Survey on algorithms for multi-index models

Joan Bruna and Daniel Hsu

Abstract. We review the literature on algorithms for estimating the index space in a multi-index model. The primary focus is on computationally efficient (polynomial-time) algorithms in Gaussian space, the assumptions under which consistency is guaranteed by these methods, and their sample complexity. In many cases, a gap is observed between the sample complexity of the best known computationally efficient methods and the information-theoretic minimum. We also review algorithms based on estimating the span of gradients using nonparametric methods, and algorithms based on fitting neural networks using gradient descent.

Key words and phrases: single-index models, multi-index models, feature learning, neural networks.

1. INTRODUCTION

In a *multi-index model* for data $(\mathbf{x}, \mathbf{y})$ from $\mathbb{R}^d \times \mathbb{R}$, the regression function $f(x) := \mathsf{E}[\mathbf{y} \mid \mathbf{x} = x]$ is assumed to depend only on a (possibly low-rank) linear transformation of the input:

$$(1) \qquad f(x) = \mathsf{E}[\mathbf{y} \mid \mathbf{x} = x] = g(Lx)$$

for some rank-r matrix $L \in \mathbb{R}^{r \times d}$ whose row space $\operatorname{ran}(L^\top)$ is called the *index space*, and some function $g \colon \mathbb{R}^r \to \mathbb{R}$ called the *link function*. When the number of variables d is large, it is common to assume that $r \ll d$, so that the linear map $x \mapsto Lx$ is viewed as a form of dimension reduction that captures the information sufficient to optimally predict the response $\mathbf{y}$ from $\mathbf{x}$ under a mean squared error criterion.[1] (We postpone issues of identifiability for now.) The special case where $r = 1$ is known as the *single-index model*.

The multi-index model is a popular model with benign high-dimensional behavior, with a rich and long history in the statistics literature, eg [BC64, BD81, CR84, Li91] and references therein. More recently, it has also gained interest as a model for studying the ability of machine learning methods to perform "feature learning", ie to automatically discover meaningful low-dimensional structure within high-dimensional data. The model in (1) is a

special case of the more general model for $(\mathbf{x}, \mathbf{y})$ where

$$(2) \qquad f(x) = g(\Phi(x))$$

and $\Phi \colon \mathbb{R}^d \to \mathbb{R}^r$ is a (possibly non-linear, but morally simpler than the original f) *feature map*. Feature learning, then, refers to the estimation of the feature map Φ from data. The feature map may serve an explanatory role in understanding a predictor of $\mathbf{y}$ from $\mathbf{x}$, in which case feature learning is an end in itself. Additionally, separating the tasks of learning Φ and learning g may be methodologically preferable; in such a scenario, feature learning is the first part of a two (or more) stage learning process.

Many recent works, under the guise of "multi-task learning" or "meta-learning", instantiate the model in (2) separately for each of multiple data sources, but constrain the feature map Φ to be shared across all models. For instance, in a computer vision context, the data sources may correspond to different object detection tasks in images; the shared feature map may give a semantic representation of images that is broadly useful for vision. The data for a single task may be too specialized and/or insufficient to identify/estimate such a feature map, but the pooling of data sources across a diverse collection of tasks may pin down the desired image representation Φ. It has been hypothesized that large neural networks trained on diverse data sets may be encode such feature maps in the intermediate layers of the network, and hence feature learning has been considered as a possible explanation for the success of neural networks in practice. See, e.g., [DHK$^+$21] for more discussion of this compelling motivation.

The goal of this article is to survey algorithms for estimating the index space $\operatorname{ran}(L^\top)$ in the multi-index model (1). We do not attempt to be exhaustive, nor do we attempt to present the strongest or most general possible methods or results. Rather, we aim to highlight some key

Courant Institute of Mathematical Sciences and Center for Data Science, New York University (e-mail: bruna@cims.nyu.edu). Department of Computer Science and Data Science Institute, Columbia University (e-mail: djhsu@cs.columbia.edu).

[1] A potentially broader aim is to identify a linear map L such that $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid L\mathbf{x}$. Goals such as this (as well as estimation goals concerning the model in (1)) are the subject of the subfield of statistics known as *Sufficient Dimension Reduction* [Li91], from which many of the methods and ideas discussed in this article are derived.

ideas and methods from the literature, and to draw connections to the recent literature on gradient-based training of neural networks and statistical-computational gaps in high-dimensional inference.

Acknowledgements: The authors would like to thank Alex Damian, Adam Klivans, Florent Krzakala, Jason Lee, Luis Rademacher, Denny Wu and the anonymous referees for useful feedback during the completion of this work.

2. PRELIMINARIES

2.1 Identifiability

We first observe that the model (1) needs an additional 'minimality' condition to be well-defined; hereafter, we will view (1) as the representation of f with *smallest* rank r, which will be denoted the *intrinsic dimension* of f. We can assume without loss of generality that L is orthogonal, i.e., $L^\intercal \in \mathrm{Stiefel}(d, r)$, the (compact) Stiefel manifold of $d \times r$ matrices with orthonormal columns.

It will be also convenient to formally define a multi-index model in terms of its generative process: the joint distribution π of $(\mathbf{x}, \mathbf{y})$ can be factorised into 'uninformative' and 'informative' components $\pi(x, y) = \pi_0(L_\perp x)\mathsf{G}(Lx, y)$, where $\mathsf{G} \in \mathcal{P}(\mathbb{R}^r \times \mathbb{R})$ is a squared-integrable $(r + 1)$-dimensional probability measure satisfying $\mathsf{G} \neq \mathsf{G}_z \otimes \mathsf{G}_y$, and where G_z and G_y denote respectively the marginal of G along the first r variables and the last variable. The link function g is thus the conditional expectation $g(z) = \mathbb{E}_\mathsf{G}[\mathbf{y}|\mathbf{z} = z]$. For instance, a deterministic multi-index model is defined as $\mathsf{G}(z, y) = \mathsf{G}_z(z)\delta(y - g(z))$.

We say a linear subspace $W \subseteq \mathbb{R}^d$ is a *mean dimension-reduction subspace* for $(\mathbf{x}, \mathbf{y})$ if $\mathbf{y} \perp\!\!\!\perp f(\mathbf{x}) \mid \mathsf{P}_W \mathbf{x}$ where P_W is a linear projector for W. (Recall that $f(x) := \mathsf{E}[\mathbf{y} \mid \mathbf{x} = x]$ is the regression function.) Let $W_\cap$ be the intersection of all mean dimension-reduction subspaces for $(\mathbf{x}, \mathbf{y})$. If $W_\cap$ itself is a mean dimension-reduction subspace, then we say that $W_\cap$ is the *central mean subspace (CMS)* for $(\mathbf{x}, \mathbf{y})$. The existence of the CMS for $\mathbf{y} \mid \mathbf{x}$ is guaranteed under rather mild conditions on the support of $\mathbf{x}$ [CL02]. For example, the CMS exists if the support of $\mathbf{x}$ is open and convex. It is therefore natural to consider the multi-index model (1) under conditions that ensure the existence of the CMS, and then to let L be a matrix whose rows form an orthonormal basis for the CMS. This setup ensures the identifiability of $W := \mathrm{ran}(L^\intercal)$. We adopt this setup throughout the survey (possibly with additional assumptions that also imply the existence of the CMS).

We will be interested in this question from two different perspectives: first, by viewing the multi-index model as a goal on itself, we will study dedicated algorithms, culminating in 'optimal' methods, in a sense that will be precised later. Next, by viewing multi-index models as a template for 'feature learning', we will describe the behavior of a 'canonical' high-dimensional learning algorithm, notably gradient-descent methods on simple neural networks, when fed data generated by a multi-index model.

2.2 Information-Theoretic Estimation Limits

Under general conditions, and assuming that the previous identifiability condition, one expects that the required number of samples to estimate $W = \mathrm{ran}(L^\perp)$ up to error ϵ will be of order $n = O(dr/\epsilon^2)$, by standard covering arguments [DH24, DPVLB24]. Indeed, this estimator is constructed by building an appropriate ϵ-net over the manifold of r-dimensional subspaces $W_1, \ldots, W_N$ in $\mathbb{R}^d$, of dimension $\sim dr$, and choosing the subspace of highest likelihood based on the observed data. This estimator is however generally intractable, since it amounts to optimizing a non-convex objective. The natural question is therefore how to design efficient algorithms to estimate the subspace, and what is their required sample complexity. This will be the focus of the next sections.

2.3 Notations

We use bold face symbols (e.g., $\mathbf{x}$, $\mathbf{z}$) to denote random variables. Let γ denote the standard Gaussian distribution on the real line, and γ_d denote the standard Gaussian distribution in $\mathbb{R}^d$. We consider the Lebesgue space $L^2(\mathbb{R}^d, \gamma_d)$ and write inner products as $\langle f, g \rangle_{\gamma_d} = \mathsf{E}_{\mathbf{z} \sim \gamma_d}[f(\mathbf{z})g(\mathbf{z})]$. Let $H_k \colon \mathbb{R}^d \to (\mathbb{R}^d)^{\otimes k}$ denote the order-k (normalized probabilist's) Hermite tensor [McC18], given by

$$H_k(u) = \frac{(-1)^k}{\sqrt{k!}} \frac{\nabla^k \gamma_d(u)}{\gamma_d(u)}, \quad u \in \mathbb{R}^d.$$

When $d = 1$, H_k becomes the degree-k (normalized probabilist's) Hermite polynomial; the normalization is in $L^2(\mathbb{R}, \gamma)$, ensuring $\mathsf{E}_{\mathbf{z} \sim \gamma}[h_k(\mathbf{z})h_l(\mathbf{z})] = \mathbb{1}\{k = l\}$. We use $\mathrm{Stiefel}(d, r)$ to denote the Stiefel Manifold of $d \times r$ orthogonal matrices, and $\mathrm{Grassman}(d, r)$ for the Grassmann Manifold of r-dimensional subspaces of $\mathbb{R}^d$.

3. THE GAUSSIAN SETTING

We will first describe a general framework to estimate the support that makes strong assumptions on the input data distribution π, but in exchange enables very general choices for the link function. In particular, the Gaussian setting where $\mathbf{x} \sim \gamma_d$ will take centerpiece, even if some of the methods described below extend to more general settings.[2] As discussed in Section 2.1, this guarantees the identifiability of $W := \mathrm{ran}(L^\intercal)$. For simplicity, we also

---

[2]The remarkable work of [CKK$^+$24a] uses a smoothed analysis framework to analyze algorithms for learning multi-index models under *worst-case* distributions on $\mathbf{x}$. In their framework, the goal of es-

assume g is smooth (though it will be clear that only weak derivatives are needed). These assumptions enable fairly simple moment-based estimators of W that are computationally tractable.

3.1 Linear estimator

We initiate our survey with arguably the simplest estimator for the support in the single-index case (where $r = \dim(W) = 1$). The following linear estimator was studied by [Bri82]: the estimate of W from an i.i.d. sample $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)$ is the line spanned by

$$(3) \qquad \hat{\mathbf{v}} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i \mathbf{x}_i.$$

(This estimator is related to the "Average Derivative Estimator" of [HS89], as the analysis below will make clear.) In this case, we may write $L = u^\intercal$ for a unit vector $u \in S^{d-1} := \{v \in \mathbb{R}^d : \|v\|_2 = 1\}$ that spans W. By Stein's lemma [Ste81] and the chain rule,

$$
\begin{aligned}
\mathsf{E}[\hat{\mathbf{v}}] = \mathsf{E}[\mathbf{y}\mathbf{x}] &= \mathsf{E}[f(\mathbf{x})\mathbf{x}] \\
&= \mathsf{E}[\nabla f(\mathbf{x})] = \mathsf{E}[g'(u^\intercal\mathbf{x})]u \\
&= \mathsf{E}[g'(\mathbf{z})]u
\end{aligned}
$$

where $\mathbf{z} \sim \gamma$ is a standard normal random variable. The mean of $\hat{\mathbf{v}}$ is therefore always in W. Moreover, if the expected derivative of g is non-zero—i.e., $\mathsf{E}[g'(\mathbf{z})] \neq 0$— then the law of large numbers implies the consistency of the linear estimator for estimating W. Indeed, using the rotational symmetry of the Gaussian measure, we can assume w.l.o.g. that $\mathbf{y} = g(\mathbf{x}_1)$, and thus $\mathsf{E}[\|\mathbf{y}\mathbf{x}\|^2] = O(1) + (d-1) \simeq d$, leading to

$$\frac{\mathsf{E}[\|\hat{\mathbf{v}} - \mathsf{E}[\hat{\mathbf{v}}]\|^2]}{\|\mathsf{E}[\hat{\mathbf{v}}]\|^2} \simeq \frac{d}{n} \, ,$$

indicating that $n \gg d$ samples are sufficient to obtain an accurate estimate in this setting of $\mathsf{E}[g'(\mathbf{z})] \neq 0$. For a more precise statement we refer the reader to [DPVLB24, Lemma F.11].

3.2 Noisy one-bit compressed sensing

[PV12] studied the linear estimator in the context of noisy one-bit compressed sensing, where the aim is to recover a signal vector u using linear measurements $u^\intercal\mathbf{x}$ that are quantized, say, to values in $\{-1, 1\}$. A natural "noise-free" variant of this problem assumes $\mathbf{y} = \operatorname{sign}(u^\intercal\mathbf{x})$. More generally, the measurements may be corrupted by noise in a way such that $\mathsf{E}[\mathbf{y} \mid \mathbf{x}] = g(u^\intercal\mathbf{x})$ for some unknown function $g \colon \mathbb{R} \to [-1, 1]$, so there is

some chance that $\mathbf{y} \neq \operatorname{sign}(u^\intercal\mathbf{x})$. For example, if the sign is flipped independently with a constant probability $\eta \in [0, 1]$, then $g(z) = (1 - 2\eta)\operatorname{sign}(z)$. (This special case was studied by [Ser99] under the guise of PAC learning homogeneous half-spaces with random classification noise under spherically symmetric distributions.) The requirement that $\mathsf{E}[g'(\mathbf{z})] \neq 0$ can be regarded as a minimum signal strength condition for the linear estimator to work. For example, this is satisfied by strictly monotone link functions like $g(z) = \tanh(z)$. (Note that it is different from assuming $\Pr[\mathbf{y} \neq \operatorname{sign}(u^\intercal\mathbf{x})] \neq 1/2$.)

In compressed sensing, it is often assumed that the signal u comes from a structured set, say, $K \subseteq B^d$, where B^d is the d-dimensional Euclidean unit ball. For example, K may be the set of sparse vectors, or vectors with low ℓ^1-norm. [PV12] reinterpret the linear estimator as being the line spanned by

$$(4) \qquad \hat{\mathbf{u}} := \arg\max_{u \in K} \left\langle u, \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i \mathbf{x}_i \right\rangle,$$

with $K = B^d$; they propose using the structured set K in (4) when one has the prior knowledge $u \in K$. Under the assumption $\mathsf{E}[g'(\mathbf{z})] \neq 0$, they show that the sample size m needed by this modified linear estimator to accurately estimate u scales only with (the square of) the *Gaussian mean width* $\mathsf{w}(K)$ *of* K, which can be much smaller than the dimension d. For example, if $K = \{v \in \mathbb{R}^d : \|v\|_2 \leq 1, \|v\|_1 \leq \sqrt{s}\}$, then $\mathsf{w}(K) = O(\sqrt{s\log(2d/s)})$. Furthermore, if K is convex, then optimization problem in (4) is a convex optimization problem, which can be solved efficiently under fairly mild conditions on K.

In summary, the assumption $\mathsf{E}[g'(\mathbf{z})] \neq 0$ enables efficient estimation methods with a rate matching the information-theoretic bound, and are thus 'optimal' in this sense of sample complexity. The natural question is then to understand how to proceed when this property does not hold.

3.3 Principal Hessian Directions

For the general multi-index case (where $r > 1$ is allowed), [Li92] proposed the "Principal Hessian Directions" (PHD) estimator, which in the present setting with normal $\mathbf{x}$, is defined by the span of the eigenvectors corresponding to the r largest (in magnitude) eigenvalues of

$$(5) \qquad \widehat{\mathbf{M}} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i (\mathbf{x}_i \mathbf{x}_i^\intercal - I_d).$$

Using (the second-order version of) Stein's lemma and the chain rule again,

$$
\begin{aligned}
\mathsf{E}[\widehat{\mathbf{M}}] = \mathsf{E}[\mathbf{y}(\mathbf{x}\mathbf{x}^\intercal - I_d)] &= \mathsf{E}[f(\mathbf{x})(\mathbf{x}\mathbf{x}^\intercal - I_d)] \\
&= \mathsf{E}[\nabla^2 f(\mathbf{x})] = L^\intercal \mathsf{E}[\nabla^2 g(L^\intercal\mathbf{x})]L \\
&= L^\intercal \mathsf{E}[\nabla^2 g(\mathbf{z})]L
\end{aligned}
$$

---

timation is to compete against smoothed target functions. Also, the works of [KSV24, CKK$^+$24b] consider learning in the presence of distribution-shift, where the distribution of $\mathbf{x}$ may differ from Gaussian at "test time".

where $\mathbf{z} \sim \gamma_r$ is now an r-dimensional standard normal random vector. The range of the mean of $\widehat{\mathbf{M}}$ is therefore a subspace of W. If, additionally, the expected Hessian $\mathsf{E}[\nabla^2 g(\mathbf{z})]$ is non-singular, then the range of $\mathsf{E}[\widehat{\mathbf{M}}]$ is, in fact, equal to W; we say that the estimator is *exhaustive* in this case. When PHD is exhaustive, the law of large numbers implies that it is a consistent estimator of W. Again, one can verify [MM18, LL20] that under this non-singular assumption (playing the analog of the moment assumption $\mathsf{E}[g'(\mathbf{z})] \neq 0$ in the linear case), and in the single-index setting, the required sample complexity to recover an accurate estimate is $n = \Theta(d)$, with a constant that depends on the link function.

3.4 Application: Learning Convex Concepts

A dimension-reduction technique of [Vem10] for PAC learning low-dimensional convex concepts turns out to be a special case of PHD. Suppose $f(x) = \mathbb{1}\{x \in K\}$ for some convex set $K \subseteq \mathbb{R}^d$ whose supporting hyperplanes' normal vectors span W. In other words, determining membership of x in K is equivalent to checking membership of $\mathsf{P}_W x$ in W. The goal is to learn a hypothesis $h \colon \mathbb{R}^d \to \{0,1\}$ with low error rate $\Pr[h(\mathbf{x}) \neq f(\mathbf{x})]$. If K is a full-dimensional convex set, then $W = \mathbb{R}^d$, and the problem appears to be computationally intractable in general (see, e.g., [KOS08]). However, if $\dim(W)$ is small, then dimension reduction can be used to reduce the computational difficulty.

Assume for simplicity that $K \cap W$ is a symmetric convex body in W with positive probability mass. First, since $\mathbf{y}$ is $\{0,1\}$-valued, we can write $\mathsf{E}[\widehat{\mathbf{M}}]$ as

$$\mathsf{E}[\widehat{\mathbf{M}}] = \mathsf{E}[\mathbf{y}(\mathbf{x}\mathbf{x}^\intercal - I_d)] = \mathsf{E}[\mathbf{y}]\mathsf{E}[\mathbf{x}\mathbf{x}^\intercal - I_d \mid \mathbf{y} = 1].$$

Furthermore, we have $\mathsf{E}[\mathbf{y}] = \Pr[f(\mathbf{x}) = 1] \neq 0$ by assumption. If $v \in W^\perp \cap S^{d-1}$, then

$$v^\intercal \mathsf{E}[\widehat{\mathbf{M}}]v = \mathsf{E}[\mathbf{y}](\mathsf{E}[(v^\intercal \mathbf{x})^2 \mid \mathbf{y} = 1] - 1)$$
$$= \mathsf{E}[\mathbf{y}](\mathsf{E}[(v^\intercal \mathbf{x})^2] - 1) = 0$$

since $v^\intercal \mathbf{x} \perp\!\!\!\perp \mathbf{y}$. Moreover, for $v \in W \cap S^{d-1}$,

$$v^\intercal \mathsf{E}[\widehat{\mathbf{M}}]v = \mathsf{E}[\mathbf{y}](\mathsf{E}[(v^\intercal \mathbf{x})^2 \mid \mathbf{y} = 1] - 1)$$
$$= \mathsf{E}[\mathbf{y}](\mathrm{var}(v^\intercal \mathbf{x} \mid \mathbf{y} = 1) - 1) < 0$$

since the variance of a truncated standard normal distribution is strictly less than one; see [KSV24, Appendix B and Appendix C] for quantitative bounds. This shows that the range of $\mathsf{E}[\mathbf{y}(\mathbf{x}\mathbf{x}^\intercal - I_d)]$ is precisely W. Projecting the data to the PHD subspace reduces the computational difficulty of learning convex concepts (say, using the general polynomial regression technique of [KOS08]), because one now only has to work in r-dimensional space as opposed to the original d-dimensional space.

3.5 The Information Exponent for Single-Index Models

An important limitation of the linear and PHD estimators is that they may be non-exhaustive depending on the link function. For example, $g(z) = z^2$ (as considered in the "phase retrieval" problem) has $\mathsf{E}[g'(\mathbf{z})] = \mathsf{E}[2\mathbf{z}] = 0$, so the linear estimator is non-exhaustive in this case. And $g(z) = z^3 - 3z$ has $\mathsf{E}[g'(\mathbf{z})] = \mathsf{E}[3\mathbf{z}^2 - 3] = 0$ and $\mathsf{E}[\nabla^2 g(\mathbf{z})] = \mathsf{E}[6\mathbf{z}] = 0$, so both the linear and PHD estimators are non-exhaustive in this case.

At this point, it is apparent that there should be a general principle at play that relates a generic estimation procedure with a structural property of the link function, the number of vanishing moments of the form $\mathsf{E}[\mathbf{y}p(\mathbf{x})]$ where p is a certain polynomial family.

Let us now illustrate this relationship by considering a natural generic strategy, namely performing Maximum-Likelihood Estimation via gradient-ascent. For that purpose, let $\mathbf{x} \sim \gamma_d$, and $\mathbf{y} = g(\theta_*^\intercal \mathbf{x}) + \xi$, where $\theta_* \in \mathcal{S}_{d-1}$ is the planted direction and ξ is a Gaussian noise independent of $\mathbf{x}$. Given iid samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_i$ from this model, the MLE estimator in the parametric class $\theta \mapsto \mathbf{y}|\mathbf{x} = \mathcal{N}(g(\theta^\intercal \mathbf{x}), 1)$ is proportional to

$$(6) \qquad \hat{L}(\theta) = \frac{1}{n}\sum_i (\mathbf{y}_i - g(\theta^\intercal \mathbf{x}_i))^2 \ .$$

It is instructive to consider the population limit of this empirical landscape, given by

$$(7) \qquad L(\theta) := \mathsf{E}[|g(\theta_*^\intercal \mathbf{x}) - g(\theta^\intercal \mathbf{x})|^2] + \mathsf{E}[\xi^2] \ .$$

Denoting $g_\theta(x) = g(\theta^\intercal x)$ and using the rotation symmetry of the Gaussian measure, this landscape is, up to a constant, equivalent to the correlation $\langle g_\theta, g_{\theta_*} \rangle_{\gamma_d}$. Clearly, this Gaussian inner product is only a function of two (correlated) scalar Gaussian variables, in the span of θ, θ_*. By parametrizing them as $z \sim \gamma_1$ and $\tilde{z} = mz + \sqrt{1 - m^2}w$, with $z, w \sim \gamma_1$ independent and $m = \theta^\intercal \theta_*$, we obtain that

$$(8) \qquad \langle g_\theta, g_{\theta_*} \rangle_{\gamma_d} = \langle g, \mathsf{A}_m g \rangle_{\gamma_1} \ ,$$

where, for $m \in [-1, 1]$, A_m is the Ornstein-Uhlenbeck (OU) semigroup given by

$$(9) \qquad \mathsf{A}_m g(z) = \mathsf{E}_w[g(mz + \sqrt{1 - m^2}w)] \ .$$

This commutative semigroup jointly diagonalises in $L^2(\mathbb{R}, \gamma)$, with eigenfunctions given by *Hermite* polynomials $\{h_k\}_k$ satisfying $\mathsf{A}_m h_k = m^k h_k$. This orthogonal structure turns out to be sufficient to provide an explicit description of the geometry of the MLE landscape above. Indeed, by decomposing the link function $g \in L^2(\mathbb{R}, \gamma)$ using this Hermite basis, $g = \sum_k \alpha_k h_k$, we obtain

$$(10) \qquad \langle g_\theta, g_{\theta_*} \rangle = \sum_k \alpha_k^2 m^k \ .$$

This representation of the loss reveals two important properties: on the one hand, thanks to the rotational invariance of the gaussian measure, this high-dimensional landscape is in fact only a function of a single scalar, the *overlap* $m = \theta^\intercal \theta_*$. Additionally, not only is this landscape secretly one-dimensional, it also has a particularly simple topology: thanks to the fact that $\ell(m) := \sum_k \alpha_k^2 m^k$ satisfies $\ell'(m) > 0$ for $m > 0$, there is a single saddle point at the equator $\{\theta; m = 0\}$ within the upper hemisphere $\{\theta; m \geq 0\}$; in particular, the loss has a single maximiser (which is global) . From (10), one can also deduce that the order of the equatorial saddle is given by the index $l^\star$ of the first non-zero coefficient α_k — precisely the number of vanishing moments of g, and referred as the *information exponent* of g in [AGJ21] (or *order of degeneracy* in [DH18]).

As a result, one should expect that from a random initialisation $\theta_0 \sim \text{Unif}(S_{d-1})$ satisfying $m_0 \geq 0$ (which holds with probability $1/2$), a local method such as (projected) gradient ascent should converge towards the 'North pole' θ_* in the population limit. Beyond this qualitative behavior, the effect of the high-dimension is felt once one tries to quantify this convergence in presence of only a finite number of samples. Indeed, a typical initialisation θ_0 has correlation $m_0 \simeq 1/\sqrt{d}$, indicating that the gradient dynamics will be initialized at a neighborhood of the saddle point — the so-called *mediocrity zone*. Now, we can view the empirical landscape (7) as a 'noisy' perturbation of L, in which $\|\nabla \hat{L}(\theta) - \nabla L(\theta)\| \simeq \sqrt{\frac{d}{n}}$ uniformly over θ under mild smoothness assumptions [MBM17]. The empirical landscape will thus be successfully optimized whenever the signal gradient 'strength', of order $\simeq m_0^{l^\star - 1} \simeq d^{-(l^\star - 1)/2}$, dominates the empirical fluctuations, of order $\sqrt{d/n}$. In other words, whenever $n \gg d^{l^\star}$, gradient methods will successfully optimise the MLE objective.

This analysis was put forward in the seminal works of Ben Arous, Gheissari and Jagannath [AGJ21], as well as Dudeja and Hsu [DH18], using slightly better algorithms. Specifically, [AGJ21] considered online SGD, which replaces the uniform gradient concentration with a sharper martingale analysis along the trajectory, leading to a slight improvement in the rate, to $O(d^{l^\star - 1})$ [3]. On the other hand, [DH18] show that only two steps of a "gradient iteration" suffices:

$$\hat{\theta}_0 \xrightarrow{\mathbf{G}} \hat{\theta}_1 \xrightarrow{\mathbf{G}} \hat{\theta}_2 \ ,$$

where

$$\theta \mapsto \mathbf{G}(\theta) := \frac{1}{\|\nabla \mathbf{F}_{l^\star}(\theta)\|_2} \nabla \mathbf{F}_{l^\star}(\theta) \ ,$$

---

[3] The cases $l^\star = 1$ and $l^\star = 2$ yield rates of $O(d)$ and $O(d \log d)$ respectively. This rate follows by studying the growth of a polynomial ODE of the form $\dot{m} = (1 - m^2) m^{l^\star - 1}$, which behaves differently for $l^\star > 2$.

and $\mathbf{F}_l$ is the orthogonal projection of the correlation onto the l-th harmonic, ie $\mathbf{F}_l(\theta) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i h_l(\theta^\intercal \mathbf{x}_i)$. We remark that the previous estimation procedure can be extended even in the setting where the information exponent $l^\star$ is unknown, via a correlation-based goodness-of-fit criterion [DH18].

Finally, these improvements culminated in Damian et al. [DNGL23], who used a landscape smoothing procedure first introduced in the physics literature [BCRT20], akin to the partial trace estimation from tensor methods [Hop18]. While the PHD estimator (5) consists in extracting the principal eigenvectors of the matrix $\widehat{\mathbf{M}}$, Damian et al consider instead the empirical tensor

$$(11) \qquad \widehat{\mathbf{T}} := \frac{1}{n} \sum_i \mathbf{y}_i H_{l^\star}(\mathbf{x}_i) \ ,$$

where $H_k(x)$ is the k-th order Hermite tensor. In the single-index setting, one easily verifies that $\mathsf{E}[\widehat{\mathbf{T}}] \propto \theta_*^{\otimes l^\star}$ is a rank-one tensor, and thus one can view the estimation of θ as a non-iid version of Tensor PCA, where from the observed data one builds a tensor $\widehat{\mathbf{T}}$ with planted rank-one structure — but where entries are correlated, as opposed to Tensor PCA. Nonetheless, [DNGL23] leverage the partial-trace estimator, an efficient spectral method that boosts the signal-to-noise ratio by averaging the tensor in directions where the signal is constant, to estimate the planted direction with a sample complexity of $O(d^{l^\star/2})$, which turns out to be optimal amongst the class of correlation-based statistical query algorithms [DLS22, ABAM23].

Finally, let us mention that the geometric picture brought by the information exponent is robust to small perturbations of the Gaussian data distribution. Indeed, [BPVZ23] demonstrates that spherical symmetry is sufficient, and that distributions whose sliced Wasserstein distance is of order $O(1/\sqrt{d})$ from the Gaussian reference also result in an efficient MLE gradient estimation whenever the information exponent is at most 2.

3.6 Leap Exponents for Multi-Index Models

The previous section outlined a framework for single-index models that exploits the Hilbertian structure of the correlation loss to obtain an explicit 'decoupling' of the two ingredients of single-index models, namely the hidden direction θ_* and the link function g. In essence, this is achieved thanks to the joint diagonalization of the OU semigroup via Hermite polynomials.

A natural question is then how to extend this framework to the general multi-index setting. Denoting again $g_W(x) = g(W^\intercal x)$ for $W \in \text{Stiefel}(d, r)$, the first step is to generalise the representation (8); it now writes

$$(12) \qquad \langle g_W, g_{W_*} \rangle_{\gamma_d} = \langle g, \mathsf{A}_M g \rangle_{\gamma_r} \ ,$$

where now A_M is the matrix semigroup given by

$$\text{(13)} \qquad \mathsf{A}_M g(z) = \mathsf{E}_{w \sim \gamma_r}[g(Mz + \sqrt{I - M^\intercal M} w)] \,,$$

defined for correlation matrices $M = W^\intercal W_* \in \mathbb{R}^{r \times r}$ such that $\|M\| \leq 1$. An important point to realize is that now $\{\mathsf{A}_M\}_M$ is no longer commutative, and thus one cannot hope to recreate the satisfying Hermite orthonormal decomposition from the single-index setting straight away. Instead, the correlation $\langle g_W, \tilde{g}_{\tilde{W}} \rangle_{\gamma_d}$ between two multi-index functions is now expressed as

$$\text{(14)} \quad \langle g_W, \tilde{g}_{\tilde{W}} \rangle_{\gamma_d} = \sum_\beta \langle g, H_\beta(U) \rangle \langle \tilde{g}, H_\beta(V) \rangle \prod_{j=1}^r \lambda_j^{\beta_j} \,,$$

where $M = W^\intercal \tilde{W} = U\Lambda V^\intercal$ is the Singular-Value Decomposition (SVD) of M, $\beta = (\beta_1, \ldots, \beta_r) \in \mathbb{N}^r$ are multi-indices, and $H_\beta(U)$ is the tensorised Hermite polynomial $H_\beta(U)(x) = \prod_{j=1}^r h_{\beta_j}(U_j^\intercal x)$ associated with the orthonormal basis U of $\mathbb{R}^r$.

While the squared-loss still admits a representation in terms of low-dimensional summary statistics $M \in \mathbb{R}^{r \times r}$, the monotonicity of the gradients is now lost: the effect of subspace internal misalignments (captured by the fact that $U \neq V$) can create spurious maxima in the MLE landscape. Such difficulty can be overcome by considering instead a two-timescale bilevel algorithm [BBPV23], that learns the link function g in a non-parametric class of r-dimensional functions in the inner loop. This effectively replaces the correlation matrix M by its Grammian $G = M^\intercal M$, leading to a 'symmetrised' optimisation landscape

$$\text{(15)} \qquad L(W) = 2\|g\|^2 - 2\sum_\beta \langle g, H_\beta(V) \rangle^2 \prod_{j=1}^r \lambda_j^{2\beta_j}$$

that recovers the benign geometrical properties of its single-index counterpart.

In particular, [ABAM23, BBPV23] show that local gradient methods on this landscape evolve along *saddle-to-saddle* dynamics, as opposed to the single 'escape from mediocrity saddle' of the single-index setting. The dynamics are still initialized at a neighborhood of a saddle point, satisfying $\|M_0\| \simeq 1/\sqrt{d}$. Next, instead of directly escaping towards the global minimiser $W = W_*$, the gradient dynamics will visit a (finite) sequence of saddle points before reaching the global optimum. Remarkably, one can characterise these intermediate saddle points in terms of a hierarchical decomposition of the link function based on multivariate Hermite expansions. Informally speaking, the span of the lowest 'frequencies' of g is learnt first; subsequently, the gradient flow dynamics reveal the directions in the orthogonal complement carrying the lowest remaining frequencies, and so on. Each step in this sequential decomposition is characterized by a *leap dimension* and *leap exponent*, corresponding respectively to the index and the order of the associated saddle point.

An important special setting where this sequential decomposition is particularly accessible is given by *staircase* functions, first introduced by Abbe et al. in [ABAB+21] in the context of Boolean functions (see also the earlier work of [KST09] in a related context). In short, a function has the staircase property if the support of their high-frequency coefficients can be 'reached' from the support of low-frequencies by adding a single coordinate, e.g., $g(z) = z_1 + z_1 z_2 + z_1 z_2 z_4$, for $z \in \{\pm 1\}^d$. The staircase property was further developed in [ABAM22, ABAM23], where a sample complexity of order $d^{l^\star - 1}$ is established for a certain class of generalised staircase functions, where $l^\star$ is the largest leap exponent (by slightly abusing notation), using local gradient methods on a suitable NN architecture. [BBPV23] shows that $l^\star$ indeed controls the complexity in the general setting, focusing on population dynamics.

Finally, another important instance of multi-index models with explicit algorithmic performance is given by *committee machines*, corresponding to a link function of the form $g(z) = \text{sign}\left(\sum_{j=1}^r \text{sign}(z_j)\right)$ [AMB+19]. In this setting, and in the language of the leap exponents, one can verify that there is energy in the linear harmonic, and thus there is no saddle-point structure, leading to a guarantee of the form $n = \Theta(d)$. Such guarantees were recently extended to a broader class of multi-index functions having non-zero linear terms in [DKL+23].

In conclusion, algorithms based on local gradient ascent on the correlation are shown to be a powerful tool to estimate the index space, with sample complexity driven by the number of vanishing Hermite moments. One can now wonder to what extent these correlation methods are 'optimal', at least in terms of said sample complexity.

3.7 Beyond Information and Leap Exponents via Inverse Regression

As it turns out, correlation-based algorithms are not optimal in the previous sense. One of the clearest instances of this lack of optimality is the seminal work of Chen and Meka [CM20]: they showed that when g is a r-dimensional, degree-s polynomial, there exists an efficient algorithm that only needs

$$n = O_{r,s}(d \log^2(1/\epsilon)(\log d)^s)$$

samples to recover the subspace up to error ϵ in a noise-free setting.

The sample complexity is thus *linear* in the dimension, matching the optimal information-theoretic rate. Taking for instance $g = h_s$ above shows an inherent advantage of this method over the previous correlation-based methods, which require $\Theta(d^{s/2})$ samples. In essence, Chen and

Meka exploit the *inverse regression* idea, first put forward in Li's seminal work [Li91]: rather than focusing solely in the regression task of modeling $\mathbf{y}|\mathbf{x}$, one can gain information by looking instead at the inverse (or 'generative') model $\mathbf{x}|\mathbf{y}$. The key observation is that, under Gaussian assumptions on $\mathbf{x}$, the moments $\mathsf{E}[\mathbf{x}^{\otimes k}|\mathbf{y}=y]$, seen as tensor-valued functions of y, are only non-constant within the index space.[4] This idea has led to several algorithmic solutions that exploit this moment structure [Coo00, LW07, KLV20].

Another seemingly separate line of work, notably [MM18, BKM$^+$19, MLKZ20], aimed to characterize the class of single-index models such that one can efficiently recover the hidden direction θ in the so-called *proportional regime*, where the number of samples satisfies $n = \alpha d$ with fixed $\alpha > 0$, and d goes to infinity. Specifically, these authors showed that whenever the joint distribution $(\mathbf{x}, \mathbf{y})$ satisfies

$$(16) \quad \max\left\{ \mathsf{E}[\mathsf{E}[(\theta_*^\intercal \mathbf{x})^2 - 1|\mathbf{y}]^2], \mathsf{E}[\mathsf{E}[\theta_*^\intercal \mathbf{x}|\mathbf{y}]^2] \right\} > 0 \,,$$

then there are efficient algorithms that estimate θ provided $n \gtrsim d$. We identify the flavor of inverse regression in (16), through the conditioning on the label $\mathbf{y}$. Let us add that these works go beyond this guarantee, and even identify fundamental thresholds $\alpha_{\mathrm{I}} \leq \alpha_{\mathrm{C}}$ capturing the required sample complexity for brute-force and efficient estimators, respectively [5]. In particular, [LL20, MM18] considered a variant of the PHD estimator, where $\hat{\theta}$ is the principal eigenvector of

$$(17) \qquad \widetilde{\mathbf{M}} = \frac{1}{n} \sum_i \mathcal{T}(\mathbf{y}_i)(\mathbf{x}_i \mathbf{x}_i^\intercal - I_d) \,,$$

and where $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ is a suitable label transformation. On the other hand, [BKM$^+$19] developed an approach based on Approximate-Message-Passing (AMP) under the same assumption (16), in which there is also an underlying label transformation — as is also the case in the aforementioned [CM20]. Finally, in [MKLZ22] it was shown that the spectral estimator associated with (17) is the linearisation of the AMP through the so-called Bethe Hessian.

---

[4]Similar results hold under slightly milder conditions on $\mathbf{x}$; see [KLV20].

[5]These results are an instance of the so-called *Computational-Statistical Gaps*, which study the interplay between statistical and computational aspects in high-dimensional inference. In particular, some inference tasks are known to exhibit a phase diagram with three distinct regimes: for $n \ll n_{\mathrm{I}}$, no estimator can recover the signal of interest, for $n \gg n_{\mathrm{C}}$, one can exhibit efficient estimation procedures that succeed, but for $n_{\mathrm{I}} \ll n \ll n_{\mathrm{C}}$, it is conjectured that *no* efficient procedure can exist. Such conjecture is often formalized by restricting the class of estimation procedures to belong to a certain computational class, e.g., Statistical Queries [Kea98], or Low-degree polynomials [HSSS16]

In retrospect, it should not be surprising that a label transformation is able to overcome the information exponent barriers from the previous section; indeed, recalling that $l^\star(\mathsf{G}) = \min\{l; \mathsf{E}[\mathbf{y}h_l(\mathbf{x})] \neq 0\}$, one can easily verify that this definition is not invariant to compositions, ie it can happen that the change of variables $\tilde{\mathsf{G}} := (\mathrm{Id} \otimes \mathcal{T})_\# \mathsf{G}$, where $(\mathbf{x}, \mathbf{y})$ is mapped to $(\mathbf{x}, \mathcal{T}(\mathbf{y}))$, satisfies $l^\star(\tilde{\mathsf{G}}) < l^\star(\mathsf{G})$!

Such lack of composition invariance suggests that one should revisit the information exponent definition to ensure it cannot be decreased by label transformation, and thus that it captures the intrinsic difficulty of estimation beyond correlation-based methods. This can be achieved by considering

$$(18) \qquad k^\star(\mathsf{G}) := \inf_{\substack{\mathcal{T}:\mathbb{R}\to\mathbb{R} \\ \text{measureable}}} l^\star(\tilde{\mathsf{G}}) \,.$$

This defines a new *generative exponent* of the single-index model [DPVLB24]. By simple consequence of (18), we have that $k^\star(\mathsf{G}) \leq l^\star(\mathsf{G})$.

This exponent can also be viewed as the leading term of an expansion. While the information exponent appears as the leading order term of the expansion of the L^2 energy $\mathsf{E}[\mathsf{E}[\mathbf{y}|\mathbf{x}]^2] = \|g\|^2 = \sum_l \mathsf{E}[\mathbf{y}h_l(\mathbf{x})]^2 = \sum_{l \geq l^\star} \mathsf{E}[\mathbf{y}h_l(\mathbf{x})]^2$ (and thus associated with correlation), the generative exponent is associated with an expansion of another metric, given by the χ^2 information: assuming that $\mathsf{G} \ll \mathsf{G}_0$, with $\mathsf{G}_0 = \mathsf{G}_z \otimes \mathsf{G}_y$, ie that G is absolutely continuous with respect to the product of its marginals, the χ^2 information is defined as $I(\mathsf{G}) = D_{\chi^2}(\mathsf{G}||\mathsf{G}_0) = \left\|\frac{\mathrm{d}\mathsf{G}}{\mathrm{d}\mathsf{G}_0}\right\|_{\mathsf{G}_0}^2 - 1$. Defining for each k the conditional moment $\zeta_k(y) = \mathsf{E}[h_k(\mathbf{z})|\mathbf{y}=y]$, one can verify [DPVLB24] that

$$(19) \qquad I(\mathsf{G}) = \sum_k \|\zeta_k\|_{\mathsf{G}_y}^2 = \sum_{k \geq k^\star} \|\zeta_k\|_{\mathsf{G}_y}^2 \,.$$

We now recognize that (16) corresponds to the assumption $k^\star \leq 2$. The conditional moments in ζ_k fully realize the vision of Li's Inverse Regression: indeed, we now extract *all* the moments of the conditional distribution $\mathbf{x}|\mathbf{y}$ rather than focusing only on the first few moments.

The representation (19) leads to a natural generalisation of the L_2 decomposition (10) in terms of the likelihood ratios $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_0}$. Equation (19) thus reveals that the likelihood ratio is in the span of polynomials of degree at least $k^\star$ in $\mathbf{x}$. As a consequence, one can deduce [DPVLB24] a lower bound for the required sample complexity of $n \gtrsim d^{k^\star/2}$, both under the SQ (Statistical-Query) and the Low-degree polynomial frameworks. These computational classes encompass a wide range of algorithms, including spectral and general moment methods, and (stochastic) gradient descent.

Moreover, the function $\zeta_k(y)$ defines the optimal label transformation $\mathcal{T}$ that reduces the generative to the information exponent. Therefore, the natural confluence of

(11) and (17) is to consider the following tensor:

$$(20) \qquad \widetilde{\mathbf{T}} := \frac{1}{n} \sum_i \zeta_{k^\star}(\mathbf{y}_i) H_{k^\star}(\mathbf{x}_i) \ .$$

Indeed, we verify that, again, this tensor has a planted spike precisely given by θ_*, ie $\mathsf{E}[\widetilde{\mathbf{T}}] \propto \theta_*^{\otimes k^\star}$. As a result, one can establish [DPVLB24] that a partial trace estimator will successfully estimate θ_* whenever $n \gtrsim d^{k^\star/2}$, thus tightly matching the computational lower bounds. As with previous algorithms, one can also extend this method to misspecified link distributions G.

Equation (16) has been recently extended to the multi-index setting in [TDD$^+$24], where AMP-based methods are shown to (weakly)[6] recover the range of the following matrix in the proportional regime $n = \Theta(d)$:

(21)
$$W_* \mathsf{E}[\zeta_1(\mathbf{y})\zeta_1(\mathbf{y})^\intercal + \zeta_2(\mathbf{y})\zeta_2(\mathbf{y})^\intercal] \ , \ \text{with}$$
$$\zeta_1(y) = \mathsf{E}[H_1(\mathbf{z})|\mathbf{y} = y] \ , \ \zeta_2(y) = \mathsf{E}[H_2(\mathbf{z})|\mathbf{y} = y] \ .$$

Similarly, [KZM25] recently extended the analysis of [MM18] to the multi-index setting, establishing weak recovery of at least one direction inside the above index space using a spectral method.

Finally, let us mention that lower bounds for multi-index models against several computational classes, including SQ and differentiable queries have been recently established in [JMS24] for product measures admitting a predefined basis aligned with the unknown subspace, and in [DLB25] for Gaussian data for the low-degree polynomial framework. Notably, these SQ/Low-degree lower bounds identify an appropriate extension of the generative exponent to the multi-index setting — the *leap generative* exponent. By adapting the partial trace estimator to this setting, [DLB25] provide an algorithm matching again this sample complexity, therefore providing a 'complete' statistical/computational description of the multi-index estimation problem in the high-dimensional regime (under Gaussian data); see also [DIKR25, DIKZ25] for concurrent works that establish both an SQ lower bound, and also provide an upper bound using a subspace conditioning algorithm.

4. GRADIENT SPAN

One of the critical drawbacks of the moment methods in Section 3 is the normality assumption on $\mathbf{x}$. A standard approach to addressing this limitation is to employ score functions for non-normal distributions on $\mathbf{x}$ [e.g., YBL17], but otherwise the techniques are mostly unchanged. Some moment methods have also been shown

to correctly estimate a subspace of the index space under weaker assumptions [e.g., LD89], but these conditions are still fairly restrictive, and exhaustiveness is also only established under rather *ad hoc* assumptions.

In this section, we describe a rather different approach based on using the span of gradients, which does not rely on the normality assumption and ensures exhaustiveness under fairly weak assumptions. The primary drawback of this method is the difficulty of efficient estimation in high dimensions.

4.1 Exhaustiveness of the gradient span

The key idea of using the gradient span comes from the chain rule (as already exploited in Section 3): the gradient of the regression function ∇f can be written as

$$\nabla f(x) = L^\intercal \nabla g(Lx),$$

so for any $x \in \mathbb{R}^d$, we have $\nabla f(x) \in W := \mathrm{ran}(L^\intercal)$. This suggests a natural strategy for estimating (at least a subspace of) W: estimate the gradient ∇f at several points, and then fit an r-dimensional subspace to these vectors. Or more simply: estimate the *expected gradient outer product (EGOP)*[7]
(22)
$$\Gamma := \mathsf{E}[\nabla f(\mathbf{x})\nabla f(\mathbf{x})^\intercal] = L^\intercal \mathsf{E}[\nabla g(L\mathbf{x})\nabla g(L\mathbf{x})^\intercal]L,$$

and then take the subspace spanned by the eigenvectors corresponding to the top r eigenvalues of the estimate.

The gradient span approach guarantees exhaustiveness under fairly weak assumptions [Sam93]. If there is a nonzero vector $w \in W$ orthogonal to $\nabla f(\mathbf{x})$ almost surely, then $(Lw)^\intercal \nabla g(L\mathbf{x}) = 0$ almost surely. This means that the link function g does not vary in the direction of $v := Lw$, contradicting the minimality of W (recalling the definition of CMS)[8]. Therefore, we must have $W = \mathrm{ran}(\Gamma)$. The exhaustiveness of (the range of) the EGOP makes it a natural target of estimation: a consistent estimate of Γ yields a consistent estimate of W.

4.2 Plug-in estimation

A plug-in approach to estimating the EGOP is:

1. Estimate the evaluation of ∇f on an i.i.d. sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (perhaps using separate data independent of these n points).
2. Form the empirical average the outer products of the estimates of $\nabla f(\mathbf{x}_i)$:
$$\frac{1}{n} \sum_{i=1}^n \widehat{\nabla f}(\mathbf{x}_i)\widehat{\nabla f}(\mathbf{x}_i)^\intercal.$$

---

[6]Weak recovery refers to an estimate $\hat{W}$ such that $\lambda_{\min}(\hat{W}^\top W_*) = \Theta(1)$, recalling that a subspace W_0 drawn uniformly at random satisfies $\lambda_{\min}(\hat{W}^\top W_*) \simeq 1/\sqrt{d}$.

[7]Also called *Outer Product of Gradients (OPG)* [XTLZ02].

[8]To make this argument precise, it suffices to assume convexity of the support of $L\mathbf{x}$.

To continue the plug-in stratregy, one can estimate ∇f by constructing an estimate $\hat{f}$ of f, and then using $\nabla \hat{f}$ as an estimate for ∇f. For example, [Sam93] instantiates this approach with Nadaraya-Watson kernel regression methods [Nad64, Wat64], and [KS22] (as with [MZ06] in passing) do so with Reproducing Kernel Hilbert Space (RKHS) methods. In both cases, the estimator $\hat{f}$ has an analytic form for which the gradient can be explicitly computed.

4.3 Local estimates of gradients

[HJPS01] suggest local linear regression [Sto77] for estimating $\nabla f(x)$ in the present context. Specifically, the estimate of the gradient at $x \in \mathbb{R}^d$ is

$$\widehat{\nabla f}(x) := b(x)$$

where $b(x) \in \mathbb{R}^d$ is obtained from a weighted linear least squares fit to an i.i.d. sample $(\tilde{\mathbf{x}}_1, \tilde{\mathbf{y}}_1), \ldots, (\tilde{\mathbf{x}}_{\tilde{n}}, \tilde{\mathbf{y}}_{\tilde{n}})$:

$$(a(x), b(x)) := \underset{(a,b) \in \mathbb{R} \times \mathbb{R}^d}{\arg\min} \sum_{i=1}^{\tilde{n}} \mathbf{w}_i(x)(a + b^\top(\tilde{\mathbf{x}}_i - x) - \tilde{\mathbf{y}}_i)^2$$

with weights $\mathbf{w}_i(x) = K(\tilde{\mathbf{x}}_i, x)$ for some kernel function $K \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (e.g., a radial kernel like $K(x, x') = \exp(-\|x - x'\|_2^2/h^2)$ for a suitable bandwidth $h > 0$). Related estimation strategies based on Nadaraya-Watson kernel methods, RKHS methods and finite differences were proposed and analyzed, respectively, by [Sam93], [MZ06], and [TWKS14]. The main downside of these approaches is that they suffer from the usual curse of dimension in nonparametric regression [Sto80]: exponentially many (in d) data may be needed for accurate estimation. [Sam93] notes, however, that estimating the EGOP may be easier than estimating the gradient field, and gives conditions under which consistency at the parametric rate may be achieved [see also YXKH23].

[HJPS01] (building on earlier work in the single-index case by [HJS01]) show how a pilot estimate of the EGOP can be used to avoid the curse of dimension in a refined estimate. In particular, in the refined estimate, the (elliptical) smoothing kernel K is chosen so as to over-smooth in directions that the pilot estimate suggests are orthogonal to W. Remarkably, this is enough to overcome the curse of dimension at least in the rate of convergence for the refined estimate: an exponential dependence on the ambient dimension d is replaced by an exponential dependence only on the intrinsic dimension $r = \dim(W)$. In fact, the rate is shown to be the parametric rate when $r \leq 3$ under some mild assumptions on the link function.

4.4 Practical implementation and other settings

The recent work of [RBPB24] combines the RKHS approach of [KS22] and the refinement technique of [HJPS01] to obtain a practical algorithm. Specifically,

they employ RKHS-based regression with an isotropic Gaussian or Laplace kernel to for an initial estimate $\hat{f}$ of f, and then use $\nabla \hat{f}$ to estimate the EGOP as above. Then the estimate of the EGOP is used to change the isotropic kernel to an anisotropic kernel in a manner very similar to that suggested by [HJPS01] to get updated estimates of f and also of the EGOP. They suggest repeating this process several times to obtain a final estimate of f and/or the EGOP.

Manual inspection of the final estimated EGOP in several datasets shows the ability of this approach to identify interpretable "features" that are intuitively relevant for the prediction task at hand. For example, on a dataset for predicting whether an image depicts a smiling face or non-smiling face, the top eigenvectors of the EGOP estimate emphasize pixel features that roughly correspond to those where the eyes and mouth are located in the image. The EGOP estimate obtained this way is also compared to the first layer weight matrices of deep neural networks trained using gradient descent on the same data, and the similarities are striking enough for the authors to posit a hypothesis connecting these methods.

Another remarkable non-parametric approach is the recent work [FB24]; the authors replace the gradient outer product estimation by a more general non-parametric method based on regularized empirical risk minimization, that promotes the 'sparsity' of the estimated function gradients using a multivariate Hermite expansion, under minimal distributional assumptions.

Finally, let us mention [HC12], in which the authors replace the iid sampling assumption by an active learning strategy to estimate the index space, leading to a sample complexity of order $O(d^2)$. In the active (or query) learning setting, it is possible to adaptively sample $\mathbf{y}$ from its conditional distribution given $\mathbf{x} = x$ for any x. This naturally gives a path towards avoiding the curse of dimension suffered by local linear regression and other non-parametric methods (cf. [HJS01, HJPS01] and Section 4.3). Other active learning methods for learning or testing single-index and multi-index models include [DMN19, DMN21, GTX$^+$24, DKK$^+$24].

5. NEURAL NETWORKS

We have so far discussed dedicated algorithms that exploit the specific structure of multi-index models. An alternative line of work considers instead a 'generic' learning algorithm, based on gradient-descent on a differentiable model, e.g., a neural network: given n iid samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \leq n}$ of a multi-index model P, a differentiable function $f_\theta \colon \mathbf{x} \mapsto f_\theta(\mathbf{x})$, and a point-wise loss $\ell(\cdot, \cdot)$, the Empirical-Risk-Minimisation (ERM) consists of

$$(23) \qquad \min_\theta \frac{1}{n} \sum_i \ell(f_\theta(\mathbf{x}_i), \mathbf{y}_i).$$

A canonical choice of loss is the mean-squared error $\ell(x, y) = (x - y)^2$, which is amenable to powerful analysis techniques in $L^2(\mathbb{R}^d, \pi_x)$. The general question in this context is thus to what extent the minimisers $\hat{\theta}$ of (23) carry information on W. When f_θ is a Neural Network (NN) implementing a composition of the form $f_\theta(\mathbf{x}) = \Phi_{\Theta_2}(\Theta_1^\top \mathbf{x})$ for a given $\Theta_1 \in \mathbb{R}^{d \times d'}$ and generic Φ, a natural proxy for $\hat{W}$ is then the span of the first layer Θ_1. As discussed earlier, the motivation in this case is to provide a quantitative description of *feature learning*, namely the ability of the NN to automatically discover the relevant compositional structure for the task — which in the case of multi-index models is directly given by the low-rank projection onto W.

At first glance, it does not seem obvious that a generic gradient-based analysis of (23) should lead to any quantitative guarantees, given the non-convexity of the associated energy landscape, even for the simplest shallow NNs. Faced with this difficulty, [MMN18, CB18, RVE18, SS20] put forward a mean-field formulation of such shallow learning. By expressing the optimization problem in terms of the empirical measure over parameters, one obtains a 'lifted' functional in the space of measures which is convex. Such convexity can then be leveraged to establish global convergence guarantees of gradient dynamics in the infinitely-wide limit. In the setting where the data is generated by a multi-index model, [HC23] showed that the associated dynamics become dimension-free, as with the summary statistics we saw in Section 3.5. However, while providing important qualitative insights about feature learning, an important limitation of these mean-field formulations is their lack of quantitative guarantees as one tries to operate with finite-width networks.

An alternative route is then to 'face' the non-convexity of the optimization problem and attempt to exploit structural properties to overcome the worst-case scenarios. Indeed, in Section 3.5 we observed that the MSE loss $L(\theta) = \|g_\theta - g_{\theta^*}\|^2$ can be efficiently optimized with gradient methods from a random initialization. One can view such model as a contrived neural network consisting of a single neuron, $x \mapsto g(\theta^\top x)$. One could thus hope that gradient-based learning on a more general NN might be able to estimate the index space. Let us now review the main results in this direction.

5.1 Information Exponent and NNs

One of the early attempts that demonstrated the ability of NNs to estimate the index space is [CBL+21]. Focusing on the setting where g is a certain class of degree-s polynomial, the authors show that a partially-trained 3-layer neural network can learn the target function with sample complexity $n = \widetilde{O}(d^{\lceil s/2 \rceil})$, consistent with the CSQ-optimal rate of [DNGL23]. More generally, under the Hessian non-degeneracy assumption from Section 3.3, and assuming Gaussian inputs, [DLS22, BES+22, DKL+23, CPD+24] consider a shallow NN $f_\theta(x) = a^\top \sigma(Wx + b)$ and squared-loss, and show that a *single* gradient step with respect to the input weights W is aligned with $W^\star$. Indeed, picking $\sigma(u) = \max(0, u)$ and denoting $W = [w_1, \ldots, w_m]$, the population gradient of each weight vector w_j writes

$$\nabla_{w_j} \|f_\theta - g_{W^\star}\|^2 \propto \mathsf{E}\left[g_{W^\star}(\mathbf{x})\mathbf{x}\mathbf{1}(\mathbf{x} \cdot w_j \geq 0)\right] .$$

Expanding again this inner product in an orthonormal Hermite basis and identifying the leading order term yields [DLS22, Lemma 1]

$$\nabla_{w_j} \|f_\theta - g_{W^\star}\|^2 = -\sqrt{2/\pi} a_j \mathsf{E}[\nabla^2 g_{W^\star}]w_j + \widetilde{O}(\sqrt{r}/d)) .$$

In words, the gradient at initialization reveals the Principal Hessian Directions. Under the assumption that these directions span the whole index space — which in the language of Section 3.6 corresponds to information exponent $l^\star \leq 2$ and a single leap, this turns out to be sufficient to establish a learning guarantee of order $n = O(d^2)$, via a hybrid gradient-descent strategy whereby second-layer features are kept frozen at initialization. By making the stronger assumption that $l^\star = 1$, one can get sharper rates of $n = O(d)$ [BES+22, MHPG+22], in accordance with Section 3.5.

Beyond the setting of $l^\star \leq 2$, several works also analyse the performance of several NN architectures with gradient-based learning. As mentioned in Section 3.6, Abbe, Boix-Adsera and Misiakiewicz study in [ABAM22, ABAM23] the ability of shallow NNs to learn a class of multi-index models characterized by the *staircase* property. In essence, multi-index functions g with the staircase property (or the 'merged' staircase, an extension put forward in [ABAM22]) admit a spectral decomposition in a certain Hermite basis $H_\beta(V)$ whereby the support of each harmonic is obtained from the lower harmonics by adding additional coordinates. As in [DLS22, DKL+23, BBSS22], the analysis of gradient-descent follows a similar layer-wise strategy, with the second-layer weights frozen at initialization, resulting in a sample complexity guarantee of the form $n = \widetilde{O}(d^{L^\star})$, where $L^\star$ is the *leap* complexity of the model, ie the largest number of revealed coordinates when going from one harmonic to the next.

Besides learning the hidden direction, another important question is to understand how the link function is being learnt. In [BMZ24], the authors study standard shallow NNs in the mean-field regime, and demonstrate that the network naturally learns different harmonics of the target link function at different timescales. Alternatively, [BBSS22] focuses on single-index models, and considers a 'specialized' shallow NN $f_\theta(x) = a^\top \sigma(Wx + b)$, with W constrained to be a rank-one matrix of the form

$W = \mathbf{1}w^\top$, and frozen biases drawn from a normal distribution. In other words, neurons share their input weights and only differ by a random (and frozen) bias. By adapting the analysis of [BAGJ21, DH18], the authors establish non-parametric rates for general link functions with a sample complexity $n = O(d^{l^\star})$ to learn both the index direction and the link function. Remarkably, the efficient algorithm of [CKM22] uses an ingenious enumeration technique based on lattice polynomials to learn any link function specified by a (homogeneous) deep ReLU network.

More recently, [DPC$^+$24] leveraged random matrix theory to study the ability of a single gradient step on a shallow NN to identify the direction in a Single-Index model, in the high-dimensional proportional regime. Finally, while the majority of the analyses focus on Gaussian or Boolean inputs, some authors have explored the effect of anisotropy in the data [MHWSE23, BES$^+$24], highlighting the importance of symmetries in establishing learning guarantees.

5.2 Breaking the Information Exponent with NNs

So far, all methods described require a sample complexity that scales with the information exponent. Based on the arguments from Section 3.7, it is natural to ask whether NNs can go beyond correlation-based learning and 'break' the information-exponent barrier.

As it turns out, the answer, put forward in [DTA$^+$24, ADK$^+$24, LOSW24], is yes. The key idea is that *reusing* data batches to estimate empirical gradients is akin to performing a label transformation that, under appropriate scaling of the training hyperparameters, realises the lowering of the exponent from (18). Indeed, by considering the correlation loss $\ell(\hat{y}, y) = 1 - \hat{y}y$ in (23), and a shallow NN of the form $f_\theta(x) = a^\top \sigma(Wx)$, the gradient with respect to each input weight w_j evaluated at a single datapoint $(\mathbf{x}, \mathbf{y})$ is

$$(24) \qquad \nabla_{w_j}\ell(f_\theta(\mathbf{x}), \mathbf{y}) = -a_j\mathbf{y}\sigma'(w_j \cdot \mathbf{x})\mathbf{x} \ .$$

Now, if we take a gradient step

$$\tilde{W} := W - \rho\nabla_W\ell(f_\theta(\mathbf{x}), \mathbf{y})$$

and evaluate the gradient, using *again* the same datapoint $(\mathbf{x}, \mathbf{y})$, we obtain [ADK$^+$24, Eq 13]

$$\nabla_{w_j}\ell(f_{a,\tilde{W}}(\mathbf{x}, \mathbf{y}), \mathbf{y})$$
$$= a_j\mathbf{y}\sigma'\left(w_j \cdot \mathbf{x} + a\rho\mathbf{y}\|\mathbf{x}\|^2\sigma'(w_j \cdot \mathbf{x})\right)\mathbf{x} \ .$$

An important property of this update is that now this gradient is outside the class of correlational queries, since it is a non-linear function of y. Since the generative exponent from (19) describes a 'closed' property, namely that certain conditional moments $k < k^\star$ are zero, a generic choice of learning rate ρ such that $\rho\|\mathbf{x}\|^2 \approx \rho d = \Theta(1)$ ensures that almost surely the resulting algorithm will now be driven by the generative exponent, rather than the information exponent of single-pass gradient methods. The resulting algorithm is analyzed in [ADK$^+$24, LOSW24] in the setting where $k^\star \leq 2$ [9], leading to a linear sample complexity guarantee.

In the context of multi-index models, the batch reuse is also shown to improve the sample complexities of the correlational queries framework. In particular, the staircase property characterizing efficiently learnable functions using correlation-based (S)GD via the leap exponent is now upgraded to the so-called 'grand staircase' property [TDD$^+$24].

Another strategy to overcome the sample complexity barriers dictated by the information and generative exponents is to consider shallow neural networks in the mean-field regime [MMN18, RVE18, CB18, SS20], trained with noisy gradient descent, referred as mean-field Langevin dynamics. In this setting, [MHWE24] demonstrates information-theoretically optimal sample complexity to learn multi-index models, at the expense of exponential memory and/or runtime.

Finally, let us mention that, thus far, all the works described above consider a certain layer-wise training scheme, whereby only first-layer weights move while the rest are frozen [ABAM23, BBSS22, ADK$^+$24, DLS22, LOSW24], or by introducing timescale separation between layers [BBPV23, BMZ24]. A notable exception to this is the work of Glasgow [Gla23], which, focusing on Boolean data and the 2-parity XOR function, established that $n = \widetilde{O}(d)$ samples are sufficient to learn, using a standard shallow NN and standard mini-batch SGD training.

REFERENCES

[ABAB$^+$21] Emmanuel Abbe, Enric Boix-Adsera, Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. The staircase property: How hierarchical structure can guide deep learning. *arXiv preprint arXiv:2108.10573*, 2021.

[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.

[ABAM23] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. *arXiv preprint arXiv:2302.11055*, 2023.

[ADK$^+$24] Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions. *arXiv preprint arXiv:2405.15459*, 2024.

---

[9] Under a slightly more contrived defintion of generative exponent in [ADK$^+$24], where label transformations are limited to polynomials, and assuming polynomial link functions in [LOSW24]

[AGJ21]   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*, 22(106):1–51, 2021.

[AMB+19]   Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: computational to statistical gaps in learning a two-layers neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124023, 2019.

[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.

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

[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.

[BC64]   G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[BCRT20]   Giulio Biroli, Chiara Cammarota, and Federico Ricci-Tersenghi. How to iron out rough landscapes and get optimal performances: averaged gradient descent and its application to tensor pca. *Journal of Physics A: Mathematical and Theoretical*, 53(17):174003, 2020.

[BD81]   Peter J. Bickel and Kjell A. Doksum. An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311, 1981.

[BES+22]   Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *arXiv preprint arXiv:2205.01445*, 2022.

[BES+24]   Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

[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, 2019.

[BMZ24]   Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2024.

[BPVZ23]   Joan Bruna, Loucas Pillaud-Vivien, and Aaron Zweig. On single index models beyond gaussian data. *arXiv preprint arXiv:2307.15804*, 2023.

[Bri82]   David R. Brillinger. A generalized linear model with 'gaussian' regressor variables. In P. J. Bickel, K. A. Doksum, and J. A. Hodges, editors, *A Festschrift for Erich L. Lehmann*, pages 97–114. CRC Press, 1982.

[CB18]   Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, pages 3036–3046, 2018.

[CBL+21]   Minshuo Chen, Yu Bai, Jason D. Lee, Tuo Zhao, Huan Wang, Caiming Xiong, and Richard Socher. Towards understanding hierarchical learning: Benefits of neural representations. *arXiv preprint arXiv:2006.13436*, 2021.

[CKK+24a] Gautam Chandrasekaran, Adam Klivans, Vasilis Kontonis, Raghu Meka, and Konstantinos Stavropoulos. Smoothed analysis for learning concepts with low intrinsic dimension. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 876–922. PMLR, 2024.

[CKK+24b] Gautam Chandrasekaran, Adam Klivans, Vasilis Kontonis, Konstantinos Stavropoulos, and Arsen Vasilyan. Efficient discrepancy testing for learning with distribution shift. *Advances in Neural Information Processing Systems*, 37:137263–137308, 2024.

[CKM22]   Sitan Chen, Adam R. Klivans, and Raghu Meka. Learning deep relu networks is fixed-parameter tractable. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 696–707, 2022.

[CL02]   R Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002.

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

[Coo00]   R. Dennis Cook. Save: a method for dimension reduction and graphics in regression. *Communications in Statistics-theory and Methods*, 29:2109–2121, 2000.

[CPD+24]   Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue M Lu, Lenka Zdeborová, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. *arXiv preprint arXiv:2402.04980*, 2024.

[CR84]   Raymond J. Carroll and David Ruppert. Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*, 79(386):321–328, 1984.

[DH18]   Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Thirty-First Annual Conference on Learning Theory*, 2018.

[DH24]   Rishabh Dudeja and Daniel Hsu. Statistical-computational trade-offs in tensor PCA and related problems via communication complexity. *The Annals of Statistics*, 52(1):131–156, 2024.

[DHK+21]   Simon Shaolei Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.

[DIKR25]   Ilias Diakonikolas, Giannis Iakovidis, Daniel M. Kane, and Lisheng Ren. Algorithms and sq lower bounds for robustly learning real-valued multi-index models. *arXiv preprint arXiv:2505.21475*, 2025.

[DIKZ25]   Ilias Diakonikolas, Giannis Iakovidis, Daniel M. Kane, and Nikos Zarifis. Robust learning of multi-index models via iterative subspace approximation. *arXiv preprint arXiv:2502.09525*, 2025.

[DKK+24]   Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostically learning multi-index models with queries. In *Sixty-fifth Annual Symposium on Foundations of Computer Science*, pages 1931–1952, 2024.

[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. *arXiv preprint arXiv:2305.18270*, 2023.

[DLB25] Alex Damian, Jason D Lee, and Joan Bruna. The generative leap: Sharp sample complexity for efficiently learning gaussian multi-index models. *arXiv preprint arXiv:2506.05500*, 2025.

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

[DMN19] Anindya De, Elchanan Mossel, and Joe Neeman. Is your function low dimensional? In *Conference on Learning Theory*, pages 979–993, 2019.

[DMN21] Anindya De, Elchanan Mossel, and Joe Neeman. Robust testing of low dimensional functions. In *Fifty-third Annual Symposium on Theory of Computing*, pages 584–597, 2021.

[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.

[DPC$^+$24] Yatin Dandi, Luca Pesce, Hugo Cui, Florent Krzakala, Yue M Lu, and Bruno Loureiro. A random matrix theory perspective on the spectrum of learned features and asymptotic generalization capabilities. *arXiv preprint arXiv:2410.18938*, 2024.

[DPVLB24] Alex Damian, Loucas Pillaud-Vivien, Jason D Lee, and Joan Bruna. Computational-statistical gaps in gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024.

[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.

[FB24] Bertille Follain and Francis Bach. Nonparametric linear feature learning in regression through regularisation. *Electronic Journal of Statistics*, 18(2):4075–4118, 2024.

[Gla23] Margalit Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. *arXiv preprint arXiv:2309.15111*, 2023.

[GTX$^+$24] Aarshvi Gajjar, Wai Ming Tai, Xu Xingyu, Chinmay Hegde, Christopher Musco, and Yi Li. Agnostic active learning of single index models with linear sample complexity. In *Thirty-Seventh Annual Conference on Learning Theory*, pages 1715–1754, 2024.

[HC12] Tyagi Hemant and Volkan Cevher. Active learning of multi-index function models. *Advances in Neural Information Processing Systems*, 25, 2012.

[HC23] Karl Hajjar and Lenaic Chizat. On the symmetries in the dynamics of wide two-layer neural networks. *arXiv preprint arXiv:2211.08771*, 2023.

[HJPS01] Marian Hristache, Anatoli Juditsky, Jorg Polzehl, and Vladimir Spokoiny. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001.

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

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

[HS89] Wolfgang Härdle and Thomas M Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408):986–995, 1989.

[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.

[JMS24] Nirmit Joshi, Theodor Misiakiewicz, and Nathan Srebro. On the complexity of learning sparse functions with statistical and gradient queries. *arXiv preprint arXiv:2407.05622*, 2024.

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

[KLV20] Timo Klock, Alessandro Lanteri, and Stefano Vigogna. Estimating multi-index models with response-conditional least squares. *arXiv preprint arXiv:2003.04788*, 2020.

[KOS08] Adam R Klivans, Ryan O'Donnell, and Rocco A Servedio. Learning geometric concepts via gaussian surface area. In *49th Annual IEEE Symposium on Foundations of Computer Science*, pages 541–550, 2008.

[KS22] Jongkyeong Kang and Seung Jun Shin. A forward approach for sufficient dimension reduction in binary classification. *Journal of Machine Learning Research*, 23(199):1–31, 2022.

[KST09] Adam Tauman Kalai, Alex Samorodnitsky, and Shang-Hua Teng. Learning and smoothed analysis. In *Fiftieth Annual Symposium on Foundations of Computer Science*, pages 395–404, 2009.

[KSV24] Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Learning intersections of halfspaces with distribution shift: Improved algorithms and sq lower bounds. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2944–2978. PMLR, 2024.

[KZM25] Filip Kovačević, Yihan Zhang, and Marco Mondelli. Spectral estimators for multi-index models: Precise asymptotics and optimal weak recovery. *arXiv preprint arXiv:2502.01583*, 2025.

[LD89] Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *The Annals of Statistics*, pages 1009–1052, 1989.

[Li91] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[Li92] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

[LL20] Yue M Lu and Gen Li. Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference: A Journal of the IMA*, 9(3):507–541, 2020.

[LOSW24] Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *arXiv preprint arXiv:2406.01581*, 2024.

[LW07] Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

[MBM17] Song Mei, Yu Bai, and Andrea Montanari. The land-scape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2017.

[McC18] Peter McCullagh. *Tensor methods in statistics: Monographs on statistics and applied probability*. Chapman and Hall/CRC, 2018.

[MHPG$^+$22] Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. *arXiv preprint arXiv:2209.14863*, 2022.

[MHWE24] Alireza Mousavi-Hosseini, Denny Wu, and Murat A Erdogdu. Learning multi-index models with neural networks via mean-field langevin dynamics. *arXiv preprint arXiv:2408.07254*, 2024.

[MHWSE23] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. *Advances in Neural Information Processing Systems*, 36:71449–71485, 2023.

[MKLZ22] Antoine Maillard, Florent Krzakala, Yue M Lu, and Lenka Zdeborová. Construction of optimal spectral methods in phase retrieval. In *Mathematical and Scientific Machine Learning*, pages 693–720. PMLR, 2022.

[MLKZ20] Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. *arXiv preprint arXiv:2006.05228*, 2020.

[MM18] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In *Conference On Learning Theory*, pages 1445–1450. PMLR, 2018.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[MZ06] Sayan Mukherjee and Ding-Xuan Zhou. Learning coordinate covariances via gradients. *Journal of Machine Learning Research*, 7(3), 2006.

[Nad64] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[PV12] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012.

[RBPB24] Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024.

[RVE18] Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, pages 7146–7155, 2018.

[Sam93] Alexander M Samarov. Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, 88(423):836–847, 1993.

[Ser99] Rocco A Servedio. On PAC learning using Winnow, Perceptron, and a Perceptron-like algorithm. In *Conference on Computational Learning Theory*, pages 296–307, 1999.

[SS20] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[Ste81] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151, 1981.

[Sto77] Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

[Sto80] Charles J Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, pages 1348–1360, 1980.

[TDD$^+$24] Emanuele Troiani, Yatin Dandi, Leonardo Defilippis, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Fundamental limits of weak learnability in high-dimensional multi-index models. *arXiv preprint arXiv:2405.15480*, 2024.

[TWKS14] Shubhendu Trivedi, Jialei Wang, Samory Kpotufe, and Gregory Shakhnarovich. A consistent estimator of the expected gradient outerproduct. In *Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 819–828, 2014.

[Vem10] Santosh S Vempala. Learning convex concepts from gaussian distributions with pca. In *51st Annual IEEE Symposium on Foundations of Computer Science*, pages 124–130, 2010.

[Wat64] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[XTLZ02] Yingcun Xia, Howell Tong, Wai Keung Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):363–410, 2002.

[YBL17] Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. High-dimensional non-Gaussian single index models via thresholded score function estimation. In *International Conference on Machine Learning*, 2017.

[YXKH23] Gan Yuan, Mingyue Xu, Samory Kpotufe, and Daniel Hsu. Efficient estimation of the central mean subspace via smoothed gradient outer products. *arXiv preprint arXiv:2312.15469*, 2023.