Learning a Single Index Model from Anisotropic Data with Vanilla Stochastic Gradient Descent

Guillaume Braun
RIKEN AIP

Ha Quang Minh
RIKEN AIP

Masaaki Imaizumi
RIKEN AIP
University of Tokyo

Abstract

We investigate the problem of learning a Single Index Model (SIM)—a popular model for studying the ability of neural networks to learn features—from anisotropic Gaussian inputs by training a neuron using vanilla Stochastic Gradient Descent (SGD). While the isotropic case has been extensively studied, the anisotropic case has received less attention and the impact of the covariance matrix on the learning dynamics remains unclear. For instance, Mousavi-Hosseini et al. (2023b) proposed a spherical SGD that requires a separate estimation of the data covariance matrix, thereby oversimplifying the influence of covariance. In this study, we analyze the learning dynamics of vanilla SGD under the SIM with anisotropic input data, demonstrating that vanilla SGD automatically adapts to the data's covariance structure. Leveraging these results, we derive upper and lower bounds on the sample complexity using a notion of effective dimension that is determined by the structure of the covariance matrix instead of the input data dimension. Finally, we validate and extend our theoretical findings through numerical simulations, demonstrating the practical effectiveness of our approach in adapting to anisotropic data, which has implications for efficient training of neural networks.

Proceedings of the 28$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 Introduction

In many high-dimensional applications, data sets are often assumed to have an underlying low-dimensional structure (e.g., images or text can be embedded in low-dimensional manifolds). This assumption provides a way to circumvent the curse of dimensionality.

The Single Index Model (SIM) provides a simple yet powerful statistical framework to evaluate the ability of an algorithm to adapt to the latent dimension of the data. In SIM, the response variable $y \in \mathbb{R}$ is linked to the covariates $x \in \mathbb{R}^d$ through a *link function* f that depends only on a rank-one projection of x. More formally, $y = f(\langle w^*, x \rangle)$, where $w^* \in \mathbb{R}^d$ represents the direction of the latent low-dimensional space and is referred to as the *single-index*. This versatile model generalizes the well-known Generalized Linear Model (GLM) when the link function is known. Additionally, SIM can be extended to capture multiple directions, leading to the Multi-Index Model (Abbe et al., 2023; Oko et al., 2024).

In recent years, SIM has become a popular generative model for studying the ability of neural networks trained with SGD or a variant. This is because the usage of SIM effectively adapts to the latent dimension of the data, in contrast to the kernel method (Ghorbani et al., 2020; Damian et al., 2022). In particular, the isotropy of the covariates x plays an important role: when x are sampled from an isotropic Gaussian distribution, it is well-known that the difficulty of estimating w^* depends on d and the information exponent associated with f (Arous et al., 2020; Bietti et al., 2022) in the online setting, or the generative exponent if one can reuse samples (Damian et al., 2024b; Arnaboldi et al., 2024; Lee et al., 2024).

However, while anisotropy of input data is common

in real-world applications such as classification, only a few works extend beyond the simplifying assumption of isotropic Gaussian data, limiting their applicability to more realistic, complex scenarios. To overcome this limitation, several works handle the anisotropy of x, for instance, Zweig et al. (2023) extends the SIM analysis to inputs generated from an approximately spherically symmetric distribution. Ba et al. (2023); Mousavi-Hosseini et al. (2023b) consider a setting where the inputs are sampled from a Gaussian distribution with a covariance matrix having a spike aligned with the single-index w^*. The first work proposes a layer-wise training method, whereas the second work studies a version of spherical SGD that requires an estimate of the covariance matrix of x. To the best of our knowledge, the only algorithm that comes with general theoretical guarantees is the mean-field Langevin dynamic analyzed by Mousavi-Hosseini et al. (2024). Unfortunately, it is impractical due to computational inefficiencies.

In practice, the structure of the covariance matrix is often unknown, and rather than specialized algorithms, simple and generic methods such as vanilla SGD are typically employed. This observation motivates our central research question:

"Can vanilla SGD learn a SIM model under a general class of covariance structures?"

Answering this question will not only demonstrate the learnability of non-isotropic covariance structures but also highlight that widely-used, simple algorithms like vanilla SGD can successfully achieve this goal. Moreover, this work represents a first step toward extending the analysis to more complex input data, such as Gaussian mixtures or functional data (Balasubramanian et al., 2024).

In this study, we analyze the vanilla SGD to learn a SIM from general anisotropic inputs and show its adaptability to general covariance Q. This contrasts with the result of Mousavi-Hosseini et al. (2023b) that shows spherical SDG fails if the algorithm is not modified to incorporate information about Q. Specifically, we show that our estimator has a constant correlation with the single-index w^* after T SGD iterations and characterize T as a function of Q, the alignment of Q with w^* and the information exponent of the link function f. Interestingly, our bound depends on an effective dimension determined by Q, instead of the input data dimension d. We also establish a Correlated Statistical Query (CSQ) lower bound, suggesting that our effective measure of the dimension is correct on average over w^*. We illustrate and complement our theoretical findings through numerical simulations.

One of the main technical challenges is that, unlike spherical SGD, the evolution of the correlation with w^* also depends on the evolution of the norm of weights. We tackle this problem by developing a method that simultaneously controls the evolution of all the parameters. Our analysis provides insight into the interplay between correlation and weight evolution, which could be instrumental in understanding the training dynamics of wider neural networks.

1.1 Related work

Single Index Model. As an extension of the Generalized Linear Model (GLM) (Nelder and Wedderburn, 1972), the Single Index Model (SIM) is a versatile and widely used statistical framework. It has been applied in various domains, including longitudinal data analysis (Jiang and Wang, 2011), quantile regression (Ma and He, 2016), and econometrics (Hardle et al., 1993). In recent years, SIMs have attracted increasing attention from the theoretical deep learning community, particularly as a tool to evaluate the ability of neural networks (NNs) to learn low-dimensional representations, often referred to as features, in contrast to kernel methods (Ghorbani et al., 2020). This distinction is highlighted by the *lazy regime* (Jacot et al., 2018), where neural networks exhibit kernel-like behavior but cannot learn features.

Several works demonstrate that a Single or Multi-Index Model can be learned from isotropic Gaussian inputs by training a two-layer neural network in a layer-wise manner (Damian et al., 2022; Mousavi-Hosseini et al., 2023a; Bietti et al., 2022, 2023; Abbe et al., 2023; Zhou and Ge, 2024). However, fewer studies have addressed the more challenging anisotropic case. The most closely related work is Mousavi-Hosseini et al. (2023b), which analyzes a specific covariance structure $Q = \frac{I_d+\kappa\theta\theta^\top}{1+\kappa}$ where $\theta \in \mathbb{S}^{d-1}$ is correlated with target index w^* and κ measures the intensity of the spike in the direction θ. In this setting, they show that spherical Gradient Flow (GF) fails to estimate w^*, but a modified algorithm using normalization depending on the covariance matrix Q succeeds. In contrast, our work analyzes vanilla SGD and demonstrates that this simpler algorithm succeeds in learning w^* without any prior estimation of Q, making it agnostic to the covariance matrix. Also, our analysis sheds light on how the learning rate η should be chosen, whereas the GF framework is not directly implementable.

Training dynamic. The dynamics of SGD and its variants in training shallow neural networks have been extensively studied (Mei et al., 2018; Jacot et al., 2018). Our analysis builds on the framework introduced by Arous et al. (2020) for spherical SGD in the online setting, which provides insight into the behavior of the estimator's weights over time. Additionally, our approach is related to the recent work by Glasgow (2024), which studies the simultaneous training of a two-layer neural network to learn the XOR function. Similar to their method, we project the weight vector onto signal and noise components to control their respective growth during training. Furthermore, recent studies on the edge of stability (EoS) phenomenon in quadratic models Zhu et al. (2024a,b); Chen et al. (2024) show that two-layer neural networks trained with a larger learning rate than the prescribed one can generalize well after an unstable training phase.

One of the primary goals of our study is to focus on vanilla SGD, a practical algorithm widely used in real-world scenarios, rather than on specialized theoretical variants like spherical SGD. We believe that this work contributes to a deeper understanding of SGD's behavior in non-convex optimization problems and highlights its effectiveness in solving complex learning tasks involving anisotropic data.

1.2 Notations

We use $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ to denote the Euclidean norm and scalar product, respectively. When applied to a matrix, $\|\cdot\|$ refers to the operator norm. Any positive semi-definite matrix Q induces a scalar product defined by $\langle x, y\rangle_Q = x^\top Q y$. The Frobenius norm of a matrix A is denoted by $\|A\|_F$. The $d \times d$ identity matrix is represented by I_d. The standard Gaussian measure on $\mathbb{R}$ is denoted by γ and the corresponding Hilbert space $L^2(\mathbb{R}, \gamma)$ is referred to as $\mathcal{H}$. The $(d-1)$-dimensional unit sphere is denoted by $\mathbb{S}^{d-1}$. We use the notation $a_n \lesssim b_n$ (or $a_n \gtrsim b_n$) for sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$ (or $a_n \geq C b_n$) for all n. If the inequalities hold only for sufficiently large n, we write $a_n = O(b_n)$ (or $a_n = \Omega(b_n)$).

2 Learning a SIM

Model. We observe for $i = 1 \dots T$ i.i.d. samples $(x^{(i)}, y^{(i)}) \in \mathbb{R}^d \times [-1, 1]$ generated by the following process: the inputs $x^{(i)} \sim \mathcal{N}(0, Q)$ are generated from a Gaussian distribution with covariance matrix Q. Also, $y^{(i)}$ follows the SIM with given $x^{(i)}$ as

$$y^{(i)} = f\left(\left\langle x^{(i)}, \frac{w^*}{\left\|Q^{1/2} w^*\right\|}\right\rangle\right), \tag{2.1}$$

where $w^* \in \mathbb{S}^{d-1}$ and $f : \mathbb{R} \to [-1, 1]$ is an unknown link function. The normalization factor $\left\|Q^{1/2} w^*\right\|$ is introduced so that $\langle x^{(i)}, \frac{w^*}{\|Q^{1/2} w^*\|}\rangle \sim \mathcal{N}(0, 1)$ and only affects the definition of f.

Assumption A1. *We assume that* $\|Q\| = 1$ *and that* $\left\|Q^{1/2} w^*\right\| = c^* \leq 1$ *is of constant order.*

This assumption is satisfied for both $Q = I$ and a spiked covariance matrix of the form $Q = (1 + \kappa)^{-1}(I + \kappa\theta\theta^\top)$, where $\theta \in \mathbb{S}^{d-1}$ is such that $\langle\theta, w^*\rangle$ is of constant order. While we believe our analysis could be extended to the setting where $\left\|Q^{1/2} w^*\right\| = o(1)$, the scarcity of input in the direction of w^* makes this more challenging. In such cases, it would be more efficient to correct the sample by estimating Q.

Learning method. Following the approach in Arous et al. (2020), we employ the correlation loss function

$$L(y, \hat{y}) = 1 - y\hat{y}.$$

Our method involves training a single neuron, defined as

$$f_w(x) = \sigma(w^\top x)$$

with $w \in \mathbb{R}^d$ and σ denotes the ReLU activation function, using vanilla gradient descent. The steps are as follows:

- Sample $w' \sim \mathcal{N}(0, I_d)$ and initialize $w^{(0)} = r\frac{w'}{\|w'\|}$ where $r > 0$ is a scaling parameter.
- Update the weight by SGD: $w^{(t+1)} = w^{(t)} - \eta\nabla_{w^{(t)}} L(y^{(t)}, f_{w^{(t)}}(x^{(t)}))$ where $\eta > 0$ is the learning rate.

Vanilla gradient descent without constraining the weights provides two key advantages: (i) computational efficiency, as it eliminates the need to estimate the covariance matrix Q and compute terms like $\left\|Q^{1/2} w^*\right\|$, and (ii) simplicity, aligning more closely with standard practices in neural network training.

Remark 1. *Previous works such as Arous et al. (2020); Bietti et al. (2022); Abbe et al. (2023) leverage spherical SGD to control the norm of the weights* $w^{(t)}$ *at each iteration, simplifying theoretical analysis. However, in the anisotropic setting, spherical SGD must be modified with knowledge of the covariance matrix* Q *to succeed, as shown by Mousavi-Hosseini et al. (2023b).*

3 Main Results

Before stating our main results, we will introduce additional notations and discuss the required assumptions.

3.1 Notation and Assumptions

First, we assume that f has an information exponent $k^* \geq 1$ and is normalized and bounded.

Assumption A2. *We assume that the link function $f : \mathbb{R} \to [-1, 1]$ is such that $\mathbb{E}(f) = 0$, $\mathbb{E}(f^2) = 1$ and f has information exponent k^* defined as*

$$k^* := \min\{k \geq 1 : \mathbb{E}(fH_k) \neq 0\}$$

where H_k is the order k Hermite polynomial (see Section A for more details).

Remark 2. *We assumed for simplicity that f takes value in $[-1, 1]$. Our analysis could be extended to the class of functions f growing polynomially, i.e., there exists a constant $C > 0$ and an integer p such that for all $x \in \mathbb{R}$, $|f(x)| \leq C(1+|x|^p)$. For example, using Lemma 16 in Mousavi-Hosseini et al. (2023b), one could obtain a uniform upper bound on $|y^{(i)}|$ and reduce to the case where $f(x) \in [-1, 1]$ for all x. We leave the full proof to future work.*

The following measure of correlation is defined in terms of the scalar product induced by Q

$$m_t := \left\langle \frac{Q^{1/2}w^{(t)}}{\left\|Q^{1/2}w^{(t)}\right\|}, \frac{Q^{1/2}w^*}{\left\|Q^{1/2}w^*\right\|} \right\rangle.$$

This is in contrast to the isotropic setting, in which the correlation between the weights $w^{(t)}$ and the signal w^* is measured by $\left\langle \frac{w^{(t)}}{\left\|w^{(t)}\right\|}, w^* \right\rangle$.

As in previous work, we also assume that the correlation is positive at initialization.

Assumption A3. *At initialization, $m_0 > 0$.*

Furthermore, we will need the following assumption to control the population gradient (see the proof of Lemma 1).

Assumption A4. *Let $xf(x) = \sum_{k \geq k^*-1} c_k \frac{H_k(x)}{\sqrt{k!}}$ and $\sigma'(x) = \sum_{k \geq 0} b_k \frac{H_k}{\sqrt{k!}}$ be the Hermite basis decomposition of $xf(x)$ and $\sigma'(x)$. Assume that there are constants $\gamma' > 0$, $c > 0$ such that for all $x \in [0, \gamma']$,*

$$\sum_{k \geq k^*-1} b_k c_k x^k \leq -cx^{k^*-1}.$$

This assumption allows us to approximate the Hermite decomposition of the gradient by its first non-zero term. A similar assumption was used in Arous et al. (2020) and Mousavi-Hosseini et al. (2023b). The relation between the information exponent of f and $x \to xf(x)$ is derived in Proposition 3.

Example of link function. Recall that $b_0 = 0.5$, $b_{2m} = 0$ and $b_{2m+1} = \frac{(-1)^m}{\sqrt{2\pi} m! 2^m (2m+1)}$ (Damian et al., 2022). If k^* is even, we can choose $f^* = \frac{H_{k^*}(x)}{\sqrt{k^*!}}$. The proof of Lemma 3 shows that $xf^*(x) = \frac{H_{k^*+1}(x)}{\sqrt{k^*!}} + \frac{kH_{k^*-1}(x)}{\sqrt{(k^*)!}}$. Consequently, there are only two non-zero coefficients in the Hermite decomposition of $xf(x)$, corresponding to odd Hermite polynomial. Hence Assumption A4 is satisfied for small enough x, up to a sign factor.

Remark 3. *One could possibly remove Assumptions A4 and A3 by considering a two-layer neural network. Since the assumption on initialization is satisfied with probability $1/2$, it will be satisfied by a constant proportion of the neurons. Moreover, by training the second layer, we could approximate f, hence controlling the sign of the coefficients appearing in Assumption A4. However, to show that Assumption A4 is satisfied in the two-layer neural network setting, previous work (Bietti et al., 2022; Lee et al., 2024) rely on specific (randomized) link functions, while our analysis relies crucially on the homogeneity of the ReLU activation function.*

3.2 Upper-bound on the required sample complexity

We analyze the upper bound on the sample complexity required to recover the single-index w^*. To simplify its statement, we introduce the following notation for a ratio:

$$\Theta := \Theta(Q, w^*) := \frac{\left\|Q^{1/2}w^*\right\|}{\left\|Q^{1/2}\right\|_F}. \tag{3.1}$$

Theorem 1. *Assume that Assumptions A1, A2, A3 and A4 hold and the initialization scaling r is such that $\left\|Q^{1/2}w^{(0)}\right\| = c_r \left\|Q^{1/2}w^*\right\|$ for some constant $c_r \in (0, 1]$.*

(1) When $k^* \geq 3$, choose $\eta = \epsilon_d^2 m_0^{k^*-2}\Theta^2$ where $\epsilon_d \to 0$ as $d \to \infty$. Then, after $T = \epsilon_d^{-2} m_0^{2(2-k^*)}\Theta^{-2}$ iterations, $w^{(T)}$ weakly recovers w^*, i.e. with probability $1 - o(1)$, $m_T \geq \delta$ for some constant $\delta > 0$.

(2) If $k^* = 1$, the same result holds with the choices $\eta = \epsilon_d^2\Theta^2$ and $T = \epsilon_d^{-2}\Theta^{-2}$.

(3) If $k^* = 2$, the result holds with the choices $\eta = \epsilon_d^2(\log m_0)^{-1}\Theta^2$ and $T = \epsilon_d^{-2}\log^2(m_0)\Theta^{-2}$.

This theorem gives conditions on the sample complexity and the learning rate to ensure that vanilla SGD weakly recovers w^*. The proof of this theorem can be divided into two parts. First, we analyze the population dynamic in Section 4.1. Then, we control the effect of the noise in Section 4.3. We discuss the extension to strong consistency, i.e., obtaining $m_t \to 1$ in Section E.3 in the appendix.

Remark 4. *The typical order of magnitude of m_0 is $\|Qw^*\| \|Q^{1/2}w^*\|^{-1}\|Q^{1/2}\|_F^{-1}$, as shown by concentration inequalities. See Section E.1 in the appendix. In particular, when $Q = I_d$, $m_0 \approx \sqrt{d}^{-1}$ and our upper bounds matches the one obtained by Arous et al. (2020). When the covariance matrix is aligned with w^*, one can have $\sqrt{d}^{-1} \ll m_0$. This accelerates the convergence, as experimentally shown in Section 5.*

Remark 5. *Our bound is, in general, weaker than the one obtained by Mousavi-Hosseini et al. (2023b) that is of order $dm_0^{(2-2k^*)}$. However, their analysis is based on gradient flow and hence cannot be implemented directly, whereas we analyzed a simple and popular algorithm used in practice. The discretization of the gradient flow is not a straightforward task. As highlighted by our analysis, the choice of the learning rate is crucial and has an effect on the required sample complexity. We believe this is the main reason one can obtain a better bound with gradient flow. Notice, however, that when Q is approximately low-rank, one could have $T \ll d$ whereas the bounds of Mousavi-Hosseini et al. (2023b) are always at least linear in d.*

3.3 Correlated Statistical Query (CSQ) lower-bound

A common way to provide a lower bound on the required sample complexity for SGD-like algorithms is to rely on the Correlated Statistical Query (CSQ) framework. It is described in Section C.

Theorem 2 (CSQ lower-bound)**.** *Assume that $\left\|Q^{1/2}\right\|_F^2 \gtrsim \|Q\|_F \sqrt{\log d}$ and $\|Q\|_F \gtrsim \sqrt{\log d}$. Let us denote*

$$v = \min\left(\frac{\|Q\|_F}{\left\|Q^{1/2}\right\|_F^2}, \frac{1}{\sqrt{d}}\right).$$

Then, for any integer $k \geq 1$, there exists a class $\mathcal{F}_k$ of polynomial functions of degree k such that any CSQ algorithm using a polynomial number of queries $q = O(d^C)$ requires a tolerance of order at most

$$\tau^2 \leq \epsilon^{k/2}$$

where $\epsilon = v\sqrt{\log(qv^{k/2})}$.

Proof. The main difficulty is constructing a large vector family with small correlations measured by the scalar product induced by Q. It is detailed in Section C. □

Remark 6. *The quantity v can be interpreted as the typical value of m_0 at initialization. Since one could always use an oracle knowledge of Q to reduce to the isotropic case (where $v = \sqrt{d}^{-1}$), our bound is only useful when $\frac{\|Q\|_F}{\left\|Q^{1/2}\right\|_F^2} \geq \frac{1}{\sqrt{d}}$. Note that the term $\frac{\|Q\|_F}{\left\|Q^{1/2}\right\|_F^2}$ corresponds to the average value of m_0 when $w^* \sim \mathcal{N}(0, I_d)$. Hence, our bound is only meaningful for values of w^* close to the average alignment with Q.*

Remark 7. *By using the heuristic $\tau = \frac{1}{\sqrt{n}}$ we obtain $n = \Omega(\log d^{k/2} d\left(\frac{\|Q\|_F}{\left\|Q^{1/2}\right\|_F^2}\right)^{k/2})$. Similarly to previous work (Damian et al., 2022), there is a gap in the dependence in k between the upper-bound provided by Theorem 1 and the lower bound. Damian et al. (2024a) show this gap can be removed in the isotropic case by using a smoothing technique.*

Remark 8. *Theorem 2 does not cover the cases where $\|Q\|_F \ll \sqrt{\log d}$ or $\left\|Q^{1/2}\right\|_F^2 \gtrsim \|Q\|_F \sqrt{\log d}$, i.e. the cases where the eigenvalues of Q are quickly decreasing. In these settings, estimating the matrix Q and incorporating it into the algorithm could lead to qualitatively better bounds.*

4 Proof Outline of Theorem 1

First, we analyze the population dynamic in Section 4.1. Then, in Section 4.3, we control the impact of the noise.

4.1 Analysis of the Population Dynamics

In this section, we assume direct access to the population gradient, meaning that the weights are updated by:

$$w^{(t+1)} = w^t - \eta \mathbb{E}_x \nabla_{w^{(t)}} L(y^{(t)}, f_{w^{(t)}}(x^{(t)})). \quad (4.1)$$

First, we will show that contrary to spherical SGD, the evolution dynamic of m_t also depends on $w^{(t)}$ and Q, making the analysis more difficult. More specifically, Lemma 1 shows that

$$m_{t+1} \approx \frac{\left\|Q^{1/2}w^{(t)}\right\|}{\left\|Q^{1/2}w^{(t+1)}\right\|} m_t + \eta \frac{\left\|Q^{1/2}w^*\right\|^2}{\left\|Q^{1/2}w^{(t+1)}\right\|} c m_t^{k^*-1}.$$

In Section 4.2, we will show that $w_{\text{sig}}^{(t)} := \langle w^{(t)}, w^* \rangle w^*$, the projection of the weight onto the

signal component, grows as m_t whereas the growing rate of $w_\perp^{(t)} := w^{(t)} - w_{\text{sig}}^{(t)}$, the projection of the weight onto the component orthogonal to the signal, is slower. As a consequence, as long as $m_t \leq \gamma_1$ for some constant $0 < \gamma_1 < 1$, $\left\|Q^{1/2}w^{(t+1)}\right\|$ remains of the same order than $\left\|Q^{1/2}w^{(0)}\right\|$. By further using the approximation

$$\frac{\left\|Q^{1/2}w^{(t)}\right\|}{\left\|Q^{1/2}w^{(t+1)}\right\|} \approx 1$$

that will be formally justified in Section 4.3 (control of E_1) we hence obtain the simplified dynamic

$$m_{t+1} \approx m_t + \tilde{\eta} m_t^{k^*-1} \tag{4.2}$$

where $\tilde{\eta} = c\eta \frac{\left\|Q^{1/2}w^*\right\|^2}{2\left\|Q^{1/2}w^{(0)}\right\|}$.

From equation (4.2), one can show by using Proposition 5.1 in Arous et al. (2020) that, if $m_0 > 0$, one needs $\frac{k^*-2}{\tilde{\eta} m_0^{k^*-2}}$ iterations to obtain a correlation m_t of constant order when $k^* \geq 3$ (the other cases can be treated separately). This result can also be derived heuristically by solving the associated differential equation $f' = \tilde{\eta} f^{k^*-1}$. The population analysis suggests that we should use a large learning rate $\tilde{\eta}$ to accelerate the convergence. However, we will see in Section 4.3 that in order to control the noise, $\tilde{\eta}$ should be small enough.

4.1.1 Evolution dynamic of m_t

By projecting equation (4.1) onto $\frac{Qw^*}{\left\|Q^{1/2}w^*\right\|}$ and dividing by $\left\|Q^{1/2}w^{(t+1)}\right\|$ we obtain:

$$m_{t+1} = \frac{\left\|Q^{1/2}w^{(t)}\right\|}{\left\|Q^{1/2}w^{(t+1)}\right\|} m_t - \frac{\eta}{\left\|Q^{1/2}w^{(t+1)}\right\|} \mathbb{E}_x y \sigma'(\langle \frac{w^{(t)}}{\left\|Q^{1/2}w^{(t)}\right\|}, x\rangle) \langle x, \frac{Qw^*}{\left\|Q^{1/2}w^*\right\|}\rangle. \tag{4.3}$$

Unlike spherical SGD optimization, which constrains the norm of the weights at each iteration, the evolution of m_t here depends crucially on the evolution of $\left\|Q^{1/2}w^{(t)}\right\|$. For instance, if $\left\|Q^{1/2}w^{(t)}\right\|$ becomes too large, then the information provided by the gradient may be lost. Furthermore, in contrast to the isotropic setting, it is not straightforward how to express the expectation of the gradient as a function of m_t. Lemma 1 addresses this point.

Lemma 1. *Assume that Assumption A4 is satisfied. Then, there exists a constant $\gamma_1 > 0$ such that as long as the sequence $(m_t)_t$ is bounded from above by γ_1, it satisfies the following relation*

$$m_{t+1} \geq \frac{\left\|Q^{1/2}w^{(t)}\right\|}{\left\|Q^{1/2}w^{(t+1)}\right\|} m_t + \eta \frac{\left\|Q^{1/2}w^*\right\|^2}{\left\|Q^{1/2}w^{(t+1)}\right\|} c m_t^{k^*-1}. \tag{4.4}$$

Proof. By writing $Qw^* = \lambda w^* + \lambda' w_\perp^*$ where $w_\perp^*$ is a unit vector orthogonal to w^*, $\lambda = \langle Qw^*, w^*\rangle = \left\|Q^{1/2}w^*\right\|^2$ and $\lambda' = \sqrt{\|Qw^*\|^2 - \left\|Q^{1/2}w^*\right\|^4}$, we can decompose the population gradient as

$$\mathbb{E}_x y \sigma'\left(\left\langle \frac{w^{(t)}}{\|Q^{1/2}w^{(t)}\|}, x\right\rangle\right)\left\langle x, \frac{Qw^*}{\|Q^{1/2}w^*\|}\right\rangle = G_1 + G_2$$

where

$$G_1 = \lambda \mathbb{E}_x y \sigma'\left(\left\langle \frac{w^{(t)}}{\|Q^{1/2}w^{(t)}\|}, x\right\rangle\right)\left\langle x, \frac{w^*}{\|Q^{1/2}w^*\|}\right\rangle$$
$$G_2 = \lambda' \mathbb{E}_x y \sigma'\left(\left\langle \frac{w^{(t)}}{\|Q^{1/2}w^{(t)}\|}, x\right\rangle\right)\left\langle x, \frac{w_\perp^*}{\|Q^{1/2}w^*\|}\right\rangle.$$

Control of G_1. To simplify the notations, let us write $z^* = \langle x, \frac{w^*}{\left\|Q^{1/2}w^*\right\|}\rangle$ and $z_t = \langle x, \frac{w^{(t)}}{\left\|Q^{1/2}w^{(t)}\right\|}\rangle$. Notice that $z^*, z_t \sim \mathcal{N}(0,1)$. Recall that $y = f(z^*)$. So we need to evaluate $\mathbb{E}_{z^*,z_t} z^* f(z^*)\sigma'(z_t)$. By using the Hermite decomposition of these functions and Proposition 2, we can see that this expectation depends essentially on the information exponent of $x \to xf(x)$ and the correlation between z^* and z_t. But by Proposition 2 with $n = 1$ (see Section A in the appendix) we have $\mathbb{E}z^* z_t = m_t$.

By using Hermite decomposition, Proposition 2, Lemma 3 , and Assumption A4, we obtain

$$G_1 = \lambda \sum_{l \geq k^*-1} c_l b_l m_t^l \leq -\lambda c m_t^{k^*-1}. \tag{4.5}$$

Control of G_2. Denote $z_\perp^* = \langle x, \frac{w_\perp^*}{\left\|Q^{1/2}w_\perp^*\right\|}\rangle$ and notice that $\mathbb{E}z_t z_\perp^* = q_t$ where

$$q_t = \langle \frac{w^{(t)}}{\left\|Q^{1/2}w^{(t)}\right\|}, \frac{w_\perp^*}{\left\|Q^{1/2}w_\perp^*\right\|}\rangle_Q.$$

We can write $z_t = m_t z^* + \sqrt{1-m_t^2} z_t^\perp$, where $z_t^\perp \sim \mathcal{N}(0,1)$ is independent from z^*. Similarly, we can decompose $z_\perp^* = q_t z_t^\perp + \sqrt{1-q_t^2}\xi$ where $\xi \sim \mathcal{N}(0,1)$ is independent from z^* and $z_t^\perp$.

Let $p_t = \frac{m_t^2}{2(1-m_t^2)}$. By combining Lemma 4 and 5 (see Section B), we obtain

$$\mathbb{E}f(z^*)\sigma'(m_t z^* + \sqrt{1-m_t^2} z_t^\perp) z_\perp^* = \frac{q_t}{\sqrt{2\pi(2p_t+1)}} \mathbb{E}_{z\sim\mathcal{N}(0,1)} f(\frac{z}{\sqrt{2p_t+1}}).$$

We can evaluate $\mathbb{E}f(\frac{z}{\sqrt{2p_t+1}})$ by applying the multiplicative property of Hermite polynomials recalled in Lemma 6 in the appendix with $\gamma = \sqrt{2p_t+1}^{-1}$.

Notice that for odd n, $\langle H_n(\gamma x), H_0(x)\rangle = 0$ and for $n = 2m$, we have

$$\langle H_n(\gamma x), H_0(x)\rangle = (\gamma^2-1)^m \frac{(2m)!}{m!} 2^{-m}$$

Since by definition $f(x) = \sum_{k\geq k^*} \frac{a_k}{\sqrt{k!}} H_k(x)$ we get

$$\begin{aligned}
&\left|\langle f(\frac{x}{\sqrt{2c_t+1}}), H_0(x)\rangle\right| \\
&= \left|\sum_{k\geq k^*/2} \frac{a_{2k}}{\sqrt{(2k)!}} (\frac{1}{2p_t+1}-1)^k \frac{(2k)!}{k!} 2^{-k}\right| \\
&\lesssim \sqrt{\left|\sum a_{2k}^2 \frac{(2k)!}{4^k (k!)^2}\right|} \sqrt{\left|\sum p_t^{2k}\right|} \\
&\qquad\qquad \text{(by Cauchy-Schwartz)} \\
&\lesssim p_t^{k^*/2} \lesssim m_t^{k^*}.
\end{aligned}$$

Here we used the fact that by Stirling formula $\frac{(2k)!}{4^k(k!)^2} \sim 1$ so the sequence is bounded, and the fact that by definition of f, $\sum a_k^2 = O(1)$. This shows that G_2 is of order at most $\lambda' m_t^{k^*}$ and is negligible compared to G_1 as long as m_t is small enough, since $\lambda' \leq 1$. □

The lower bound obtained in Lemma 1 is only useful when $m_t > 0$. This is ensured by Assumption A3.

In the next two sections, we are going to control the growth of $\left\|w^{(t)}\right\|$. We are going to show that as long as $m_t \leq \gamma_1$ for some constant $\gamma_1 > 0$, the weights remain bounded and do not evolve quickly, i.e. $\left\|Q^{1/2}w^{(t)}\right\| \approx \left\|Q^{1/2}w^{(0)}\right\|$ so that equation (4.4) is equivalent to

$$m_{t+1} \geq m_t + \tilde{\eta} m_t^{k^*-1} \tag{4.6}$$

where $\tilde{\eta} = c\eta \left\|Q^{1/2}w^*\right\|^2 / 2 \left\|Q^{1/2}w^{(0)}\right\|$. This last relation is similar to the one derived in the isotropic case by Arous et al. (2020).

4.2 Control of the growth of $\left\|Q^{1/2}w^{(t)}\right\|$

In this section, we justify the approximation $\left\|Q^{1/2}w^{(t)}\right\| \approx \left\|Q^{1/2}w^{(0)}\right\|$ for $t \leq T$. By recursion, we obtain

$$Q^{1/2}w^{(t)} = Q^{1/2}w^{(0)} + \eta \sum_{l\leq t-1} \mathbb{E}y\sigma'(\langle w^{(l)}, x\rangle) Q^{1/2}x. \tag{4.7}$$

Let $w_\perp^{(l)}$ be the projection of $w^{(l)}$ onto the space orthogonal to w^*. There are only two directions, $Q^{3/2}w^*$ and $Q^{3/2}w_\perp^{(l)}$, in which the projections of the vector $\mathbb{E}y\sigma'(\langle w^{(l)}, x\rangle)Q^{1/2}x$ are non zero. Indeed, if v is orthogonal to $Q^{3/2}w^*$ then

$$\mathbb{E}_x\langle Q^{1/2}x, v\rangle\langle x, w^*\rangle = \langle Q^{1/2}v, w^*\rangle_Q = 0$$

and similarly for $w_\perp^{(l)}$. As a consequence $\langle Q^{1/2}x, v\rangle$ is independent of z and z_t, and the resulting expectation is zero. To evaluate $\left\|Q^{1/2}w^{(t)}\right\|$ it is sufficient to evaluate the projection of the expectation in the directions identified previously. By a similar analysis as in Lemma 1 we obtain

$$\mathbb{E}f(z^*)\sigma'(z_t)\langle x, \frac{Q^2w^*}{\left\|Q^{3/2}w^*\right\|}\rangle \approx \lambda^2 c m_t^{k^*-1}.$$

We can also show that

$$\mathbb{E}f(z^*)\sigma'(z_t)\langle x, \frac{Q^2w_\perp^{(l)}}{\left\|Q^{3/2}w_\perp^{(l)}\right\|}\rangle \leq C m_t^{k^*}.$$

The details of the calculations can be found in Section D.1. This shows that as long as $\lambda^2 c\eta \sum_{t\leq T} m_t^{k^*-1}$ and $C\eta \sum_{t\leq T} m_t^{k^*}$ remains smaller to $0.5\left\|Q^{1/2}w^{(0)}\right\|$, we have

$$0.5\left\|Q^{1/2}w^{(0)}\right\| \leq \left\|Q^{1/2}w^{(t)}\right\| \leq 1.5\left\|Q^{1/2}w^{(0)}\right\|.$$

The previous conditions are satisfied by choice of the initialization scale, η and T: the contribution in the direction $Q^{3/2}w^*$ grows slower than m_{t+1}, and similarly for the contribution in the other direction.

4.3 Analysis of the noisy dynamic

In this section, we will describe how the noise can be controlled so that after T iterations, m_t is well predicted by the population dynamic analysis performed in Section 4.1.

We decompose the gradient into two components: the population version and the stochastic noise V_t

$$\nabla_{w^{(t)}} L = \mathbb{E}(\nabla_{w^{(t)}} L) + V_t.$$

4.3.1 Control of $\left\|Q^{1/2}w^{(t)}\right\|$

As shown in the analysis of the population dynamic, it is critical to control $\left\|Q^{1/2}w^{(t)}\right\|$. In the noisy setting, we obtain the following counterpart of equation (4.7)

$$\begin{aligned}
Q^{1/2}w^{(t)} = Q^{1/2}w^{(0)} &+ \eta \sum_{l\leq t-1} \mathbb{E}y\sigma'(\langle w^{(l)}, x\rangle) Q^{1/2}x \\
&+ \eta Q^{1/2} \sum_{t\leq T} V_t.
\end{aligned} \tag{4.8}$$

By using Doob's maximal inequality (see Lemma 10 in appendix), we can show that $\eta \left\|\sum_{t\le T} V_t\right\| = o(1)$. Hence, the result follows the population dynamic analysis.

4.3.2 Evolution of m_t

Instead of controlling directly m_t, it is more convenient to study the dynamic of the related quantity $\tilde{m}_t := \langle w^{(t)}, \frac{w^*}{\|Q^{1/2}w^*\|}\rangle_Q$ that avoids dividing by the random quantity $\|Q^{1/2}w^t\|$. Since for $t \le T$ we have $0.5\|Q^{1/2}w^0\| \le \|Q^{1/2}w^t\| \le 1.5\|Q^{1/2}w^0\|$, one can easily relate $\tilde{m}_t$ to m_t. By definition, we have

$$\tilde{m}_{t+1} = \tilde{m}_t - \eta \mathbb{E} y\sigma'(\langle w^{(t)}, w^*\rangle)\langle x, \frac{Qw^*}{\|Q^{1/2}w^*\|}\rangle + \eta\langle V_t, \frac{Qw^*}{\|Q^{1/2}w^*\|}\rangle.$$

The expectation term corresponds to G_1 analyzed in Lemma 1 and the stochastic term forms a martingale that Doob's Lemma can control, see Lemma 9 in the appendix. Hence, we have obtained a recursion of the form

$$\tilde{m}_{T+1} \ge \eta' \sum_{t\le T} \tilde{m}_t^{k^*-1} + \eta H_T$$

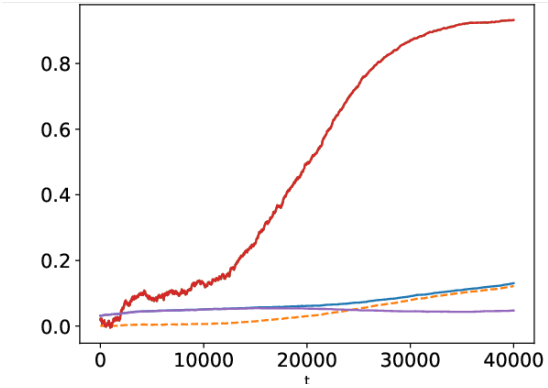
where $H_T = \sum_{t\le T}\langle V_t, \frac{Qw^*}{\|Q^{1/2}w^*\|}\rangle$ and $\eta' = c'\left\|Q^{1/2}w^{(0)}\right\|^{k^*-1}\lambda\eta$. We used the fact that $m_t \ge \frac{2}{3}\left\|Q^{1/2}w^{(0)}\right\|\tilde{m}_t$ for $t \le T$.

To conclude the proof of Theorem 1, it remains to understand how many iterations T are necessary so that m_t becomes of constant order. Sequence satisfying $c_{t+1} \ge c_t + \eta m_t^l$ have been analyzed formally in Arous et al. (2020) based on Bihari–LaSalle inequality. Here, we present a heuristic way to recover the result. The continuous analogous of the relation $c_{t+1} \ge c_t + \eta m_t^l$ is $f'(t) = \eta f^l(t)$. By integrating between 0 and T, we obtain $\frac{1}{f^{l-1}(0)} - \frac{1}{f^{l-1}(T)} = \eta T$ for $l \ge 2$. Since $f(T)$ should be of constant order, it is negligible compared to $\frac{1}{f^{l-1}(0)} = m_0^{-l+1}$. Given the choice $\eta = \frac{\epsilon}{\sqrt{T}\|Q^{1/2}\|_F}$ (necessary to control the stochastic error), solving the equation leads to $T = \left\|Q^{1/2}\right\|_F^2 \epsilon^{-2} m_0^{-2(l-1)}$.

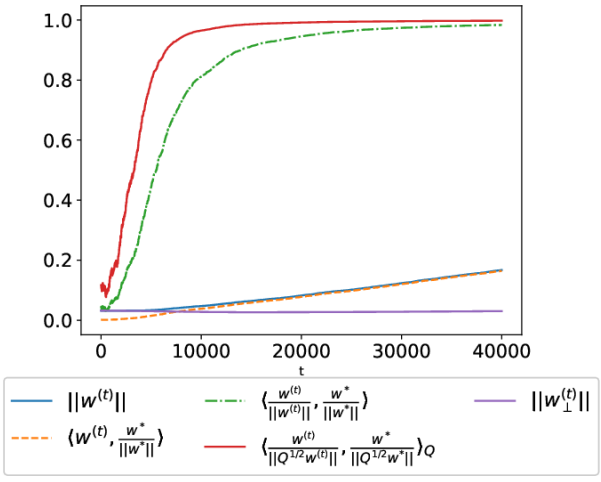
5 Numerical Experiments

In this section, we illustrate our theoretical results through numerical simulations [1]. The implementation details are provided in Section F.

Anisotropy can help. We consider the following setting: $y = H_2(\langle x, \frac{w^*}{\|Q^{1/2}w^*\|}\rangle)$ where $H_2(x) = x^2 - 1$, $w^* \in \mathbb{S}^{d-1}$, and $x \sim \mathcal{N}(0, Q)$ with a covariance matrix of the form $Q = \frac{I_d + \kappa w^*(w^*)^\top}{1+\kappa}$, parametrized by $\kappa > 0$. The information exponent of H_2 is 2. We set the dimension $d = 1000$, the sample size $T = 40000$, and the learning rate 0.00002. The learning dynamics when $\kappa = 0$ (isotropic case) is plotted in Figure 1a, while those for $\kappa = 6$ in Figure 1b. The improved alignment at initialization when $\kappa = 6$ significantly accelerates learning.



(a) Isotropic setting $\kappa = 0$



(b) Anisotropic setting $\kappa = 6$

Figure 1: Comparison of learning dynamics in isotropic and anisotropic settings.

Comparison between SGD and Spherical SGD. We compare the performance of vanilla SGD (`SGD`) with spherical SGD (`SpheSGD`). We used an oracle knowledge of the covariance matrix to implement the algorithm. Figure 2 shows that the two

[1] The code is available at `https://glmbraun.github.io/AniSIM`

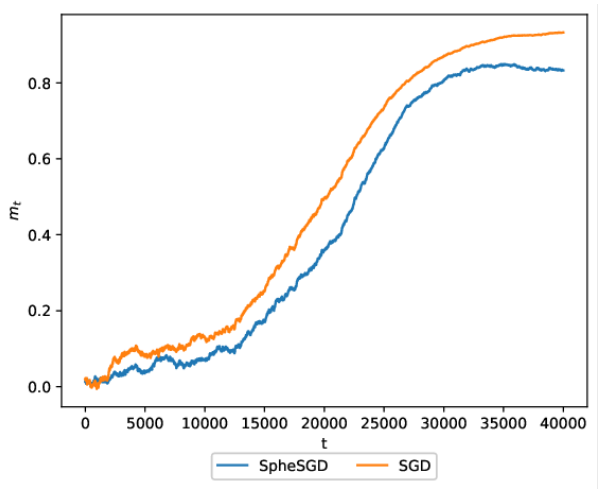


Figure 2: Comparison of learning dynamics between vanilla SGD and spherical SGD.

`SGD` algorithms behave similarly.

Adaptive Learning Rate. As shown in Section F, progressively increasing the learning rate is beneficial to accelerate the learning dynamic. This is consistent with the theoretical insight that the learning rate should be small enough to control noise at each iteration; however, as the signal increases during training, higher noise can be tolerated.

Batch Reuse. We demonstrate in Appendix F that reusing the same batch can significantly reduce the required sample complexity.

6 Conclusion

We analyzed the problem of learning a SIM from anisotropic Gaussian inputs using vanilla SGD. Unlike previous approaches relying on spherical SGD, which require prior knowledge of the covariance structure, our analysis shows that vanilla SGD can naturally adapt to the anisotropic geometry without estimating the covariance matrix. Our theoretical contributions include an upper bound on the sample complexity and a CSQ lower bound, depending on the covariance matrix structure instead of the input data dimension. Numerical simulations validated these theoretical findings and demonstrated the practical effectiveness of vanilla SGD.

This work opens up several avenues for future research. First, our analysis has focused on the training dynamics of a single neuron, but extending these insights to deeper or wider neural networks would be valuable for a broader understanding of how vanilla SGD performs in more complex architectures. Additionally, achieving the generative exponent in sample complexity by reusing data remains an open question.

Bibliography

E. Abbe, E. B. Adserà, and T. Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195, pages 2552–2623. PMLR, 2023.

L. Arnaboldi, Y. Dandi, F. Krzakala, L. Pesce, and L. Stephan. Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions, 2024.

G. B. Arous, R. Gheissari, and A. Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *J. Mach. Learn. Res.*, 22:106:1–106:51, 2020.

J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, and D. Wu. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

K. Balasubramanian, H.-G. Müller, and B. K. Sriperumbudur. Functional linear and single-index models: A unified approach via gaussian stein identity, 2024.

A. Bietti, J. Bruna, C. Sanford, and M. J. Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems*, 2022.

A. Bietti, J. Bruna, and L. Pillaud-Vivien. On learning gaussian multi-index models with gradient flow, 2023.

X. Chen, K. Balasubramanian, P. Ghosal, and B. Agrawalla. From stability to chaos: Analyzing gradient descent dynamics in quadratic regression. *Trans. Mach. Learn. Res.*, 2024, 2024.

A. Damian, J. Lee, and M. Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 5413–5452, 2022.

A. Damian, E. Nichani, R. Ge, and J. D. Lee. Smoothing the landscape boosts the signal for sgd optimal sample complexity for learning single index models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2024a.

A. Damian, L. Pillaud-Vivien, J. D. Lee, and J. Bruna. Computational-statistical gaps in gaussian single-index models, 2024b.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems*, volume 33, pages 14820–14830, 2020.

M. Glasgow. SGD finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the XOR problem. In *The Twelfth International Conference on Learning Representations*, 2024.

W. Hardle, P. Hall, and H. Ichimura. Optimal Smoothing in Single-Index Models. *The Annals of Statistics*, 21(1):157 – 178, 1993.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

C.-R. Jiang and J.-L. Wang. Functional single index models for longitudinal data. *The Annals of Statistics*, 39(1):362 – 388, 2011.

J. D. Lee, K. Oko, T. Suzuki, and D. Wu. Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

S. Ma and X. He. Inference for single-index quantile regression models with profile optimization. *The Annals of Statistics*, 44(3):1234 – 1268, 2016.

S. Mei, A. Montanari, and P.-M. 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.

A. Mousavi-Hosseini, S. Park, M. Girotti, I. Mitliagkas, and M. A. Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2023a.

A. Mousavi-Hosseini, D. Wu, T. Suzuki, and M. A. Erdogdu. Gradient-based feature learning under structured data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

A. Mousavi-Hosseini, D. Wu, and M. A. Erdogdu. Learning multi-index models with neural networks via mean-field langevin dynamics, 2024. URL `https://arxiv.org/abs/2408.07254`.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, 135:370–384, 1972.

D. Nualart and E. Nualart. *Introduction to Malliavin Calculus*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2018.

K. Oko, Y. Song, T. Suzuki, and D. Wu. Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations. In S. Agrawal and A. Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4009–4081. PMLR, 30 Jun–03 Jul 2024.

B. Szörényi. Characterizing statistical query learning: simplified notions and proofs. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory*, ALT'09, page 186–200, 2009.

M. Zhou and R. Ge. How does gradient descent learn features – a local analysis for regularized two-layer neural networks, 2024. URL `https://arxiv.org/abs/2406.01766`.

L. Zhu, C. Liu, A. Radhakrishnan, and M. Belkin. Catapults in sgd: spikes in the training loss and their impact on generalization through feature learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR, 2024a.

L. Zhu, C. Liu, A. Radhakrishnan, and M. Belkin. Quadratic models for understanding catapult dynamics of neural networks. In *The Twelfth International Conference on Learning Representations*, 2024b.

A. Zweig, L. Pillaud-Vivien, and J. Bruna. On single-index models beyond gaussian data. In *Advances in Neural Information Processing Systems*, volume 36, pages 10210–10222, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material

We provide background on Hermite polynomials, including key properties that are crucial to our analysis. Technical lemmas involving the evaluation of Gaussian integrals are collected in Section B. In Section C, we introduce the CSQ framework and prove Theorem 2. In Section E, we complete the proof of Theorem 1. Finally, in Section F, we present additional numerical experiments along with implementation details.

A Hermite polynomials

Consider the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \gamma)$ where γ denotes the standard Gaussian measure on $\mathbb{R}$ and let $\mathcal{H} = L^2(\mathbb{R}, \gamma)$ the associated Hilbert space of squared integrable function with respect to γ.

We will define Hermite polynomial following the approach of Nualart and Nualart (2018). Toward this end, let us define two differential operators. For any $f \in C^1(\mathbb{R})$ we define the *derivative operator* $Df(x) = f'(x)$ and the *divergence operator* $\delta f(x) = xf(x) - f'(x)$. The following lemma show that the operators D and δ are adjoint.

Lemma 2. *Denote by $C_p^1(\mathbb{R})$ the space of continuously differentiable functions that grows at most polynomially, i.e., there exists some integer $N \geq 1$ and constant $C > 0$ such that $|f'(x)| \leq C(1 + |x|^N)$. For any $f, g \in C_p^1(\mathbb{R})$, we have*

$$\langle Df, g\rangle_{\mathcal{H}} = \langle f, \delta g\rangle_{\mathcal{H}}.$$

Proof. The result is derived directly by using integration by parts. □

The Hermite polynomials are defined as follows:

$$H_0(x) = 1,$$
$$H_n(x) = \delta^n 1.$$

Proposition 1. *The sequence of normalized Hermite polyniomials $(\frac{1}{\sqrt{n!}} H_n)_{n\geq 0}$ is an orthonormal basis of $\mathcal{H}$.*

Another particularly useful property of the Hermite polynomial that we will rely on heavily is the simple characterization of the correlation between two Hermite polynomials with correlated Gaussian inputs.

Proposition 2. *Let $x \sim \mathcal{N}(0, I_d)$. For any $w, w' \in \mathbb{S}^{d-1}(\mathbb{R})$ we have*

$$\mathbb{E}_x H_n\left(\langle x, w\rangle\right) H_{n'}\left(\langle x, w'\rangle\right) = \mathbf{1}_{\{n=n'\}} n! \langle w, w'\rangle^n.$$

This result can be extended straightforwardly to anisotropic inputs.

Corollary 1. *Let $x \sim \mathcal{N}(0, Q)$ for some general covariance matrix Q. For any $w, w' \in \mathbb{R}^d$ we have*

$$\mathbb{E}_x H_n\left(\langle x, \frac{w}{\|Q^{1/2}w\|}\rangle\right) H_{n'}\left(\langle x, \frac{w'}{\|Q^{1/2}w'\|}\rangle\right) = \mathbf{1}_{\{n=n'\}} n! \left(\langle \frac{w}{\|Q^{1/2}w\|}, \frac{w'}{\|Q^{1/2}w'\|}\rangle_Q\right)^n.$$

The following lemma relates the information exponent of f to $x \to xf(x)$, the function naturally appearing in the population gradient.

Lemma 3. *Assume that $f \in \mathcal{H}$ has information exponent $k \geq 1$. Then, the function $x \to xf(x) \in \mathcal{H}$ has information exponent $k - 1$.*

Proof. Assume that $k \geq 2$. Recall that the Hermite's polynomials satisfy $H_n'(x) = nH_{n-1}$ (e.g. it derives easily as an application of Lemma 2) and $H_{n+1}(x) = xH_n(x) - H_n'(x)$ (by definition $H_{n+1} = \delta^{n+1}\mathbf{1} = \delta H_n$).

By consequence, for all $n \in \mathbb{N}^*$, we have

$$\begin{aligned}\langle xf(x), H_n(x)\rangle_{\mathcal{H}} &= \langle f(x), xH_n(x)\rangle_{\mathcal{H}} \\ &= \langle f(x), H_{n+1}(x)\rangle_{\mathcal{H}} + \langle f(x), H_n'(x)\rangle_{\mathcal{H}} \\ &= \langle f(x), H_{n+1}(x)\rangle_{\mathcal{H}} + n\langle f(x), H_{n-1}(x)\rangle_{\mathcal{H}}.\end{aligned}$$

If $n < k-1$ all the terms are null. But for $n = k-1$, $\langle f(x), H_{n+1}(x)\rangle_{\mathcal{H}} \neq 0$, while $n\langle f(x), H_{n-1}(x)\rangle_{\mathcal{H}} = 0$ by definition of k. The case where $k = 1$ can be treated similarly. □

B Technical lemmas

Recall that σ' is the sign function, formally defined as

$$\sigma'(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

Lemma 4. *Let $p \in [0,1]$ and X, Y be two independent standard Gaussian r.v. We have*

$$\mathbb{E}_Y \sigma'\left(pX + \sqrt{1-p^2}Y\right) Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{p^2}{2(1-p^2)}X^2}.$$

Proof. By symmetry, we have

$$\begin{aligned}2\mathbb{E}_Y \sigma'(pX + \sqrt{1-p^2}Y)Y &= \mathbb{E}_Y\left(\sigma'(pX + \sqrt{1-p^2}Y) - \sigma'(pX - \sqrt{1-p^2}Y)\right) Y \\ &= \mathbb{E}_Y \mathbf{1}_{\{|Y| \geq \frac{p}{\sqrt{1-p^2}}|X|\}}|Y| \\ &= \frac{2}{\sqrt{2\pi}} e^{-\frac{p^2}{2(1-p^2)}X^2}.\end{aligned}$$

□

Lemma 5. *Let $c > 0$, $X \sim \mathcal{N}(0,1)$ and $f \in \mathcal{H}$. We have*

$$\mathbb{E}_X f(X) e^{-cX^2} = \frac{1}{\sqrt{2c+1}} \mathbb{E}_X f\left(\frac{X}{\sqrt{2c+1}}\right).$$

Proof. Use the change of variable $u = \sqrt{1+2c}x$. □

Lemma 6. *For every $\gamma > 0$, $n \in \mathbb{N}^*$ we have*

$$H_n(\gamma x) = \sum_{k=0}^{\frac{n}{2}} \gamma^{n-2k}(\gamma^2 - 1)^k \binom{n}{2k} \frac{(2k)!}{k!} 2^{-k} H_{n-2k}(x).$$

Proof. This idendity is classical, but since we didn't find a proper reference, we provide a simple proof. Recall that the Hermite polynomials satisfy the following identity for all $t, x \in \mathbb{R}$ (see Nualart and Nualart (2018))

$$e^{-\frac{t^2}{2} + tx} = \sum_{n \geq 0} H_n(x) \frac{t^n}{n!}.$$

So we have

$$\sum_{n \geq 0} H_n(\gamma x) \frac{t^n}{n!} = e^{-\frac{t^2}{2} + \gamma t x} = e^{-\frac{\gamma^2 t^2}{2} + \gamma t x} e^{\frac{(\gamma^2 - 1)t^2}{2}}.$$

By using the series development of the previous exponential functions we get

$$\sum_{n\geq 0} H_n(\gamma x)\frac{t^n}{n!} = \sum_{j\geq 0} H_j(x)\frac{(\gamma t)^j}{j!}\sum_{k\geq 0}\frac{(\gamma^2-1)^k t^{2k}}{2^k k!}$$
$$= \sum_j \sum_k \frac{t^{j+2k}}{(j+2k)!} H_j(x)\gamma^j(\gamma^2-1)^k 2^{-k}\frac{(j+2k)!}{j!k!}.$$

Let $n = j + 2k$. By identifying the coefficient associated in the serie expansion, we obtain the stated formula. □

C CSQ lower-bound

The Correlational Statistic Query framework is a restricted computational model where we access knowledge of the data distribution $(x, y) \sim P$ by addressing query $\phi : \mathbb{R}^d \to \mathbb{R}$ to an oracle that returns $\mathbb{E}_{(x,y)\sim P}(y\phi(x))+\epsilon$ where ϵ is some noise term bounded by τ, the tolerance parameter. SGD is an algorithm belonging to this framework (note, however, that in the CSQ framework, the noise can be adversarial).

The classical way to obtain a lower bound (Damian et al., 2022) is to construct a large class of function $\mathcal{F}$ with small correlations. The following lemma provides a lower bound.

Lemma 7 (Szörényi (2009), Damian et al. (2022))**.** *Let $\mathcal{F}$ be a class of function and $\mathcal{D}$ be a data distribution such that*

$$\mathbb{E}_{x\sim\mathcal{D}} f^2(x), \quad |\mathbb{E}_{x\sim\mathcal{D}} f(x)g(x)| \leq \epsilon, \quad \forall f \neq g \in \mathcal{F}.$$

Then any CSQ algorithm requires at least $\frac{|\mathcal{F}|(\tau^2-\epsilon)}{2}$ queries of tolerance τ to output a function in $\mathcal{F}$ with $L^2(\mathcal{D})$ loss at most $2 - 2\epsilon$.

We then usually use the heuristic $\tau = \frac{1}{\sqrt{n}}$ to derive a lower bound on the sample complexity.

Lemma 8. *Assume that the covariance matrix Q satisfies $\|Q\| = 1$, $\left\|Q^{1/2}\right\|_F^2 \gtrsim \|Q\|_F \sqrt{\log d}$, and $\|Q\|_F \geq C\sqrt{\log d}$ for some sufficiently large constant $C > 0$. Let*

$$m = C\frac{\|Q\|_F}{\left\|Q^{1/2}\right\|_F^2}.$$

Then, for $\epsilon = m\sqrt{\log(qm^{k/2})}$, where $q = O(d^c)$ (for some constant $c > 0$) is the number of queries, there exists an absolute constant C_1 and a set $\mathcal{E}$ with cardinality at least $0.5e^{C_1\log(qm^{k/2})}$ such that $\forall w \neq v \in \mathcal{E}$ we have

$$\left|\left\langle \frac{w}{\left\|Q^{1/2}w\right\|}, \frac{v}{\left\|Q^{1/2}v\right\|}\right\rangle_Q\right| \leq \epsilon.$$

Proof. This is an adaptation of Lemma 3 in Damian et al. (2022) to the anisotropic case. Let $w_1, \ldots, w_p$ be i.i.d. Gaussian random variables $w_i \sim \mathcal{N}(0, I_d)$. By the Hanson-Wright inequality, for all $i \in [p]$, we have:

$$\mathbb{P}\left(\left|w_i^\top Q w_i - \left\|Q^{1/2}\right\|_F^2\right| \geq t\right) \leq e^{-c\min\left(\frac{t^2}{\|Q\|_F^2}, \frac{t}{\|Q\|}\right)}.$$

Since $w_i^\top Q w_i = \left\|Q^{1/2}w_i\right\|^2$, choosing $t = C\|Q\|_F\sqrt{\log(qm^{k/2})}$, we get that this probability is bounded by $e^{-c'\log(qm^{k/2})} \cup e^{-C\log d}$, where $c' = C/c$. This holds because, by assumption, $\|Q\|_F \gtrsim \sqrt{\log(d)}$ and $\log(qm^{k/2}) \gtrsim \log d$.

Similarly, for every $i \neq j \in [p]$, we have:

$$\mathbb{P}\left(\left|w_i^\top Q w_j\right| \geq t\right) \leq e^{-c\min\left(\frac{t^2}{\|Q\|_F^2}, \frac{t}{\|Q\|}\right)} \leq e^{-c'\log(qm^{k/2})}.$$

Using a union bound over all pairs $i, j \in [p]$, we obtain that, with probability at least $1 - 2p^2 e^{-c' \log(qm^{k/2})}$:

$$\forall i \in [p], \quad c_1 \left\| Q^{1/2} \right\|_F^2 \leq \left\| Q^{1/2} w_i \right\|^2,$$

and for all $i \neq j \in [p]$:

$$|\langle w_i, w_j \rangle_Q| \leq C \|Q\|_F \sqrt{\log(qm^{k/2})}.$$

This implies that for $i \neq j$:

$$\left| \left\langle \frac{w_i}{\left\| Q^{1/2} w_i \right\|}, \frac{w_j}{\left\| Q^{1/2} w_j \right\|} \right\rangle_Q \right| \leq \epsilon,$$

where $\epsilon = C \frac{\|Q\|_F}{\left\| Q^{1/2} \right\|_F^2} \sqrt{\log(qm^{k/2})}$, completing the proof. □

Theorem 3. *Assume that the assumptions of Lemma 8 are satisfied. For any integer $k \geq 1$, there exists a class $\mathcal{F}_k$ of polynomial functions of degree k such that any CSQ algorithm using a polynomial number of queries q requires a tolerance τ of order at most*

$$\tau^2 \leq \epsilon^{k/2}.$$

Remark 9. *By using the heuristic $\tau^2 = \frac{1}{\sqrt{n}}$ we obtain $n = \Omega(\log d^{k/2} d \left(\frac{\|Q\|_F}{\left\| Q^{1/2} \right\|_F^2} \right)^{k/2})$. The term $\frac{\|Q\|_F}{\left\| Q^{1/2} \right\|_F^2}$ corresponds to the average value of m_0 when $w^* \sim \mathcal{N}(0, I_d)$. Similar to previous work (Damian et al., 2022), there is a gap in the dependence in k between the upper-bound provided by Theorem 1 and the lower bound. Damian et al. (2024a) show this gap can be removed in the isotropic case by using a smoothing technique.*

Remark 10. *Lemma 8 doesn't cover the cases where $\|Q\|_F \ll \sqrt{\log d}$ or $\left\| Q^{1/2} \right\|_F^2 \gtrsim \|Q\|_F \sqrt{\log d}$, i.e. the cases where the eigenvalues of Q are quickly decreasing. In these settings, estimating the matrix Q and incorporating it into the algorithm could lead to qualitatively better bounds.*

Proof. Recall that $\mathcal{E}$ is the set constructed in Lemma 8 and consider the class of functions

$$\mathcal{F}_k = \left\{ x \to \frac{H_k\left(\left\langle x, \frac{w}{\left\| Q^{1/2} w \right\|} \right\rangle \right)}{\sqrt{k!}} \;\middle|\; w \in \mathcal{E} \right\}.$$

For any $w \neq w' \in \mathcal{E}$ we have

$$\left| \left\langle \frac{H_k\left(\langle x, \frac{w}{\left\| Q^{1/2} w \right\|} \rangle \right)}{\sqrt{k!}}, \frac{H_k\left(\langle x, \frac{w'}{\left\| Q^{1/2} w' \right\|} \rangle \right)}{\sqrt{k!}} \right\rangle_Q \right| \leq \epsilon^k.$$

We obtain the result from Lemma 7 and elementary algebra. □

D Additional proofs

D.1 Proof of the claims in Section 4.2

Recall the decomposition $Qw^* = \lambda w^* + \lambda' w_\perp^*$ of Lemma 1. We have $Q^2 w^* = Q(\lambda w^* + \lambda' w_\perp^*) = \lambda^2 w^* + \lambda\lambda' w_\perp^* + \lambda' Q w_\perp^*$. Since $\langle Q w_\perp^*, w^* \rangle = \langle w_\perp^*, Q w^* \rangle = \lambda'$, we obtain

$$Q^2 w^* = (\lambda^2 + (\lambda')^2) w^* + \lambda\lambda' w_\perp^* + \lambda' \lambda'' \tilde{w}_\perp^*$$

where $\lambda'' \tilde{w}_\perp^* = Qw_\perp^* - \lambda' w^*$. As a consequence, we can decompose

$$\mathbb{E} f(z^*)\sigma'(z_t)\langle x, \frac{Q^2 w^*}{\left\|Q^{3/2}w^*\right\|}\rangle = G_1' + G_2' + G_3'$$

where

$$G_1' = (\lambda^2 + (\lambda')^2)\frac{\left\|Q^{1/2}w^*\right\|}{\left\|Q^{3/2}w^*\right\|}\mathbb{E} z^* f(z^*)\sigma'(z_t)$$
$$G_2' = \lambda\lambda'\frac{\left\|Q^{1/2}w_\perp^*\right\|}{\left\|Q^{3/2}w^*\right\|}\mathbb{E} f(z^*)\sigma'(z_t) z_\perp^*$$
$$G_3' = \lambda'\lambda''\frac{\left\|Q^{1/2}\tilde{w}_\perp^*\right\|}{\left\|Q^{3/2}w^*\right\|}\mathbb{E} f(z^*)\sigma'(z_t)\tilde{z}_\perp^*$$

where $\tilde{z}_\perp^* = \langle x, \frac{Q^{1/2}\tilde{w}_\perp^*}{\left\|Q^{1/2}\tilde{w}_\perp^*\right\|}\rangle$. Notice that $G_1' = (\lambda^2 + (\lambda')^2)\frac{\left\|Q^{1/2}w^*\right\|}{\left\|Q^{3/2}w^*\right\|}G_1 \approx -c\lambda(\lambda^2 + (\lambda')^2)\frac{\left\|Q^{1/2}w_\perp^*\right\|}{\left\|Q^{3/2}w^*\right\|}m_t^{k^*-1}$ by the proof of Lemma 1. The terms G_2' and G_3' can also be analyzed as in Lemma 1.

E Proof of Theorem 1

In this section, we complete the proof of Theorem 1 sketched in the main text.

E.1 Initialization

Here, we justify the claim that m_0 is of order $\|Qw^*\|\,\|Q^{1/2}w^*\|^{-1}\|Q^{1/2}\|_F^{-1}$ with positive probability.

Recall that $w' \sim \mathcal{N}(0, I_d)$. Hence, $\langle w', Qw^*\rangle \sim \mathcal{N}(0, \|Qw^*\|^2)$. This implies that $m_0 > 0$ with probability $1/2$ and

$$\mathbb{P}(c_1\sigma \le |\langle w', Qw^*\rangle| \le c_2\sigma) = 1 - \mathbb{P}(|\langle w', Qw^*\rangle| \ge c_2\sigma) - \mathbb{P}(|\langle w', Qw^*\rangle| \le c_1\sigma).$$

But

$$\mathbb{P}(|\langle w', Qw^*\rangle| \ge c_2\sigma) \le e^{-c_2^2/2}$$

and

$$\mathbb{P}(|\langle w', Qw^*\rangle| \le c_1\sigma) \le 1 - e^{-c_1^2/2}.$$

So, if c_1 is chosen small enough, and c_2 large enough

$$\mathbb{P}(c_1\sigma \le |\langle w', Qw^*\rangle| \le c_2\sigma) \ge 1 - \epsilon.$$

Now, let us control $\left\|Q^{1/2}w'\right\|^2 = w'^\top Q w'$. This is a quadratic form in w' that has expectation $\mathbb{E}\left\|Q^{1/2}w'\right\|^2 = \left\|Q^{1/2}\right\|_H^2$. By Hanson-Wright inequality, we have

$$\mathbb{P}\left(\left|\left\|Q^{1/2}w'\right\|^2 - \left\|Q^{1/2}\right\|_H^2\right| \ge t\right) \le e^{-c\min(\frac{t^2}{\|Q\|_H^2}, \frac{t}{\|Q\|})}.$$

By choosing $t = C\|Q\|_H \lesssim \left\|Q^{1/2}\right\|_H^2$ since $\|Q\| = 1$, we obtain that with positive probability

$$0.5\left\|Q^{1/2}\right\|_H^2 \le \left\|Q^{1/2}w'\right\|^2 \le 1.5\left\|Q^{1/2}\right\|_H^2.$$

We obtaine the claimed result by taking the quotient.

E.2 Control of the noise

First, let us recall Doob's maximal inequality that will be used frequently to control the noise.

Theorem 4 (Doob's Maximal Inequality)**.** *Let $(X_t)_{t\leq T}$ be a martingale or positive submartingale belonging to L^p for some $p \geq 1$. Then for every $\lambda > 0$ we have*

$$\mathbb{P}\left(\sup_{t\leq T}|X_t| \geq \lambda\right) \leq \frac{\mathbb{E}\,|X_T|^p}{\lambda^p}.$$

Lemma 9. *For all $\epsilon > 0$ we have*

$$\mathbb{P}\left(\sup_{t\leq T}\left|\sum_{l\leq t}\langle V_l, w^*\rangle\right| \geq \frac{\sqrt{T}}{\epsilon}\right) \leq \epsilon^2 \left\|Q^{1/2}w^*\right\|^2.$$

Proof. Let us define $M_t = \sum_{l\leq t}\langle V_l, w^*\rangle$. Notice that since $y\sigma'(.)$ is always bounded by one, we have

$$\mathbb{E}(M_T^2) \leq T\mathbb{E}\langle x, w^*\rangle^2 \leq \left\|Q^{1/2}w^*\right\|^2 T.$$

By consequence, Theorem 4 applied with $p = 2$ leads to the result. Notice that the bound is uniform is the initial value $w^{(0)}$, as in Arous et al. (2020). □

Lemma 10. *Let us denote $M'_t = \left\|\sum_{l\leq t} V_l\right\|^2$. This is a submartingale and for all $\epsilon > 0$ we have*

$$\mathbb{P}\left(\sup_{t\leq T} M'_t \geq \frac{\left\|Q^{1/2}\right\|_F^2}{\epsilon}T\right) \leq \epsilon.$$

In particular, it implies that with probability at least $1-\epsilon$, for all $t \leq T$, we have

$$\left\|\sum_{l\leq t} V_l\right\| \leq \epsilon^{-1/2}\sqrt{T}\left\|Q^{1/2}\right\|_F.$$

Proof. By definition $M'_t = \sum_i \langle \sum_l V_l, e_i\rangle^2$. Since any convex function of a martingale is a submartingale, $\langle \sum_l V_l, e_i\rangle^2$ is a submartingale and M'_t is a submartingale as a sum of submartingale.

Now observe that

$$\begin{aligned}
\mathbb{E}M'_T &= \mathbb{E}\sum_{t\leq T}\|V_t\|^2 + \mathbb{E}\sum_{t\neq t'}\langle V_t, V_{t'}\rangle \\
&= \mathbb{E}\sum_{t\leq T}\|V_t\|^2 && \text{(since } \mathbb{E}(V_{t+1}|H_t) = 0.\text{)} \\
&\leq T\left\|Q^{1/2}\right\|_H^2.
\end{aligned}$$

□

E.3 The Descent Phase

Assume that Assumption A4 is valid for $\gamma' = 1$.

It is clear from the previous analysis in Section 4.3 that the directional martingale error term E_2 is negligible compared to m_t. However, $\left\|Q^{1/2}w^{(t+1)}\right\|$ is no longer necessarily bounded and the approximate dynamic (4.2) is no longer valid. The analysis done in section 4.1 suggests that $\left\|Q^{1/2}w^{(t)}\right\|$ grows at a similar rate

than m_t and if the initial scaling of the weights r is small enough, one should have $\left\|Q^{1/2}w^{(t)}\right\| \approx m_t$. Hence, from Lemma 1 we obtain the following approximated dynamic

$$m_{t+1} \approx \frac{m_t}{m_{t+1}} m_t + \frac{\eta}{m_{t+1}} m_t^{k^*-1}$$

that is equivalent to

$$m_{t+1}^2 \approx m_t^2 + \eta m_t^{2\frac{(k^*-1)}{2}}$$

that can be solved similarly as (4.2) with the change of variable $u_t = m_t^2$. All these approximations remain to be made rigorous.

F Additional numerical experiments

The experiments were conducted using Python on a CPU Intel Core i7-1255U. The code is available at `https://glmbraun.github.io/AniSIM`.

F.1 Description of `SpheSGD`

The spherical gradient ∇^s with respect to the geometry induced by Q is defined by

$$\nabla^s_w L = \nabla_w L - \langle \nabla_w L, w \rangle_Q w.$$

We update the weights as follows

$$\tilde{w}^{(t+1)} = w^{(t)} - \eta \nabla^s_{w^{(t)}} L$$
$$w^{(t+1)} = \frac{\tilde{w}^{(t+1)}}{\left\|Q^{1/2}\tilde{w}^{(t+1)}\right\|}.$$

F.2 Adaptative learning rate.

The theoretical analysis shows that η should be chosen small enough to control the impact of noise. However, as the signal increases, more noise can be tolerated. Here, we consider a SIM of the form of the form $y = \texttt{Sign}(\langle x, w^* \rangle)$ where $x \sim \mathcal{N}(0, I_d)$ with $d = 4000$, $n = 8000$ and $w^* \in \mathbb{S}^{d-1}$. We run vanilla SGD with a learning rate $\eta = 0.000001$ and `AdaptLR-SGD` where at each gradient step the learning rate is increased: $\eta_{t+1} = \eta_t(1 + 0.000001)$. As shown in Figure 3, increasing the learning rate accelerates the algorithm's convergence. Determining a data-driven method to select an appropriate learning rate is left for future work.

F.3 Data reuse

We consider the following algorithm referred to as `RepSGD`, that use the same data batch two times:

$$\tilde{w}^{(t)} = w^t - \eta_1 \nabla^s_{w^{(t)}} L$$
$$w^{(t+1)} = w^t - \eta_2 \nabla^s_{\tilde{w}^{(t)}} L.$$

This is similar to the algorithm analyzed in Arnaboldi et al. (2024), except that we do not use spherical gradient update nor use a retractation to ensure that the norm of $w^{(t)}$ remains equal to one.

We consider the learning the following single index model $y = H_3(\langle w^*, x \rangle)$ with $x \sim \mathcal{N}(0, I_d)$, $d = 4000$ and $n = 80000$. We fix the learning rates $\eta_1 = \eta = \eta_2 = -0.0001$.

Figure 4 shows that while vanilla `SGD` is unable to learn the single index w^*, `RepSGD` achieves weak recovery with the same sample complexity.

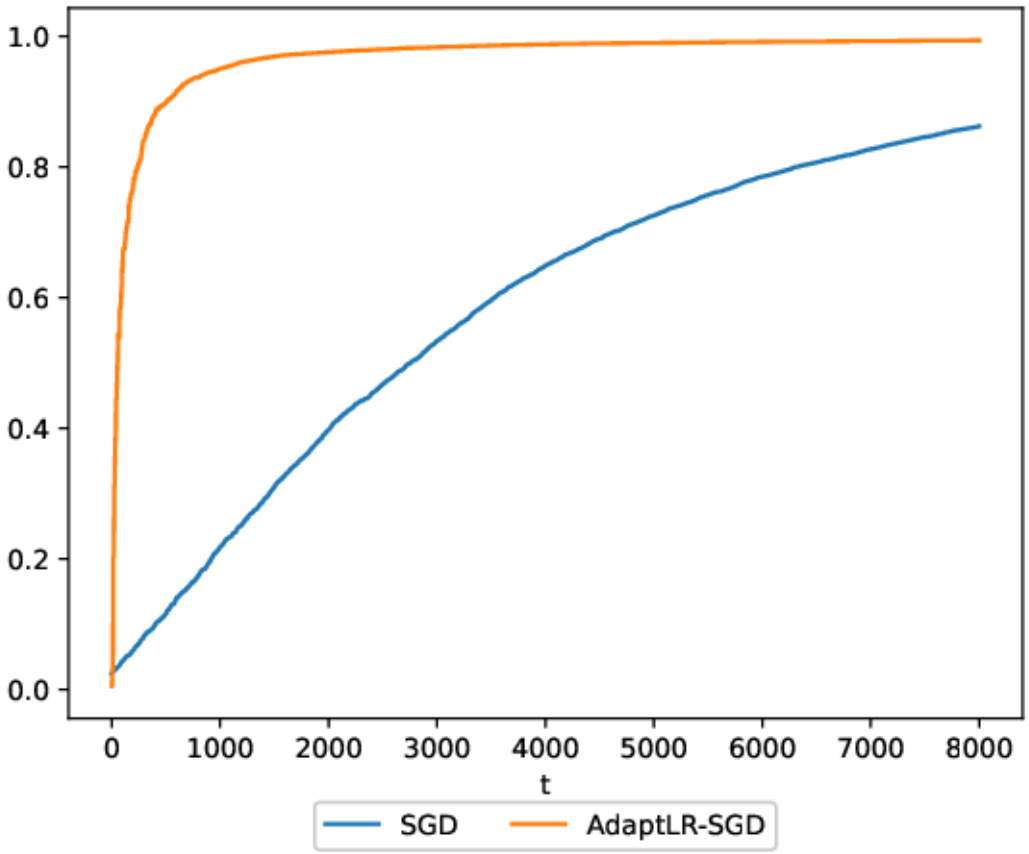


Figure 3: Comparison of learning dynamics between `Vanilla SGD` and `RepSGD`.

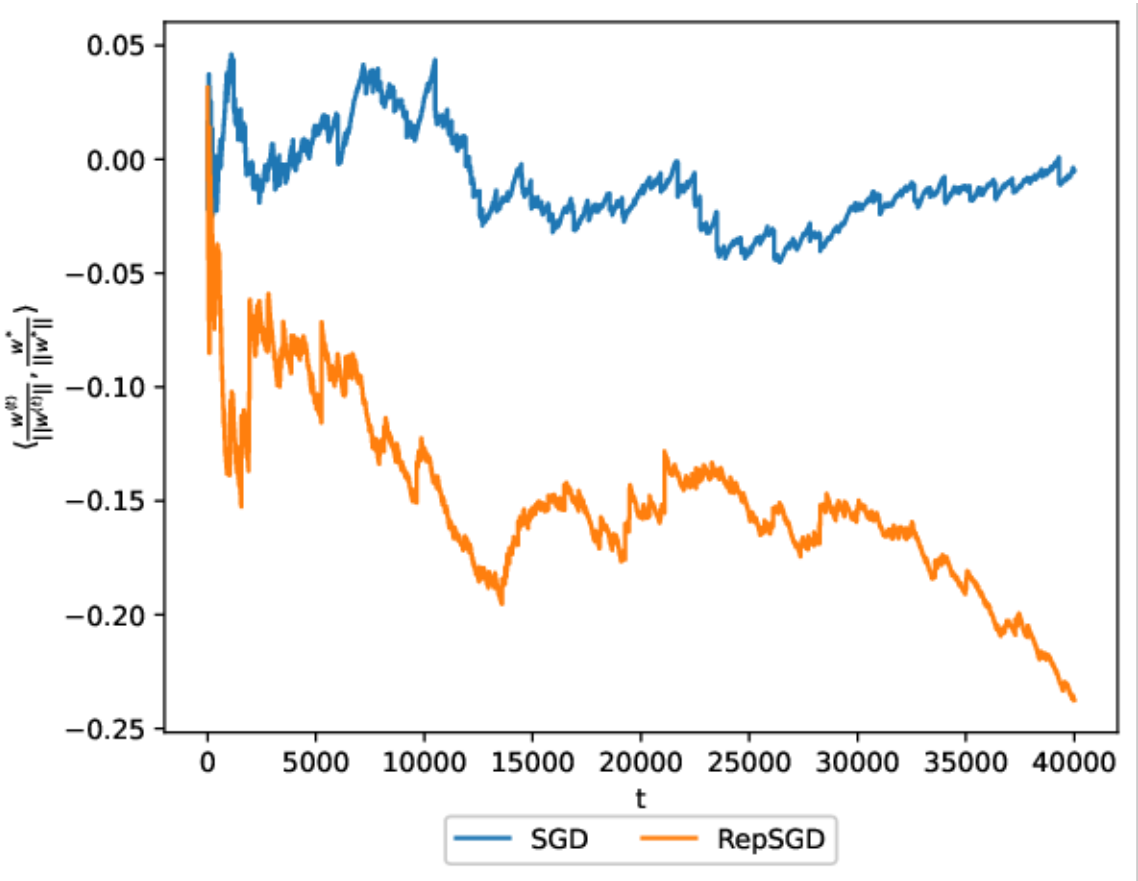


Figure 4: Comparison of learning dynamics between `Vanilla SGD` and `RepSGD`.