From Information to Generative Exponent: Learning Rate Induces Phase Transitions in SGD

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2025

Abstract

To understand feature learning dynamics in neural networks, recent theoretical works have focused on gradient-based learning of Gaussian single-index models, where the label is a function of a latent one-dimensional projection of the input. While the sample complexity of online SGD is determined by the *information exponent* of the link function, recently proposed variants of SGD that introduce *non-correlational* updates are instead limited by the *generative exponent*. However, this picture is only valid for sufficiently large learning rate. In this paper, we characterize the relationship between learning rate and sample complexity for a general class of gradient-based algorithms, and demonstrate a phase transition from an "information exponent regime" with small learning rate to a "generative exponent regime" with large learning rate. Our framework covers prior analyses of online SGD and SGD with batch reuse, while also introducing a new layer-wise training algorithm. Our theoretical study demonstrates that the choice of learning rate is as important as the design of the algorithm in achieving statistical and computational efficiency.

1. Introduction

A key aspect of deep learning theory is to understand how neural networks can adapt to underlying data structure and achieve desirable statistical and computational complexity through their optimization dynamics. Towards this goal, several works have focused on learning target functions that depend on low-dimensional projections of data, such as single- and multi-index models. For Gaussian data and online SGD on the squared loss, the number of training samples/iterations needed to learn a single-index model depends on the *information exponent* of the target function [9].

Variants of SGD that reuse batches [4, 23, 27] can have sample complexity controlled by the *generative exponent* [20] of the target, which is at most as large as the information exponent, and can be significantly smaller. However, a puzzling observation around reusing batches is that the sample complexity of full-batch gradient flow on the squared loss, through the best known upper bounds, still depends on the information exponent [10, 30]. This suggests that the role of the learning rate, while ignored in the current literature, is also crucial in determining sample complexity.

In this work, we characterize the regimes of complexity emerging from the choice of learning rate. In Section 3, we introduce our general framework and provide a careful learning-rate-dependent analysis of the sample complexity of learning single-index models, resulting in bounds that explicitly demonstrate phase transitions induced by the choice of learning rate. We show that our framework is expressive enough to capture both vanilla online SGD (Appendix D.1) and algorithms with non-correlational update rules such as SGD with batch reuse (Section 4.2). Additionally, we introduce a new layer-wise training algorithm that uses a different scaling of learning rate for the first and second layers of the network, thus using a two-timescales dynamics. We demonstrate that the performance of this algorithm also depends critically on the learning rate of the second layer.

Notation. For $k \in \mathbb{N}$, we use $[k]$ to denote the set $\{1, \ldots, k\}$. All asymptotic notation is with respect to the input dimension d. We use $\tilde{O}(\cdot)$ and $\tilde{\Theta}(\cdot)$ to denote $O(\cdot)$ and $\Theta(\cdot)$ up to polylogarithmic factors, respectively. Similarly, the relations $\lesssim$ and $\gtrsim$ denote bounds up to polylogarithmic factors. We write $a \asymp b$ when $a \lesssim b$ and $a \gtrsim b$. An event is said to occur *with high probability* if its probability is at least $1 - o_d(1)$. For any $g \in L^2(\mathcal{N}(0,1))$, we write its Hermite expansion as $g(z) = \sum_{k=0}^{\infty} u_k(g)\mathsf{He}_k(z)$, where He_k denotes the k-th probabilist's Hermite polynomial [33] and $u_k(g) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[g(z)\mathsf{He}_k(z)]$ is the k-th Hermite coefficient of g.

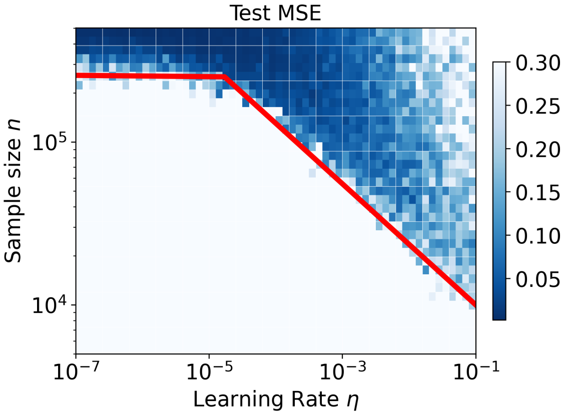

Figure 1: Test MSE of alternating SGD for different choices of (η, n) in the setting $\sigma_* = \sigma = \mathsf{He}_3$, $d = 50$. See Appendix E for complete details.

2. Problem Setup

We consider a supervised regression setting where the inputs are drawn from the Gaussian distribution and the labels are generated according to the *single-index model*, i.e.

$$y_i = \sigma_*(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_* \rangle) + \zeta_i, \quad \boldsymbol{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \tag{2.1}$$

where $\boldsymbol{\theta}_* \in \mathbb{S}^{d-1}$ is the ground truth direction, $\sigma_* : \mathbb{R} \to \mathbb{R}$ is a (nonlinear) link function, and ζ_i is i.i.d. symmetric sub-Weibull [40] label noise.

We learn the above model with a two-layer neural network f with N hidden neurons, first-layer weights $\boldsymbol{w}_j \in \mathbb{S}^{d-1}$, biases $b_j \in \mathbb{R}$, second layer weights $a_j \in \mathbb{R}$, and polynomial activations $\sigma_j : \mathbb{R} \to \mathbb{R}$ as in [27]. The network outputs a weighted average of the hidden layer activations:

$$f(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}) = \frac{1}{N}\sum_{j=1}^{N} a_j \sigma_j\big(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle + b_j\big). \tag{2.2}$$

Our objective is to characterize the number of iterations/samples required for *weak recovery* of $\boldsymbol{\theta}_*$ as a function of the learning rate for online iterative algorithms. That is, starting from an initialization $\boldsymbol{w}_j^{(0)}$ with $\langle \boldsymbol{\theta}, \boldsymbol{w}_j^{(0)} \rangle \asymp d^{-\frac{1}{2}}$ — which occurs with high probability for $\boldsymbol{w}_j^{(0)} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ — we seek T such that $\langle \boldsymbol{\theta}_*, \boldsymbol{w}_j^{(T)} \rangle \gtrsim \frac{1}{\operatorname{polylog} d}$ with high probability. Once this is achieved, *strong recovery* (i.e., recovery of $\langle \boldsymbol{\theta}_*, \boldsymbol{w} \rangle \geq 1 - \varepsilon$ for some $\varepsilon > 0$) and approximation of the target via ridge regression on $\boldsymbol{a}$ proceed with smaller sample complexity (see Appendices C.6, C.7).

We introduce two properties of σ_* that are known to control the complexity of gradient-based learning.

Definition 1 (Information Exponent [9]) *For any $g \in L^2(\mathcal{N}(0,1))$, let $u_k(g)$ denote the kth coefficient in its Hermite expansion. The information exponent of g is defined as*

$$\mathrm{IE}(g) := \min\{k > 0 : u_k(g) \neq 0\}. \tag{2.3}$$

Throughout this paper, we denote the information exponent of the link function σ_* in (2.1) by p, and we use the notation $p_i := \mathrm{IE}(\sigma_*^i)$ for $i \geq 2$. Ben Arous et al. [9] show that online SGD with the square loss has sample complexity $n = \tilde{\Theta}(d^{(p-1)\vee 1})$.

Definition 2 (Generative Exponent [20]) *For any $g \in L^2(\mathcal{N}(0,1))$, the generative exponent is defined as the smallest information exponent over all L^2 transformations of g, i.e.,*

$$\mathrm{GE}(g) = \inf_{\mathcal{T} \in L^2(g_{\#}\mathcal{N}(0,1))} \mathrm{IE}(\mathcal{T} g). \tag{2.4}$$

Note that $\mathrm{GE}(g) \leq \mathrm{IE}(g)$ for all g. Throughout this paper, we denote the generative exponent of σ_* in (2.1) by p_*. Arnaboldi et al. [4], Lee et al. [27] show that SGD has sample complexity $n \gtrsim d^{(p_*-1)\vee 1}$ when going over each sample twice.

3. Sample Complexity of a Generic Online Algorithm

To generalize several notions of gradient-based learning of single-index models, we consider updates to a first-layer weight $\boldsymbol{w}$ of the form

$$\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} + \gamma \psi_\eta(y^{(t)}, \langle \boldsymbol{x}^{(t)}, \boldsymbol{w}^{(t)} \rangle) \boldsymbol{P}^{\perp}_{\boldsymbol{w}^{(t)}} \boldsymbol{x}^{(t)}, \quad \boldsymbol{w}^{(t+1)} \leftarrow \frac{\boldsymbol{w}^{(t+1)}}{||\boldsymbol{w}^{(t+1)}||}, \tag{3.1}$$

where $(\boldsymbol{x}^{(t)}, y^{(t)})$ is an i.i.d. draw from the target single-index model (2.1), $\gamma > 0$, $\eta \geq 0$, $\boldsymbol{P}^{\perp}_{\boldsymbol{w}^{(t)}} = \boldsymbol{I}_d - \boldsymbol{w}\boldsymbol{w}^\top$, and ψ_η is an update function based on a "general gradient oracle". This formulation is similar to that of Chen et al. [16], who also use generalized gradients, but additionally incorporate weight perturbation and averaging.

Importantly, our framework contains the batch-reuse SGD of [4, 27], in which $\eta > 0$ has the interpretation of a second learning rate. This hyperparameter will dictate the phase transition of interest by controlling the scale of non-correlational terms in the oracle and placing a constraint on the largest possible learning rate γ. The key quantities elucidating the effect of η on the sample complexity are the Hermite coefficients

$$\mu_i(\eta) := u_i\big(a \mapsto u_{i-1}\big(b \mapsto \psi_\eta(\sigma_*(a), b)\big)\big), \quad i \in [r]. \tag{3.2}$$

We make two key assumptions. The first ensures all noise terms are sub-Weibull, which allows us to make concentration arguments. The second provides some degree of alignment between σ and σ_*, without which the model misspecification is so severe that weak recovery may not be achieved (see e.g., [16, Assumption 4.1(b)]).

Assumption 3 *The link function σ_* is a polynomial of degree $q = \Theta(1)$, and the update function ψ_η is a polynomial of degree at most $r = \Theta(1)$ in each of its arguments with $O(1)$ coefficients.*

Assumption 4 *For any $i^* \in \arg\min_{1 \leq i \leq r : \mu_i \neq 0} |\mu_i(\eta)|^{-1}(d^{\frac{i-2}{2} \vee 0})$, we have $\mu_{i^*}(\eta) > 0$.*

Below, we state our main result for a generic gradient-based algorithm (proof in Appendix C).

Theorem 5 *Suppose Assumptions 3 and 4 hold. Let $\boldsymbol{w}^{(0)} \in \mathbb{S}^{d-1}$ such that $\langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(0)} \rangle \asymp d^{-\frac{1}{2}}$ and $\gamma \lesssim \max_{1 \leq i \leq r} \mu_i(\eta) d^{-(\frac{i}{2} \vee 1)}$. Then, with high probability,*

$$T(\eta) = \min_{\substack{1 \leq i \leq r \\ \mu_i > 0}} \tilde{\Theta}\big(\gamma^{-1}(\mu_i(\eta))^{-1} d^{\frac{i-2}{2} \vee 0}\big) \tag{3.3}$$

iterations of (3.1) are necessary and sufficient to achieve $\langle \boldsymbol{\theta}_*, \boldsymbol{w} \rangle \gtrsim \frac{1}{\operatorname{polylog} d}$.

```
Algorithm 1: Alternating SGD
Input: Learning rates $\eta, \gamma > 0$, sample size $T$
Initialize $\boldsymbol{w}^{(0)} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, $a = 1$
for $t = 0$ to $t = T - 1$ do
    Draw i.i.d. sample $(\boldsymbol{x}, \boldsymbol{y})$
    Update $\tilde{a}^{(t+1)} \leftarrow a + \eta y \sigma(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)} \rangle)$
    Update $\boldsymbol{w}^{(t+1)} \leftarrow$
      $\boldsymbol{w}^{(t)} + \gamma y \tilde{a}^{(t+1)} \sigma'(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)} \rangle) \boldsymbol{P}^{\perp}_{\boldsymbol{w}^{(t)}} \boldsymbol{x}$
    Normalize $\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t+1)} / ||\boldsymbol{w}^{(t+1)}||$
end
Output $\boldsymbol{w}^{(T)}$
```

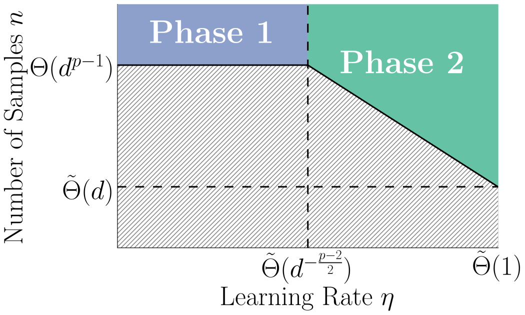

Figure 2: Illustration of the sample complexity of Algorithm 1 as a function of η in the case $p_2 = 2$.

4. Applications

We illustrate the applicability of Theorem 5 with two algorithms that incorporate non-correlational updates: a layer-wise algorithm we term *alternating SGD* and batch reuse SGD. In this work, we focus on gradients of the *correlation loss* $\ell(y, y') = 1 - yy'$.

4.1. Alternating SGD

Algorithm 1, which we call *alternating SGD*, employs a two-step process to update a weight $\boldsymbol{w}$. First, it computes a gradient update for $\boldsymbol{a}$ with learning rate η. Then, it uses the updated value $\tilde{a}$ in a gradient update on $\boldsymbol{w}$ with learning rate γ. Concretely, in the notation of our framework, we show in Appendix D.2 that

$$\psi(y, z) = ya\sigma'(z) + \eta y^2 \sigma(z)\sigma'(z). \tag{4.1}$$

The first term corresponds to a vanilla SGD update, while the second term is the non-correlational term that arises from using the updated second layer parameter. Intuitively, if η is sufficiently large and $p_2 < p$, this term will speed up training. The following makes this rigorous.

Corollary 6 *Assume $\mu_p, \mu_{p_2} > 0$, $\eta \lesssim 1$, $\gamma \asymp \max\{d^{-(\frac{p}{2} \vee 1)}, \eta d^{-(\frac{p_2}{2} \vee 1)}\}$, and $\boldsymbol{w}^{(0)} \in \mathbb{S}^{d-1}$ such that $\langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(0)} \rangle \asymp d^{-\frac{1}{2}}$. Then, with high probability,*

$$T(\eta) = \tilde{\Theta}\big(d^{(p-1)\vee 1}\big) \wedge \tilde{\Theta}\big(\eta^{-2} d^{(p_2-1)\vee 1}\big). \tag{4.2}$$

iterations of Algorithm 1 are necessary and sufficient to achieve weak recovery.

The assumption $\mu_p, \mu_{p_2} > 0$ is derived from our more general Assumption 4 and holds with $\Theta(1)$ probability if σ follows the randomized construction in [27, Appendix B.1]. The sample complexity result implies a phase transition between the regime where the correlational term dominates and one where the non-correlational term dominates, occurring at (when $p \geq 2$)

$$d^{p-1} \asymp \eta^{-2} d^{(p_2-1)\vee 1} \iff \eta \asymp d^{-\frac{1}{2}[(p-p_2)\vee(p-2)]}. \tag{4.3}$$

For example, if $\sigma_* = \mathsf{He}_3$, then $p = 3$ and $p_2 = 2$. The phase transition occurs at $\eta \asymp d^{-\frac{1}{2}}$. At or below this threshold, alternating SGD has quadratic complexity, while $\eta \gtrsim \frac{1}{\mathrm{polylog}\, d}$ gives $\tilde{\Theta}(d)$ complexity. Intermediate values of η interpolate between these two regimes. We illustrate this in Figure 1.

Algorithm 2: Batch Reuse SGD

```
Input: Learning rates $\eta, \gamma > 0$, sample size $T$
Initialize $\boldsymbol{w}^{(0)} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$
for $t = 0$ to $T-1$ do
    Draw i.i.d. sample $(\boldsymbol{x}, \boldsymbol{y})$
    $\tilde{\boldsymbol{w}}^{(t)} \leftarrow \boldsymbol{w}^{(t)} + \eta y \sigma'(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)} \rangle) \boldsymbol{P}^{\perp}_{\boldsymbol{w}^{(t)}} \boldsymbol{x}$
    $\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} + \gamma y \sigma'(\langle \boldsymbol{x}, \tilde{\boldsymbol{w}}^{(t)} \rangle) \boldsymbol{P}^{\perp}_{\boldsymbol{w}^{(t)}} \boldsymbol{x}$
    Normalize $\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t+1)} / \|\boldsymbol{w}^{(t+1)}\|$
end
Output $\boldsymbol{w}^{(T)}$
```

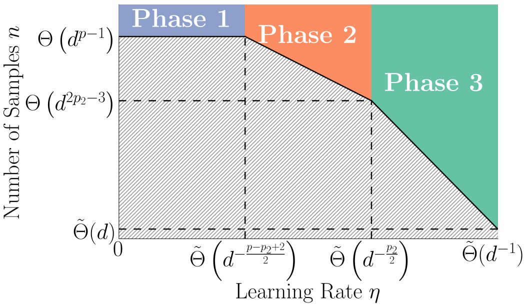

Figure 3: Sample complexity of Algorithm 2 in the case of two phase transitions and $p_3 \leq 2$.

4.2. Batch Reuse SGD

The analysis of the batch reuse SGD algorithm (Algorithm 2) proceeds similarly to the previous subsection, with the key difference being that it can exhibit up to r phase transitions. By a Taylor expansion argument (Appendix D.4), we see that the algorithm implements monomial transformations of the label up to degree r.

Corollary 7 *Suppose Assumptions 3 and 4 hold, $\eta \lesssim d^{-1}$, $\gamma \lesssim \max_{1 \leq i \leq r} (\eta d)^{i-1} d^{-(\frac{p_i}{2} \vee 1)}$, and $\boldsymbol{w}^{(0)} \sim \mathbb{S}^{d-1}$ such that $\langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(0)} \rangle \asymp d^{-\frac{1}{2}}$. Then, with high probability,*

$$T(\eta) = \min_{1 \leq i \leq r} \tilde{\Theta}\big((\eta d)^{-2(i-1)} d^{(p_i - 1) \vee 1}\big). \tag{4.4}$$

iterations of Algorithm 2 are necessary and sufficient to achieve weak recovery.

For any two distinct i, j with $\mu_i, \mu_j > 0$, η induces the phase transition:

$$(\eta d)^{-2(i-1)} d^{(p_i-1) \vee 1} \leq (\eta d)^{-2(j-1)} d^{(p_j-1) \vee 1} \iff \eta \leq d^{\frac{[(p_j-1) \vee 1] - [(p_i-1) \vee 1]}{2(j-i)} - 1}. \tag{4.5}$$

In particular, suppose that $u_{p_*-1}(\sigma^{(I)}(\sigma')^{I-1}) u_{p_*}(\sigma_*^I) > 0$ and $u_p(\sigma_*) u_p(\sigma) > 0$ hold, which can be achieved with $\Theta(1)$ probability by a randomized activation agnostic to σ_* as in [27]. Taking $\eta \lesssim d^{-\frac{p+1}{2}}$ gives the sample complexity $T = \Theta(d^{(p-1) \vee 1})$, which matches the online SGD bound [9]. On the other hand, when $r \geq I$, taking $\eta \gtrsim \frac{1}{d}$ as in [27] matches their sample complexity bound $T = \tilde{\Theta}(d)$. Appendix E details an experiment with batch reuse SGD exhibiting this phase transition.

5. Conclusion

This work demonstrates that learning rate is a fundamental factor in determining the sample complexity of gradient-based algorithms for learning single-index models with neural networks. We show that algorithms that employ a combination of correlational and non-correlational update terms (with distinct learning rates) exhibit phase transitions between two or more sample complexity regimes as a function of the relative scaling of the non-correlational term. Natural directions for future work include an extension of our framework to multi-index models, more general input distributions, non-polynomial activation functions, and non-constant learning rates.

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] Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925. PMLR, 2023.

[3] Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Online learning and information exponents: On the importance of batch size, and time/complexity tradeoffs. *arXiv preprint arXiv:2406.02157*, 2024.

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

[5] 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.

[6] 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.

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

[8] David Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.

[9] 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.

[10] 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.

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

[12] Guillaume Braun, Minh Ha Quang, and Masaaki Imaizumi. Learning a single index model from anisotropic data with vanilla stochastic gradient descent. *arXiv preprint arXiv:2503.23642*, 2025.

[13] Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. *Advances in Neural Information Processing Systems*, 37:71306–71351, 2024.

[14] Seok-Ho Chang, Pamela C Cosman, and Laurence B Milstein. Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

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

[16] Siyu Chen, Beining Wu, Miao Lu, Zhuoran Yang, and Tianhao Wang. Can neural networks achieve optimal computational-statistical tradeoff? an analysis on single-index model. In *The Thirteenth International Conference on Learning Representations*, 2025.

[17] Dean S Clark. Short proof of a discrete gronwall inequality. *Discrete applied mathematics*, 16 (3):279–281, 1987.

[18] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

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

[20] 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.

[21] 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.

[22] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

[23] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborova, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: breaking the curse of information and leap exponents. In *Proceedings of the 41st International Conference on Machine Learning*, pages 9991–10016, 2024.

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

[25] Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pages 4772–4784. PMLR, 2021.

[26] 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.

[27] Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny 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.

[28] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

[29] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in neural information processing systems*, 32, 2019.

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

[31] Alireza Mousavi-Hosseini, Adel Javanmard, and Murat A Erdogdu. Robust feature learning for multi-index models in high dimensions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aKkDY1Wca0.

[32] Alireza Mousavi-Hosseini, Denny Wu, and Murat A Erdogdu. Learning multi-index models with neural networks via mean-field langevin dynamics. In *The Thirteenth International Conference on Learning Representations*, 2025.

[33] Ryan O'Donnell. Analysis of boolean functions. *arXiv preprint arXiv:2105.10386*, 2021.

[34] Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4009–4081. PMLR, 2024.

[35] Yunwei Ren and Jason D Lee. Learning orthogonal multi-index models: A fine-grained information exponent analysis. *arXiv preprint arXiv:2410.09678*, 2024.

[36] Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D Lee. Emergence and scaling laws in sgd learning of shallow neural networks. *arXiv preprint arXiv:2504.19983*, 2025.

[37] Berfin Şimşek, Amire Bendjeddou, and Daniel Hsu. Learning gaussian multi-index models with gradient flow: Time complexity and directional convergence. *arXiv preprint arXiv:2411.08798*, 2024.

[38] 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.

[39] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[40] Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

[41] 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.

[42] 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.

Appendix A. Related Work

Feature Learning and Single-Index Models. There is a vast body of literature on algorithms for learning Gaussian single-index models, see e.g. [15, 24]. Here, we focus on more recent works that use gradient-based training. Ben Arous et al. [9] studied online SGD for learning high-dimensional single-index models with known non-linearity, where they introduced the information exponent as the quantity controlling the number of samples needed to learn the model. The representation learned by a network on single-index models with information exponent 1 was studied in [5, 30], while Bietti et al. [10] considered gradient flow for learning functions with higher information exponent. On multi-index models, Damian et al. [21] considered one gradient step for learning polynomials, and Abbe et al. [1] studied learning general multi-index models where a saddle-to-saddle dynamics can emerge. General multi-index models remain difficult to analyze [11, 22, 32]. However, several works have studied the simpler case of additive models [34–37].

Going beyond unstructured isotropic Gaussian data, many works considered the existence of additional input structure or modifications of the single-index model, such as a spiked covariance [6, 7, 12, 26, 30, 42], sparsity in the input [41], or a perturbation of the target [19]. The recovery of the low-dimensional multi-index subspace has been used to go beyond standard learning frameworks, e.g. to obtain better theoretical guarantees for adversarial robustness [31].

On the other hand, CSQ and SQ lower bounds for learning single-index models where developed in [21] and [20] respectively, where the former depends on the information and the latter depends on the generative exponent. Similar lower bounds were derived in [1] for multi-index models, where a leap exponent controls the complexity, and Troiani et al. [38] studied approximate message passing as a proxy for computational lower bounds.

Learning Rate and Generalization. Numerous works have studied the effect of learning rate on optimization and generalization in deep learning. Notably, deep networks with large learning rate can operate near the "edge of stability" [18], where it has been empirically observed that such large learning rates improve generalization by preferring flat minima [8, 25, 28, 29, and references therein], learning sparse features [2], or obtaining larger margin [13]. Closer to our setting, Arnaboldi et al. [3] study the optimal choice of learning rate for online SGD. However, while their algorithm always remains in a correlational regime, we consider a wide range of learning rates to understand the effect of non-optimal choices in practice, and demonstrate phase transitions in the behavior of the SGD depending on stepsize, going from correlational regimes dominated by information exponent to full statistical query regimes dominated by generative exponent.

Appendix B. Key Lemmas

Lemma 8 (Discrete Grönwall Inequality [17]) *Let $\{m_t\}_{t=0}^{\infty}$ be a sequence such that $m_0 = a$ and $m_t \le a + c\sum_{j=0}^{t-1} m_j$ for all $t \ge 1$, where $a, c > 0$. Then, for all $t \ge 0$,*

$$m_t \le a(1+c)^t \le ae^{ct}. \tag{B.1}$$

Moreover, if instead $m_t \ge a + c\sum_{j=0}^{t-1} m_j$ for all $t \ge 1$, then $m_t \ge a(1+c)^t$ for all $t \ge 0$.

Proof The result easily follows by induction. The statement is trivial for $t = 0$. Suppose now that it holds for some $t \geq 0$. Then,

$$m_{t+1} \leq a + c\sum_{j=0}^{t} m_j \leq a + c\sum_{j=0}^{t} a(1+c)^j = a + ca\left(\frac{(1+c)^{t+1}-1}{(1+c)-1}\right) = a(1+c)^{t+1}. \quad \text{(B.2)}$$

The same argument can be used for the reversed inequality. ■

Lemma 9 (Discrete Bihari-LaSalle Inequality [9, Appendix C]; [27, Lemma 18]) *Let $\{m_t\}_{t=0}^{\infty}$ be a sequence such that $m_0 = a$ and $m_t \leq a + c\sum_{j=0}^{t-1} m_j^{k-1}$ for all $t \geq 1$, where $a, c > 0$, and $k \geq 3$. Then,*

$$m_t \leq \frac{a}{(1-(k-2)ca^{k-2}t)^{\frac{1}{k-2}}}, \quad \forall\, 0 \leq t \leq \frac{1}{c(k-2)a^{k-2}}. \quad \text{(B.3)}$$

Moreover, if instead $m_t \geq a + c\sum_{j=0}^{t-1} m_j^{k-1}$, then

$$m_t \geq \frac{a}{(1-\frac{c}{2}a^{k-2}t)^{\frac{1}{k-2}}}, \quad \forall\, 0 \leq t \leq \tfrac{2}{c}(a^{-(k-2)} - c). \quad \text{(B.4)}$$

Proof

Let $\{a_t\}_{t=0}^{\infty}$ be such that $a_0 = a$ and $a_t = a + c\sum_{j=0}^{t-1} a_j^{k-1}$.

Upper Bound. Define $\{b_t\}_{t=0}^{\infty}$ by $b_0 = a$ and $b_t = a + \sum_{j=0}^{t-1} c(m_j)^{k-1}$. Then, $m_t \leq b_t$ by definition. We prove that $b_t \leq a_t$ by induction. Clearly, $b_0 = a_0$. Now,

$$b_{t+1} = a + \sum_{j=0}^{t} c(m_j)^{k-1} = b_t + c(m_t)^{k-1} \leq b_t + c(b_t)^{k-1} \leq a_t + c(a_t)^{k-1} = a_{t+1}, \quad \text{(B.5)}$$

where the last inequality follows from the induction hypothesis. Hence, $m_t \leq b_t \leq a_t$ for all $t \geq 0$.

Notice for all $t \geq 1$ that

$$c = \frac{a_{t+1} - a_t}{a_t^{k-1}} = \int_{a_t}^{a_{t+1}} \frac{1}{a_t^{k-1}}\, dx \geq \int_{a_t}^{a_{t+1}} \frac{1}{x^{k-1}}\, dx = \frac{1}{k-2}\left(\frac{1}{a_t^{k-2}} - \frac{1}{a_{t+1}^{k-2}}\right). \quad \text{(B.6)}$$

Rearranging the above, we have

$$a_{t+1}^{-(k-2)} \geq a_t^{-(k-2)} - c(k-2). \quad \text{(B.7)}$$

Unrolling the recurrence, we obtain

$$a_t^{-(k-2)} \geq a_0^{-(k-2)} - c(k-2)t. \quad \text{(B.8)}$$

So long as $a^{-(k-2)} - c(k-2)t > 0$, we can rearrange to obtain the desired upper bound

$$a_t \leq \frac{1}{\left(a_0^{-(k-2)} - c(k-2)t\right)^{\frac{1}{k-2}}} = \frac{a}{\left(1-(k-2)ca^{k-2}t\right)^{\frac{1}{k-2}}}. \quad \text{(B.9)}$$

The condition $a^{(k-2)} - c(k-2)t > 0$ holds so long as

$$t < \frac{1}{c(k-2)a^{k-2}} = \Theta(a^{-(k-2)}), \tag{B.10}$$

matching the condition in (B.3).

Lower Bound. A similar induction argument to the one in the upper bound proof shows that $m_t \geq a_t$ for all $t \geq 0$.

For each $t \geq 0$, let $b_t = a_t^{-(k-2)}$. Rewriting a step of the recurrence as

$$a_{t+1} = a_t\left(1 + \frac{c}{a_t^{-(k-2)}}\right) \tag{B.11}$$

allows us to write a recurrence for $\{b_t\}_{t=0}^{\infty}$:

$$b_{t+1} = b_t\left(1 + \frac{c}{b_t}\right)^{-(k-2)} \leq b_t\left(\frac{1}{1+\frac{c}{b_t}}\right) = \frac{b_t}{\frac{b_t+c}{b_t}} = \frac{b_t^2}{b_t+c} = b_t - \frac{cb_t}{b_t+c}. \tag{B.12}$$

Now, so long as $b_t \geq c$, we have $b_{t+1} \leq b_t - \frac{c}{2}$. Unrolling the recurrence and rewriting in terms of the a_t gives

$$\begin{aligned} & b_t \leq b_0 - \tfrac{c}{2}t \\ & \Rightarrow a_t^{-(k-2)} \leq a_0^{-(k-2)} - \tfrac{c}{2}t \\ & \Rightarrow a_t \geq \frac{1}{(a_0^{-(k-2)} - \frac{c}{2}t)^{\frac{1}{k-2}}} = \frac{a}{(1-\frac{c}{2}a^{k-2}t)^{\frac{1}{k-2}}}. \end{aligned} \tag{B.13}$$

It remains to characterize t for which $b_t \geq c$ holds. Notice

$$b_0 - \frac{c}{2}t \geq c \iff t \leq \frac{2}{c}(b_0 - c) = \Theta(a^{-(k-2)}), \tag{B.14}$$

which matches the condition on t in (B.4). ■

Appendix C. Proof of Main Result

We follow a very similar line of reasoning to the proofs of other sample complexity bounds involving the information and generative exponent in the literature, e.g., [9, 27]. Given a sample $(\boldsymbol{x}, y)$ from the target single-index model (2.1), recall from Section 3 that the update equation for $\boldsymbol{w}$ is

$$\boldsymbol{w}^{(t+1)} = \frac{\boldsymbol{w}^{(t)} + \gamma\psi_\eta(\boldsymbol{y}, \langle \boldsymbol{x}, \boldsymbol{w}^{(t)}\rangle)\boldsymbol{P}^\perp_{\boldsymbol{w}^{(t)}}\boldsymbol{x}}{||\boldsymbol{w}^{(t)} + \gamma\psi_\eta(\boldsymbol{y}, \langle \boldsymbol{x}, \boldsymbol{w}^{(t)}\rangle)\boldsymbol{P}^\perp_{\boldsymbol{w}^{(t)}}\boldsymbol{x}||}, \tag{C.1}$$

where $\boldsymbol{P}^\perp_{\boldsymbol{w}} = \boldsymbol{I}_d - \boldsymbol{w}\boldsymbol{w}^\top$. Throughout this section, we adopt the notation $\kappa^{(t)} = \langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(t)}\rangle$ and $\boldsymbol{g}(\boldsymbol{w}; \boldsymbol{x}, y) = \psi_\eta(y, \langle \boldsymbol{x}, \boldsymbol{w}\rangle)\boldsymbol{P}^\perp_{\boldsymbol{w}}\boldsymbol{x}$. We are interested in the dynamics of the alignment with the ground truth

$$\kappa^{(t+1)} = \frac{\kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle}{||\boldsymbol{w}^{(t)} + \gamma\boldsymbol{g}^{(t)}||}. \tag{C.2}$$

In Section C.1, using standard Gaussian tail bounds and tools from high-dimensional probability, we characterize the concentration of the initial alignment $\kappa^{(0)}$ about $d^{-\frac{1}{2}}$. Next, in Section C.2, we describe the "slowdown" in the alignment dynamics due to normalization. In Section C.3, we lower bound the expected update after one step. In Section C.4, we expand the expected dynamics over t steps and employ a standard martingale bound to control the noise, leading to a high probability upper bound on sample complexity when the initial alignment is of order $d^{-\frac{1}{2}}$. This is complemented by a matching lower bound proven in the same way in Section C.5. The upper and lower bound immediately imply Theorem 5, our main result. Subsequently, we discuss how weak recovery leads to strong recovery (Section C.6) and approximation of the target to arbitrary accuracy (Section C.7). This will elucidate the fact that achieving weak recovery is the sample complexity bottleneck for any generic online algorithm satisfying our formalism in Section 3.

C.1. Initial Alignment

We follow [27] in showing a high-probability lower bound for the alignment between a hidden neuron's weight vector $\boldsymbol{w}$ and the ground truth direction $\boldsymbol{\theta}_*$. We make a small modification to remove the dependence on step size from the bound.

Lemma 10 *Let $\boldsymbol{w}^{(0)} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Then, $\mathbb{P}(\kappa^{(0)} \geq C_0 d^{-\frac{1}{2}}) = \Omega(1)$ for any constant $C_0 > 0$. Moreover, for any $\delta' > 0$ there exists $\tilde{C}_0 \geq r$ such that $\mathbb{P}(\kappa^{(0)} \geq \tilde{C}_0 d^{\frac{1}{2}}) \leq \delta'$.*

Proof We may write

$$\kappa^{(0)} = \langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(0)} \rangle \overset{d}{=} \frac{\langle \boldsymbol{e}_1, \boldsymbol{g} \rangle}{||\boldsymbol{g}||}, \tag{C.3}$$

where $\boldsymbol{e}_1 \in \mathbb{R}^d$ is the first standard basis vector and $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$.

To proceed, as in [27], we require the following lemma.

Lemma 11 (Theorem 2 in [14]) *For any $\beta > 1$ and $s \in \mathbb{R}$, we have*

$$\frac{\sqrt{2e(\beta-1)}}{2\beta\sqrt{\pi}} e^{-\frac{\beta s^2}{2}} \leq \int_s^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt. \tag{C.4}$$

Then,

$$\begin{aligned}\mathbb{P}\big(\kappa^{(0)} \geq C_0 d^{-\frac{1}{2}}\big) &\geq \mathbb{P}\big(\langle \boldsymbol{e}_1, \boldsymbol{g} \rangle \geq 2C_0 \wedge ||\boldsymbol{g}|| \leq C_0 d^{\frac{1}{2}}\big) \\ &\geq \mathbb{P}(\langle \boldsymbol{e}_1, \boldsymbol{g} \rangle \geq 2C_0) - \mathbb{P}(||\boldsymbol{g}|| \geq C_0 d^{-\frac{1}{2}}) \\ &\geq \frac{\sqrt{2e(\beta-1)}}{2\beta\sqrt{\pi}} e^{-2C_0^2\beta} - e^{-\Omega(d)},\end{aligned} \tag{C.5}$$

where the second term follows from Gaussian concentration of the norm. Taking $\beta = 2$, we see that the above is $\Theta(1)$.

We can derive a high probability bound using Lipschitz concentration [39, Theorem 5.1.4] to obtain

$$\mathbb{P}\big(|\kappa^{(0)}| \geq \tilde{C}_0 d^{\frac{1}{2}}\big) \leq 2\exp(-\tilde{c}\tilde{C}_0^2). \tag{C.6}$$

for some $\tilde{c} > 0$. Arguing by symmetry and taking $\tilde{C}_0$ sufficiently large gives the second part of the result. ■

C.2. Normalization Error

Lemma 12 *Suppose $\kappa^{(t)} \geq 0$. The update (C.2) satisfies the lower bound*

$$\kappa^{(t+1)} \geq \kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle - \gamma^2\kappa^{(t)}||\boldsymbol{g}^{(t)}||^2 - \gamma^3\big|\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle\big|||\boldsymbol{g}^{(t)}||^2. \tag{C.7}$$

Proof When $\kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle \geq 0$, we have

$$\begin{aligned}\kappa^{(t+1)} &= \frac{\kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle}{||\boldsymbol{w}^{(t)} + \gamma \boldsymbol{g}^{(t)}||}\\ &= \frac{\kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle}{\sqrt{1+\gamma^2||\boldsymbol{g}^{(t)}||^2}}\\ &\geq (\kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle)(1-\gamma^2||\boldsymbol{g}^{(t)}||^2)\\ &\geq \kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle - \kappa^{(t)}\gamma^2||\boldsymbol{g}^{(t)}||_2^2 - \gamma^3|\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle|\,||\boldsymbol{g}^{(t)}||_2^2.\end{aligned} \tag{C.8}$$

The second line follows from the facts $\langle \boldsymbol{w}^{(t)}, \boldsymbol{g}^{(t)}\rangle = 0$ (due to $\boldsymbol{P}_{\boldsymbol{w}}^{\perp}$) and $\boldsymbol{w}^{(t)} \in \mathbb{S}^{d-1}$. The third line is trivial if $\gamma^2||\boldsymbol{g}^{(t)}||^2 \geq 1$. Otherwise, observe that when $\gamma^2||\boldsymbol{g}^{(t)}||^2 < 1$,

$$\begin{aligned}&1-\gamma^2||\boldsymbol{g}^{(t)}||^2 \leq \frac{1}{\sqrt{1+\gamma^2||\boldsymbol{g}^{(t)}||^2}}\\ \iff\ &\big(1-\gamma^2||\boldsymbol{g}^{(t)}||^2\big)^2\big(1+\gamma^2||\boldsymbol{g}^{(t)}||^2\big) \leq 1\\ \iff\ &\big(1-\gamma^4||\boldsymbol{g}^{(t)}||^2\big)\big(1-\gamma^2||\boldsymbol{g}^{(t)}||^2\big) \leq 1,\end{aligned} \tag{C.9}$$

where the last line clearly holds. Now, when $\kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle < 0$, the same lower bound can be shown via

$$\begin{aligned}\kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle - \kappa^{(t)}\gamma^2||\boldsymbol{g}^{(t)}||_2^2 - \gamma^3|\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle|\,||\boldsymbol{g}^{(t)}||_2^2 &\leq \kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle\\ &\leq \frac{\kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)}\rangle}{(1+\gamma^2||\boldsymbol{g}^{(t)}||^2)^{1/2}}.\end{aligned} \tag{C.10}$$

■

C.3. One-Step Population Dynamics

We extend the definition of the coefficients μ_i from (3.2) to handle label noise. Define

$$\mu_i := \mathbb{E}_\zeta[\hat{\mu}_i(\zeta)] := \mathbb{E}_\zeta\Big[u_i\Big(a \mapsto u_{i-1}\big(b \mapsto \psi_\eta(\sigma_*(a)+\zeta, b)\big)\Big)\Big],\ i \in [r]. \tag{C.11}$$

Lemma 13 *Assume that for some $t \geq 0$ we have $d^{-\frac{1}{2}} \leq \kappa^{(t)} \lesssim \frac{1}{\operatorname{polylog} d}$. Moreover, suppose that Assumption 4 holds. Then, there exists $C > 0$ such that taking $\gamma \leq C \max_{1\leq i\leq r}\mu_i d^{-(\frac{i}{2}\vee 1)}$ yields the following lower bound for the one-step dynamics of the alignment $\kappa^{(t)} := \langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(t)}\rangle$:*

$$\kappa^{(t+1)} \geq \kappa^{(t)} + \gamma C_1 \sum_{i=1}^{r} \mu_i (\kappa^{(t)})^{i-1}(1-(\kappa^{(t)})^2) + \gamma\nu^{(t)}, \tag{C.12}$$

where $\nu^{(t)}$ is a mean-zero sub-Weibull random variable with $\Theta(1)$ tail parameter and $C_1 > 0$ is a constant.

Proof Omitting the superscript t, the expected update to the alignment with the ground truth $\boldsymbol{\theta}_*$ (given the previous iterate) is

$$\begin{aligned}
&\mathbb{E}[\langle \boldsymbol{\theta}_*, \boldsymbol{g}\rangle] = \boldsymbol{\theta}_*^\top \boldsymbol{P}_{\boldsymbol{w}}^\perp \mathbb{E}_{\boldsymbol{x},y}\big[\psi_\eta\big(y, \langle \boldsymbol{x}, \boldsymbol{w}\rangle)\boldsymbol{x}\big)\big] \\
&= \boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{w}}^\perp \mathbb{E}_{\boldsymbol{x},y}\left[\sum_{j=0}^{r} u_j\big(b_j \mapsto \psi_\eta(y,b)\big)\mathsf{He}_j(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\boldsymbol{x}\right] \\
&= \boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{w}}^\perp \mathbb{E}_{\boldsymbol{x},\zeta}\left[\sum_{j=0}^{r}\sum_{i=0}^{qr} u_i\Big(a \mapsto u_j\big(b \mapsto \psi_\eta(\sigma_*(a)+\zeta, b)\big)\Big)\mathsf{He}_i(\langle \boldsymbol{x}, \boldsymbol{\theta}_*\rangle)\mathsf{He}_j(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\boldsymbol{x}\right] \\
&= \sum_{i=1}^{qr}\sum_{j=0}^{r} \mathbb{E}_{\boldsymbol{x},\zeta}\left[u_i\Big(a \mapsto u_j\big(b \mapsto \psi_\eta\big(\sigma_*(a)+\zeta, b\big)\big)\Big) i\mathsf{He}_{i-1}(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\mathsf{He}_j(\langle \boldsymbol{x}, \boldsymbol{\theta}_*\rangle)\right]\langle \boldsymbol{\theta}_*, \boldsymbol{P}_{\boldsymbol{w}}^\perp \boldsymbol{\theta}_*\rangle \\
&= \sum_{i=1}^{r} \mu_i \langle \boldsymbol{\theta}_*, \boldsymbol{w}\rangle^{i-1}\big(1 - \langle \boldsymbol{\theta}_*, \boldsymbol{w}\rangle^2\big),
\end{aligned} \tag{C.13}$$

where the fourth line uses Stein's Lemma and the fact $\boldsymbol{P}_{\boldsymbol{w}}^\perp \boldsymbol{w} = \mathbf{0}$. Thus, the size of the update will be dictated by the first index i_* such that $|\mu_i|\langle \boldsymbol{\theta}_*, \boldsymbol{w}\rangle^{i-1}$ is largest. Moreover, the centred random variable $\langle \boldsymbol{\theta}_*, \boldsymbol{g}\rangle - \mathbb{E}[\langle \boldsymbol{\theta}_*, \boldsymbol{g}\rangle]$ is sub-Weibull with constant order tail parameter since Gaussian random variables are sub-Weibull and the latter class is closed under polynomial transformation. (See [40] for more details on sub-Weibull random variables).

We must also control the normalization error from Lemma 12:

$$\gamma^2 \kappa^{(t)} ||\boldsymbol{g}||^2 + \gamma^3 |\langle \boldsymbol{\theta}_*, \boldsymbol{g}\rangle|\, ||\boldsymbol{g}||^2. \tag{C.14}$$

Note that

$$\begin{aligned}
||\boldsymbol{g}||^2 &= ||\boldsymbol{P}_{\boldsymbol{w}}^\perp \boldsymbol{x}||^2 \big|\psi_\eta(y, \langle \boldsymbol{x}, \boldsymbol{w}\rangle)\big|^2 \\
&= \left(\sum_{j=0}^{r} u_j\big(b \mapsto \psi_\eta(y,b)\big)\mathsf{He}_j(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\right)(||\boldsymbol{x}||^2 - \langle \boldsymbol{x}, \boldsymbol{w}\rangle^2),
\end{aligned} \tag{C.15}$$

and therefore,

$$\mathbb{E}\big[||\boldsymbol{g}||^2\big] = \mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{j=0}^{r} u_j\big(b \mapsto \psi_\eta(y,b)\big)\mathsf{He}_j(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\right)(d - \langle \boldsymbol{x}, \boldsymbol{w}\rangle^2)\right] \lesssim d. \tag{C.16}$$

By the same token, we use our derivation in (C.13) to argue $\mathbb{E}[|\langle \boldsymbol{\theta}_*, \boldsymbol{g}\rangle|||\boldsymbol{g}||^2] \lesssim d$. The error (C.15) is a sub-Weibull random variable with tail parameter proportional to $\gamma^2 d$.

By Lemma 12, this implies that the one step dynamics take the form

$$\kappa^{(t+1)} \geq \kappa^{(t)} + \gamma \mathbb{E}[\langle \boldsymbol{\theta}_*^{(t)}, \boldsymbol{g}^{(t)}\rangle] + \gamma \nu^{(t)} - C_2 \gamma^2 \kappa^{(t)}(d + \xi^{(t)}). \tag{C.17}$$

for some positive constant C_2 and sub-Weibull random variables (with constant order parameter) $\nu^{(t)}$, $\xi^{(t)}$ that are independent of previous iterations. Now, choosing $\gamma \leq C \max_{1\leq i\leq r} \mu_i d^{-(\frac{i}{2}\vee 1)}$ and recalling $\kappa^{(t)} \geq d^{-\frac{1}{2}}$ leads to

$$C_2 \gamma^2 \kappa^{(t)} d \leq C_2 C \gamma \kappa^{(t)} \max_{1\leq i\leq r} \mu_i d^{-(\frac{i-2}{2}\vee 0)} \leq C_2 C \gamma \max_{1\leq i\leq r} \mu_i (\kappa^{(t)})^{(i-1)\vee 1}, \tag{C.18}$$

which can be made a sufficiently small constant multiple of $\gamma \max_{1\le i\le r} \mu_i (\kappa^{(t)})^{i-1}$ with an appropriate choice of C. Hence, the expected one-step normalization error can be absorbed into the expected one-step population dynamics (C.13). Furthermore, since the constraint on γ also implies $\gamma d \lesssim 1$, we may absorb the noise $\gamma^2 d \xi^{(t)}$ into the $\gamma \nu^{(t)}$ noise term.

This leaves us with the alignment dynamics

$$\kappa^{(t+1)} \ge \kappa^{(t)} + \gamma C_1 \sum_{i=1}^{r} \mu_i (\kappa^{(t)})^{i-1} \big(1 - (\kappa^{(t)})^2\big) + \gamma \nu^{(t)}. \tag{C.19}$$

■

C.4. Sample Complexity Upper Bound

Proposition 14 (Generic Sample Complexity Upper Bound) *Fix $c \gtrsim \frac{1}{\operatorname{polylog} d}$. Suppose $\langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(0)} \rangle \ge C_0 d^{-\frac{1}{2}}$ for some $C_0 > 0$. Then, there exists $C \gtrsim \frac{1}{\operatorname{polylog} d}$ such that for any $\delta \in (0,1)$, setting $\gamma \le C\delta \max_{1\le i\le r} \mu_i d^{-(\frac{i}{2}\vee 1)}$ gives $\langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(t)} \rangle \gtrsim \frac{1}{\operatorname{polylog} d}$ within*

$$T(\eta) = \min_{\substack{1\le i\le r\\ \mu_i>0}} \tilde{\Theta}\Big(\gamma^{-1}\big(\mu_i(\eta)\big)^{-1} d^{\frac{i-2}{2}\vee 0}\Big) \tag{C.20}$$

iterations with probability at least $1-\delta$.

Proof Unrolling the recurrence from Lemma 13,

$$\kappa^{(t)} \ge \kappa^{(0)} + \gamma C_1 \sum_{i=1}^{r} \sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1} \big(1 - (\kappa^{(s)})^2\big) - \gamma \Bigg| \sum_{s=0}^{t-1} \nu^{(s)} \Bigg|. \tag{C.21}$$

Let T denote the weak recovery time in Theorem 5. Since $\nu^{(t)}$ is sub-Weibull, we have, for some constant $C_3 > 0$,

$$\mathbb{E}\Bigg[\Bigg|\sum_{s=0}^{T-1} \nu^{2s}\Bigg|^2\Bigg] = \sum_{s=0}^{T-1} \mathbb{E}\big[|\nu^{2s}|^2\big] \le C_3 T. \tag{C.22}$$

Moreover,

$$\begin{aligned}
\mathbb{P}\Bigg(\max_{0\le t\le T-1} \Bigg|\sum_{s=0}^{t} \nu^{2s}\Bigg|^2 \ge 4C_3\delta^{-1}T\Bigg) &\le \frac{\delta}{4C_3T}\mathbb{E}\Bigg[\max_{0\le t\le T-1}\Bigg|\sum_{s=0}^{t} \nu^{2s}\Bigg|^2\Bigg] \quad \text{by Markov's inequality} \\
&\le \frac{\delta}{C_3T}\mathbb{E}\Bigg[\Bigg|\sum_{s=0}^{T-1} \nu^{2s}\Bigg|^2\Bigg] \quad \text{by Doob's maximal inequality.}
\end{aligned} \tag{C.23}$$

Assume without loss of generality that $\kappa \ge 2d^{-\frac{1}{2}}$. Our bound on the dynamics after t updates becomes, with high probability,

$$\begin{aligned}
\kappa^{(t)} &\ge 2d^{-\frac{1}{2}} + \gamma C_1 \sum_{i=1}^{r}\sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1}(1-(\kappa^{(s)})^2) - 2\gamma C_3^{\frac{1}{2}}\delta^{-\frac{1}{2}}T^{\frac{1}{2}} \\
&= 2d^{-\frac{1}{2}} + \gamma C_1 \sum_{i=1}^{r}\sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1}(1-(\kappa^{(s)})^2) - \gamma^{\frac{1}{2}} C_4 \delta^{-\frac{1}{2}} \min_{\substack{1\le i\le r\\ \mu_i>0}} \mu_i^{-\frac{1}{2}} d^{-(\frac{i-2}{4}\vee 0)}.
\end{aligned} \tag{C.24}$$

for some $C_4 = \tilde{\Theta}(1)$. Now, recalling that $\gamma \leq C\delta \max_{1\leq i\leq r} \mu_i d^{-(\frac{i}{2}\vee 1)}$, we have

$$\begin{aligned}\gamma^{\frac{1}{2}} C_4 \delta^{-\frac{1}{2}} \min_{\substack{1\leq i\leq r\\ \mu_i>0}} \mu_i^{-(\frac{i-2}{4}\vee 0)} &\leq C^{\frac{1}{2}} \left(\min_{\substack{1\leq i\leq r\\ \mu_i>0}} \mu_i^{-\frac{1}{2}} d^{-(\frac{i-2}{4}\vee 0)} \right) \max_{1\leq i\leq r} \mu_i^{\frac{1}{2}} d^{-(\frac{i}{4}\vee\frac{1}{2})} \\ &\leq C^{\frac{1}{2}} C_4 d^{-\frac{1}{2}},\end{aligned} \tag{C.25}$$

which can be made less than $d^{-\frac{1}{2}}$ by choosing C sufficiently small. Hence, our final (high probability) upper bound for the multi-step dynamics is

$$\kappa^{(t)} \geq d^{-\frac{1}{2}} + \gamma C_1 \sum_{i=1}^{r} \sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1} \big(1 - (\kappa^{(s)})^2\big). \tag{C.26}$$

Now, we unroll each of the r terms in the expected dynamics and determine how quickly each one reaches $c \asymp \frac{1}{\operatorname{polylog} d}$. Consider terms where $\mu_i > 0$. Since $\kappa^{(t)} \lesssim \frac{1}{\operatorname{polylog} d}$ by assumption, we may absorb the factor $(1 - (\kappa^{(t)})^2)$ into the constant C_1 (abusing notation). On the other hand, the contributions of terms with $\mu_i \leq 0$ will be negligible by Assumption 4.

For the $i = 1$ term, the noiseless dynamics give

$$\begin{aligned} & d^{-\frac{1}{2}} + \gamma C_1 \mu_i t \geq 2c \\ \iff \; & t \geq \gamma^{-1} C_1^{-1} \mu_i^{-1} (2c - d^{-\frac{1}{2}}) = \Theta(\gamma^{-1}\mu_i^{-1}). \end{aligned} \tag{C.27}$$

When $i = 2$, we have, by Grönwall's Inequality (Lemma 8)

$$\begin{aligned} d^{-\frac{1}{2}} + \gamma C_1 \mu_i \sum_{s=0}^{t-1} \kappa^{(s)} \geq d^{-\frac{1}{2}} \big(1 + \gamma C_1 \mu_i\big)^t \geq 2c \\ \iff t \log\big(1 + \gamma C_1 \mu_i\big) \geq \log 2c + \frac{1}{2} \log d \\ \iff t \geq \frac{\log 2c + \frac{1}{2}\log d}{\log(1 + \gamma C_1 \mu_i)} = \tilde{\Theta}(\gamma^{-1}\mu_i^{-1}), \end{aligned} \tag{C.28}$$

where the final equality follows from the fact $x - \frac{x^2}{2} \leq \log(1 + x) \leq x$ for $x \in (0, 1)$.

Lastly, for $i \geq 3$, we have, from the Bihari-LaSalle inequality (Lemma 9),

$$\begin{aligned} d^{-\frac{1}{2}} + \gamma C_1 \mu_i \sum_{s=0}^{t-1} (\kappa^{(s)})^{i-1} &\geq \frac{d^{-\frac{1}{2}}}{(1 - \frac{1}{2}\gamma C_1 \mu_i d^{-\frac{i-2}{2}} t)^{\frac{1}{i-2}}} \geq 2c \\ \iff d^{-\frac{i-2}{2}} &\geq (2c)^{i-2} - \tfrac{1}{2}(2c)^{i-2} \gamma C_1 \mu_i d^{-\frac{i-2}{2}} t \\ \iff t &\geq 2\gamma^{-1} C_1^{-1} \mu_i^{-1} d^{\frac{i-2}{2}} \big((2c)^{i-2} - d^{-\frac{i-2}{2}}\big) = \Theta(\gamma^{-1}\mu_i^{-1} d^{\frac{i-2}{2}}). \end{aligned} \tag{C.29}$$

Thus, the maximum weak recovery time is indeed

$$T = \min_{\substack{1\leq i\leq r\\ \mu_i>0}} \Theta(\gamma^{-1}\mu_i^{-1} d^{\frac{i-2}{2}\vee 0}). \tag{C.30}$$

■

C.5. Sample Complexity Lower Bound

The proof of the matching sample complexity lower bound proceeds much in the same way as that of the upper bound.

Proposition 15 (Generic Sample Complexity Lower Bound) *Fix* $c \gtrsim \frac{1}{\operatorname{polylog} d}$*. Suppose that* $\langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(0)} \rangle \leq \tilde{C}_0 d^{-\frac{1}{2}}$ *for some* $\tilde{C}_0 > 0$*. Then, there exists a constant* $\tilde{C} \gtrsim \frac{1}{\operatorname{polylog} d}$ *such that for all* $\delta \in (0,1)$*, setting* $\gamma \leq \tilde{C}\delta \max_{1\leq i\leq r} \mu_i d^{-(\frac{i}{2}\vee 1)}$ *gives* $\langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(t)} \rangle < c$ *for all iterations* $t \leq T$ *of* (3.1)*, where*

$$T(\eta) = \min_{\substack{1\leq i\leq r \\ \mu_i > 0}} \tilde{\Theta}\big(\gamma^{-1}\big(\mu_i(\eta)\big)^{-1} d^{\frac{i-2}{2}\vee 0}\big), \tag{C.31}$$

with probability at least $1-\delta$*.*

Proof [Proof of Proposition 15] The projection error is trivial to handle, as we obtain

$$\kappa^{(t+1)} = \frac{\kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)} \rangle}{||\boldsymbol{w}^{(t)} + \gamma \boldsymbol{g}^{(t)}||} \leq \kappa^{(t)} + \gamma\langle \boldsymbol{\theta}_*, \boldsymbol{g}^{(t)} \rangle, \tag{C.32}$$

since $\langle \boldsymbol{w}^{(t)}, \boldsymbol{g}^{(t)} \rangle = 0$ and $||\boldsymbol{w}|| = 1$. Moreover, from Section C.3, the expected one-step update to the alignment is given by

$$\mathbb{E}[\langle \boldsymbol{\theta}_*, \boldsymbol{g} \rangle] = \sum_{k=1}^{r} \mu_k \langle \boldsymbol{\theta}_*, \boldsymbol{w} \rangle^{k-1} \big(1 - \langle \boldsymbol{\theta}_*, \boldsymbol{w} \rangle^2\big). \tag{C.33}$$

Therefore, the full dynamics are

$$\begin{aligned} \kappa^{(t+1)} &\leq \kappa^{(t)} + \gamma \sum_{i=1}^{r} \mu_i (\kappa^{(t)})^{i-1} (1 - (\kappa^{(t)})^2) + \gamma \nu^{(t)} \\ &\leq \kappa^{(0)} + \gamma \sum_{i=1}^{r} \sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1} (1 - (\kappa^{(s)})^2) + \gamma \Bigg| \sum_{s=0}^{t-1} \nu^{(s)} \Bigg| \\ &\leq \tilde{C}_0 d^{-\frac{1}{2}} + \gamma \sum_{i=1}^{r} \sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1} (1 - (\kappa^{(s)})^2) + \gamma C_3^{\frac{1}{2}} \delta^{-\frac{1}{2}} T^{\frac{1}{2}} \\ &\leq 2\tilde{C}_0 d^{-\frac{1}{2}} + \gamma \sum_{i=1}^{r} \sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1} (1 - (\kappa^{(s)})^2), \end{aligned} \tag{C.34}$$

where the third line follows from the martingale bound in the previous subsection, and the last line follows from the constraint $\gamma \leq \tilde{C}\delta \max_{1\leq i\leq r} \mu_i d^{-(\frac{i}{2}\vee 1)}$ with $\tilde{C}$ taken sufficiently small. Then, finding the minimum weak recovery time proceeds exactly as in the previous section, using Grönwall's inequality for $i = 2$ and the Bihari-LaSalle inequality for $i \geq 3$, once again giving

$$T = \min_{\substack{1\leq i\leq r \\ \mu_i > 0}} \tilde{\Theta}\big(\gamma^{-1} \mu_i^{-1} d^{\frac{i-2}{2}\vee 0}\big). \tag{C.35}$$

■

Together, the sample complexity upper and lower bounds imply Theorem 5.

C.6. Strong Recovery

Now, starting with $\boldsymbol{w}$ that has achieved weak recovery, we characterize the maximum number T' of subsequent updates of the form (3.1) required to achieve strong recovery with high probability.

Proposition 16 (Strong Recovery Given Weak Recovery) *Let $\varepsilon > 0$. Suppose that $\langle \boldsymbol{\theta}_*, \boldsymbol{w}^{(0)} \rangle \geq 2c$ for some $c \gtrsim \frac{1}{\operatorname{polylog} d}$. Then, there exists a constant $C > 0$ such that for all $\delta \in (0,1)$, setting $\gamma \leq C\delta d^{-1}\varepsilon \max_{1\leq i\leq r} \mu_i c^{i-1}$ implies that the update rule (3.1) achieves $\langle \boldsymbol{\theta}_*, \boldsymbol{w} \rangle \geq 1-\varepsilon$ within*

$$T' = \min_{\substack{1\leq i\leq r\\ \mu_i>0}} \tilde{\Theta}(\gamma^{-1}\varepsilon^{-1}\mu_i^{-1}) \tag{C.36}$$

iterations with probability at least $1-\delta$.

Remark 17 *If $\varepsilon = \tilde{\Theta}(1)$, then $T' \lesssim T$. That is, achieving weak recovery is the bottleneck during training.*

Remark 18 *In the algorithms we consider in Section 4, we have $\max_{1\leq i\leq r} \mu_i = \Theta(1)$. Therefore, given that weak recovery has already been achieved, then strong recovery for $\varepsilon = \tilde{\Theta}(1)$ proceeds after at most $\tilde{\Theta}(d)$ additional iterations with high probability when $\gamma \asymp d^{-1}$.*

Proof Similarly to Section C.3, we have a lower bound on the one-step dynamics:

$$\kappa^{(t+1)} \geq \kappa^{(t)} + \gamma \sum_{i=1}^{r} \mu_i (\kappa^{(t)})^{i-1}\big(1-(\kappa^{(t)})^2\big) + \gamma\nu^{(t)} - C_2\gamma^2 d. \tag{C.37}$$

Since $\langle \boldsymbol{\theta}, \boldsymbol{w}^{(t)} \rangle \leq 1-\varepsilon$ and Assumption 4 holds, we can re-write this as

$$\kappa^{(t+1)} \geq \kappa^{(t)} + \gamma C_1 \varepsilon \sum_{i=1}^{r} \mu_i (\kappa^{(t)})^{i-1} + \gamma\nu^{(t)} - C_2\gamma^2 d \tag{C.38}$$

for some constant $C_1 > 0$. Setting $\gamma \leq C\delta d^{-1}\varepsilon$ leads to

$$C_2\gamma^2\kappa^{(t)} d \leq C_2 C\delta\varepsilon\gamma \max_{1\leq i\leq r} \mu_i c^{i-1} \leq C_2 C\gamma\delta\varepsilon \max_{1\leq i\leq r} \mu_i (\kappa^{(t)})^{i-1}. \tag{C.39}$$

Thus, taking C sufficiently small ensures that this is a fraction of the dominant term in the population update.

Unrolling this over t steps, we obtain

$$\begin{aligned}
\kappa^{(t)} &\geq 2c + \gamma C_1 \varepsilon \sum_{i=1}^{r}\sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1} - \gamma \left| \sum_{s=0}^{t-1} \nu^{(s)} \right| \\
&\geq 2c + \gamma C_1 \varepsilon \sum_{i=1}^{r}\sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1} - \gamma C_3^{\frac{1}{2}} \delta^{-\frac{1}{2}} T' \\
&= 2c + \gamma C_1 \varepsilon \sum_{i=1}^{r}\sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1} - \gamma^{\frac{1}{2}} C_4 \delta^{-\frac{1}{2}} \varepsilon^{-\frac{1}{2}} \min_{\substack{1\leq i\leq r\\ \mu_i>0}} \mu_i^{-\frac{1}{2}}
\end{aligned} \tag{C.40}$$

using the same martingale bound as in C.4. Now recalling $\gamma \leq C\delta d^{-1}\varepsilon \max_{1\leq i\leq r} \mu_i c^{i-1}$, we have

$$\gamma^{\frac{1}{2}} C_4 \delta^{-\frac{1}{2}} \varepsilon^{-\frac{1}{2}} \min_{\substack{1\leq i\leq r\\ \mu_i>0}} \mu_i^{-\frac{1}{2}} \leq C^{\frac{1}{2}} C_4 d^{-\frac{1}{2}}, \tag{C.41}$$

which is of lower order than c. Hence, our final (high probability) upper bound for the multi-step dynamics is

$$\kappa^{(t)} \geq c + \gamma C_1 \varepsilon \sum_{i=1}^{r} \sum_{s=0}^{t-1} \mu_i (\kappa^{(s)})^{i-1}. \tag{C.42}$$

We analyze how quickly this exceeds $1-\varepsilon$. For the $i=1$ term, we obtain

$$\begin{aligned} &\kappa^{(t)} \geq c + \gamma C_1 \varepsilon \mu_1 t \geq 1-\varepsilon \\ \iff &t \geq (1-\varepsilon-c) C_1^{-1} \gamma^{-1} \varepsilon^{-1} \mu_1^{-1} = \Theta(\gamma^{-1}\varepsilon^{-1}\mu_1^{-1}). \end{aligned} \tag{C.43}$$

For the $i=2$ term, we have, by Grönwall's inequality

$$\begin{aligned} &\kappa^{(t)} \geq c + \gamma C_1 \varepsilon \mu_2 \sum_{s=0}^{t-1} \kappa^{(s)} \geq c(1+\gamma C_1 \varepsilon \mu_2)^t \geq 1-\varepsilon \\ \iff &t \geq \frac{\log(1-\varepsilon) - \log c}{\log(1+\gamma C_1 \varepsilon \mu_2)} = \tilde{\Theta}(\gamma^{-1}\varepsilon^{-1}\mu_2^{-1}). \end{aligned} \tag{C.44}$$

For the $i \geq 3$ terms, we have, by the Bihari-LaSalle inequality,

$$\begin{aligned} &\kappa^{(t)} \geq c + \gamma C_1 \varepsilon \mu_i \sum_{s=0}^{t-1} \left(\kappa^{(s)}\right)^{i-1} \geq \frac{c}{(1-\frac{1}{2}\gamma C_1 \varepsilon \mu_i c^{i-2} t)^{\frac{1}{i-2}}} \geq 1-\varepsilon \\ \iff &t \geq 2\big(1-(\tfrac{c}{1-\varepsilon})^{i-2}\big) \gamma^{-1} C_1^{-1} \varepsilon^{-1} \mu_i^{-1} c^{-(i-2)} = \Theta(\gamma^{-1}\varepsilon^{-1}\mu_i^{-1}). \end{aligned} \tag{C.45}$$

Hence, the (high probability) maximum strong recovery time given weak recovery is indeed

$$T' = \min_{\substack{1\leq i\leq r\\ \mu_i>0}} \Theta(\gamma^{-1}\varepsilon^{-1}\mu_i^{-1}). \tag{C.46}$$

■

C.7. Ridge Regression on the Second Layer

For completeness, we state the following result from [27] that outlines the sample complexity of proceeding from strong recovery to approximation of the target to arbitrary accuracy via ridge regression. In particular, if the error tolerance is of constant order, then the sample complexity obtaining strong recovery (from weak recovery) is strictly larger.

Proposition 19 (Second Layer Training, [27, Lemma 20]) *Let $\varepsilon > 0$ and $N = \tilde{\Theta}(\varepsilon^{-1})$. Suppose that $\tilde{\Theta}(N)$ neurons in (2.2) satisfy $\langle \boldsymbol{\theta}_*, \boldsymbol{w}_j \rangle \geq 1-\varepsilon$. Let $b_j \sim \mathrm{Unif}([-C_b, C_b])$ such that $C_b =$*

$\tilde{O}(1)$. *Then, there exists a choice of penalty parameter* $\lambda = \tilde{\Theta}(1)$ *such that the solution* $\hat{\boldsymbol{a}} = (\hat{a}_1, \ldots, \hat{a}_N)$ *of ridge regression with* $\tilde{\Theta}(N^{-4} + \varepsilon^{-4})$ *samples satisfies*

$$\mathbb{E}_{x \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)}\left[\left|\frac{1}{N}\sum_{j=1}^{N} \hat{a}_j \sigma_j(\langle \boldsymbol{x}, \boldsymbol{w}_j\rangle + b_j) - \sigma_*(\langle \boldsymbol{x}, \boldsymbol{\theta}_*\rangle)\right|^2\right] \lesssim \varepsilon^2. \tag{C.47}$$

with high probability.

The key assumption in the above is that a constant proportion (up to polylogarithmic factors) of the neurons achieve strong recovery. Recall that the initial alignment is sufficiently large with constant order probability (Section C.1), and that each of weak (Section C.4) and strong (Section C.6) recovery occur with high probability given such an initialization.

Appendix D. SGD Variants

Recall that, from our proof of Theorem 5 in the previous section, the coefficients

$$\mu_i := \mathbb{E}_\zeta[\hat{\mu}_i(\zeta)] := \mathbb{E}_\zeta\left[u_i\Big(a \mapsto u_{i-1}\big(b \mapsto \psi_\eta(\sigma_*(a) + \zeta, b)\big)\Big)\right], \; i \in [r]. \tag{D.1}$$

are the key quantities governing the sample complexity of an online algorithm that fits in our framework. We detail the computation of these coefficients for each of the three SGD variants we consider in this work: online SGD (Section D.1), alternating SGD (Section D.2), and batch reuse SGD (Section D.4). This along with Theorem 5 immediately imply the corollaries in Section 4 on the sample complexity of these algorithms. Additionally, in Section D.3, we investigate how alternating SGD can be generalized to an online algorithm for a D-layer neural network and calculate the μ_i.

D.1. Online SGD

Given a fresh data point $(\boldsymbol{x}, y)$, the spherical vanilla online SGD update has the form

$$\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} + \gamma y \sigma'(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\boldsymbol{P}^{\perp}_{\boldsymbol{w}^{(t)}}\boldsymbol{x}, \quad \boldsymbol{w}^{(t+1)} \leftarrow \frac{\boldsymbol{w}^{(t+1)}}{||\boldsymbol{w}^{(t+1)}||}. \tag{D.2}$$

Therefore, under our general framework introduced in Section 3, the update oracle is

$$\psi_\eta(y, z) = y\sigma'(z). \tag{D.3}$$

Hence, for $i \in [r]$,

$$u_{i-1}\big(b \mapsto \psi_\eta(y, b)\big) = \mathbb{E}_{b \sim \mathcal{N}(0,1)}[y\sigma'(b)\mathsf{He}_{i-1}(b)] = yu_{i-1}(\sigma') = yiu_i(\sigma), \tag{D.4}$$

which leads to

$$\begin{aligned}\hat{\mu}_i(\zeta) &= u_i\big(a \mapsto (\sigma_*(a) + \zeta)iu_i(\sigma)\big)\\ &= iu_i(\sigma)\Big(\mathbb{E}_{a \sim \mathcal{N}(0,1)}[\sigma_*(a)\mathsf{He}_i(a)] + \mathbb{E}_{a \sim \mathcal{N}(0,1)}[\zeta\mathsf{He}_i(a)]\Big)\\ &= iu_i(\sigma)u_i(\sigma_*).\end{aligned} \tag{D.5}$$

Note that the second expectation in the second line is zero since $\mathbb{E}_{a \sim \mathcal{N}(0,1)}[\mathsf{He}_i(a)] = 0$ for all $i \in \mathbb{N}$. Since the above has no dependence on ζ, it is immediate that $\mu_i = iu_i(\sigma)u_i(\sigma_*)$.

D.2. Alternating SGD

The alternating SGD (Algorithm 1) update for a single neuron is

$$\tilde{a}^{(t+1)} \leftarrow a + \eta y \sigma(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)} \rangle), \quad \boldsymbol{w}^{(t+1)} \leftarrow \frac{\boldsymbol{w}^{(t)} + \gamma y \tilde{a}^{(t+1)} \sigma'(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)} \rangle)}{||\boldsymbol{w}^{(t)} + \gamma y \tilde{a}^{(t+1)} \sigma'(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)} \rangle)||}. \tag{D.6}$$

Note that we only use the second layer update $\tilde{a}$ in order to update the first layer parameters $\boldsymbol{w}$. We do not replace the second layer parameter with $\tilde{a}$ at the subsequent iteration, but instead keep a. For simplicity, assume that we have $a = 1$. Then,

$$\psi_\eta(y, z) = y\sigma'(z) + \eta y^2 \sigma(z)\sigma'(z). \tag{D.7}$$

Hence, for $i \in [r]$,

$$\begin{aligned} u_{i-1}\big(b \mapsto \psi_\eta(y, b)\big) &= y\mathbb{E}_{b\sim\mathcal{N}(0,1)}[\sigma'(b)\mathsf{He}_{i-1}(b)] + \eta y^2 \mathbb{E}_{b\sim\mathcal{N}(0,1)}[\sigma(b)\sigma'(b)\mathsf{He}_{i-1}(b)] \\ &= yiu_i(\sigma) + \eta y^2 u_{i-1}(\sigma\sigma'), \end{aligned} \tag{D.8}$$

which leads to (using the result of our calculation in the previous subsection)

$$\begin{aligned} \hat{\mu}_i(\zeta) &= iu_i(\sigma)u_i(\sigma_*) + u_i\big(a \mapsto \eta(\sigma_*(a) + \zeta)^2 u_{i-1}(\sigma\sigma')\big) \\ &= iu_i(\sigma)u_i(\sigma_*) + \eta u_{i-1}(\sigma\sigma')\big(\mathbb{E}_{a\sim\mathcal{N}(0,1)}[\sigma_*^2(a)\mathsf{He}_i(a)] + 2\zeta\mathbb{E}_{a\sim\mathcal{N}(0,1)}[\sigma_*(a)\mathsf{He}_i(a)]\big). \end{aligned} \tag{D.9}$$

Simplifying and taking expectation with respect to ζ (which is mean-zero) gives

$$\mu_i = iu_i(\sigma)u_i(\sigma_*) + \eta u_{i-1}(\sigma\sigma')u_i(\sigma_*^2). \tag{D.10}$$

D.3. "Deep" Alternating SGD

We describe a natural extension of alternating SGD to neural networks with $D > 2$ layers. We define a D-layer neural network student by the recurrence

$$f(\boldsymbol{x}) = f_{D-1}(\boldsymbol{x}), \quad f_0(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}, \quad f_i(\boldsymbol{x}) = \boldsymbol{A}_i\sigma(f_{i-1}(\boldsymbol{x})), i \in [D-1], \tag{D.11}$$

where $\boldsymbol{W}_0 \in \mathbb{R}^{N\times d}$ as before and $\boldsymbol{A}_i \in \mathbb{R}^{N_{i+1}\times N_i}$ such that $N_1 = N$ and $N_D = 1$. We are still interested in recovery of the ground truth direction $\boldsymbol{\theta}$ by the first-layer weights $\boldsymbol{W}$. To make the theoretical analysis tractable, we consider the simplified sparse network where $N_1 = N_2 = \cdots = N_{D-1} = N$, $\boldsymbol{A}_1$ is a $N_2 \times N_1$ matrix of ones[1], and $\boldsymbol{A}_2 = \boldsymbol{A}_3 = \cdots = \boldsymbol{A}_{D-1} = \boldsymbol{I}_N$ with off-diagonal entries frozen at zero during training (i.e., they do not receive gradients). Hence, the output of the network is of the form

$$f(\boldsymbol{x}) = \sum_{j=1}^{N} a_j^{(D-1)}\sigma\big(\circ\cdots\circ\sigma(a_j^{(1)}\sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j\rangle))\big). \tag{D.12}$$

Hence, to analyze weak recovery, it suffices to focus on a single summand (where we drop the subscript j for convenience):

$$a^{(D-1)}\sigma\big(\circ\cdots\circ\sigma(a^{(1)}\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle))\big), \tag{D.13}$$

[1] Note that we chose the same initialization for our two-layer network in the previous subsection.

which we express as the recurrence

$$F(z) = F_{D-1}(z), \quad z = F_0(z) = \langle \boldsymbol{x}, \boldsymbol{w} \rangle, \quad F_i(z) = a^{(i)}\sigma\big(F_{i-1}(z)\big), i \in [D-1]. \tag{D.14}$$

We propose the following update rule inspired by our alternating SGD algorithm:

$$\begin{aligned} z^{(t)} &\leftarrow \langle \boldsymbol{x}, \boldsymbol{w}^{(t)} \rangle \\ \tilde{a}^{(i)} &\leftarrow a^{(i)} + \eta y \Bigg(\prod_{j=i+1}^{D-1} a^{(i)} \sigma'(F_{j-1}(z^{(t)})) \Bigg) \sigma\big(F_{i-1}(z^{(t)})\big) \\ \boldsymbol{w}^{(t+1)} &\leftarrow \boldsymbol{w}^{(t)} + \gamma y \Bigg(\prod_{i=1}^{D-1} \tilde{a}^{(i)} \sigma'\big(F_{i-1}(z^{(t)})\big) \Bigg) \boldsymbol{P}^{\perp}_{\boldsymbol{w}^{(t)}} \boldsymbol{x}, \quad \boldsymbol{w}^{(t+1)} \leftarrow \frac{\boldsymbol{w}^{(t+1)}}{||\boldsymbol{w}^{(t+1)}||}. \end{aligned} \tag{D.15}$$

Expanding the update for $\boldsymbol{w}$ (before normalization) gives

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w} + \gamma y \prod_{i=1}^{D-1} \Bigg[\Bigg(a^{(i)} + \eta y \Bigg(\prod_{j=i+1}^{D-1} a^{(j)} \sigma'(F_{j-1}(z)) \Bigg) \sigma(F_{i-1}(z)) \Bigg) \sigma'(F_{i-1}(z)) \Bigg] \boldsymbol{P}^{\perp}_{\boldsymbol{w}} \boldsymbol{x}, \tag{D.16}$$

where we have omitted the superscript (t) on the right-hand side for readability. This fits into our framework (3.1) since the $a^{(i)}$ remain constant. In fact, for simplicity, we may fix all $a_i = 1$ for all $i \in [D-1]$. Our update oracle is then

$$\begin{aligned} \psi_\eta(y, z) = \sum_{i=0}^{D-1} \Bigg[\eta^i y^{i+1} &\sum_{S \in \mathcal{P}_i([D-1])} \Bigg(\prod_{j \notin S} \sigma'(\sigma^{\circ(j-1)}(z)) \Bigg) \\ &\cdot \Bigg(\prod_{k \in S} \Bigg(\prod_{l=k+1}^{D-1} \sigma'(\sigma^{\circ(l-1)}(z)) \Bigg) \sigma^{\circ k}(z) \sigma'(\sigma^{\circ(k-1)}(z)) \Bigg) \Bigg], \end{aligned} \tag{D.17}$$

where $\mathcal{P}_i([D-1])$ denotes the set of all subsets of $[D-1]$ of cardinality i. Now, assuming that the Hermite coefficients of the relevant compositions and products of σ and σ' are positive, this gives

$$\mu_i \asymp \sum_{j=1}^{D} \eta^{j-1} u_i(\sigma_*^j). \tag{D.18}$$

Under an optimal choice of γ, Theorem 5 implies a sample complexity of

$$T = \max_{1 \le i \le D} \tilde{\Theta}\big(\eta^{-2(i-1)} d^{(p_i - 1) \vee 1}\big) \tag{D.19}$$

for deep alternating SGD to attain weak recovery.

Now, the positivity of the relevant Hermite coefficients is a nontrivial assumption. We focus on a tractable special case where $D = 3$, and $\sigma(z) = z^2$. In this case, we have

$$\begin{aligned} \psi_\eta(y, z) &= y\sigma'(z)\sigma'\big(\sigma(z)\big) + \eta y^2 \Big([\sigma'\big(\sigma(z)\big)]^2 \sigma(z)\sigma'(z) + \sigma'(z)\sigma\big(\sigma(z)\big)\sigma'\big(\sigma(z)\big) \Big) \\ &\quad + \eta^2 y^3 \Big([\sigma'(\sigma(z))]^2 \sigma(\sigma(z))\sigma(z)\sigma'(z) \Big) \\ &\asymp yz^3 + \eta y^2 z^7 + \eta^2 y^3 z^{11} \end{aligned} \tag{D.20}$$

and therefore

$$u_{i-1}\big(b \mapsto \psi_\eta(y,b)\big) = y u_{i-1}(b^3) + \eta y^2 u_{i-1}(b^7) + \eta^2 y^3 u_{i-1}(b^{11}), \tag{D.21}$$

which leads to

$$\mu_i \asymp u_{i-1}(b^3) u_i(\sigma_*) + \eta u_{i-1}(b^7) u_i(\sigma_*^2) + \eta^2 u_{i-1}(b^{11}) u_i(\sigma_*^3). \tag{D.22}$$

It is immediate that weak recovery is achieved with $\tilde{\Theta}(d)$ complexity if $\eta = \tilde{\Theta}(1)$ and at least one of the following holds: $u_2(\sigma_*) > 0$, $u_2(\sigma_*^2) > 0$, $u_2(\sigma_*^3) > 0$. This is a strict improvement over alternating SGD on a two-layer neural network when p and p_2 are larger than 2 but $p_3 = 2$ *and* over batch reuse SGD under the same conditions and quadratic σ.

Of course, the above has the important limitation that it will not recover in $\tilde{\Theta}(d)$ time if $u_2(\sigma_*^k) = 0$ for all $k \in \{1,2,3\}$ but $u_1(\sigma_*^k) > 0$ for at least one such k. This can be partially resolved by considering the activation $\sigma(z) = z^3$, in which case $u_0([\sigma'(\sigma(z))]^2 \sigma(\sigma(z))\sigma(z)\sigma'(z)) > 0$ and targets satisfying $p_3 = 1$ can be recovered in $\tilde{\Theta}(d)$ time. However, we cannot assume to know the target a priori, and a more generally applicable choice of σ is preferable (perhaps a randomized approach as in [27]). We leave a more thorough examination of potential choices of activation function to future work.

D.4. Batch Reuse SGD

The update for Batch Reuse SGD (Algorithm 2) takes the form

$$\tilde{\boldsymbol{w}}^{(t)} \leftarrow \boldsymbol{w}^{(t)} + \eta y \sigma'(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)}\rangle) \boldsymbol{P}_{\boldsymbol{w}^{(t)}} \boldsymbol{x}, \quad \boldsymbol{w}^{(t+1)} \leftarrow \frac{\boldsymbol{w}^{(t)} + \gamma y \sigma'(\langle \boldsymbol{x}, \tilde{\boldsymbol{w}}\rangle) \boldsymbol{P}^\perp_{\boldsymbol{w}^{(t)}} \boldsymbol{x}}{||\boldsymbol{w}^{(t)} + \gamma y \sigma'(\langle \boldsymbol{x}, \tilde{\boldsymbol{w}}\rangle) \boldsymbol{P}^\perp_{\boldsymbol{w}^{(t)}} \boldsymbol{x}||}. \tag{D.23}$$

Combining the two steps (and disregarding normalization for the time being), we have

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \gamma y \sigma'\big(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)}\rangle + \eta y \sigma'(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)}\rangle)\langle \boldsymbol{x}, \boldsymbol{P}^\perp_{\boldsymbol{w}^{(t)}} \boldsymbol{x}\rangle\big) \boldsymbol{P}^\perp_{\boldsymbol{w}^{(t)}} \boldsymbol{x} \tag{D.24}$$

The presence of $||\boldsymbol{x}||^2_{\boldsymbol{P}^\perp_{\boldsymbol{w}^{(t)}}}$ in the update prevents us from immediately casting this into our formalism. We handle this as follows. Using a Taylor expansion, we have

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \gamma y \sum_{k=1}^{r} \frac{\sigma^{(k)}(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)}\rangle) y^{k-1} \eta^{k-1} \sigma'(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)}\rangle)^{k-1} ||\boldsymbol{x}||^{2(k-1)}_{\boldsymbol{P}_{\boldsymbol{w}^{(t)}}}}{(k-1)!}. \tag{D.25}$$

Note that $||\boldsymbol{x}||^2_{\boldsymbol{P}_{\boldsymbol{w}^{(t)}}} \sim \chi^2_{d-1}$ and therefore $\mathbb{E}[||\boldsymbol{x}||^{2(i-1)}_{\boldsymbol{P}_{\boldsymbol{w}^{(t)}}}] = \Theta(d^{i-1})$. This, along with the assumption $\eta d \lesssim 1$, allows us to replace $||\boldsymbol{x}||^{2(i-1)}_{\boldsymbol{P}^{(t)}}$ in each term of the Taylor expansion with d^{i-1} and add a sub-Weibull remainder term $\xi^{(t)}$ with $O(1)$ tail parameter:

$$\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)} \asymp \gamma y \sum_{k=1}^{r} \sigma^{(k)}(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)}\rangle) y^{k-1} (\eta d)^{k-1} \big(\sigma'(\langle \boldsymbol{x}, \boldsymbol{w}^{(t)}\rangle)\big)^{k-1} \boldsymbol{P}_{\boldsymbol{w}^{(t)}} \boldsymbol{x} + \gamma \xi^{(t)} \boldsymbol{P}_{\boldsymbol{w}^{(t)}} \boldsymbol{x}. \tag{D.26}$$

The $\xi^{(t)}$ term can then be absorbed into the noise that appears in the multi-step analysis in Sections C.4, C.5, and C.6. Hence, we can take

$$\psi_\eta(y, z) = \sum_{k=1}^{r} (\eta d)^{k-1} \sigma^{(k)}(z) \big(\sigma'(z)\big)^{k-1} y^k. \tag{D.27}$$

Hence, for $i \in [r]$,

$$\begin{aligned} u_{i-1}\big(b \mapsto \psi_\eta(y, b)\big) &= \sum_{k=1}^{r} (\eta d)^{k-1} y^k \mathbb{E}_{b\sim\mathcal{N}(0,1)}[\sigma^{(k)}(b)\big(\sigma'(b)\big)^{k-1} \mathsf{He}_{i-1}(b)] \\ &= \sum_{k=1}^{r} (\eta d)^{k-1} y^k u_{i-1}(\sigma^{(k)}(\sigma')^{k-1}), \end{aligned} \tag{D.28}$$

which leads to

$$\begin{aligned} \hat{\mu}_i(\zeta) &= \sum_{k=1}^{r} (\eta d)^{k-1} \sigma^{(k)} (\sigma')^{k-1} u_i\big(a \mapsto (\sigma_*(a) + \zeta)^k\big) \\ &= \sum_{k=1}^{r} (\eta d)^{k-1} u_{i-1}\big(\sigma^{(k)}(\sigma')^{k-1}\big) \sum_{l=0}^{k} \binom{l}{k} u_i\big(a \mapsto (\sigma_*(a))^l \zeta^{k-l}\big) \\ &= \sum_{k=1}^{r} (\eta d)^{k-1} u_{i-1}\big(\sigma^{(k)}(\sigma')^{k-1}\big) \sum_{l=0}^{k} \binom{l}{k} \zeta^{k-1} u_i(\sigma_*^l). \end{aligned} \tag{D.29}$$

Then, taking expectation with respect to ζ gives

$$\mu_i \asymp \sum_{k=1}^{r} (\eta d)^{k-1} u_{i-1}\big(\sigma^{(k)}(\sigma')^{k-1}\big) u_i(\sigma_*^k). \tag{D.30}$$

Appendix E. Experiment Details

In this section, we provide the details on the experiment that generated Figure 1 in the main text and discuss an additional experiment on batch reuse SGD in the same setting. Throughout, we consider a noiseless single-index teacher (2.1) with $\sigma_* = \mathsf{He}_3$ and a two-layer neural network student (2.2) with initialization $a_j = 1$ and $\boldsymbol{w}_j \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ for all $j \in [N]$, where $d = 50$ and $N = 1024$. We consider logarithmically spaced meshes of 70 η values between 10^{-7} and 1 and and 50 n values between 10^3 and $10^{5.7}$ (taking larger n is prohibitively expensive on our single GPU). We train in a single-pass over the data (i.e., one epoch) with fixed batch size $B = 32$. In other words, we perform $\lfloor n/B \rfloor$ online updates. Subsequently, we solve for the exact ridge regression solution $\hat{\boldsymbol{a}}$ with penalty parameter $\lambda = 10^{-2}$. We evaluate the resulting two-layer network on a test dataset of size $n' = 4096$ generated by the teacher and compute the test MSE for each combination of (η, n).

For our experiment depicted in Figure 1, we implement alternating SGD as specified in Algorithm 1 with the only change being that we do not implement the projection $\boldsymbol{P}_{\boldsymbol{w}}^{\perp}$. We update each a_j to obtain $\tilde{a}_j$, then update $\boldsymbol{w}_j$ and normalize. At each step, we keep $a_j = 1$. We choose $\gamma = \max\{d^{-\frac{3}{2}}, \eta d^{-1}\}$ as per Corollary 6. Collecting the test MSEs for each (η, n) combination

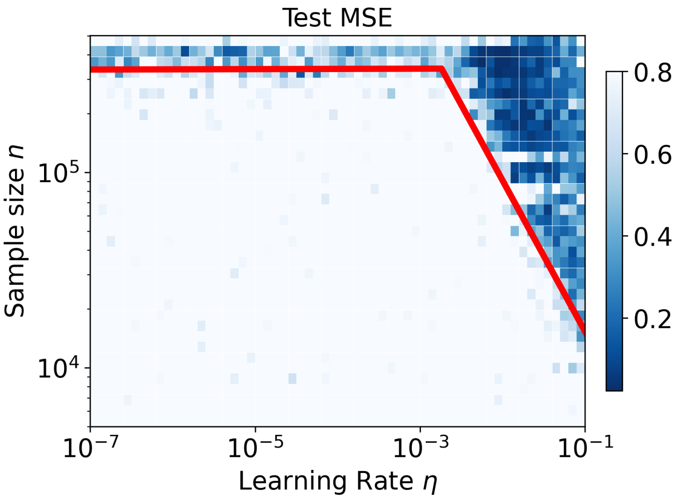

Figure 4: Test MSE of batch reuse SGD for different choices of η and n. The hyperparameter γ is chosen according to Corollary 7.

in our mesh gives the colorbar in Figure 1. We apply a threshold so that only combinations that achieve test MSE below 0.3 are displayed. We then use these points to fit a linear spline with a single knot (in the log-log scale) to better illustrate the phase transition. From this, we can see that the sample complexity remains flat for small η before decaying after η reaches a critical value. On the other hand, if η is too large (beyond 10^{-1}; not shown in the plot), small test MSE is no longer reliably achieved.

We perform a similar experiment for batch reuse SGD, the result of which is displayed in Figure 4. We implement Algorithm 2, with the only modification being that we omit $\boldsymbol{P}_{\boldsymbol{w}}^{\perp}$. We choose $\gamma = \max\{d^{-\frac{3}{2}}, \eta\}$ as per Corollary 7. For the purposes of visualization, we apply a more tolerant threshold of 0.8 to the test MSEs and fit a linear spline using the points with test MSE less than 0.6. Once again, we observe a “flat” sample complexity for small η, followed by a decay after a critical value is reached.