde{Y} = \sum_i \tilde{X}_i$. Then,

$$\begin{aligned}
\left\|\mathbb{E}[\tilde{Y}] - \mathbb{E}[Y]\right\| &\leq \sum_{i=1}^n \left\|\mathbb{E}[X_i] - \mathbb{E}[\tilde{X}_i]\right\| \\
&= \sum_{i=1}^n \left\|\mathbb{E}[X_i \mathbf{1}_{\|X\|\geq R}]\right\| \\
&\leq \sum_{i=1}^n \sqrt{\mathbb{E}[\|X_i\|^2]\mathbb{P}[\|X\| \geq R]} \\
&\leq \sigma \sum_{i=1}^n \mathbb{P}[\|X_i\| \geq R] \\
&\leq \sigma\delta.
\end{aligned} \tag{210}$$

Finally, because $\mathbb{E}[\tilde{X}_i^2] \leq \mathbb{E}[X_i^2]$, we have by Bernstein's inequality that with probability at least $1-\delta$,

$$\begin{aligned}
\left\|\tilde{Y} - \mathbb{E}[\tilde{Y}]\right\| &\lesssim \sigma\sqrt{\log(1/\delta)} + R\log(1/\delta) \\
&\lesssim \sigma\sqrt{\log(1/\delta)} + B\log^{k/2}(n/\delta)\log(1/\delta).
\end{aligned} \tag{211}$$

Combining everything gives with probability at least $1-2\delta$,

$$\begin{aligned}\|Y\| &\lesssim \sigma\delta + \sigma\sqrt{\log(1/\delta)} + B\log^{k/2}(n/\delta)\log(1/\delta)\\ &\lesssim \sigma\sqrt{\log(1/\delta)} + B\log^{k/2}(n/\delta)\log(1/\delta). \end{aligned} \tag{212}$$

■

We will often use the following lemma from [DNGL23, Lemma 23] when we have tight bounds on $\|X\|_1$ and all moments of Y but only a very loose bound on $\|X\|_2$:

Lemma I.4. *Let X, Y be random variables with $\|Y\|_p \le Bp^{k/2}$ for*

$$p = \min\left(2, \frac{1}{k}\cdot\log\left(\frac{\|X\|_2}{\|X\|_1}\right)\right).$$

Then,

$$\mathbb{E}[XY] \le \|X\|_1 \cdot B \cdot (ep)^{k/2}. \tag{213}$$

For matrix concentration we will use the following corollary of [BvH23, Theorem 2.7]:

Lemma I.5. *Let $Y = \sum_{i=1}^n Z_i$ where $\{Z_i\}_{i=1}^n$ are self-adjoint, mean zero, and independent random matrices. Define:*

$$\sigma := \|\mathbb{E}[Y^2]\|_2^{1/2}, \quad \sigma_* := \sup_{v,w\in S^{d-1}} \mathbb{E}[(v^T Y w)^2]^{1/2}, \quad \overline{R} := \mathbb{E}\left[\max_{i\in[n]}\|Z_i\|_2^2\right]^{1/2}. \tag{214}$$

Then for any $R \ge \overline{R}^{1/2}\sigma^{1/2} + \overline{R}\sqrt{2}$ and any $t \ge 0$, if $\delta = \mathbb{P}[\max_{i\in[n]}\|Z_i\|_2 \ge R]$, then with probability at least $1-\delta-de^{-t}$,

$$\|Y - \mathbb{E}Y\|_2 - 2\sigma \lesssim \sigma_* t^{1/2} + R^{1/3}\sigma^{2/3}t^{2/3} + Rt. \tag{215}$$

To compute $\overline{R}$ and the tail probability δ, we will use the following lemma:

Lemma I.6. *Let $\{Z_i\}_{i=1}^n$ be a sequence of independent random variables with polynomial tails, i.e. there exists B, k such that $\mathbb{E}[|Z_i|^p]^{1/p} \le Bp^{k/2}$. Define $R = \max_{i=1}^n Z_i$. Then for any $p \le \log(n)/k$, $E[|R|^p]^{1/p} \lesssim B\log^{k/2}(n)$ and for any $\delta \ge 0$, with probability at least $1-\delta$, $R \lesssim B\log^{k/2}(n/\delta)$.*

Proof.

$$\mathbb{E}\left[|R|^p\right]^{1/p} = \mathbb{E}\left[\max_{i\in[n]}|Z_i|_2^p\right]^{1/p} \le \mathbb{E}\left[\sum_{i=1}^n |Z_i|_2^p\right]^{1/p} \le n^{1/p}Bp^{k/2}. \tag{216}$$

Now plugging in $p = \log(n)/k$ gives:

$$\mathbb{E}[R^p]^{1/p} = \mathbb{E}\left[\left(\max_{i\in[n]}|Z_i|_2\right)^p\right]^{1/p} \le B\left(\frac{e^2\log(n)}{k}\right)^{\frac{k}{2}} \lesssim B\log^{\frac{k}{2}}(n). \tag{217}$$

In addition by Markov's inequality we have that when $p = \frac{t^{2/k}}{e}$,

$$\mathbb{P}\left[R \ge tB\right] \le \left(\frac{n^{1/p}p^{k/2}}{t}\right)^p = n\exp\left(-\frac{k}{2e}t^{2/k}\right). \tag{218}$$

■