Fundamental Limits of Learning Single-Index Models under Structured Data

Jivan Waber JIVAN.WABER@GMAIL.COM
Vector Institute

Alireza Mousavi-Hosseini MOUSAVI@CS.TORONTO.EDU
Vector Institute and University of Toronto

Murat A. Erdogdu ERDOGDU@CS.TORONTO.EDU
Vector Institute and University of Toronto

Abstract

Recent works have developed a comprehensive picture of gradient-based learning of isotropic Gaussian single-index models by developing computational lower bounds along with optimal algorithms. In this work, we demonstrate that the picture can change significantly when the data covariance is structured and contains some degree of information about the target. Through studying a spiked covariance model, we show that for the class of Correlational Statistical Query (CSQ) learners, a simple preconditioning of online SGD already achieves an almost optimal sample complexity. Unlike the isotropic case, further smoothing the landscape does not improve this complexity. We prove similar lower bounds in the Statistical Query (SQ) class, where we demonstrate a gap between the SQ lower bound and the performance of the algorithms that are known to be optimal in the isotropic setting. Finally, we show a stark contrast in the information-theoretic limit, where the tight lower bound goes through a sudden phase transition from d to 1 depending on covariance structure, where d is the dimension of the input. Overall, our analysis provides a clear characterization of when and how the spike simplifies learning by improving over isotropic covariance.

1. Introduction

The single-index model with isotropic input data has been thoroughly studied in recent years [1, 5, 12], but it has remained relatively underexplored in the context of structured input data. In practice, data contain structure that simplifies learning [23, 25]. In this work, we consider the problem of learning a single-index model in high dimensions under structured input data:

$$y = f^\star\Big(\frac{\langle \boldsymbol{u}^\star, \boldsymbol{x}\rangle}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}^\star\|}\Big) + \xi, \quad \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \xi \sim \mathcal{N}(0, \sigma^2). \tag{1.1}$$

where $f^\star : \mathbb{R} \to \mathbb{R}$ is the link function, $\boldsymbol{x} \in \mathbb{R}^d$ are the inputs with covariance $\boldsymbol{\Sigma}$ and $\xi \in \mathbb{R}$ corresponds to label noise with finite variance $\sigma^2 > 0$. Learning the model consists in estimating the target to achieve small population loss. In particular, it can be achieved by approximately recovering the unknown link function and the hidden dimension $\boldsymbol{u}^\star$. To avoid identifiability (scaling) issues, we standardize the projected inputs by $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}^\star\|$. To see concretely how covariance structure affects sample complexity, we focus on the *spiked single-index model* that was studied in [24]:

$$\boldsymbol{\Sigma} = \frac{1}{1+\kappa}(\mathbf{I}_d + \kappa\boldsymbol{\theta}\boldsymbol{\theta}^\top), \quad \kappa \asymp d^{r_2}, \quad \langle \boldsymbol{u}^\star, \boldsymbol{\theta}\rangle \asymp d^{-r_1}, \quad r_1 \in [0, 1/2], \quad r_2 \in [0, 1]. \tag{1.2}$$

Here, $\kappa \asymp d^{r_2}$ captures the magnitude of the spike, and $\langle \boldsymbol{u}^\star, \boldsymbol{\theta} \rangle \asymp d^{-r_1}$ the alignment of the signal direction with the spike. Our main theorems show that it is $r := \max\{r_1 - r_2/2, 0\}$ that accurately captures their interplay.

1.1. Contributions

We prove fundamental limits for learning the spiked single-index model with different classes of algorithms. More precisely:

- **Computational lower bounds.** We prove two Statistical Query (SQ) lower bounds. Notably, the sample complexity of online SGD with squared loss is captured by the Correlational Statistical Query (CSQ) framework. Additionally, online SGD coupled with ”label transformations” falls within the general SQ framework. Our lower bounds provide an extension to the recent treatment of the isotropic single-index model in [12]. More precisely, for $l^\star$, $k^\star$—the information exponent, respectively the generative exponent—of the single-index model, the lower bounds depend on r as soon as $r < 1/4$: we get the informal sample complexity from the CSQ lower bound $n = \Omega(d^{2rl^\star})$ and the SQ lower bound based on VSTAT $n = \Omega(d^{2rk^\star})$. When $r \geq 1/4$, our lower bounds coincide with the isotropic ones [11, 12]: for CSQ, $n = \tilde{\Omega}(d^{l^\star/2})$ and for SQ, $n = \tilde{\Omega}(d^{k^\star/2})$. This suggests that the spike is helpful if and only if $r < 1/4$.

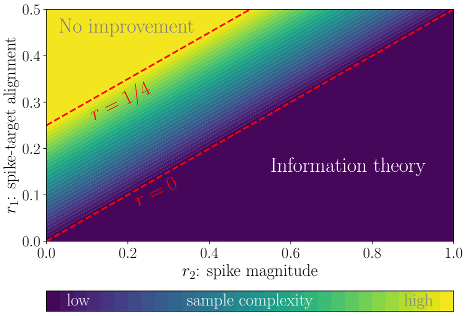

Figure 1: Sample complexity to learn the spiked SIM. Smaller r_1 denotes a better spike-target alignment, while larger r_2 denotes a larger spike magnitude.

- **Information-theoretic phase transition.** We also prove an information-theoretic lower bound, which is the same for the isotropic and spiked settings as long as $r > 0$. Namely, $n = \Omega(d)$. Although this bound was already informally known in the isotropic setting, we believe that it has not been proved before for the single-index model. Note that it matches the information-theoretic upper bound in [12]. Moreover, for $r = 0$, and $r_1 < r_2/2$, the sample complexity undergoes a stark phase transition as $n = \Omega(1)$ samples are sufficient in this regime.

- **Near-matching upper bounds.** Finally, we provide upper bounds that nearly match our statistical query lower bounds by showing how initializing existing algorithms such as online SGD [5] and tensor power iteration [12] at the covariance spike reduces their sample complexity. These algorithms require $n = \tilde{\mathcal{O}}(d^{2r(l^\star-2)+1})$ samples. In addition, by coupling tensor power iteration with label transformations to reduce the information to the generative exponent as in [12], we get the SQ upper bound $n = \tilde{\mathcal{O}}(d^{2r(k^\star-2)+1})$.

2. Related work

The problem of learning Gaussian single-index models has been extensively studied, see e.g. [9, 14]. [5] characterized the number of samples needed to learn the model with online SGD in terms of the information exponent, a single number which encapsulates the hardness of the problem. In [2, 24],

the authors studied the feature learning mechanism in neural networks for single-index models with information exponent 1, and higher information exponent was treated in [6] where they analyzed the behavior of gradient flow. In the context of multi-index models, [13] studied learning polynomials with one gradient step, while [1] observed saddle-to-saddle dynamics.
A series of works studied variations of the single-index model such as the presence of additional input structure [3, 4, 19, 30]. In particular, [28] considered sparsity in the input, while [10] studied a perturbation of the target. Most relevant to our setting, [7, 24] considered a structured covariance. Although [7] analyzes single-index learning under anisotropic covariance, their spectral assumptions fail in our spiked-covariance regime, since our eigenvalues decay too quickly.
Finally, there has been marked interest in providing computational lower bounds on learning single-index models. For isotropic Gaussian inputs, the work [13] constructed a CSQ lower bound depending on the information exponent, while [12] proved a SQ lower bound in terms of the generative exponent, as well as a matching upper bound relying on tensor power iteration and partial trace. Very recently, [21] generalized these lower bounds by considering rotationally symmetric input distributions. Extensions to multi-index models identified a leap exponent that governs the complexity [1], and [27] probed the fundamental computational limits of learning with approximate message passing.

3. Preliminaries

3.1. Statistical queries

Given a regression problem where we have a joint distribution $\mathbb{P}_{\boldsymbol{X},Y}$ on $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$, the goal is to predict Y given $\boldsymbol{X}$ by choosing a predictor $h : \mathbb{R}^d \to \mathbb{R}$ among a class of predictors $\mathcal{H}$. The statistical query framework [22] contains the following class of algorithms. At each iteration, the algorithm can (possibly adaptively) choose a query $q : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ from a class of queries $\mathcal{Q}$, and the oracle returns a response $\hat{\mathbb{E}}[q]$ with the following guarantee

$$\left|\hat{\mathbb{E}}[q] - \mathbb{E}_{\boldsymbol{X},Y}[q(\boldsymbol{X}, Y)]\right| \leq \tau,$$

where τ is the noise tolerance of the oracle. In particular, note that the tolerance can be chosen adversarially by the oracle. One widely considered restriction of the class of queries is known as Correlational Statistical Queries (CSQ) [8], where given some class of real-valued functions on $\mathbb{R}^d$ denoted by $\tilde{\mathcal{Q}}$, the class $\mathcal{Q}$ is given by

$$\mathcal{Q} = \{q(\boldsymbol{x}, y) = \tilde{q}(\boldsymbol{x})y \,:\, \tilde{q} \in \tilde{\mathcal{Q}}\}.$$

Example 1. The online SGD update with respect to the squared error loss fits the CSQ framework, apart from the noise not being adversarial. Indeed, for an estimator $f_{\boldsymbol{w}}, \boldsymbol{w} \in \mathcal{W}$ (for example a neural network), the gradient of the population risk admits the decomposition

$$\nabla\mathcal{R}(\boldsymbol{w}) = \frac{1}{2}\,\mathbb{E}[\nabla(Y - f_{\boldsymbol{w}}(\boldsymbol{X})^2] = \mathbb{E}[Y\nabla f_{\boldsymbol{w}}(\boldsymbol{X})] - \mathbb{E}[f_{\boldsymbol{w}}(\boldsymbol{X})\nabla f_{\boldsymbol{w}}(\boldsymbol{X})],$$

from which we see that the target and the inputs only interact through a correlational query (the first term on the right-hand side).

3.2. Reformulation as an isotropic SIM

Lemma 2 (Single-index model translation)**.** *The spiked single-index model defined in Equations (1.1), (1.2) is equivalent to the isotropic single-index model*

$$y = f^\star(\langle \overline{\boldsymbol{u}}^\star, \boldsymbol{z}\rangle) + \xi, \quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_d), \quad \overline{\boldsymbol{u}}^\star \in \mathbb{S}^{d-1} \quad \textit{subject to} \quad |\langle \boldsymbol{\theta}, \overline{\boldsymbol{u}}^\star\rangle| = cd^{-r},$$

for some absolute constant $c > 0$. Furthermore, this imposes a restriction on the unknown direction:

$$\overline{\boldsymbol{u}}^\star \in S_{\boldsymbol{\theta},r} := \{\overline{\boldsymbol{u}} \in \mathbb{S}^{d-1} : |\langle \overline{\boldsymbol{u}}, \boldsymbol{\theta}\rangle| = cd^{-r}\}.$$

Proof. *The spiked single-index model can be reformulated as an isotropic single-index model by whitening the data. For this purpose, let*

$$\overline{\boldsymbol{u}}^\star := \frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}^\star}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}^\star\|}, \quad \textit{which yields} \quad \frac{\langle \boldsymbol{u}^\star, \boldsymbol{x}\rangle}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}^\star\|} = \langle \overline{\boldsymbol{u}}^\star, \boldsymbol{z}\rangle,$$

where $\boldsymbol{z}$ is a standard Gaussian random vector in $\mathbb{R}^d$. Then, we can relate the constraints on the alignment between $\boldsymbol{u}^\star$ and $\boldsymbol{\theta}$ and the alignment between $\overline{\boldsymbol{u}}^\star$ and $\boldsymbol{\theta}$ as follows. For any absolute constant $c_1 > 0$, there exist absolute constants $c_2, c > 0$ such that

$$|\langle \boldsymbol{\theta}, \boldsymbol{u}^\star\rangle| = c_1 d^{-r_1} \Leftrightarrow |\langle \boldsymbol{\theta}, \overline{\boldsymbol{u}}^\star\rangle| = c_2 \frac{d^{-r_1}\sqrt{1+d^{r_2}}}{\sqrt{d^{r_2}(1+d^{-2r_1})}} = cd^{-r}. \tag{3.1}$$

□

Remark. [Invariance of SQ under whitening] For any bounded query $q : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, setting

$$q_{\boldsymbol{x}}(\boldsymbol{x}, y) = q\big(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}, y\big) \quad \text{and} \quad q_{\boldsymbol{z}}(\boldsymbol{z}, y) = q\big(\boldsymbol{\Sigma}^{1/2}\boldsymbol{z}, y\big)$$

gives a one-to-one correspondence between queries with respect to $\boldsymbol{x}$ and queries with respect to $\boldsymbol{z}$. In particular, any SQ algorithm for one model can be simulated with identical query complexity and tolerance for the other.

4. Main Results: Computational and Statistical Bounds for Learning the Spiked SIM

We prove computational lower bounds for learning the spiked single-index model on algorithms having access to statistical queries, which informally include online SGD. We also prove an information-theoretic lower bound. Additionally, we prove complementary upper bounds, by characterizing the sample complexity required for online SGD and tensor power iteration to learn the spiked SIM. The general strategy for constructing these types of lower bounds is to exhibit a class of almost orthogonal hypotheses and to determine how hard they are to distinguish under our statistical and computational constraints. In their basic form, statistical query lower bounds are solving an easier question, which is to quantify the hardness of distinguishing between a hypothesis class with a planted structure versus a unplanted hypothesis. Instead of describing an estimation problem, they describe a testing/detection problem. Even though this problem seems much easier and hence unable to provide tight lower bounds, it turns out that these lower bounds successfully capture most of the hardness in notorious estimation problems such as Tensor PCA [15] and planted k-SAT [17].

However, there is a well-documented estimation/testing gap in some of these lower bounds, resulting in a small gap between the upper bounds and lower bounds. For a more thorough exposition, we refer the reader to [16]. Our results are subject to such a gap and we hypothesize that it stems from this estimation/testing discrepancy.

Remark. The bounds we present depend on well-known complexity measures of the single-index model, namely, the information and generative exponents. We defer their definition to the appendix.

4.1. Lower bounds

Theorem 3 (CSQ lower bound)**.** *Consider the class of bounded correlational queries $\tilde{q} \in \tilde{\mathcal{Q}}$, $\mathbb{E}\big[\tilde{q}(\boldsymbol{z})^2\big] = 1$. For $l^* \geq 3$, to learn the class of functions*

$$\mathcal{F}_{l^\star,\boldsymbol{\theta},r} := \Big\{\boldsymbol{z} \mapsto g(\langle \overline{\boldsymbol{u}}, \boldsymbol{z}\rangle) : \overline{\boldsymbol{u}} \in S_{\boldsymbol{\theta},r}, \quad \|g\|_{L^2(\gamma)} = 1, \quad \mathrm{IE}(g) = l^\star\Big\},$$

with squared L^2 error at most 1, any correlational statistical query learner with $q = \mathrm{poly}(d)$ queries requires a tolerance of at most

$$\tau \lesssim \begin{cases} d^{-rl^\star} & \text{if } r < \frac{1}{4}, \\ \left(\frac{\log(qd)}{d}\right)^{l^\star/4} & \text{if } r \geq \frac{1}{4}. \end{cases}$$

Remark. Using the informal translation to sample complexity lower bounds via $\tau \asymp 1/\sqrt{n}$, we obtain the sample complexity lower bound $n = \Omega(d^{2rl^\star})$ when $r < 1/4$ and $n = \tilde{\Omega}(d^{l^\star/2})$ when $r \geq 1/4$.

Theorem 4 (SQ lower bound)**.** *Any SQ algorithm using $\mathrm{poly}(d)$ queries to $\mathrm{VSTAT}(n)$ to learn the spiked single-index model with generative exponent $k^\star \geq 3$ requires at least the following number of samples:*

$$n = \begin{cases} \Omega(d^{2rk^\star}) & \text{if } r < \frac{1}{4}, \\ \tilde{\Omega}(d^{k^\star/2}) & \text{if } r \geq \frac{1}{4}. \end{cases}$$

Remark. Both the CSQ and SQ lower bounds suggest that the spike only helps when $r < 1/4$, since otherwise they coincide with the isotropic setting [11, 13].

Remark. Note that the polylog factors in the lower bounds of Theorems 3 and 4 can be removed by refining the hypothesis set construction, as done in [12, Lemma E.3].

Theorem 5 (Information-theoretic lower bound)**.** *Any algorithm using (possibly) infinitely many queries to learn the spiked single-index model requires at least the following number of samples:*

$$n = \begin{cases} \Omega(d) & \text{if } r_1 \geq r_2/2, \\ \Omega(1) & \text{if } r_1 < r_2/2. \end{cases}$$

4.2. Upper bounds

Theorem 6 (CSQ upper bound)**.** *A version of online SGD and tensor power iteration learn the spiked single-index model with $n = \tilde{O}(d^{2r(l^*-2)+1})$ samples, where $l^\star = IE(f^\star) \geq 3$ is the information exponent of $f^\star$.*

More precisely this sample complexity is achieved by tensor power iteration initialized at the spike $\boldsymbol{\theta}$, which directly extends the results of [12] to the spiked setting. However, note that it is not a SQ algorithm. Interestingly, online spherical SGD initialized at the spike also achieves this sample complexity, extending the results of [5]. Although not completely formal due to the adversarial noise in statistical queries, the sample complexity of online SGD on squared loss is widely assumed to be captured by the statistical query framework. Fundamentally, the added structure in the inputs helps us boost the initialization from random alignment of order $d^{-1/2}$ to alignment of order d^{-r}.

Remark. Note that this upper bound coincide with the isotropic upper bound (and lower bound) for $r = 1/4$. The fact that both our versions of tensor power iteration and online SGD coincide in their sample complexity is of particular interest. Indeed, [5] showed that online SGD can only learn the isotropic SIM with sample complexity $\Omega(d^{l^\star - 1})$, whereas tensor power iteration (with partial trace warm start) succeeds with $\mathcal{O}(d^{l^\star/2})$ ([12]), which would suggest that online SGD is not an optimal learning algorithm for the task. However, making the modeling assumptions closer to reality by adding helpful structure in the inputs enables online SGD to get closer to a theoretically optimal algorithm. This might help explain the success of gradient-based approaches in practice.

Remark. [Landscape smoothing cannot help] Note that smoothing the loss landscape [11] cannot be used in conjunction to the initialization at the spike. Informally, smoothing the loss landscape during the weak recovery phase amounts to boosting the initialization from random to $d^{-1/4}$ with the optimal smoothing parameter $d^{1/4}$. Since the initialization at the spike enjoys an alignment of order at least $d^{-1/4}$, further smoothing cannot help. More precisely, by [11, Lemma 14], the smoothing parameter λ cannot exceed $d^{1/4}$ for their results to hold, and thus does not allow to further improve the sample complexity.

Corollary 7 (SQ upper bound)**.** *A version of tensor power iteration learns the spiked single-index model with $n = \tilde{\mathcal{O}}(d^{2r(k^* - 2)+1})$ samples, where $k^\star \geq 3$ is the generative exponent.*

5. Conclusion

In this paper, we have provided a precise description of the fundamental limits of learning single-index models when the input covariance has a spiked structure. Our work illustrates two broader insights:

1. *Preconditioning suffices.* A simple whitening and a spike–based initialization of online SGD achieves nearly optimal CSQ performance, demonstrating how covariance alignment enables algorithms to learn more efficiently.

2. *Computational-statistical gap behavior.* As the covariance structure becomes more informative about the target, the computational-statistical gap tightens and the sample complexity can reach the information-theoretic limit irrespective of the information/generative exponent.

Future directions. Two natural extensions of this work are to consider general covariance structure (e.g. power-law spectrum) and multi-index models. There also remains to close the gap between the lower bounds and the upper bounds.

Acknowledgments

JW and AMH thank Denny Wu and Nuri Mert Vural for insightful discussions.

References

[1] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023.

[2] 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. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

[3] 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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[4] Lorenzo Bardone and Sebastian Goldt. Sliding down the stairs: how correlated latent variables accelerate learning with neural networks. *arXiv preprint arXiv:2404.08602*, 2024.

[5] 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.

[6] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in Neural Information Processing Systems*, 35:9768–9783, 2022.

[7] Guillaume Braun, Minh Ha Quang, and Masaaki Imaizumi. Learning a single index model from anisotropic data with vanilla stochastic gradient descent, 2025. URL https://arxiv.org/abs/2503.23642.

[8] Nader H Bshouty and Vitaly Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2(Feb):359–395, 2002.

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

[10] Elisabetta Cornacchia, Dan Mikulincer, and Elchanan Mossel. Low-dimensional functions are efficiently learnable under randomly biased distributions. *arXiv preprint arXiv:2502.06443*, 2025.

[11] 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. *Advances in Neural Information Processing Systems*, 36, 2023.

[12] Alex Damian, Loucas Pillaud-Vivien, Jason Lee, and Joan Bruna. Computational-statistical gaps in gaussian single-index models. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1262–1262. PMLR, 2024.

[13] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

[14] Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, pages 1887–1930. PMLR, 2018.

[15] Rishabh Dudeja and Daniel Hsu. Statistical query lower bounds for tensor pca. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

[16] Vitaly Feldman. A general characterization of the statistical query complexity. volume 65 of *Proceedings of Machine Learning Research*, pages 785–830, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR. URL http://proceedings.mlr.press/v65/feldman17c.html.

[17] Vitaly Feldman, Will Perkins, and Santosh Vempala. On the complexity of random satisfiability problems with planted solutions. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 77–86, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746577. URL https://doi.org/10.1145/2746539.2746577.

[18] 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), April 2017. ISSN 0004-5411. doi: 10.1145/3046674. URL https://doi.org/10.1145/3046674.

[19] Taj Jones-McCormick, Aukosh Jagannath, and Subhabrata Sen. Provable benefits of unsupervised pre-training and transfer learning via single-index models. *arXiv preprint arXiv:2502.16849*, 2025.

[20] Taj Jones-McCormick, Aukosh Jagannath, and Subhabrata Sen. Provable benefits of unsupervised pre-training and transfer learning via single-index models, 2025. URL https://arxiv.org/abs/2502.16849.

[21] Nirmit Joshi, Hugo Koubbi, Theodor Misiakiewicz, and Nathan Srebro. Learning single-index models via harmonic decomposition, 2025. URL https://arxiv.org/abs/2506.09887.

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

[23] Noam Levi and Yaron Oz. The underlying scaling laws and universal statistical structure of complex datasets, 2024. URL https://arxiv.org/abs/2306.14975.

[24] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[25] D. L. Ruderman and W. Bialek. Statistics of natural images, 1994.

[26] Balázs Szörényi. Characterizing statistical query learning: simplified notions and proofs. In *International Conference on Algorithmic Learning Theory*, pages 186–200. Springer, 2009.

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

[28] Nuri Mert Vural and Murat A Erdogdu. Pruning is optimal for learning sparse features in high-dimensions. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4787–4861. PMLR, 2024.

[29] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[30] Thomas T Zhang, Behrad Moniri, Ansh Nagwekar, Faraz Rahman, Anton Xue, Hamed Hassani, and Nikolai Matni. On the concurrence of layer-wise preconditioning methods and provable feature learning. *arXiv preprint arXiv:2502.01763*, 2025.

Contents

1 Introduction **1**
1.1 Contributions 2

2 Related work **2**

3 Preliminaries **3**
3.1 Statistical queries 3
3.2 Reformulation as an isotropic SIM 4

4 Main Results: Computational and Statistical Bounds for Learning the Spiked SIM **4**
4.1 Lower bounds 5
4.2 Upper bounds 5

5 Conclusion **6**

A Proof of Computational Lower Bounds **10**
A.1 CSQ lower bound 10
A.2 SQ lower bound 11
A.2.1 SQ framework 11
A.2.2 Instantiation for single-index model 12
A.2.3 Proof of SQ lower bound 13

B Proof of Information-Theoretic Lower Bound **14**

C Proofs of Upper Bounds **17**
C.1 Tensor Power Iteration 17
C.2 Online SGD 18
C.2.1 Proofs of Propositions 34 and 35 20

Notations. For vectors, we use $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ to denote Euclidean inner product and norm respectively. For matrices, $\|\cdot\|$ denotes the operator norm. For asymptotic orders with respect to d as $d \to \infty$, $\mathcal{O}(\cdot), \Theta(\cdot), \Omega(\cdot)$ stand for the standard Big-O notations, and $\tilde{\mathcal{O}}(\cdot), \tilde{\Theta}(\cdot), \tilde{\Omega}(\cdot)$ hide (poly-)logarithmic factors. We write $f(d) \asymp g(d)$ and $f(d) = \Theta(g(d))$ interchangeably. $o(\cdot)$ and $\omega(\cdot)$ correspond to the standard little-o, respectively little-omega notations. $\nabla_{\boldsymbol{w}}$ denotes the Euclidean gradient with respect to $\boldsymbol{w}$, and we omit the subscript when it is clear from context. We denote the expectation of a random variable $X \sim P$ by $\mathbb{E}_X$ or $\mathbb{E}_P$ and we omit the subscript when clear from context. The $s-$dimensional Gaussian measure on $\mathbb{R}^s$ is denoted by γ_s, and $\gamma := \gamma_1$.

Appendix A. Proof of Computational Lower Bounds

Lemma 8 (Number of almost orthogonal vectors in $S_{\boldsymbol{\theta},r}$)**.** *For any absolute constant $c > 0$, there exists an absolute constant $C > 0$ such that for any $M = \mathcal{O}(\mathrm{poly}(d))$, there exist M elements in $S_{\boldsymbol{\theta},r}$ that satisfy, for all $i, j \in [M], i \neq j$,*

$$\langle \overline{\boldsymbol{u}}_i, \overline{\boldsymbol{u}}_j \rangle \leq \epsilon + cd^{-r}, \tag{A.1}$$

for $\epsilon = \sqrt{\frac{C \log M}{d}}$.

Proof. *Given $M > 0$, let us construct a subset $S_\epsilon \subset S_{\boldsymbol{\theta},r}$ of cardinality $|S_\epsilon| = M$ that satisfies, for all $i, j \in [M], i \neq j$, $|\langle \overline{\boldsymbol{u}}_i, \overline{\boldsymbol{u}}_j \rangle| \leq \epsilon + c^2 d^{-2r}$. Afterwards, we find an upper bound on M with the probabilistic method. To construct S_ϵ, we use the following procedure. Let $\{\boldsymbol{t}_i\}_{i=1}^M \sim$ $\mathrm{Unif}\Big(\frac{1}{\sqrt{1-c^2d^{-2r}}}\mathbb{S}^{d-2}\Big)$. Suppose $\{\boldsymbol{\theta}, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_d\}$ form an orthonormal basis of $\mathbb{R}^d$, and construct the matrix $\boldsymbol{A} = (\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_d) \in \mathbb{R}^{d\times(d-1)}$. Note that $\langle \boldsymbol{A}\boldsymbol{t}, \boldsymbol{\theta}\rangle = 0$ for all $\boldsymbol{t} \in \mathbb{R}^{d-1}$, and that $\boldsymbol{A}^\top \boldsymbol{A} = \mathbf{I}_{d-1}$. Then, we let $\overline{\boldsymbol{u}}_i = \boldsymbol{A}\boldsymbol{t}_i + cd^{-r}\boldsymbol{\theta}$. As a result, we have*

$$\langle \overline{\boldsymbol{u}}_i, \overline{\boldsymbol{u}}_j \rangle = \langle \boldsymbol{t}_i, \boldsymbol{t}_j \rangle + c^2 d^{-2r}, \quad \textit{and} \quad \langle \boldsymbol{\theta}, \overline{\boldsymbol{u}}_i \rangle = cd^{-r}, \quad \textit{for all} \quad i \neq j, \quad i, j \in [M]. \tag{A.2}$$

Note that for fixed $i \neq j$, Hoeffding's inequality yields $\mathbb{P}(|\langle \boldsymbol{t}_i, \boldsymbol{t}_j \rangle| > \epsilon) \leq 2e^{-c_1 d\epsilon^2}$, for some absolute constant $c_1 > 0$. Thus by a union bound, the probability that $|\langle \boldsymbol{t}_i, \boldsymbol{t}_j \rangle| \leq \epsilon$ for all $i \neq j$ is at least $1 - M^2 e^{-c_1 d\epsilon^2}$. Therefore, if we take $\epsilon = \sqrt{\frac{C \log M}{d}}$, for some absolute constant $C > 0$, it is guaranteed that there exist $\{\boldsymbol{t}_i\}_{i=1}^M$ such that $|\langle \boldsymbol{t}_i, \boldsymbol{t}_j \rangle| \leq \epsilon$ for $i \neq j$. With this construction, we have

$$|\langle \overline{\boldsymbol{u}}_i, \overline{\boldsymbol{u}}_j \rangle| \leq \epsilon + c^2 d^{-2r}, \quad \textit{for all} \quad i \neq j, \, i, j \in [M].$$

Therefore, $S_{\boldsymbol{\theta},r}$ contains at least M vectors that are well-separated, and to ensure $\epsilon = o(1)$, we pick M to be at most $\mathrm{poly}(d)$. □

A.1. CSQ lower bound

The Hermite expansion is a central tool enabling the study of the Gaussian single-index model. The (normalized) Hermite polynomials $\{h_j\}_{j\geq 0}$ are defined as

$$h_j(w) := \frac{(-1)^j e^{w^2/2}}{\sqrt{j!}} \frac{d^j}{dw^j} e^{-w^2/2},$$

and form an orthonormal basis for $L^2(\gamma) := \{g : \mathbb{R} \to \mathbb{R} : \int_{\mathbb{R}} g(w)^2 d\gamma(w) < \infty\}$ leading to a measure of complexity of the link function $f^\star$.

Definition 9 (Information exponent)**.** *Given $f \in L^2(\gamma)$, with Hermite expansion $f = \sum_{j \geq 0} \alpha_j h_j$, the information exponent of f is $\mathrm{IE}(f) := \inf\{j > 0 : \alpha_j \neq 0\}$.*

The proof of the CSQ lower bound relies on constructing a subclass $\mathcal{F} \subseteq \mathcal{F}_{l,\boldsymbol{\theta},r}$ of highly uncorrelated functions. In particular, we use the following lemma, adapted from Szörényi [26, Theorem 2].

Lemma 10 (Damian et al. [13, Lemma 2])**.** *Let $\epsilon > 0$, and let $\mathcal{F}$ be a class of bounded real-valued functions such that for every pair $f, g \in \mathcal{F}$ where $f \neq g$ we have $|\mathbb{E}[f(\boldsymbol{X})g(\boldsymbol{X})]| \leq \epsilon$ and $\mathbb{E}\big[f(\boldsymbol{X})^2\big] = 1$ for all $f \in \mathcal{F}$. Then, any correlational statistical query learner whose queries are bounded in L^2 norm, i.e. $\mathbb{E}\big[\tilde{q}(\boldsymbol{X})^2\big] = 1$, requires at least $\frac{|\mathcal{F}|(\tau^2 - \epsilon)}{2}$ queries to learn $\mathcal{F}$ up to $2(1-\epsilon)$ L^2 error.*

Proof. [Proof of Theorem 3] Let $S_\epsilon \subset S_{\boldsymbol{\theta},r}$ with cardinality M be defined as in the proof of Lemma 8, and let $\mathcal{F}_\epsilon := \{\boldsymbol{z} \mapsto h_{l^\star}(\langle \boldsymbol{u}, \boldsymbol{z} \rangle) : \overline{\boldsymbol{u}} \in S_\epsilon\} \subset \mathcal{F}_{l^\star,\boldsymbol{\theta},r}$. Recall from the properties of Hermite polynomials that for every $\boldsymbol{u}, \boldsymbol{u}' \in S$, we have

$$\mathbb{E}_{\boldsymbol{Z} \sim \mathcal{N}(0,\mathbf{I}_d)}\big[h_{l^\star}(\langle \overline{\boldsymbol{u}}, \boldsymbol{Z} \rangle) h_{l^\star}(\langle \overline{\boldsymbol{u}}', \boldsymbol{Z} \rangle)\big] = \langle \overline{\boldsymbol{u}}, \overline{\boldsymbol{u}}' \rangle^{l^\star}.$$

As a result, for $f \neq g$, $f, g \in \mathcal{F}_{l^\star,\boldsymbol{\theta},r}$, we have $|\mathbb{E}[f(\boldsymbol{Z})g(\boldsymbol{Z})]| \leq C_{l^\star}(\epsilon^{l^\star} + c^{2l^\star} d^{-2rl^\star})$, for a constant $C_{l^\star}$ depending only on $l^\star$. By Lemma 10, any CSQ learner requires a tolerance of at most $\tau^2 \leq 2q/M + C_{l^\star}(\epsilon^{l^\star} + c^{2l^\star} d^{-2rl^\star})$, where q denotes the number of queries. Taking $M = 2qd$, we obtain $\tau^2 \leq 1/d + C_{l^\star}\left(\left(\frac{\log(2qd)}{d}\right)^{l^\star/2} + c^{2l^\star} d^{-2rl^\star}\right)$, which completes the proof. □

A.2. SQ lower bound

A.2.1. SQ FRAMEWORK

Following [12], we introduce the SQ framework of [18] that enables to capture the complexity of a great variety of testing and estimation problems with a single number called the statistical dimension, for algorithms having access to a certain oracle. This framework generalizes the approach used for CSQ by allowing joint queries over the target and inputs, and we instantiate it in the setting of the single-index model.

Definition 11 (Search problem over distributions)**.** *Let $\mathcal{X}$ be a domain, $\mathcal{D}$ be a set of distributions over $\mathcal{X}$, $\mathcal{F}$ 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 consists in finding a valid solution $f \in \mathcal{Z}(D)$ given oracle access to samples from an unknown distribution $D \in \mathcal{D}$. We will also use $\mathcal{Z}_f$ to denote the set of distributions $\mathcal{D}$ for which f is a valid solution.*

Definition 12 (Relative pairwise correlation)**.** *Given two distributions D_1, D_2 and a reference distribution D,*

$$\chi_D(D_1, D_2) := \int \frac{D_1(x) D_2(x)}{D(x)} dx - 1. \tag{A.3}$$

Definition 13 ($(\gamma, \beta)-$correlation)**.** *We say that a set of m distributions $\mathcal{D} = \{D_1, \ldots, D_m\}$ is $(\gamma, \beta)-$correlated relative to a distribution D_0 over $\mathcal{X}$ if $|\chi_{D_0}(D_i, D_j| \leq \gamma$ for $i \neq j$ and $|\chi_{D_0}(D_i, D_i| \leq \beta$ for all $i \in [m]$.*

Definition 14 (SQ dimension)**.** *Given a search problem $\mathcal{Z}$ and parameters $\gamma, \beta > 0$, 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 $\mathcal{X}$ and a finite set of distributions $\mathcal{D}_D \subset \mathcal{D}$ with $|\mathcal{D}_D| \geq m$ such that for any $f \in \mathcal{F}, \mathcal{D}_f := \mathcal{D}_D \setminus \mathcal{Z}_f$ is $(\gamma, \beta)-$correlated relative to D_0.*

Definition 15 (VSTAT Oracle)**.** *Let D be the input distribution over the domain $\mathcal{X}$. Given n samples from D, for any query function $h : \mathcal{X} \to [0, 1]$, $\mathrm{VSTAT}(n)$ oracle returns a value $v \in [\mathbb{E}_{x\sim D}[h(x)] - \tau, \mathbb{E}_{x\sim D}[h(x)] + \tau]$, where $\tau = \max\left\{\frac{1}{n}, \sqrt{\frac{\mathbb{E}_{x\sim D}[h(x)](1-\mathbb{E}_{x\sim D}[h(x)])}{n}}\right\}$.*

The VSTAT oracle corresponds to the response to a query having access to n i.i.d. samples with the noise tolerance τ corresponding to at most the concentration bound.

Lemma 16 (General SQ lower bound, Corollary 3.12 in [18])**.** *For any $\gamma' > 0$, an SQ algorithm requires at least $\mathcal{SD}(\mathcal{Z}, \gamma, \beta) \cdot \frac{\gamma'}{\beta - \gamma}$ queries to $\mathrm{VSTAT}(\frac{1}{3(\gamma+\gamma')})$ to solve $\mathcal{Z}$.*

Corollary 17. *For any $\beta, \gamma, \tau \geq 0$, any algorithm requires at least $\mathcal{SD}(\mathcal{Z}, \gamma, \beta) \cdot \frac{\frac{3}{n} - \gamma}{\beta - \gamma}$ queries to $\mathrm{VSTAT}(n)$ to solve $\mathcal{Z}$.*

A.2.2. Instantiation for single-index model

We follow the exposition in [12] and apply it to our specific setting to get the SQ lower bound for the spiked single-index model.

Definition 18 (Isotropic Gaussian single-index model)**.** *We say that a joint distribution $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R})$ follows an isotropic Gaussian single index model if there exists a probability measure $P \in \mathcal{G} \subset \mathcal{P}(\mathbb{R}^2)$ and $\boldsymbol{v} \in \mathbb{S}^{d-1}$ such that $\mathbb{P} = [R_{\boldsymbol{v}} \otimes I_2]_\sharp[\gamma_{d-1} \otimes P]$, where $R_{\boldsymbol{v}} \in \mathcal{O}_d$ is any orthogonal matrix whose last column is $\boldsymbol{v}$, i.e. of the form $R_{\boldsymbol{v}} = [R_\perp \boldsymbol{v}]$, and*

$$\mathcal{G} := \{\nu_{(w,y)} \in \mathcal{P}(\mathbb{R} \times \mathbb{R});\ \nu_w = \gamma_1;\ \mathbb{E}_\nu[Y^2] < \infty;\ \chi^2(\nu_{(w,y)}||\gamma_1 \otimes \nu_y) > 0\}. \tag{A.4}$$

Having restated the problem as an isotropic single-index model in Subsection 3.2 enables us to use the lower bound from [12] and apply it to our more general setting. Let $W := \langle \overline{\boldsymbol{u}}, \boldsymbol{Z} \rangle$, for $\overline{\boldsymbol{u}} \in S_{\boldsymbol{\theta},r}$ and $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$ so that $W \sim \mathcal{N}(0, 1)$. Then the search problem $\mathcal{Z}$ is of the following form:

- **Domain:** $\mathcal{X} = \mathbb{R}^d \times \mathbb{R}$ (it represents the $(\boldsymbol{Z}, Y)$ pair),
- **Distributions:** $\mathcal{D} = \{\mathbb{P}_{\overline{\boldsymbol{u}}} : \overline{\boldsymbol{u}} \in S_{\boldsymbol{\theta},r}\}$,
- **Solution Set:** $\mathcal{F} = S_{\boldsymbol{\theta},r}$,
- **Valid Solutions:** $\mathcal{Z}(\mathbb{P}_{\overline{\boldsymbol{u}}^\star}) = \{\overline{\boldsymbol{u}} \in \mathcal{F}\ :\ |\langle \overline{\boldsymbol{u}}, \overline{\boldsymbol{u}}^\star \rangle| = \Theta(1)\}$,
- **Reference Distribution:** $D_0 = \gamma_d \otimes P_y$.

Then, we can use Corollary 17 to obtain a SQ lower bound in our setting. In order to do this, we must find a lower bound on $\mathcal{SD}(\mathcal{Z}, \gamma, \beta)$ for some suitable γ, β.

A.2.3. PROOF OF SQ LOWER BOUND

Definition 19. *For a distribution $P \in \mathcal{G}$, let*

$$\zeta_k(Y) := \mathbb{E}_P[h_k(W)|Y]. \quad \textit{and} \quad \lambda_k := \|\zeta_k\|_{P_y} = \left(\mathbb{E}[\zeta_k(Y)^2]\right)^{1/2}. \tag{A.5}$$

Definition 20. *The generative exponent of P in $\mathcal{G}$ is defined as* $\mathrm{GE}(P) := \inf\{k > 0 : \lambda_k \neq 0\}$.

Remark. The generative exponent admits a variational form that makes its relationship with the information exponent clearer. By reformulating the information exponent as

$$\mathrm{IE}(P) = \inf\{l > 0 : \beta_l \neq 0\} \quad \text{where} \quad \beta_l := \mathbb{E}_P[Y h_l(W)],$$

we have that the variational formula for the generative exponent:

$$\mathrm{GE}(P) := \inf_{T \in L^2(P_y)} \mathrm{IE}((\mathbf{I}_d \otimes T)_\sharp P).$$

We introduce [12, Lemma 3.1] without proof.

Lemma 21 (Relative pairwise correlation expansion)**.** *Let $\overline{\boldsymbol{u}}, \overline{\boldsymbol{u}}' \in S_{\boldsymbol{\theta},r}$, and let $m = \langle \overline{\boldsymbol{u}}, \overline{\boldsymbol{u}}' \rangle$. Then we have*

$$\chi_0(\mathbb{P}_{\overline{\boldsymbol{u}}}, \mathbb{P}_{\overline{\boldsymbol{u}}'}) = \mathbb{E}_{\mathbb{P}_0}\left[\frac{d\mathbb{P}_{\overline{\boldsymbol{u}}}}{d\mathbb{P}_0} \cdot \frac{d\mathbb{P}_{\overline{\boldsymbol{u}}'}}{d\mathbb{P}_0}\right] - 1 = \sum_{k \leq k^*} \lambda_k^2 m^k. \tag{A.6}$$

Lemma 22 (SQ dimension lower bound)**.** *Given a positive integer $M = \mathcal{O}(\mathrm{poly}(d))$, and $\epsilon = \sqrt{\frac{C\log(M)}{d}}$ for some absolute constant $C > 0$, the statistical dimension of the search problem $\mathcal{Z}$ satisfies the following:*

$$\mathcal{SD}(\mathcal{Z}, 2C_{k^\star}(\lambda_{k^\star}^2 \epsilon^{k^*} + c^{2k^\star} d^{-2rk^\star}), 1) \geq M, \tag{A.7}$$

for some constant $C_{k^\star} > 0$ that only depends on $k^\star$.
Proof. *Let $M = \mathcal{O}(\mathrm{poly}(d))$. Let $\epsilon > 0$ and let S_ϵ be defined as in the proof of Lemma 8. By construction, the cardinality of S_ϵ is M. Let $\mathcal{D} = \{\mathbb{P}_{\overline{\boldsymbol{u}}_i} : \overline{\boldsymbol{u}}_i \in S_\epsilon\}$. For any $i \neq j$,*

$$\begin{aligned}
\chi_0(\mathbb{P}_{\overline{\boldsymbol{u}}_i}, \mathbb{P}_{\overline{\boldsymbol{u}}_j}) &= \sum_{k \geq 1} \lambda_k^2 \langle \overline{\boldsymbol{u}}_i, \overline{\boldsymbol{u}}_j \rangle^k \leq \lambda_{k^\star}^2 \langle \overline{\boldsymbol{u}}_i, \overline{\boldsymbol{u}}_j \rangle^{k^\star} + \frac{\langle \overline{\boldsymbol{u}}_i, \overline{\boldsymbol{u}}_j \rangle^{k^\star+1}}{1 - \langle \overline{\boldsymbol{u}}_i, \overline{\boldsymbol{u}}_j \rangle} \\
&\leq \lambda_{k^\star}^2 (\epsilon + c^2 d^{-2r})^{k^\star} + 2(\epsilon + c^2 d^{-2r})^{k^\star+1} \\
&\leq C_{k^\star}\left(\lambda_{k^\star}^2 \epsilon^{k^\star} + (c^2 d^{-2r})^{k^\star} + 2(\epsilon^{k^\star+1} + (c^2 d^{-2r})^{k^\star+1})\right),
\end{aligned}$$

where $C_{k^\star} > 0$ is a constant depending only on $k^\star$. Therefore, for $\lambda_{k^\star}^2 \geq 2\max\{\epsilon, c^2 d^{-2r}\}$, we can upper bound the right-hand side by $2C_{k^\star}(\lambda_{k^\star}^2 \epsilon^{k^\star} + c^{2k^\star} d^{-2rk^\star}))$. Therefore, $\mathcal{SD}(\mathcal{Z}, 2C_{k^\star}(\lambda_{k^\star}^2 \epsilon^{k^*} + c^{2k^\star} d^{-2rk^\star}), 1) \geq M$. □

Proof. [Proof of Theorem 4] By Lemma 22 and Corollary 17, the number of queries q to $\mathrm{VSTAT}(n)$ satisfies the following:

$$q \geq M \cdot \frac{\frac{3}{n} - 2C_{k^\star}(\lambda_{k^\star}^2 \epsilon^{k^*} + d^{-2rk^\star})}{1 - 2C_{k^\star}(\lambda_{k^\star}^2 \epsilon^{k^*} + d^{-2rk^\star})} \geq \frac{M}{2} \cdot \left(\frac{3}{n} - 2C_{k^\star}(\lambda_{k^\star}^2 \epsilon^{k^*} + d^{-2rk^\star})\right), \tag{A.8}$$

which implies

$$\frac{3}{n} \le \frac{2q}{M} + 2C_{k^\star}(\lambda_{k^\star}^2 \epsilon^{k^*} + c^{2k^\star} d^{-2rk^\star}). \tag{A.9}$$

Setting $M = 2qn = \mathrm{poly}(d)$ yields

$$n \ge \frac{1}{C_{k^\star}(\lambda_{k^\star}^2 \epsilon^{k^\star} + c^{2k^\star} d^{-2rk^\star})},$$

which implies that

$$n = \begin{cases} \tilde{\Omega}(d^{k^\star/2}) & \text{if } r \ge \frac{1}{4}, \\ \Omega(d^{2rk^\star}) & \text{if } r < \frac{1}{4}. \end{cases}$$

□

Appendix B. Proof of Information-Theoretic Lower Bound

In order to provide an information-theoretic lower bound, we introduce the framework and some results from Wainwright [29, Chapter 15]. By reyling on Fano's inequality and Yang-Barron's method, we can lower bound the minimax risk by using the packing number of the set of parameters $S \in \{\mathbb{S}^{d-1}, S_{\boldsymbol{\theta},r}\}$ (depending on whether we tackle the isotropic case or the spiked covariance case), as well as the covering number of the parameterized set of joint probabilities $\mathcal{P} := \{\mathbb{P}_{\boldsymbol{u}} : \boldsymbol{v} \in S\}$ associated with single-index models.

Setup Given the class of isotropic Gaussian single-index models $\mathcal{K}$, we will derive a sample complexity lower bound on the hardness of learning $\mathbb{P} \in \mathcal{K}$, by constructing a minimax lower bound for the subset of distributions $\mathcal{P} := \{\mathbb{P}_{\boldsymbol{v}} : \boldsymbol{v} \in S\} \subset \mathcal{K}$ consisting of the joint distributions corresponding to the single-index model

$$y = f(\langle \boldsymbol{x}, \boldsymbol{v} \rangle) + \xi, \quad \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_d), \quad \xi \sim \mathcal{N}(0, \sigma^2),$$

for a fixed $f \in L^2(\gamma)$ with $\|f\|_{L^2(\gamma)} = 1$ and $\mathrm{IE}(f) = l^\star$. Now, given a family $\bar{S} \subset S$ of cardinality M, let $\bar{\mathcal{P}} := \{\mathbb{P}_{\boldsymbol{v}} : \boldsymbol{v} \in \bar{S}\}$ be a family of single-index hypotheses, also of cardinality M. Let $\boldsymbol{Z} \in \mathbb{R}^d \times \mathbb{R}$ be a sample generated by uniformly sampling an index $J \sim \mathrm{Unif}([M])$ and then generating data $(\boldsymbol{X}, Y)$ according to $\mathbb{P}_{\boldsymbol{v}_J}$. In this way, the observation $\boldsymbol{Z}$ follows the mixture distribution $\mathbb{Q}_{\boldsymbol{Z}} = \bar{\mathbb{Q}} := \frac{1}{M}\sum_{j=1}^{M} \mathbb{P}_{\boldsymbol{v}_j}$, which is the average over all hypotheses. Then the mutual information between the random variables $\boldsymbol{Z}$ and J is defined as

$$I(\boldsymbol{Z}, J) := \mathrm{KL}(\mathbb{Q}_{\boldsymbol{Z},J} || \mathbb{Q}_{\boldsymbol{Z}} \mathbb{Q}_J),$$

where $\mathbb{Q}_{\boldsymbol{Z},J}$ denotes the joint distribution of $(\boldsymbol{Z}, J)$ and $\mathbb{Q}_J$ is the distribution of J.

Definition 23. *The minimax risk in our setting is defined as*

$$R^\star := \inf_{\hat{\boldsymbol{u}}} \sup_{\boldsymbol{u} \in S} \mathbb{E}_{\mathbb{P}_{\boldsymbol{u}}}(\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|^2),$$

where the infimum ranges over all possible estimators.

We present two lemmas that we will use for our minimax lower bounds, and whose proofs can be found in [29, Chapter 15].

Lemma 24 (Fano's inequality)**.** *Given $\delta > 0$, let $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_M\} \subset S$ be a $2\delta-$packing of S in the Euclidean norm $\|\cdot\|$, and suppose that J is uniformly distributed over the index set $[M]$, and $(\boldsymbol{Z}|J = j) \sim \mathbb{P}_{\boldsymbol{v}_j}$. Then the minimax risk is lower bounded as*

$$R^\star \geq \delta^2 \cdot \left(1 - \frac{I(\boldsymbol{Z}, J_\delta) + \log 2}{\log M}\right).$$

Lemma 25 (Yang-Barron method)**.** *Let $\mathcal{N}_{KL}(\eta, \mathcal{P})$ denote the $\eta-$covering number of $\mathcal{P}$ with respect to the square-root KL divergence. Then the mutual information is upper bounded as*

$$I(\boldsymbol{Z}, J_\delta) \leq \inf_{\eta > 0}\{\eta^2 + \log(\mathcal{N}_{\mathcal{KL}}(\eta, \mathcal{P}))\}.$$

Lemma 26. *Given $\boldsymbol{v}, \boldsymbol{v}' \in S$, the KL divergence between $\mathbb{P}_{\boldsymbol{v}}$ and $\mathbb{P}_{\boldsymbol{v}'}$ satisfies*

$$KL(\mathbb{P}_{\boldsymbol{v}} \| \mathbb{P}_{\boldsymbol{v}'}) \asymp \frac{1}{\sigma^2}(1 - \langle \boldsymbol{v}, \boldsymbol{v}' \rangle^{l^\star}), \tag{B.1}$$

where $k^\star$ denotes the information exponent of the single index.
Proof. *The conditional probability distribution of $Y|\boldsymbol{X}$ for $(Y, \boldsymbol{X}) \sim \mathbb{P}_{\boldsymbol{v}}$ is given by the Gaussian distribution $\mathcal{N}(f(\langle \boldsymbol{X}, \boldsymbol{v} \rangle), \sigma^2)$. Hence, by standard computations of the KL for the normal location model (see e.g. [29][Example 15.13]),*

$$\begin{aligned} KL(\mathbb{P}_{\boldsymbol{v}} \| \mathbb{P}_{\boldsymbol{v}'}) &= \frac{1}{2\sigma^2} \mathbb{E}_{\boldsymbol{X}}[\left(f(\langle \boldsymbol{X}, \boldsymbol{v} \rangle) - f(\langle \boldsymbol{X}, \boldsymbol{v}' \rangle)\right)^2] \\ &= \frac{1}{\sigma^2}(1 - (\alpha_{k^\star}^2 \langle \boldsymbol{v}, \boldsymbol{v}' \rangle^{l^\star} + o(\langle \boldsymbol{v}, \boldsymbol{v}' \rangle^{l^\star + 1})) \asymp \frac{1}{\sigma^2}(1 - \langle \boldsymbol{v}, \boldsymbol{v}' \rangle^{l^\star}), \end{aligned}$$

where f admits the Hermite expansion $f = \sum_{l \geq l^\star} \alpha_l h_l$. □

Assuming that $C_\epsilon(S)$ is an $\epsilon-$covering, that is, $\|\boldsymbol{v} - \boldsymbol{v}'\| \leq \epsilon$ for all distinct $\boldsymbol{v}, \boldsymbol{v}'$ in $C_\epsilon(S)$, then $1 - \langle \boldsymbol{v}, \boldsymbol{v}' \rangle \leq \frac{\epsilon^2}{2}$. Thus,

$$KL(\mathbb{P}_{\boldsymbol{v}} \| \mathbb{P}_{\boldsymbol{v}'}) \lesssim \frac{1}{\sigma^2}(1 - \langle \boldsymbol{v}, \boldsymbol{v}' \rangle) \leq \frac{1}{2\sigma^2} \cdot \epsilon^2. \tag{B.2}$$

Consequently, for an i.i.d. sample of n data points, we get the bound

$$\sqrt{KL(\mathbb{P}_{\boldsymbol{v}}^{\otimes n} \| \mathbb{P}_{\boldsymbol{v}'}^{\otimes n})} \lesssim \frac{\sqrt{n}}{\sigma}\epsilon =: \eta. \tag{B.3}$$

Therefore, $\log(\mathcal{N}_{\mathcal{KL}}(\eta, \mathcal{P})) \leq \log(\mathcal{C}_\epsilon(\mathcal{S}))$.
By standard properties of coverings and packings, for $P_{2\delta}(S)$ a $2\delta-$packing of size M, that is, $\|\boldsymbol{v} - \boldsymbol{v}'\| \geq 2\delta$ for all distinct $\boldsymbol{v}, \boldsymbol{v}'$ in $P_{2\delta}(S)$,

$$\log(P_{2\delta}(S) \asymp \log(C_\delta(S)). \tag{B.4}$$

In order to get a bound on the minimax risk that only depends on δ, we must select a value of δ for which

$$\frac{I(\boldsymbol{Z}, J_\delta) + \log 2}{\log M} \geq 1/2.$$

By following the recipe in [29], we first prescribe $\eta > 0$ such that

$$\eta^2 \geq \log\left(\mathcal{C}_{\frac{\sigma}{\sqrt{n}}\eta}(S)\right). \tag{B.5}$$

Then we pick the largest $\delta > 0$ that satisfies the lower bound

$$\log(M(\delta)) \geq 4\eta^2 + 2\log 2. \tag{B.6}$$

Fact 27. *For $\epsilon > 0$, the covering number of the sphere of radius ρ in Euclidean norm is given by*

$$\mathcal{N}(\epsilon, \rho\mathbb{S}^{d-1}, \|\cdot\|) \asymp \left(\frac{\rho}{\epsilon}\right)^{d-1}.$$

Let us focus on the case $S = \mathbb{S}^{d-1}$ first. In that case, Equation (B.5) yields the inequality

$$\eta^2 \gtrsim d\log\left(\frac{\sqrt{n}}{\eta}\right), \tag{B.7}$$

which is fulfilled by setting $\eta \asymp \sqrt{d}$. We immediately see from Equations (B.4)-(B.6) that the pair (δ, ϵ) must satisfy $\delta \lesssim \epsilon$. Since we must pick the largest such δ, taking $\delta \asymp \epsilon$ ensures that Equations (B.5)-(B.6) are satisfied. Hence, $R^\star \gtrsim \delta^2 \asymp \eta^2/n \asymp d/n$, which yields the sample complexity lower bound

$$n = \Omega(d).$$

Proposition 28. *For $r = 0$, there are two cases for the constraint $\langle \overline{\boldsymbol{u}}, \boldsymbol{\theta}\rangle \asymp d^{-r}$. If $r_1 = r_2/2$, then $|\langle \overline{\boldsymbol{u}}, \boldsymbol{\theta}\rangle| = c$, for some absolute constant $c \in (0, 1]$. However, if $r_1 < r_2/2$, it further holds that $|\langle \overline{\boldsymbol{u}}, \boldsymbol{\theta}\rangle| = c \approx 1$.*

Proof. For $r_1 < r_2/2$,

$$|\langle \overline{\boldsymbol{u}}, \boldsymbol{\theta}\rangle| = \frac{\left|\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\rangle\right|}{\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\right\|} = \frac{\sqrt{1+\kappa}|\langle \boldsymbol{\theta}, \boldsymbol{u}\rangle|}{\sqrt{1+\kappa\langle \boldsymbol{\theta}, \boldsymbol{u}\rangle^2}} \approx \frac{\sqrt{\kappa}|\langle \boldsymbol{\theta}, \boldsymbol{u}\rangle|}{\sqrt{\kappa}|\langle \boldsymbol{\theta}, \boldsymbol{u}\rangle|} = 1,$$

where the approximation comes from

$$\kappa\langle \boldsymbol{\theta}, \boldsymbol{u}\rangle^2 = c_1^2 c_2 d^{r_2 - 2r_1} = \omega(1), \quad \text{which implies} \quad \sqrt{1+\kappa\langle \boldsymbol{\theta}, \boldsymbol{u}\rangle^2} \approx \sqrt{\kappa}|\langle \boldsymbol{\theta}, \boldsymbol{u}\rangle|.$$

□

Remark. We will prove below that Proposition 28 implies that the information-theoretic lower bound undergoes a sharp transition from $n = \Omega(d)$ to $n = \Omega(1)$ at $r_1 = r_2/2$. Moreover, when $r_1 < r_2/2$, taking $\boldsymbol{\theta}$ as our estimator for $\overline{\boldsymbol{u}}^\star$ is enough to learn the single-index model's unknown direction.

Lemma 29 (Covering number of $S_{\boldsymbol{\theta},r}$)**.** *For $\epsilon > 0$, the covering number of $S_{\boldsymbol{\theta},r} = \{\overline{\boldsymbol{u}} \in \mathbb{S}^{d-1} : \langle \overline{\boldsymbol{u}}, \boldsymbol{\theta}\rangle = cd^{-r}\}$ satisfies*

$$\log \mathcal{N}(\epsilon, S_{\boldsymbol{\theta},r}, \|\cdot\|) \asymp \begin{cases} d\log(1/\epsilon) & \textit{if } r_1 \geq \frac{r_2}{2} \\ 1 & \textit{if } r_1 < \frac{r_2}{2} \end{cases}$$

Proof. *For $r_1 < \frac{r_2}{2}$, by Proposition 28, $S_{\theta,r} = \{\overline{\boldsymbol{u}} \in \mathbb{S}^{d-1} : \langle \overline{\boldsymbol{u}}, \boldsymbol{\theta} \rangle \approx 1\}$. For $r_1 \geq \frac{r_2}{2}$, given the orthonormal basis $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d\}$ where $\boldsymbol{\theta}_d = \boldsymbol{\theta}$,*

$$\langle \overline{\boldsymbol{u}}, \boldsymbol{\theta} \rangle = \sum_{i=1}^{d-1} \langle \overline{\boldsymbol{u}}, \boldsymbol{\theta}_i \rangle + cd^{-r}\boldsymbol{\theta},$$

and thus the covering number of $S_{\boldsymbol{\theta},r}$ satisfies

$$\log C_\epsilon(S_{\theta,r}) \asymp \log C_\epsilon\left(\frac{1}{\sqrt{1 - c^2 d^{-2r}}}\mathbb{S}^{d-2}\right) \asymp (d-2)\log\left(\frac{1}{\sqrt{1-c^2d^{-2r}}\epsilon}\right)$$
$$\asymp d \log\left(\frac{1}{\epsilon}(1 + c^2 d^{-2r})\right) \asymp d\log(1/\epsilon).$$

□

In the case of $S = S_{\boldsymbol{\theta},r}$ and $r_1 \geq r_2/2$, Equation (B.6) yields the inequality

$$\eta^2 \gtrsim d \log\left(\frac{\sqrt{n}}{\eta}\right).$$

Thus, we must pick $\eta \asymp \sqrt{d}$. Hence $R^\star \geq \delta^2 \asymp \eta^2/n$, which yields the sample complexity lower bound

$$n = \Omega(d).$$

However, when $r_1 < r_2/2$, we can pick $\eta^2 \asymp 1$ to obtain $R^\star \gtrsim 1/n$, which yields the sample complexity lower bound $n = \Omega(1)$.

Appendix C. Proofs of Upper Bounds

We derive upper bounds on learning the spiked single-index model by initializing existing algorithms at the covariance spike. Our upper bounds only consider learning the unknown direction $\overline{\boldsymbol{u}}^\star$ of the isotropic formulation of the spiked single-index model; they do not include the sample complexity to learn the link function $f^\star$. However, once $\overline{\boldsymbol{u}}^\star$ (or $\boldsymbol{u}^\star$) is approximately recovered, recovering $f^\star$ is a simple convex optimization problem. A more detailed exposition of how this is can be done with two-layer neural networks and gradient descent is provided in [13].

C.1. Tensor Power Iteration

We use the tensor power iteration algorithm from [12] initialized at the covariance spike. By simply plugging in the initial alignment $m_0 = \Theta(d^{-r})$ in [12], we obtain the upper bound on the sample complexity

$$n = \tilde{\mathcal{O}}(d^{2r(l^\star - 2)+1}) \tag{C.1}$$

to learn the spiked single-index model.

Proof. By [12, Lemma F.4], for any initial alignment $m_0 = \Theta(d^{-r})$, for any $r \in [0, 1/4]$ and until the alignment reaches $1/4$, each step of tensor power iteration with $\Theta(d^{1+2r(l^\star-2)})$ samples achieves an alignment $m_1 \geq 2m_0$ with high probability. Therefore, taking $s = \mathcal{O}(\log d)$ steps of tensor power iteration yields $m_s = \Theta(1)$ with high probability. Hence, $n = \tilde{\mathcal{O}}(d^{1+2r(l^\star-2)})$ are sufficient to weakly learn the hidden dimension $\overline{\boldsymbol{u}}^\star$. By taking another step of tensor power iteration with $n \gtrsim d/\epsilon$, where $\gtrsim$ can hide dependence on k, we can furthermore achieve the strong recovery of $\overline{\boldsymbol{u}}^\star$, that is, we can reach an alignment greater than $1-\epsilon$ with high probability. □

This upper bound coincides with the one we obtain below with Online SGD initialized at the covariance spike.

For the SQ upper bound, we can analogously use [12, Algorithm 2] with initialization at $\boldsymbol{\theta}$. At a high level, the algorithm finds a label transformation with a denoiser $T \in L^2(P_y)$ such that $\mathrm{IE}(T(P)) = \mathrm{GE}(P)$. This yields the sample complexity upper bound

$$n = \tilde{\mathcal{O}}(d^{1+2r(k^\star-2)}),$$

depending on the generative exponent $k^\star$.

C.2. Online SGD

By extending the work of [5] to consider the case where the initial alignment between the weight and the hidden direction is of order d^{-r}, for $r \in [0, 1/2]$, we can derive a sample complexity bound for online SGD initialized at the spike of the covariance matrix. This contrasts with the uninformative weight initialization which achieves an alignment of order $1/\sqrt{d}$ with high probability. In the analysis performed in [5], the alignment at timestep t is decomposed into four components: the initial alignment, a drift term, a martingale term corresponding to the sample-wise error, and higher-order terms in the learning rate, corresponding to the projection onto the sphere. Then, the game consists of finding the appropriate learning rate and number of samples, so that the initial alignment and the drift term dominate the dynamics and make online SGD achieve weak recovery of the hidden direction. Therefore, starting with a better initialization yields a better sample complexity as it allows the selection of a larger learning rate. There is no technical innovation in the proof and the improved sample complexity has been described in a concurrent work [20] in a different context and without proof, but we show here how the proof of [5] generalizes for completeness. We only adapt the weak recovery proof, as it is the phase where the information exponent appears. The descent phase/strong recovery always takes $\mathcal{O}(d)$ samples.

Setup We study the isotropic single-index model $y = f^\star(\langle \boldsymbol{z}, \overline{\boldsymbol{u}}^* \rangle) + \xi, \quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_d), \quad \xi \sim \mathcal{N}(0, \sigma^2)$. We denote the loss by $\mathcal{L} : \mathbb{S}^{d-1} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, and study the case where the population loss is of the form

$$\Phi(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{Z},Y}[\mathcal{L}] = \phi(m(\boldsymbol{w})) \quad \text{where} \quad m(\boldsymbol{w}) := \langle \boldsymbol{w}, \overline{\boldsymbol{u}}^* \rangle,$$

for some $\phi : [-1, 1] \to \mathbb{R}$. Moreover, we define the sample-wise error

$$H^t(\boldsymbol{w}) := \mathcal{L}(\boldsymbol{w}; \boldsymbol{z}_t, y_t) - \Phi(\boldsymbol{w}).$$

In this subsection, we denote the information exponent by k, and we follow the more general definition of [5].

Algorithm 1: Spherical Online SGD Initialized at Covariance Spike

```
Data: $(\boldsymbol{z}_i, y_i)_{i=1}^n \sim \mathbb{P}_{\overline{\boldsymbol{u}}^\star}$,   $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$   with   $|\langle \boldsymbol{\theta}, \overline{\boldsymbol{u}}^\star \rangle| = cd^{-r}$,   step size $\eta > 0$
Result: Final iterate $\boldsymbol{w}$ on the sphere
$\boldsymbol{w} \leftarrow \boldsymbol{\theta}$;
for $t \in [n]$ do
    $\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\eta}{d}(\mathbf{I}_d - \boldsymbol{w}\boldsymbol{w}^\top)\nabla\mathcal{L}(\boldsymbol{w}; \boldsymbol{z}_t, y_t)$;
    $\boldsymbol{w} \leftarrow \boldsymbol{w}/\|\boldsymbol{w}\|$;
end
return $\boldsymbol{w}$;
```

Definition 30 (Information exponent)**.** *The population loss Φ has information exponent k if $\phi \in C^{k+1})$ and there exist $C, c > 0$ such that*

$$\begin{cases} \frac{d^l \phi}{dm^l}(0) = 0 \quad 1 \leq l \leq k \\ \frac{d^k \phi}{dm^k}(0) \leq -c < 0 \\ \left\| \frac{d^{k+1}\phi}{dm^{k+1}}(m) \right\|_\infty \leq C. \end{cases}$$

Definition 31. *For a population loss of the form $\Phi(\boldsymbol{w}) := \mathbb{E}[\mathcal{L}] = \phi(m(\boldsymbol{w}))$ where $m(\boldsymbol{w}) := \langle \boldsymbol{w}, \overline{\boldsymbol{u}}^\star \rangle$, for some $\phi : [-1, 1] \to \mathbb{R}$, we say that **Assumption A**$_\varrho$ for some ϱ in $(0, 1]$ holds if*

1. *ϕ is differentiable, and*
2. *ϕ' is strictly negative on the interval $(0, \varrho)$.*

*When $\varrho = 1$, we say that **Assumption A** holds.*

Definition 32. *For a data distribution and loss pair $(\mathbb{P}, \mathcal{L})$, **Assumption B** holds if there exist $C_1, \iota > 0$ such that the following two moment bounds hold for all d :*

1. *We have that*

$$\sup_{\boldsymbol{w} \in \mathbb{S}^{d-1}} \mathbb{E}\left[|\langle \nabla H(\boldsymbol{w}), \overline{\boldsymbol{u}}^\star \rangle|^2 \right] \leq C_1,$$

2. *and that*

$$\sup_{\boldsymbol{w} \in \mathbb{S}^{d-1}} \mathbb{E}\left[\|\nabla H(\boldsymbol{w})\|^{4+\iota} \right] \leq C_1 d^{2+\iota/2}.$$

For every μ, define the hitting times

$$\tau_\mu^+ := \min\{t \geq 0 : m(\boldsymbol{W}_t) \geq \mu\}, \quad \text{and} \quad \tau_\mu^- := \min\{t \geq 0 : m(\boldsymbol{W}_t) \leq \mu\}.$$

Fix ι given in Assumption B and define

$$\bar{L} := \sup_{\boldsymbol{w}} \mathbb{E}\left[\left\| \frac{1}{\sqrt{d}} \nabla H(\boldsymbol{w}) \right\|^{4+\iota} \right] \vee \sup_{\boldsymbol{w}} \mathbb{E}\left[\left\| \frac{1}{\sqrt{d}} \nabla H(\boldsymbol{w}) \right\|^2 \right] \vee 1. \tag{C.2}$$

Additionally, let $a_k = ck, a_{k+1} = C(k+1)$, where C, c are as in the definition of the information exponent, and let m_t denote $m(\boldsymbol{W}_t)$.

Theorem 33. *Suppose there exists $\varrho > 0$ such that Assumptions A_ϱ and B hold and that the population loss has information exponent k. Let (α_d, η_d) be as in Proposition 34. Then there exists $\nu > 0$ such that if $\boldsymbol{W}_t = \boldsymbol{W}_t^{d,\eta}$ is the online SGD with step size η, we have for every $\gamma > 0$,*

$$\lim_{d\to\infty} \inf_{\boldsymbol{w}_0 : m(\boldsymbol{w}_0)\geq \gamma/d^r} \mathbb{P}_{\boldsymbol{w}_0}\Big(\tau_\nu^+ < \alpha d\Big) = 1.$$

where, we recall, τ_ν^+ is the stopping time $\inf\{t : m_t > \nu\}$.

Proposition 34. *Suppose that Assumptions A_ϱ and B hold and that the population loss has information exponent $k \geq 3$. Let $D = D_d$ and $\epsilon = \mathcal{O}(d^{1+(k-3)r})$, and suppose $\alpha = \alpha_d$ is of at most polynomial growth in d, and $\eta = \eta_d$ is such that $\alpha\eta^2 \leq \epsilon$ and for some $K > 0$,*

$$\eta \leq \bar{\eta}_d(k) := \frac{a_k \gamma^{k-2}}{K\bar{L} d^{r(k-2)} \log d}. \tag{C.3}$$

Then for every $\gamma > 0$ and every $T \leq M := \alpha d$ satisfying

$$T \leq \frac{d^{2(1-r)}\gamma^2}{D^2\eta^2} =: \bar{t}\,, \tag{C.4}$$

online SGD with step-size η satisfies the following as $d \to \infty$ for some $\nu > 0$, uniformly over the choice of D, ϵ, K: there exists a constant $C(C_1, a_k, a_{k+1}) > 0$ such that

$$\inf_{\boldsymbol{w}_0\in E_{\gamma/d^r}} \mathbb{P}_{\boldsymbol{w}_0}\Big(m_t \geq \frac{m_0}{2} + \frac{\eta a_k}{8d}\sum_{j=0}^{t-1} m_j^{k-1} \quad \forall t \leq \tau^-_{\gamma/2d^r} \wedge \tau_\nu^+ \wedge T\Big) \geq 1 - \frac{C}{D^2} - o(1)\,. \tag{C.5}$$

Proposition 35 (Weak recovery)**.** *Under the assumptions of Proposition 34, for $\alpha \gtrsim d^{2r(k-2)}\log d$ and $\eta \leq \bar{\eta}(k)$ satisfying $\alpha\eta^2 \leq \epsilon$ for $\epsilon = \mathcal{O}(d^{1+(k-3)r})$, there exists $\nu_0(\varrho, a_k, a_{k+1}) > 0$ such that for every $\nu < \nu_0$, for every $\gamma > 0$, we have*

$$\lim_{d\to\infty} \inf_{\boldsymbol{w}_0\in E_{\gamma/d^r}} \mathbb{P}_{\boldsymbol{w}_0}\Big(\tau_\nu^+ \leq \bar{t} \wedge M\Big) = 1\,.$$

Corollary 36. *We can weakly recover the single-index model with online SGD initialized at $\boldsymbol{\theta}$ with sample complexity*

$$n = \alpha d = \tilde{\mathcal{O}}(d^{2r(k-2)+1}).$$

C.2.1. PROOFS OF PROPOSITIONS 34 AND 35

We need to adapt the proofs of the bounds of all the different components appearing in the decomposition of the alignment m_t. Without loss of generality, we assume that $\overline{\boldsymbol{u}}^\star = \boldsymbol{e}_1$, the first vector of the canonical basis. From [5], we have the inequality

$$m_t \geq m_{t-1} - \frac{\eta}{d}\langle \nabla\Phi(\boldsymbol{W}_{t-1}), \boldsymbol{e}_1\rangle - \frac{\eta}{d}\langle \nabla H^t(\boldsymbol{W}_{t-1}), \boldsymbol{e}_1\rangle - \frac{\eta^2}{d}L_t\mathbf{1}_{\{L_t<\hat{L}\}}|m_{t-1}| \tag{C.6}$$

$$- \eta^2\left(\frac{A}{d^2} + \frac{1}{d}L_t\mathbf{1}_{\{L_t\geq\hat{L}\}}\right)|m_{t-1}| - \eta^3\left(\frac{A}{d^2} + \frac{L_t}{d}\right)\left(\left|\frac{\langle\nabla\Phi(\boldsymbol{W}_{t-1}), \boldsymbol{e}_1\rangle}{d}\right| + \left|\frac{\langle\nabla H^t(\boldsymbol{W}_{t-1}), \boldsymbol{e}_1\rangle}{d}\right|\right), \tag{C.7}$$

where A is a constant depending only on a_k, a_{k+1} and $L_t := \left\|\frac{1}{\sqrt{d}}\nabla H^t(\boldsymbol{W}_{t-1})\right\|^2$. Let us start by controlling the higher order corrections that arise from the projection on the sphere of the parameters at each iteration of the algorithm. The bounds do not change much, except carefully pick $\hat{L}$.
Observe that for every $\nu < 1/2$, for every $\boldsymbol{w} \in \left\{\boldsymbol{w} : m(\boldsymbol{w}) \in [0,\nu]\right\}$,

$$\langle \nabla m(\boldsymbol{w}), \boldsymbol{e}_1\rangle = (\boldsymbol{e}_1 - \langle \boldsymbol{w}, \boldsymbol{e}_1\rangle \boldsymbol{w}) = 1 - m(\boldsymbol{w})^2 \geq 1 - \nu^2 \geq 1/2.$$

Then there exists $\nu_0(\varrho, a_k, a_{k+1}) > 0$ such that for all $\nu < \nu_0$, for all $x \in \left\{\boldsymbol{w} : m(\boldsymbol{w}) \in [0,\nu]\right\}$,

$$\frac{1}{4}a_k m(\boldsymbol{w})^{k-1} \leq -\langle \nabla\Phi(\boldsymbol{w}), \boldsymbol{e}_1\rangle \leq \frac{3}{2}a_k m(\boldsymbol{w})^{k-1}.$$

Lemma 37. *Let $\nu, \gamma > 0$ with $\nu < 1/2$. For all $K > 0$, for all $\eta < \bar{\eta}_d(k)$, for $k \geq 2$,*

$$\forall \boldsymbol{w} \in \mathcal{A} := \left\{\boldsymbol{w} : m(\boldsymbol{w}) \in \left[\frac{\gamma}{2d^r}, \nu\right]\right\}, \quad \frac{\eta^2}{d}|m(\boldsymbol{w})| \leq \frac{\eta}{d}2^{k-2}a_k\frac{|m(\boldsymbol{w})|^{k-1}}{K\bar{L}\log d}.$$

Proof. *Since $\mathbf{w} \in \mathcal{A}$ and given Equation (C.3), we have*

$$\eta^2|m(\boldsymbol{w})| \leq \frac{a_k}{K\bar{L}\log d}\cdot\eta|m(\boldsymbol{w})|\cdot\left(\frac{\gamma}{d^r}\right)^{k-2} \leq \frac{a_k}{K\bar{L}\log d}\eta|m(\boldsymbol{w})|^{k-1},$$

where the last inequality follows from the fact that $|m(\boldsymbol{w})| > \frac{\gamma}{2d^r}$. □

Lemma 38. *Suppose that $\alpha\eta^2 \leq \epsilon$ for some $\epsilon = \mathcal{O}(d^{1+(k-3)r})$, and let $\bar{L}$ be as in Equation (C.2). There exists $C = C(\bar{L}, A, C_1, C_2) > 0$ such that the following holds uniformly over $\mathbf{w}_0 \in \mathbb{S}^{d-1}$:*

$$\mathbb{P}_{\boldsymbol{w}_0}\left(\sup_{t\leq M}\eta^3\sum_{j=0}^{t-1}\left(\frac{A}{d^2}+\frac{L_{j+1}}{d}\right)\left(\left|\frac{\langle\nabla\Phi(\boldsymbol{W}_j),\boldsymbol{e}_1\rangle}{d}\right|+\left|\frac{\langle\nabla H^{j+1}(\boldsymbol{W}_j),\boldsymbol{e}_1\rangle}{d}\right|\right) > \frac{\gamma}{10d^r}\right) \leq \frac{C\epsilon\eta}{\gamma d^{1-r}}, \tag{C.8}$$

$$\mathbb{P}_{\boldsymbol{w}_0}\left(\sup_{t\leq M}\eta^2\sum_{j=0}^{t-1}\left(\frac{A}{d^2}+\frac{L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}}{d}\right)|m(\boldsymbol{W}_j)| > \frac{\gamma}{10d^r}\right) \leq \frac{C\epsilon d^r}{\hat{L}^{1+\iota/4}\gamma}. \tag{C.9}$$

Proof. *We start by proving the first bound, by using Markov's inequality and Cauchy-Schwarz inequality.*

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{w}_0}\left(\sup_{t\leq M}\eta^3\sum_{j=0}^{t-1}\left(\frac{A}{d^2}+\frac{L_{j+1}}{d}\right)\left(\left|\frac{\langle\nabla\Phi(\boldsymbol{W}_j),\boldsymbol{e}_1\rangle}{d}\right|+\left|\frac{\langle\nabla H^{j+1}(\boldsymbol{W}_j),\boldsymbol{e}_1\rangle}{d}\right|\right) > \lambda\right) \\
&\leq \frac{M\eta^3}{\lambda d^2}\sup_{\boldsymbol{w}}\mathbb{E}\left[\frac{|\nabla H(\boldsymbol{w})|^2}{d}+\frac{A}{d}\right]\Big(|\langle\nabla\Phi(\boldsymbol{w}),\boldsymbol{e}_1\rangle|+|\langle\nabla H(\boldsymbol{w}),\boldsymbol{e}_1\rangle|\Big) \\
&\leq \frac{\alpha d\eta^3}{\lambda d^2}\sqrt{\sup_{\boldsymbol{w}}\mathbb{E}\left[\left|\frac{\nabla H(\boldsymbol{w})}{d}\right|^4\right]}\sqrt{\sup_{\boldsymbol{w}}|\langle\nabla\Phi(\boldsymbol{w}),\boldsymbol{e}_1\rangle|^2+\sup_{\boldsymbol{w}}|\langle\nabla H(\boldsymbol{w}),\boldsymbol{e}_1\rangle|^2} \\
&\leq \frac{\alpha d\eta^3}{\lambda d^2}\sqrt{\bar{L}+\frac{A^2}{d^2}}\cdot\sqrt{A+C_1}.
\end{aligned}$$

Take $\lambda = \frac{\gamma}{10d^r}$ and use $\alpha\eta^2 \le \epsilon$, which yields the upper bound

$$\frac{d^{r-1}\eta\epsilon}{\gamma}\sqrt{\left(\bar{L}+\frac{A^2}{d^2}\right)(A+C_1)}.$$

For the second bound, we also use Markov's and Cauchy-Schwarz inequalities. To ease notations, denote $L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}$ by $\tilde{L}_{j+1}$.

$$\mathbb{P}_{\boldsymbol{w}_0}\left(\sup_{t\le M}\sum_{j=0}^{t-1}\eta^2\left(\frac{A}{d^2}+\frac{\tilde{L}_{j+1}}{d}\right)|m_j|>\lambda\right)\le\mathbb{P}_{\boldsymbol{w}_0}\left(\eta^2\sum_{j=0}^{M-1}\frac{\tilde{L}_{j+1}}{d}>\lambda-\frac{A\epsilon}{d}\right).$$

By Markov's inequality,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{w}_0}\left(\sum_{j=0}^{M-1}\frac{\tilde{L}_{j+1}}{d}>\Lambda\right)&\le\frac{M}{\Lambda}\sup_{j\le M}\mathbb{E}_{\boldsymbol{w}_0}\left[\tilde{L}_{j+1}\right]\\&\le\frac{M}{\Lambda}\sqrt{\sup_{\boldsymbol{w}}\mathbb{E}\left(\left|\frac{1}{\sqrt{d}}\nabla H(\boldsymbol{w})\right|^4\right)\cdot\sup_{\boldsymbol{w}}\mathbb{P}\left(\left|\frac{1}{\sqrt{d}}\nabla H(\boldsymbol{w})\right|^2>\hat{L}\right)}\\&\le\frac{\alpha d}{\Lambda}\sqrt{\bar{L}\sup_{\boldsymbol{w}}\frac{\mathbb{E}\left(\left|\frac{1}{\sqrt{d}}\nabla H(\boldsymbol{w})\right|^{4+\iota}\right)}{\hat{L}^{2+\iota/2}}}\le\frac{\alpha d}{\Lambda}\frac{\bar{L}}{\hat{L}^{1+\iota/4}}.\end{aligned}$$

By setting $\Lambda=\frac{d}{\eta^2}\left(\lambda-\frac{A\epsilon}{d}\right)$ with $\lambda=\frac{\gamma}{10d^r}$, the left-hand side in Equation (C.9) is bounded by

$$\frac{C\epsilon d^r}{\hat{L}^{1+\iota/4}\gamma},$$

for some $C(A,\bar{L})>0$. □

By setting $\hat{L}=d^{1+(k-2)r}$, we can ensure that the term in Equation (C.9) vanishes as d goes to infinity. By summing over all times $t\ge1$, and plugging Lemma 37, as well as Lemma 38 into the inequality given by Equations (C.6) and (C.7), we obtain that uniformly over $\boldsymbol{w}_0\in\mathbb{S}^{d-1}$, with probability $1-o_d(1)$,

$$m_t\ge\frac{4}{5}m_0+\sum_{j=0}^{t-1}\frac{\eta a_k|m_j|^{k-1}}{4d}\left(1-\frac{L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}}{K\bar{L}\log d}\right)-\frac{\eta}{d}\sum_{j=0}^{t-1}\left\langle\nabla H^{j+1}(\boldsymbol{W}_j),\boldsymbol{e}_1\right\rangle,\quad\forall t\le\tau^-_{\frac{\gamma}{2d^r}}\wedge\tau^+_\nu.\tag{C.10}$$

Proposition 39. *Let $\hat{L}=d^{1+(k-2)r}$. For $k\ge2$, if $\alpha\eta^2\le\epsilon$ for some $\epsilon>0,\eta\le\bar{\eta}_d(k)$, and α is at most polynomial in d, then for every $\gamma>0$,*

$$\lim_{d\to\infty}\inf_{\boldsymbol{w}_0}\mathbb{P}_{\boldsymbol{w}_0}\left(\sum_{j=0}^{t-1}\frac{\eta a_k|m_j|^{k-1}}{4d}\left(1-\frac{L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}}{K\bar{L}\log d}\right)\ge-\frac{\gamma}{10d^r}+\sum_{j=0}^{t-1}\frac{\eta a_k|m_j|^{k-1}}{8d},\quad\forall t\le M\right)=1.$$

Moreover, this limit holds uniformly over choices of $\epsilon>0$.

Proof. The strategy of the proof is to define a submartingale, and control its conditional second moment, as well as its martingale increments, to use a martingale inequality due to Freedman. To ease notations, denote $L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}$ by $\tilde{L}_{j+1}$. It suffices to prove the following: for $\hat{L} = d^{1+(k-2)r}$, and α at most polynomial in d with $\alpha\eta^2 \le \epsilon$ for some $\epsilon > 0$, we have for every $\gamma > 0$,

$$\lim_{d\to\infty} \sup_{\boldsymbol{w}_0\in\mathbb{S}^{d-1}} \mathbb{P}_{\boldsymbol{w}_0}\left(\inf_{t\le M} \sum_{j=0}^{t-1} \frac{\eta a_k |m_j|^{k-1}}{4d}\Big(\frac{1}{2} - \frac{\tilde{L}_{j+1}}{K\bar{L}\log d}\Big) < -\frac{\gamma}{10d^r} \right) = 0.$$

Fix any $\boldsymbol{w}_0 \in \mathbb{S}^{d-1}$.

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{w}_0}\left(\inf_{t\le M} \sum_{j=0}^{t-1} \frac{\eta a_k |m_j|^{k-1}}{4d}\Big(\frac{1}{2} - \frac{\tilde{L}_{j+1}}{K\bar{L}\log d}\Big) < -\frac{\gamma}{10d^r} \right) \\
&\qquad\le M \sup_{t\le M} \mathbb{P}_{\boldsymbol{w}_0}\left(\sum_{j=0}^{t-1} \frac{\eta a_k |m_j|^{k-1}}{4d}\Big(\frac{1}{\log d} - \frac{\tilde{L}_{j+1}}{K\bar{L}\log d}\Big) < -\frac{\gamma}{10d^r} \right).
\end{aligned}$$

Let $Z_t := \sum_{j=0}^{t-1} \frac{\eta a_k |m_j|^{k-1}}{4d}\Big(\frac{1}{\log d} - \frac{\tilde{L}_{j+1}}{K\bar{L}\log d}\Big)$. Since for all j, m_j is $\mathcal{F}_j-$measurable and

$$\mathbb{E}[\tilde{L}_{j+1}|\mathcal{F}_j] \le \sup_{\boldsymbol{w}} \mathbb{E}\left[\left|\frac{1}{\sqrt{d}}\nabla H(\boldsymbol{w})\right|^2\right] \le \bar{L},$$

we have that for all $K \ge 1, Z_t$ is an $\mathcal{F}_t-$submartingale. The martingale increments are bounded as follows.

$$|Z_t - Z_{t-1}| = \frac{\eta a_k |m_{t-1}|^{k-1}}{4d}\left|\frac{1}{\log d} - \frac{\tilde{L}_t}{K\bar{L}\log d}\right| \le \frac{\eta a_k}{4d\log d}(1 \vee \frac{\hat{L}}{K\bar{L}}) \le \frac{\eta a_k}{4d\log d}\Big(1 + \frac{\hat{L}}{\bar{L}}\Big).$$

For the conditional variances, we have

$$\begin{aligned}
\mathbb{E}[(Z_t - Z_{t-1})^2|\mathcal{F}_{t-1}] &= \Big(\frac{\eta a_k}{4d\log d}\Big)^2 \mathbb{E}\left[|m_{t-1}|^{2(k-1)}\left(1 - \frac{\tilde{L}_t}{K\bar{L}}\right)^2\middle|\mathcal{F}_{t-1}\right] \\
&\le \Big(\frac{\eta a_k}{4d\log d}\Big)^2\left(1 + \frac{1}{\bar{L}^2}\mathbb{E}\left[\tilde{L}_t^2\middle|\mathcal{F}_{t-1}\right]\right) \le \Big(\frac{\eta a_k}{4d\log d}\Big)^2\left(1 + \frac{1}{\bar{L}^2}\mathbb{E}\left[L_t^2\middle|\mathcal{F}_{t-1}\right]\right) \\
&= \Big(\frac{\eta a_k}{4d\log d}\Big)^2\left(1 + \frac{1}{\bar{L}^2}\left[\left|\frac{1}{\sqrt{d}}\nabla H^t(X_{t-1})\right|^4\middle|\mathcal{F}_{t-1}\right]\right) \\
&\le \Big(\frac{\eta a_k}{4d\log d}\Big)^2\Big(1 + \frac{1}{\bar{L}}\Big) \le \Big(\frac{\eta a_k}{4d\log d}\Big)^2 \quad a.s.
\end{aligned}$$

Therefore, by Freedman's inequality:

$$\sup_{t\le M} \mathbb{P}_{\boldsymbol{w}_0}\left(\sum_{j=0}^{t-1} \frac{\eta a_k |m_j|^{k-1}}{4d}\Big(\frac{1}{\log d} - \frac{\tilde{L}_{j+1}}{K\bar{L}\log d}\Big) \ge -\frac{\gamma}{10d^r} \right) \le \exp\left(\frac{-\gamma/100d^{2r}}{M\Big(\frac{\eta a_k}{4d\log d}\Big)^2 + \frac{1}{3}\frac{\eta a_k}{4d\log d}\cdot\frac{\gamma}{10d^r}\Big(1 + \frac{\hat{L}}{\bar{L}}\Big)} \right).$$

Now there remains to choose $\hat{L}$ appropriately, to ensure that the bound is $o(\frac{1}{M})$, so that the overall bound is $o(1)$. This constrains us to take either $\epsilon = \Omega(d^{1-2r+\mu}\log d)$ or $\eta\hat{L} = \Omega(d^{1-r+\mu})$, for some $\mu > 0$. Assuming we take the largest $\eta = \Theta(d^{-r(k-2)}/\log d)$, taking $\hat{L} = d^{1+(k-2)r}$ ensures we have the right bound, irrespective of $\epsilon > 0$.

□

Proposition 40. *If C_1 is as in Assumption B, for every $\lambda > 0$, we have*

$$\sup_{T\le M}\sup_{\boldsymbol{w}_0\in\mathbb{S}^{d-1}}\mathbb{P}_{\boldsymbol{w}_0}\left(\max_{t\le T}\frac{1}{\sqrt{T}}\left|\sum_{j=0}^{t-1}\langle\nabla H^{j+1}(\boldsymbol{W}_j),\boldsymbol{e}_1\rangle\right|\ge\lambda\right)\le\frac{2C_1}{\lambda^2}\,, \tag{C.11}$$

Proof. *Let $\tilde{M}_t = \frac{dM_t}{\eta}$. It is a martingale with variance*

$$\sup_{\boldsymbol{w}_0\in\mathbb{S}^{d-1}}\mathbb{E}_{\boldsymbol{w}_0}\left[\sum_{j=0}^{t-1}\Big(\langle\nabla H^{j+1}(\boldsymbol{W}_j),\boldsymbol{e}_1\rangle\Big)^2\right]\le t\sup_{\boldsymbol{w}}\mathbb{E}\left[\Big(\langle\nabla H(\boldsymbol{w}),\boldsymbol{e}_1\rangle\Big)^2\right]\le C_1 t.$$

By Doob's maximal inequality,

$$\sup_{\boldsymbol{w}_0}\mathbb{P}_{\boldsymbol{w}_0}\Big(\sup_{t\le T}\Big|\tilde{M}_t>\lambda\sqrt{T}\Big|\Big)\le\frac{2\sup_{\boldsymbol{w}_0}\mathbb{E}_{\boldsymbol{w}_0}[\tilde{M}_t^2]}{\lambda^2 t}\le\frac{2C_1}{\lambda^2}.$$

□

Therefore, by Proposition 40, for all $b > 0$,

$$\sup_{T\le M}\sup_{\boldsymbol{w}_0\in\mathbb{S}^{d-1}}\mathbb{P}_{\boldsymbol{w}_0}\left(\max_{t\le T}\frac{\eta}{d}\left|\sum_{j=0}^{t-1}\langle\nabla H^{j+1}(\boldsymbol{W}_j),\boldsymbol{e}_1\rangle\right|\ge\frac{\eta b\sqrt{T}}{10d}\right)\le\frac{200C_1}{b^2T}. \tag{C.12}$$

We can now prove Proposition 34 by combining Equation (C.10), Proposition 39, and Proposition 40. For every $\gamma > 0$ and $\nu > \nu_0(\varrho, a_k, a_{k+1})$,

$$\lim_{d\to\infty}\inf_{\boldsymbol{w}_0\in E_{\gamma/d^r}}\mathbb{P}_{\boldsymbol{w}_0}\Big(m_t\ge\frac{7}{10}m_0+\sum_{j=0}^{t-1}\frac{\eta a_k|m_j|^{k-1}}{8d}-\frac{\eta}{d}\sum_{j=0}^{t-1}\langle\nabla H^{j+1}(\boldsymbol{W}_j),\boldsymbol{e}_1\rangle\,,\ \ \forall t\le\tau^-_{\frac{\gamma}{2d^r}}\wedge\tau^+_\nu\wedge M\Big)=1.$$

Furthermore, if $D = D_d, \eta \le \bar{\eta}_d(k)$ and $\bar{t}$ are as in Proposition 34, for all $\boldsymbol{w}_0 \in E_{\gamma/d^r}$, if $T \le \bar{t}$, then

$$\frac{\eta D\sqrt{T}}{10d}\le\frac{\gamma}{10d^r}\le\frac{m_0}{10}.$$

Therefore, by applying the directional error martingale bound (Equation (C.12)) with $b = D$, we obtain the desired bound.

Proof. [Proof of Proposition 35] Observe the following discrete analogue of Bihari-LaSalle inequality: suppose that $(m_t)_{t\in\mathbb{N}}$ is a sequence satisfying, for some $k \ge 3$ and $a, b > 0$,

$$m_t\ge a+\sum_{j=0}^{t-1}bm_j^{k-1}\quad\text{then}\quad m_t\ge\frac{a}{(1-(k-2)ba^{k-2}t)^{1/(k-2)}}.$$

Therefore, for $a = \frac{m_0}{2}, b = \frac{\eta a_k}{8d}$, we obtain

$$m_t \geq \frac{m_0}{(1-(k-2)\frac{\eta a_k}{8d}m_0^{k-2}t)^{1/(k-2)}} =: g_k(t).$$

$g_k(t) \geq \nu$ provided

$$\nu^{k-2}\Big(1-(k-2)\frac{\eta a_k}{8d}\Big(\frac{\gamma}{d^r}\Big)^{k-2}t\Big) = o\Big(\Big(\frac{\gamma}{d^r}\Big)^{k-2}\Big).$$

Thus, for $t \geq t^\star = \Big\lceil \frac{8d}{\eta a_k (k-2)\gamma^{k-2}} d^{r(k-2)}\Big(1-\frac{K}{d^{r(k-2)}}\Big)\Big\rceil$, for K large enough, we get that $g_k(t) \geq \nu$. The only remaining thing is to ensure $t^\star \leq \bar{t} \wedge M$. $t^\star \leq \bar{t}$ always holds for $\eta \leq \bar{\eta}(k)$. To ensure $t^\star \leq M = \alpha d$, we must have

$$\alpha \gtrsim d^{r(k-2)}/\eta.$$

By optimizing over the values for η, and taking into account the constraint that $\alpha\eta^2 \leq \epsilon = \mathcal{O}(d^{1+(k-3)r})$, we must pick $\eta = \Theta\Big(\frac{d^{-r(k-2)}}{\log d}\Big)$, and hence,

$$\alpha \gtrsim d^{2r(k-2)}\log d.$$

□

Corollary 41. *Under Assumptions A and B, for $\eta = \bar{\eta}(k), \alpha \gtrsim d^{2r(k-2)}\log d$ there exists $\nu_0(a_k, a_{k+1})$, such that for all $\nu < \nu_0$, for every $\gamma > 0$, we have*

$$\lim_{d\to\infty} \inf_{\boldsymbol{w}_0 \in E_{\gamma/d^r}} \mathbb{P}_{\boldsymbol{w}_0}\Big(\tau_\nu^+ \leq \bar{t} \wedge M\Big) = 1.$$

Hence the sample complexity $M = \mathcal{O}(d^{2r(k-2)+1}\log d)$ is sufficient for weak recovery of the single-index model.