FLATNESS AND GENERALIZATION: LEARNING MULTI-INDEX MODELS WITH HOMOGENEOUS NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

A common heuristic used to explain the generalization of first-order gradient methods on non-convex neural networks is that "flat interpolators generalize well" (Hochreiter & Schmidhuber, 1994; Keskar et al., 2017), where flatness can be measured by the trace of the Hessian of the empirical loss. However, Dinh et al. (2017) showed that any interpolator can be made sharper or flatter using the *symmetry of the network* that can change flatness while keeping the population and empirical losses unchanged. This result makes the earlier heuristic statement vacuous. In this paper, we show that for learning an unknown multi-index model with 2-layer non-convex homogeneous neural networks, there is a connection between flatness and generalization, despite the existence of symmetries. This connection pertains to the "flattest" interpolators, i.e., the interpolators that have orderwise minimum flatness among all interpolators. First, we show that there exists a natural class of non-generalizing interpolators whose flatness cannot be made closer to the flattest possible, even using symmetries. Second, we show that for data generated by a sum of single-index models, if the approximation error and label noise are low, all flattest interpolators achieve small population loss, i.e., the flattest interpolators always generalize. This establishes a direct link between flatness and generalization which applies to a large class of activations and realistic data distributions.

1 INTRODUCTION

In machine learning, overparameterization refers to the paradigm where there are more model parameters than available data samples, such as in modern implementations of neural networks (NN). In this paradigm, the global minimizers of empirical (training) loss can interpolate, or perfectly fit the training data. However, due to the high-dimensional and non-convex landscape of the empirical loss, there are often "bad" interpolators, i.e., interpolators that have large population (test) loss (Zhang et al., 2021; Nagarajan & Kolter, 2019). Surprisingly, stochastic gradient methods on these overparametrized models usually converge to "good" interpolators, which generalize well. This empirical observation has not yet been rigorously justified, but one possible explanation for it, from (Hochreiter & Schmidhuber, 1994) initially, is the following: *stochastic gradient methods converge to flat interpolators of the empirical loss and flat interpolators of empirical loss always generalize*. In this paper, sharpness of an interpolator is the trace of the Hessian of the empirical loss. Flat interpolators have small trace. The first part of this explanation can be justified due to stable convergence to flat minima for large step size (stochastic) gradient descent (Wu et al., 2018; 2022). In this work, we will try to theoretically justify the second part of this explanation, which states that flat interpolators generalize.

Several existing works have tried to justify this explanation. Experimentally, flatness is strongly correlated with generalization (Jiang et al., 2020; Keskar et al., 2017) and algorithms that explicitly minimize some notion of flatness generalize better (Keskar et al., 2017; yeh Chiang et al., 2023; Foret et al., 2021). However, there is also evidence refuting this claim. Andriushchenko et al. (2023) show that the correlation between flatness and generalization depends on the choice of model and Wen et al. (2023); Schliserman et al. (2025) construct flat interpolators that do not generalize.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Notably, an important setting where this claim can be refuted is in the presence of "symmetries", i.e., transformations of model weights that do not change the input-output relationship. The following well-known result from Dinh et al. (2017) summarizes the impact of such symmetries on flatness.

Proposition 1 (Symmetries for ReLU Neural Networks (Dinh et al., 2017) (see Def 1))**.** *There exist symmetries that rescale the layer weights of a NN such that the model output remains the same for a given input (thus, empirical and population losses remain unchanged), however, the flatness changes.*

In particular, Dinh et al. (2017) use this property to show that one can make any interpolator sharper using symmetry, which implies that good interpolators can be sharp. Using symmetry, as in Prop. 1, one can also make any interpolator flatter. This contradicts the claim "flat interpolators always generalize", as we can make bad interpolators flat. But, how flat can we make these bad interpolators? Can we make bad interpolators the flattest interpolators, i.e., those with the minimum trace of Hessian among all interpolators? Or is there a fundamental barrier that prohibits the symmetries from making certain bad interpolators achieve the minimum flatness? A necessary condition for this claim to be true is to show there are bad interpolators that cannot be made the flattest interpolators in the presence of symmetry. This is the first question we consider in this paper:

***Q1**: Is it possible to construct a class of bad interpolators, that cannot be made the flattest interpolators using symmetries?*

If the answer to this question is yes, then there can be a link between flatness and generalization in the presence of symmetry. Note that this does not prevent other bad interpolators from being flattest. This forms our second question, and a *sufficient* condition for our claim.

***Q2**: Do all flattest interpolators generalize?*

We consider the feature-learning setting to answer these two questions. In our setting, the true labels come from a multi-index model, and we use a 2-layer neural network with homogeneous non-linear activations to learn it. This is a popular regression problem, where the labels depend only on a few directions ($m^\star \ll d$) of the features, specified by the matrix $\Theta^\star \in \mathbb{R}^{m^\star \times d}$, via an unknown link function $\sigma^\star$. The 2-layer network has m hidden neurons with activation $\sigma : \mathbb{R} \to \mathbb{R}$ and the weight vector $\mathbf{w} \in \mathbb{R}^{m(d+1)}$ consists of outer-layer weights $\mathbf{a} \in \mathbb{R}^m$, and inner-layer weights $\{\theta_j\}_{j\in[m]}, \theta_j \in \mathbb{R}^d, \forall j \in [m]$. We sample $n \gtrsim d$ datapoints from the distribution and ensure overparametrization by setting $m \geqslant n$. Let S be the set of training samples, $F_S : \mathbb{R}^{m(d+1)} \to \mathbb{R}_+$ be the empirical loss and $F : \mathbb{R}^{m(d+1)} \to \mathbb{R}_+$ be its population loss. Then, an interpolator achieves $F_S(\mathbf{w}) = 0$ and its flatness is given by $\mathrm{Tr}(\nabla^2 F_S(\mathbf{w}))$.

Even in this simplified setting, the answer to the questions posed earlier is highly non-trivial. This setup i) has both good and bad interpolators, ii) it is non-linear and non-convex, iii) it has a symmetry that can change flatness without changing population loss, however, iv) gradient-based methods provably generalize with $\asymp d$ samples (Oko et al., 2024b; Arous et al., 2021). Further, existing flatness-based generalization bounds either do not apply to these settings (Qiao et al., 2024; Haddouche et al., 2025; Wu & Su, 2023; Liang et al., 2025a) or are too loose (Liang et al., 2025b). Additionally, this setting is a more representative example of real-world NN training, than the worst-case distributions considered as counter-examples in (Wen et al., 2023; Schliserman et al., 2025). In Section 1.2, we provide a comparison of our results to these works.

1.1 Main Contributions

Our main contributions are the answers to our two questions posed earlier along with several intermediate results. We first state all our results.

- For c-homogeneous activations (Assump. 1), we characterize the flattest that an interpolator $\mathbf{w}$ can be made using symmetry as $\Upsilon(\mathbf{w}) \propto \|\mathbf{a}\|_p^p$ where $p = \frac{2c}{c+1}$, and $c \geqslant 1$. (Lemma 3).
- For homogeneous activation and multi-index link (Assump. 1 and 2),
  1. we show that the flatness of an interpolator $\mathbf{w}$ is lower bounded by $\Upsilon^\star$, i.e., $\Upsilon(\mathbf{w}) \gtrsim \Upsilon^\star$. (Thm. 1)
  2. we characterize a set of bad interpolators in Def. 2, whose flatness is at least, $n^{\frac{c}{c+1}}$ times larger than this lower bound, i.e., $\Upsilon(\mathbf{w}) \gtrsim n^{\frac{c}{c+1}} \Upsilon^\star$. (Thm. 2).

3. we show that any interpolator that satisfies Eq (1), achieves the lower bound on flatness, only if its outer-layer weights have a polylog ℓ_1 norm, i.e., $\Upsilon(\mathbf{w}) \lesssim \Upsilon^\star \implies \|\mathbf{a}\|_1 = \tilde{\mathcal{O}}(1)$. (Prop 2).

.

- For homogeneous activation, and link function being approximately from sum of single-index models with low approximation error and low label noise (Assump. 1, 2 and 3), we show that,
  1. an interpolator achieving the lower bound on flatness exists, i.e., $\exists \mathbf{w}, F_S(\mathbf{w}) = 0, \Upsilon(\mathbf{w}) \asymp \Upsilon^\star$. (Thm. 3)
  2. all interpolators satisfying $\|\mathbf{a}\|_1 = \tilde{\mathcal{O}}(1)$, thus achieving minimum flatness, $\Upsilon(\mathbf{w}^\star) \asymp \Upsilon^\star$, generalize with sample complexity $n \gtrsim d^{\max\{1, c/2\}}$. (Thm. 3)

Answer to Q1: Under all our assumptions, we can characterize a set of bad interpolators that cannot be made flattest despite symmetry. In fact, their flatness is atleast $n^{\frac{c}{c+1}}$ times larger than the flattest. These interpolators have their inner-layer weights misaligned to the true direction $\Theta^\star$ that generated the data.

Answer to Q2: Under all our assumptions, all interpolators achieving minimum flatness and satisfying Eq (1). For $c \in [1, 2]$, our sample complexity, $n \asymp d$, achieves the optimal SQ lower bound for single-index data (Chen & Meka, 2020).

Therefore, for our setting of learning a multi-index model with 2-layer neural networks under symmetry, there is a relationship between flatness and generalization, as the flattest interpolators under certain conditions generalize. However, the original claim in the previous works (Liang et al., 2025b; Wu & Su, 2023) is not precise, as interpolators with an arbitrary value of flatness might not generalize for $n \asymp \text{poly}(d)$ due to symmetry. Our results also show benign overfitting for the flattest interpolator for 2-layer homogeneous NN with multi-index data, while previous works (Wen et al., 2023; Ding et al., 2024) could only show this for specific instances of our setting.

1.2 Comparison to Related Works

In this section, we compare our results with the most relevant existing works and defer a more comprehensive comparison to Appendix A. The closest result to ours is (Ding et al., 2024, Thm 7.1). They show that for Quadratic activations ($c = 2$), when learning a sum of single-index models, where the unknown link function is the same as the activation, the flattest interpolator generalizes. Note that this is a special case of our result (cf. Thm. 3), with both zero approximation error and label noise. However, our main results (Thms 1,2, and 3) are much stronger than theirs, since (i) we consider homogeneous activations that includes the ReLU, LeakyReLU and Quadratic activation, (ii) we allow non-zero label noise, (iii) the link function and the model's activation can be different $\sigma^\star \neq \sigma$, and (iv) we characterize both flattest good as well as bad interpolators. One of their key insights is a "balancedness condition" between layer weights. Interestingly, our results exhibit a more general form of balancedness, $\|\mathbf{a}\|_1 = \mathcal{O}(1)$ in Prop. 2. Further, their proof techniques focus on equivalence to matrix factorization, however, our proofs are significantly different since we use concentration arguments to account for general non-quadratic homogeneous activations with noise and approximation errors.

Two counter-examples where there are bad interpolators that are flattest are provided in (Schliserman et al., 2025) for smooth convex optimization, and in (Wen et al., 2023) for learning a 2-layer NN with bias for the XOR problem, a classification variant of our multi-index models. These examples are not contradictory to our findings due to their limited settings. Schliserman et al. (2025) rely on a specific discrete data distribution, specifically the features are sampled from $\{0,1\}^d$, and they require some coordinate of all features to be 0. Note that this condition is not possible for gaussian features with $n \gtrsim d$. As for (Wen et al., 2023), for ReLU networks with bias, there could be flattest interpolators that don't generalize for a large enough bias. However, their construction is an artifact of the ReLU activation which is simultaneously Lipschitz and 0 on half its domain. The only homogeneous ReLU activation satisfies this condition is ReLU. If we remove either of this condition, for instance a ReLU^2 which is not Lipschitz, or a LeakyReLU which isn't 0 on half of its domain, then their construction fails, inspite of bias. Our results consider no bias, but include ReLU, LeakyReLU and Quadratic. In fact, (Wen et al., 2023, Thm. 3.1), showing flattest interpolators of ReLU networks generalize is a special instance of our results, where they implicitly assume Eq (1).

Organization. In Section 2, we formally define our setup, including the symmetry induced by homogeneous activations. In Section 3, we define the class of bad interpolators and characterize

their flatness. In Section 4, we show that flattest interpolators generalize. We summarize our proof techniques in Section 5 and our conclusions in Section 6. We defer experiments for verifying our theoretical results to Appendix B.

2 SETUP

Notation. We use $[n]$ to denote the set $\{1,2,...,n\}$. We use $\mathcal{O},o,\Omega$ and Θ to denote the usual complexity notation with $\tilde{\mathcal{O}}$ handling polylog terms. For example, if a term $b=\mathcal{O}(1)$, then it is order-wise at most a constant with respect to m,n and d. All bounds that hold with high probability hold with probability $1-C\delta-\mathrm{poly}(m^{-b_1},n^{-b_2},d^{-b_3})$ for constants $C\geqslant 1$ and $\delta,b_1,b_2,b_3>0$. We use $\gtrsim,\lesssim$ and $\asymp$ to represent greater than, less than or equal order-wise up to poly-logarithmic factors of problem parameters and δ^{-1}. poly is used to denote any bounded-degree polynomial in its arguments. We use $\mathrm{span}(Q)$ to denote the linear span of a set of vectors $Q\subset\mathbb{R}^d$, and $\mathbb{S}^{d-1}$ to denote the unit sphere in d dimensions.

Data, Model & Losses. Each data point is $\mathbf{z}=(\mathbf{x},y)\in\mathbb{R}^d\times\mathbb{R}$, where $\mathbf{x}\sim\mathcal{N}(0,\mathbb{I}_d)$ are the features and $\mathbf{y}$ is the label. We generate the labels as $y=\sigma^\star(\Theta^\star\mathbf{x})+\xi$ for some continuous and differentiable almost everywhere function $\sigma^\star:\mathbb{R}^{m^\star}\to\mathbb{R}$, a fixed orthonormal matrix $\Theta^\star\in\mathbb{R}^{m^\star\times d}$ such that $\Theta^\star(\Theta^\star)^\top=\mathbb{I}_{m^\star}$ where the jth row of $\Theta^\star$ is denoted by the vector $\theta_j^\star\in\mathbb{S}^{d-1}$, and label noise $\xi\sim\mathcal{N}(0,\zeta^2)$ is independent of $\mathbf{x}$ with $\zeta=o(1)$. We assume that $m^\star=\mathcal{O}(1)$, as it is a multi-index model. Two common and well-known examples that are covered under our problem setup are the following –

1. Single-Index Models (Oko et al., 2024b): We set $m^\star=1$, and $y=\sigma^\star(\langle\theta^\star,\mathbf{x}\rangle)+\xi$ for $\theta^\star\in\mathbb{S}^{d-1}$.
2. Sum of Single-Index Models (Oko et al., 2024a): We set $y=\sum_{j=1}^{m^\star}a_j^\star\tilde{\sigma}^\star(\langle\theta_j^\star,\mathbf{x}\rangle)+\xi$ with $\tilde{\sigma}^\star:\mathbb{R}\to\mathbb{R}$ being a continuous and differentiable function, and $a_j^\star\in\mathbb{R}$ with $a_j^\star=\mathcal{O}(1)\ \forall j\in[m^\star]$.

The activation $\sigma:\mathbb{R}\to\mathbb{R}$ is continuous and differentiable almost everywhere, the weights of the network are $\mathbf{w}^\top:=[a_1,a_2,\dots,a_m,\theta_1^\top,\theta_2^\top,\dots,\theta_m^\top]$, and its output for a given feature $\mathbf{x}\in\mathbb{R}^d$ is $h(\mathbf{w},\mathbf{x})=\sum_{j=1}^m a_j\sigma(\langle\theta_j,\mathbf{x}\rangle)$. The width of the network satisfies $Cn\leqslant m\leqslant C'n$ for some constants $C'>C>2$. We use $\psi^\star=\mathbb{E}\big[(\sigma^\star)^2(\mathbf{b})\big]$ to denote the signal power and $S=\{\mathbf{z}_i\}_{i\in[n]}$ to denote the set of data points. Using square loss, $f(\mathbf{w},\mathbf{z})=\frac{1}{2}(h(\mathbf{w},\mathbf{x})-y)^2$, we define the empirical loss F_S, population loss F, and their minimizers as

$$F_S(\mathbf{w})=\frac{1}{n}\sum_{\mathbf{z}_i\in S}f(\mathbf{w},\mathbf{z}_i),\quad F(\mathbf{w})=\mathop{\mathbb{E}}_{\mathbf{z}}[f(\mathbf{w},\mathbf{z})],\quad F^\star=\min_{\mathbf{w}}\ F(\mathbf{w}).$$

An interpolator satisfies $F_S(\mathbf{w})=0$. Throughout this paper, we will use "good" and "bad" to quantify the excess population loss of a weight $\mathbf{w}\in\mathbb{R}^{m(d+1)}$. A model with weight vector $\mathbf{w}$, is "good"/"bad" if

$$F(\mathbf{w})-F^\star=o(1)\quad\textbf{(Good)};\qquad F(\mathbf{w})-F^\star>\kappa\cdot(\psi^\star+\zeta^2),\text{ where }\kappa\text{ is a constant.}\quad\textbf{(Bad)}.$$

From the definition of bad weights, we need their excess population loss to be lower-bounded by a constant (κ) multiple of the signal power ($\psi^\star$) plus the noise variance (ζ^2). For example, if a weight has excess population loss upper-bounded by $\mathcal{O}\big(\mathrm{poly}(n^{-b_1},d^{-b_2},m^{-b_3})\big)$ for any constants $b_1,b_2,b_3>0$ it is a good weight.

Activations & Symmetry. Note that the symmetry in Prop. 1 is a by-product of using homogeneous activations. A function $g:\mathbb{R}\to\mathbb{R}$ is c-Homogeneous if $\forall\alpha>0,b\in\mathbb{R}$, $g(\alpha b)=\alpha^c g(b)$. Our activation function σ for the 2-layer NN is a special class of homogeneous activations satisfying the following assumption.

Assumption 1 (Piece-wise Polynomial Activation)**.** *For all $b\geqslant 0$, $\sigma(b)=c'|b|^c$, and for all $b<0$, $\sigma(b)=c''|b|^c$, where $c',c''\in\mathbb{R}$ are some constants such that at most one of them is 0, $c'+c''\neq 0$, and $c\geqslant 1$.*

The above assumption forces σ to be c-Homogeneous. All even-degree monomials of degree c, as well as LeakyReLU, ReLU and ReLU^c (He et al., 2024) satisfy the above assumption. The above assumption does not include odd-degree monomials and monomials with $c<1$ due to technical difficulty in our proof, explained in Section 5. It also does not include non-polynomial activations like tanh and sigmoid. We define the symmetry induced by homogeneous activations used in Prop. 1.

Definition 1 (Rescaling Symmetry). *For a* $\mathbf{w} \in \mathbb{R}^{m(d+1)}$ *consisting of weights* $\{a_j\}_{j\in[m]}$ *and* $\{\theta_j\}_{j\in[m]}$*, let* $\mathcal{W}_{rescale}(\mathbf{w})$ *be the set of weights* $\tilde{\mathbf{w}} \in \mathbb{R}^{m(d+1)}$ *with weights* $\{a_j\alpha_j^{-c}\}_{j\in[m]}$ *and* $\{\alpha_j\theta_j\}_{j\in[m]}$ *for all* $\alpha_j > 0$*. Then,* $h(\mathbf{w},\mathbf{x}) = h(\tilde{\mathbf{w}},\mathbf{x}), \forall \mathbf{x}\in\mathbb{R}^d, \forall \tilde{\mathbf{w}} \in \mathcal{W}_{rescale}$.

This symmetry is in the space of model weights, and is quantified by the scalars $\{\alpha_j\}_{j\in[m]}$ with $\mathcal{W}_{rescale}(\mathbf{w})$ being the equivalence class of $\mathbf{w}$. Existing works (Neyshabur et al., 2015; Zhao et al., 2026) have extensively studied the impact of this symmetry on loss landscapes. As the output of the model does not change with this symmetry, both population and empirical loss remain the same. In the next section, we will see how this symmetry changes the flatness of an interpolator.

2.1 Flatness under Rescaling Symmetry

The flatness of an interpolator $\mathbf{w}$ is defined as $\mathrm{Tr}(\nabla^2 F_S(\mathbf{w}))$. For non-smooth activations like ReLU and LeakyReLU, we only consider $\mathbf{w}$ that are away from the kinks, so that the Hessian can be defined. Note that several definitions have been used in literature for flatness (Wu & Su, 2023; Dinh et al., 2017; Schliserman et al., 2025), among which Wu et al. (2022); Ding et al. (2024); Gatmiry et al. (2023); Wen et al. (2023) use our definition. For our setting, the flatness of an interpolator $\mathbf{w}$ simplifies to $\mathrm{Tr}(\nabla^2 F_S(\mathbf{w})) = \frac{1}{n}\sum_{i=1}^n \|\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}_i)\|_2^2$.

Due to Rescaling Symmetry (Def. 1), although $h(\tilde{\mathbf{w}},\mathbf{x})$ remains the same, $\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x})$ changes, thus changing flatness. We first define the flattest interpolator obtained by applying rescaling symmetry to a specific interpolator $\mathbf{w}$. To define this, we need to set a reference point for a given $\mathbf{w}\in\mathbb{R}^{m(d+1)}$. Both the inner layer weights $\{\theta_j\}_{j\in[m]}$ and the outer layer weights $\{a_j\}_{j\in[m]}$ can be changed using rescaling symmetry. For our reference, we will fix the inner-layer weights to be unit norm, $\theta_j \in \mathbb{S}^{d-1}, \forall j \in [m]$, and the outer-layer weights $\mathbf{a}$ chosen such that $\mathbf{w}$ interpolates. The following Lemma quantifies $\Upsilon(\mathbf{w})$, the flatness of the flattest interpolator obtained by applying the rescaling symmetry to $\mathbf{w}$.

Lemma 1 (Flattest Interpolator under Rescaling Symmetry). *Under Assump. 1, for an interpolator* $\mathbf{w}$ *such that* $\theta_j \in \mathbb{S}^{d-1}, \forall j \in [m]$*, with high probability,*

$$\Upsilon(\mathbf{w}) \coloneqq \sum_{j=1}^m B(\theta_j)|a_j|^{\frac{c}{c+1}} = \inf_{\hat{\mathbf{w}}\in\mathcal{W}_{rescale}(\mathbf{w})} \mathrm{Tr}(\nabla^2 F_S(\hat{\mathbf{w}})).$$

$$\text{where } B(\theta_j) = \left(\frac{1}{n}\sum_{i=1}^n \sigma^2(\langle\theta_j,\mathbf{x}_i\rangle)\right)^{\frac{1}{c+1}} \left(\frac{1}{n}\sum_{i=1}^n (\sigma')^2(\langle\theta_j,\mathbf{x}_i\rangle)\|\mathbf{x}_i\|_2^2\right)^{\frac{c}{c+1}}.$$

Further, if $\{\theta_j\}_{j\in[m]}$ *are independent of* S*, then, with high probability,* $\Upsilon(\mathbf{w}) \asymp d^{\frac{c}{c+1}} \|\mathbf{a}\|_{\frac{2c}{c+1}}^{\frac{2c}{c+1}}$.

$\Upsilon(\mathbf{w})$ is the minimum flatness in the equivalence class. The proof of this lemma, provided in App. D.1, is quite direct, as the objective is a convex and separable function in the scalars $\{\alpha_j\}_{j\in[m]}$ that define the equivalence class. The implications of this lemma are substantial. First, the quantity $\Upsilon(\mathbf{w})$ removes the impact of symmetry. Second, it separates the contributions of inner and outer-layer weights. Here, the additional $d^{\frac{c}{c+1}}$ term is obtained from the inner-layer weights $\{\theta_j\}_{j\in[m]}$ being unit norm and independent of samples S. The remaining term in $\Upsilon(\mathbf{w})$ is $\|\mathbf{a}\|_p^p$, which is a convex ℓ_p-norm with $p = \frac{2c}{c+1} \in [1,2)$ for $c \geqslant 1$. Third, Lemma 1 does not put any assumption on $\mathbf{w}$ or the data distribution, apart from the fact that $\mathbf{w}$ interpolates on the training set S.

As we have removed the contribution of rescaling symmetry, to check if a bad interpolator can be made flattest by rescaling, we need to evaluate $\Upsilon(\mathbf{w})$ from Lemma 1 for a bad interpolator. In the next section, we consider a specific class of bad interpolators, where we can appropriately bound $\Upsilon(\mathbf{w})$ to answer **Q1**.

3 Bad Interpolators That are Not Flattest

In this section, we answer **Q1**. The proof involves two steps, the first being a bound on the minimum flatness of an interpolator (Thm 1) and the second being a lower bound on the flatness of bad interpolators (Thm. 2). We first define a natural class of bad interpolators for multi-index data distributions.

To obtain a lower bound on $\Upsilon(\mathbf{w})$ for an interpolator $\mathbf{w}$, we state additional assumptions required on the unknown link function $\sigma^\star$ that generates the labels y and the width of the network.

Assumption 2 (Link Function). *For Gaussian random vector, $\mathbf{b}\sim\mathcal{N}(\mu,\varsigma)$ where $\mu\in\mathbb{R}^{m^\star}, \|\mu\|_2=\tilde{\mathcal{O}}(1)$ and $\varsigma\in\mathbb{R}^{m^\star\times m^\star}$, $c_\varsigma\mathbb{I}_{m^\star}\preceq\varsigma\preceq C_\varsigma\mathbb{I}_{m^\star}$ for constants $c_\varsigma, C_\varsigma>0$, the link function satisfies:*

$$\textbf{\textit{Mean}}: \|\mathbb{E}[\sigma^\star(\mathbf{b})]\|_2=\mathcal{O}\Big(\|\mu\|_2^c+\|\varsigma\|_F^{\frac{c}{2}}\Big), \qquad \textbf{\textit{Variance}}: Var[\sigma^\star(\mathbf{b})]=\Omega\Big(\|\varsigma\|_F\|\mu\|_2^{2c-2}+\|\varsigma\|_F^c\Big),$$

$$\textbf{\textit{Tails}}: \sigma^\star(\mathbf{b}) \text{ is a } (\frac{c}{2},\mathcal{O}(\|\mu\|_2^c+\|\varsigma\|_F^{\frac{c}{2}}))\text{-Sub-Weibull random vector.}$$

Note that a (ϱ,K)-Sub-Weibull random variable has its kth central moment bounded by $\mathcal{O}(Kk^\varrho)$ (Vladimirova et al., 2020; Sambale, 2023; Zhang & Wei, 2022) for $\rho,K>0$. Sub-Gaussian and Sub-exponential distributions are also (ϱ,K)-Sub-Weibull with $\varrho=\frac{1}{2}$ and $\varrho=1$ respectively, with K being their Sub-exponential or Sub-Gaussian parameters. We provide an overview of Sub-Weibull distributions in App. E.2.1. The above assumption forces a polynomial bound on the link function, which is in fact satisfied by our activations in Assump. 1. With this assumption, our first result is a lower bound on $\Upsilon(\mathbf{w})$ for a fixed interpolator.

Theorem 1 (Flattest Interpolator). *If Assump. 1 and 2 hold, then, for any interpolator $\mathbf{w}$ with same $\{\theta_j\}_{j\in[m]}$ selected independent of S, with high probability, $\Upsilon(\mathbf{w})\gtrsim\Upsilon^\star:=d^{\frac{c}{c+1}}m^{-\frac{c-1}{c+1}}$.*

Note that $\Upsilon^\star$ is a lower bound, however, for noiseless case $\zeta=0$ and when learning activation $\sigma^\star=\sigma$ for single-index or sum of single-index models, this lower bound is achievable (cf. Thm 4). Note that the flattest interpolator must have its flatness lower than this quantity. We will compare $\Upsilon^\star$ derived above with $\Upsilon(\mathbf{w})$ for a bad interpolator as defined below. We require additional notation for this definition. Note that, $\Theta\in\mathbb{R}^{m\times d}$ is the matrix of inner-layer weights, $vU\in\mathbb{R}^{n\times m}$ is the matrix of activations, i.e. $\mathbf{U}_{i,j}=\sigma(\langle\theta_j,\mathbf{x}_i\rangle)$, and $\bar{\Phi}=\frac{1}{n}\mathbb{E}\big[(\mathbf{U}-\mathbb{E}[\mathbf{U}])^\top(\mathbf{U}-\mathbb{E}[\mathbf{U}])\big]$ is the covariance of activations.

Definition 2 ($(\rho^\star,\rho',\kappa)$-Bad Interpolator). *For $\rho^\star,\rho'\in[0,\frac{1}{2}]$, and constants $\kappa,\in(0,\frac{1}{2})$, $\kappa_1,\kappa_2>0$, we say that an interpolator $\mathbf{w}\in\mathbb{R}^{m(d+1)}$ is $(\rho^\star,\kappa,\kappa_1,\kappa_2)$-bad if, $\{\theta_j\}_{j\in[m]}$ are selected independent of S, and*

1. $F(\mathbf{w})-F^\star\geqslant\kappa(\zeta^2+\psi^\star)$.
2. $|\langle\mathbf{v}_1,\mathbf{v}_2\rangle|\leqslant\rho^\star, \quad \forall\,\mathbf{v}_1\in\mathrm{span}(\{\theta_j\}_{j\in[m]})\cap\mathbb{S}^{d-1}, \forall\,\mathbf{v}_2\in\mathrm{span}(\{\theta_j^\star\}_{j\in[m^\star]})\cap\mathbb{S}^{d-1}$.
3. $\lambda_{\min}(\bar{\Phi})\geqslant\kappa_1$ *and* $\sigma_{\max}(\Theta)\leqslant\kappa_2$.

Let us understand each condition in the above definition. The first condition ensures that the population loss is large, by a constant κ for this interpolator, so that it is indeed a bad interpolator. The second condition requires any unit vector in the span of the inner-layer weights, $\mathrm{span}(\{\theta_j\}_{j\in[m]})$, to be misaligned with all vectors in the span of true directions, $\mathrm{span}(\{\theta_j^\star\}_{j\in[m]})$. A consequence of this condition is that each layer weight is also misaligned with the true direction, i.e., $\big|\big\langle\theta_j,\theta_{j'}^\star\big\rangle\big|\leqslant\rho^\star, \forall j\in[m], j'\in[m^\star]$. For multi-index models, this condition also ensures that the population loss is large. The third condition requires the bad interpolators to have well-conditioned inner-layer weights and activations, thus not all of them align to the same direction. For a good interpolator, we would want all inner-layer weights aligned to the true direction. As $\mathrm{Tr}(\bar{\Phi}), \|\Theta\|_F^2=\Theta(m)$, the third condition is only valid if $m\leqslant C''d$ for some large constant $C''>0$. This condition is required to ensure interpolation. Note that the bad interpolators defined above always exist as described with an example in App. D.3.

Theorem 2 (Flattest Bad Interpolator). *Suppose Assump. 1 and 2 hold. Then, for a $(\rho^\star,\kappa,\kappa_1,\kappa)$-bad interpolator $\mathbf{w}$ (Def. 2), with high probability,*

$$\Upsilon(\mathbf{w})\gtrsim(dn)^{\frac{c}{c+1}}m^{-\frac{c-1}{c+1}}, \text{ if } \rho^\star=o((m^\star)^{-\frac{1}{4}}).$$

When the correlation $\rho^\star$ is strictly smaller than a constant, the flatness of a bad interpolator in Def. 2 is at least $n^{\frac{c}{c+1}}$ times larger than the $\Upsilon^\star$. This answers **Q1**, thus satisfying the necessary condition for our claim "flat interpolators generalize". Since $n=\Omega(1)$, these bad interpolators are indeed very sharp, and even using rescaling symmetry, we cannot make them as flat as possible. This completes the picture for the argument of (Dinh et al., 2017) with respect to rescaling symmetry: while good interpolators can be made as sharp as possible, not all bad interpolators can be made as flat as possible.

Note that Thms. 1 and 2 do not prevent other bad interpolators, for instance those with large $\rho^\star$ or not satisfying Def. 2, from achieving the minimum flatness $\Upsilon^\star$. To answer **Q2**, we need to fully characterize the class of interpolators achieving the minimum flatness $\Upsilon^\star$. We do this in the next section.

4 FLATTEST INTERPOLATORS GENERALIZE

In this section, we will answer **Q2**, showing that under some assumptions, flattest interpolators always generalize. To prove this result, we provide a tight characterization of the set of flattest interpolators in terms of the outer-layer weights, $\mathbf{a}$.

Proposition 2 (Necessary Condition for Flattest Interpolator)**.** *Suppose Assump. 1 and 2 hold. Then, for an interpolator* $\mathbf{w}$ *with* $\theta_j \in \mathbb{S}^{d-1}, \forall j \in [m]$*, with high probability,*

$$\min_{j\in[m]} B(\theta_j) \gtrsim d^{\frac{c}{c+1}} \text{ and } \Upsilon(\mathbf{w}^\star) \lesssim \Upsilon^\star \implies \|\mathbf{a}\|_1 = \tilde{\mathcal{O}}(1). \tag{1}$$

This condition on $\|\mathbf{a}\|_1$ allows us a clean characterization of $\Upsilon^\star$ as long as the condition on $B(\theta_j)$, comparable to that in Lem. 3, holds. As $\Upsilon^\star$ is a lower bound on $\Upsilon(\mathbf{w})$, we do not know if there are interpolators that actually achieve this lower bound. To show that such interpolators do exist, we need additional assumptions on the data distribution, namely small label noise and approximation error, quantified by the following assumption.

Assumption 3 (Low Approximation Error and Label Noise)**.** *For some constants,* $\epsilon_1, \epsilon_2 > 0$*,*

- ***Approx. Error:*** *there exists a 2-layer NN of width* $\breve{m} = \mathcal{O}(1)$*, with unit norm inner-layer weights* $\{\breve{\theta}_j\}_{j\in[\breve{m}]}$*, and bounded outer-layer weights,* $\breve{\mathbf{a}} \in \mathbb{R}^{\breve{m}}$, $\|\breve{\mathbf{a}}\|_\infty = \tilde{\mathcal{O}}(1)$*, such that,* $\forall \mathbf{x} \in \mathbb{R}^d$ *with* $\max\{\|\Theta^\star \mathbf{x}\|_\infty , \max_{j\in\breve{m}} \left|\left\langle \breve{\theta}_j, \mathbf{x} \right\rangle\right|\} \leqslant C\sqrt{\log(d)}$*,* $\left|\sigma^\star(\Theta^\star \mathbf{x}) - \sum_{j\in[\breve{m}]} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x}\rangle)\right| \leqslant \Delta = \tilde{\mathcal{O}}(m^{-\frac{1}{2}} n^{-\frac{1}{2}-\epsilon_1})$.
- ***Label Noise:*** $\zeta = \tilde{\mathcal{O}}(m^{-\frac{1}{2}} n^{-\frac{1}{2}-\epsilon_2})$.

This assumption forces the link function, $\sigma^\star$, to be well-approximated by a 2-layer network of width $\breve{m}$, activation σ, and bounded outer-layer weights. A straightforward example for the above assumption is when $\Delta = 0$, and the unknown link function is a sum of activations. This is often referred to as the teacher-student setup (Goldt et al., 2020). The above assumption can handle cases beyond the teacher-student example, as long as the approximation error, Δ, is sufficiently small. Further, the bounds on $\|\breve{\mathbf{a}}\|_\infty$ ensure that outer-layer weights of this approximation are not super-large and satisfies Eq (1). Note that the restriction to a set of size $\mathcal{O}(\sqrt{\log(d)})$ ensures that for reasonable choice of $\mathbf{x}$, like our gaussian features, this approximation bound is valid. Existing works have shown approximation error of $\mathcal{O}(\text{poly}(\breve{m}^{-1}))$ (Siegel & Xu, 2022; Yang & Zhou, 2024), however in contrast, we need a much stronger approximation error. The label noise requirement implies the SNR, $\psi^\star \zeta^{-2}$, grows with n. Similar conditions between SNR and n are often required in prior works for benign overfitting of overparametrized NNs with isotropic $\mathbf{x}$ and single-index labels, see (Frei et al., 2022, (A1), (A4)).

We now present our main result answering **Q2**. This is divided into 2 results– proving the existence of an interpolator achieving flatness $\Upsilon^\star$, and bounding the population loss of any interpolator achieving flatness $\Upsilon^\star$.

Theorem 3 (Interpolators with constant $\|\mathbf{a}\|_1$ Generalize)**.** *Suppose Assump. 1, 2 and 3 hold. Then, for* $n \asymp d$*, there exists an interpolator* $\mathbf{w} \in \mathbb{R}^{m(d+1)}$ *with unit-norm inner-layer weights and outer-layer weights* $\|\mathbf{a}\|_1 = \mathcal{O}(1)$ *and flatness* $\Upsilon(\mathbf{w}) \asymp \Upsilon^\star$ *with high probability. Further, for* $n \gtrsim d$ *and* $R = \mathcal{O}(\text{polylog}(d))$*, with high probability,*

$$\mathcal{W}_{flat} := \{\mathbf{w} : F_S(\mathbf{w}) = 0, \|\mathbf{a}\|_1 \leqslant R, \|\theta_j\|_2 = 1, \forall j \in [m]\}, \quad \sup_{\mathbf{w}\in\mathcal{W}_{flat}} F(\mathbf{w}) - F^\star \lesssim \sqrt{\frac{d}{n}} + \frac{d^{\frac{c}{2}}}{n^{\min\{1,\frac{c}{2}\}}}.$$

Summarizing the results of the above Thm., Thm. 4 and Prop. 2, we first consider only interpolators satisfying Eq (1). Note that this includes the flattest interpolator. Under Assump. 3, this set is non-empty for $\Delta = \zeta = 0$ and all elements of this set generalize, with sample complexity, $n \gtrsim d^{\max\{1,\frac{c}{2}\}}$. For $\Delta, \zeta \neq 0$ but satisfying the bounds in Assump. 3, this set is again non-empty and all its elements generalize, but for $n \asymp d$, with $c \in [1,2]$. For $\Delta, \zeta \neq 0$ and $c > 2$, we cannot guarantee that this set is non-empty, but if it is, it will generalize with $n \gtrsim d^{\frac{c}{2}}$. This Theorem answers **Q2**, as under Eq (1), this set contains flattest interpolators, and they generalize. Further, our sample complexity of d for $c \leqslant 2$ is optimal, as (Damian et al., 2023) show that single-index models require d samples to learn polynomials of arbitrary degree. For $c > 2$, our sample complexity is not optimal, however, it is still much better than

the $\exp(d)$ samples required by existing generalization bounds on arbitrary flat interpolators (Liang et al., 2025b;a). Further, (Wen et al., 2023)'s result for ReLU networks is a special instance of the above Theorem for $c=1$. In the next section, we provide brief proof sketches for our main results.

5 Proof Sketches

In this section, we now explain the key theoretical tools used to prove all the main results. The complete proofs are provided in App. D. At a high level, we need to quantify the relationship between flatness and population loss. For Thms. 1 and 2, this is via lower bounds on $\Upsilon(\mathbf{w})$ in terms of the correlation $\rho^\star$ (Def. 2) between inner-layer weights and the true direction $\Theta^\star$. For multi-index models, this correlation is a strong indicator of population loss. For Thm. 3, this is via upper bounds on population loss based on $\|\mathbf{a}\|_1$, without using any intermediate connections to the correlation. At a high-level, all these results rely on a few fundamental techniques – to handle terms containing $\mathbf{a}$, we use convex duality (Boyd & Vandenberghe, 2004, Chapter 5) and for all other terms, we appropriately use concentration of measure.

For all our proofs, the core ingredient is Lemma 1, that encodes the contribution of symmetry in terms of a convex ℓ_p norm of $\mathbf{a}$, with $p = \frac{2c}{c+1}$, decided by the symmetry (Def. 1). If $\mathbf{y} \in \mathbb{R}^n$ is the vector of labels, then, $\mathbf{w}$ is an interpolator, iff $\mathbf{U}\mathbf{a} = \mathbf{y}$. Note that this is a linear constraint in terms of $\mathbf{a}$. Therefore, flatness is a convex function in $\mathbf{a}$ with a linear constraint in $\mathbf{a}$ for interpolation.

Proof Sketch of Thms. 1 and 2. We present this proof sketch in Fig. 1. At the core of this proof, we want to find a lower bound on $\Upsilon(\mathbf{w})$ that is independent of $\mathbf{a}$. Duality can exactly compute the expression $\min_{\mathbf{a}} \Upsilon(\mathbf{w})$, however, there is no closed-form expression for it, except in the case of quadratic activations (Ding et al., 2024). To obtain a closed-form expression for all activations in Assump. 1, we use a tight lower bound by a novel connection to overparametrized linear regression. Consider the interpolation constraint $\mathbf{U}\mathbf{a} = \mathbf{y}$. Fixing the inner-layer weights fixes the activation, and for $m \geqslant n$, this problem resembles overparametrized linear regression in $\mathbf{a}$. From (Bartlett & Mendelson, 2003), we know the minimum ℓ_2 norm interpolator for overparametrized linear regression is $\mathbf{a}_{\min,\ell_2} = \mathbf{U}^\top(\mathbf{U}\mathbf{U}^\top)^{-1}\mathbf{y}$. As flatness $\Upsilon(\mathbf{w})$ is an ℓ_p norm, we can obtain a tight lower bound on it via the ℓ_2 norm of the minimum ℓ_2 norm interpolator in Lemma 5.

The remainder of the proof utilizes two different bounds on $\|\mathbf{a}_{\min,\ell_2}\|_2$ for all interpolators and bad interpolators respectively. For bad interpolators that are not aligned with the true direction, $\Theta^\star$, the labels $\mathbf{y}$ are almost independent of the activation matrix $\mathbf{U}$. Hence, a lower bound on $\|\mathbf{a}_{\min,\ell_2}\|$ depends on $\mathrm{Tr}(\mathbf{U}\mathbf{U}^\top)^{-1})$, by a Hanson-Wright bound (Lemma 16). In contrast, the flattest interpolator is achieved when $\mathbf{y}$ and activation are completely dependent variables, obtaining a bound of $\sigma_{\max}^{-2}(\mathbf{U})$, by Cauchy-Schwarz.

Finally, Lemma 2 establishes a difference in the eigenvalues of the activation matrix. From Assump. 1, the activation matrix $\mathbf{U}$ has non-zero mean. Hence, $\mathbf{U}\mathbf{U}^\top$ has a very large maximum eigenvalue, $\sigma_{\max}^2(\mathbf{U}) \asymp mn$, while the rest of its eigenvalues are much smaller, $\asymp m$. This separation between eigenvalues of $\mathbf{U}$ allows us to establish a separation between bad interpolators and the flattest interpolator. To establish these bounds on eigenvalues of the activation matrix $\mathbf{U}$, whose columns might be correlated, note that $\sigma_{\max}$ doesn't require any assumptions, but we do need Condition 3 in Def. 2 for all other σ_i. We only use σ_i for bad interpolators in our proof to ensure interpolation.

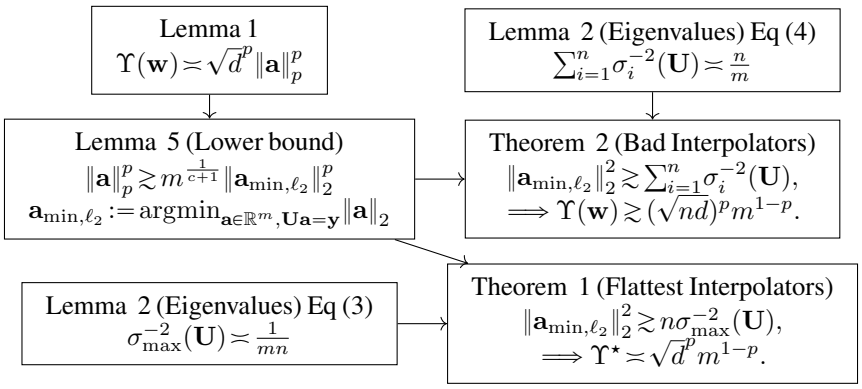


Figure 1: Proof Sketch of Thms. 1 and 2.

Proof Sketch of Thm. 3. For the existence proof in Theorem 3, note that $\|\mathbf{a}\|_\infty = \mathcal{O}(m^{-1})$ is a sufficient condition to achieve $\Upsilon(\mathbf{w}) \asymp \Upsilon^\star$. Repeating the true 2-layer network in Assumption 3 produces an outer-weight vector $\mathbf{a}$ with $\|\mathbf{a}\|_\infty = \mathcal{O}(m^{-1})$. However, due to approximation error and noise, interpolation forces $\|\mathbf{a}\|_1 \gtrsim \sqrt{n}(\zeta+\Delta)$. To ensure we still satisfy $\Upsilon(\mathbf{w}) \asymp \Upsilon^\star$, small noise and approximation error in Assump. 3 becomes necessary. In fact, the constraint of $n \asymp d$ comes from interpolating these noise and approximation terms with n neurons using Condition 4 of Def. 2.

For the generalization bound, we quantify the Rademacher complexity of the class of 2-layer networks that have $\|\mathbf{a}\|_1 = \tilde{\mathcal{O}}(1)$. Existing works (Neyshabur et al., 2015; Golowich et al., 2018; Wen et al., 2023) have shown similar bounds under this condition for ReLU networks. This norm condition yields width-independent Rademacher complexity bounds. The following bound summarizes this, and is a consequence of Cauchy-Schwarz.

$$F(\mathbf{w}) - F^\star \lesssim \hat{\mathfrak{R}}_n(\sigma) = \mathop{\mathbb{E}}_{\varepsilon_i \sim \mathrm{Unif}\{\pm 1\}} \left[\sup_{\|\theta\|_2 = 1} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \sigma(\langle \theta, \mathbf{x}_i \rangle) \right| \right].$$

After this, our proof differs from all existing works, as we need to bound the Rademacher complexity, $\hat{\mathfrak{R}}_n(\sigma)$, for arbitrary c-homogeneous activations, not just ReLU. A naive truncation based bound to make a c-homogeneous activation Lipschitz incurs a $\frac{d^c}{n}$ error which is not tight. To ensure our bounds are tighter for $c > 1$, we use Dudley's chaining (Wainwright, 2019, Section 5.3.3), (Dirksen, 2015) from scratch to precisely handle the Sub-Weibull tails of c-homogeneous activation. This is summarized in Lemma 13, which ensures the $d^{\frac{c}{2}}/n^{\min\{1, \frac{c}{2}\}}$ bound.

We conclude our paper with a summary of our main contributions, promising extensions and limitations in the next Section. We defer experimental validation of our theoretical results to App. B.

6 Conclusion

In this paper, we have shown that despite symmetry that can change the flatness of an interpolator, for multi-index data, some bad interpolators are not the flattest interpolators. Further, if labels are generated as a sum of single-index models with low approximation error and low label noise, flattest interpolators generalize. Therefore, the original claim of "flat interpolators generalize" might be false with symmetry, but "flattest interpolators still generalize". This provides a connection between flatness and generalization which was unknown for general noisy data distributions in the presence of symmetry with Homogeneous NNs.

We conclude our paper with a discussion of its limitations that form promising directions for future work:

Data-dependent θ_j and Eq (1): If $\{\theta_j\}_{j\in[m]}$ were close to a constant dimensional subspace of $\mathbb{S}^{d-1}$ for the flattest interpolator, then Eq (1) automatically holds, and flattest interpolator generalizes with $n \asymp d$ samples for $c > 2$ in Thm. 3. The main reason for these two issues is not optimizing the original flatness expression $\Upsilon(\mathbf{w})$ over inner-layer weights $\{\theta_j\}_{j\in[m]}$, which we believe can show that the flattest interpolator does lie close to this subspace.

Higher label noise: A high label-noise level makes the flatness of every interpolator the same as that of a bad interpolator in Thm. 2, as the label $\mathbf{y}$ effectively has a large component independent of activation. To extend our results to both constant ζ^2 and constant SNR, which is the case for gradient-based methods on single-index and sum of single-index data Oko et al. (2024b;a), we need to consider weights beyond interpolators, $F_S(\mathbf{w}) = 0$, to approximate interpolators, $F_S(\mathbf{w}) \lesssim \zeta^2$, as none of the gradient methods converge to interpolators in finite steps.

Broader necessary and sufficient conditions: While our setting is more general than that of existing works Ding et al. (2024), it is still focused on a particular form of symmetry, a particular class of multi-index data, and a particular NN architecture. In light of negative results from Schliserman et al. (2025); Wen et al. (2023), there is still hope for discovering a more general set of necessary and sufficient conditions for flattest interpolators to generalize, of which our problem setup is a special instance.

Algorithms that converge to flattest interpolators: Our work does not provide any algorithms that can converge to these flattest interpolators. Existing optimizers like SAM (Foret et al., 2021; Kwon et al., 2021) and its variants, or (S)GD with large step-sizes (Cohen et al., 2021), attempt to reduce flatness; however, these are not theoretically guaranteed to reach the flattest interpolators (Schliserman et al., 2025; Wen et al., 2023) except in specific models (Wu et al., 2023).

AI USE STATEMENT

In this work, we used generative AI tools to verify the correctness of our proofs and to improve the writing of the paper. We have not used AI tools to generate datasets, or in developing theoretical models or conceptual frameworks, or in formulating mathematical claims, or in mathematical proofs or experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For all our theoretical results, we have included a proof sketch in Section 5. Further, Appendix D includes the proofs for all the main results (Lem. 3, Thms. 1, 2 and 3). The remaining proofs of all theoretical results and the technical tools required for them are described in Appendix E. The experiments section in Appendix B is self-contained with a full experimental setup to reproduce our experiments, with the code for them provided in the supplementary material.

REFERENCES

Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pp. 2552–2623, 12–15 Jul 2023.

Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162(3-4):531–586, November 2014.

Linara Adilova, Amr Abourayya, Jianning Li, Amin Dada, Henning Petzka, Jan Egger, Jens Kleesiek, and Michael Kamp. Fam: Relative flatness aware minimization, 2023.

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 840–902. PMLR, 23–29 Jul 2023.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021.

Daniel Bartl and Shahar Mendelson. Random embeddings with an almost gaussian distortion. *Advances in Mathematics*, 400, May 2022.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482, March 2003. ISSN 1532-4435.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow part i: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, 78(12):2354–2435, 2025.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Joan Bruna and Daniel Hsu. Survey on Algorithms for Multi-Index Models. *Statistical Science*, 40 (3):378 – 391, 2025.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: biasing gradient descent into wide valleys*. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, dec 2019. doi: 10.1088/1742-5468/ab39d9.

Dennis Chemnitz and Maximilian Engel. Characterizing Dynamical Stability of Stochastic Gradient Descent in Overparameterized Learning, September 2024.

Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1161–1227. PMLR, 09–12 Jul 2020.

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.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Alex Damian, Jason D. Lee, and Joan Bruna. The generative leap: Tight sample complexity for efficiently learning gaussian multi-index models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Leonardo Defilippis, Yatin Dandi, Pierre Mergny, Florent Krzakala, and Bruno Loureiro. Optimal spectral transitions in high-dimensional multi-index models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Lijun Ding, Dmitriy Drusvyatskiy, Maryam Fazel, and Zaid Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 13(2):iaae009, 05 2024. ISSN 2049-8772. doi: 10.1093/imaiai/iaae009.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 06–11 Aug 2017.

Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(none), 2015.

Gintare Karolina Dziugaite and Daniel Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of entropy-SGD and data-dependent priors. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1377–1386. PMLR, 10–15 Jul 2018.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1), February 2010. ISSN 0090-5364.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2668–2703. PMLR, 02–05 Jul 2022.

Khashayar Gatmiry, Zhiyuan Li, Tengyu Ma, Sashank J. Reddi, Stefanie Jegelka, and Ching-Yao Chuang. What is the inductive bias of flatness regularization? a study of deep matrix factorization models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Sebastian Goldt, Madhu S Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher–student setup*. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124010, dec 2020.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299. PMLR, 06–09 Jul 2018.

Maxime Haddouche, Paul Viallard, Umut Simsekli, and Benjamin Guedj. A pac-bayesian link between generalisation and flat minima. In *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 481–511. PMLR, 24–27 Feb 2025.

Juncai He, Tong Mao, and Jinchao Xu. Expressivity and approximation properties of deep neural networks with relu^k activation, 2024.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, NIPS'94, pp. 529–536, Cambridge, MA, USA, 1994. MIT Press.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5905–5914. PMLR, 18–24 Jul 2021.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.

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

Tongtong Liang, Alexander Cloninger, Rahul Parhi, and Yu-Xiang Wang. Generalization below the edge of stability: The role of data geometry, 2025a.

Tongtong Liang, Dan Qiao, Yu-Xiang Wang, and Rahul Parhi. Stable minima of reLU neural networks suffer from the curse of dimensionality: The neural shattering phenomenon. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

MOSEK ApS. *The MOSEK optimization toolbox for Python, version 11.1*, 2026. URL `https://www.mosek.com`.

Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7108–7118. PMLR, 13–18 Jul 2020.

Rotem Mulayoff and Tomer Michaeli. Exact Mean Square Linear Stability Analysis for SGD. In *Proceedings of Thirty Seventh Conference on Learning Theory*, pp. 3915–3969. PMLR, June 2024. ISSN: 2640-3498.

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The Implicit Bias of Minima Stability: A View from Function Space. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17749–17761. Curran Associates, Inc., 2021.

Mor Shpigel Nacson, Rotem Mulayoff, Greg Ongie, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability in multivariate shallow relu networks. In *ICLR*, 2023.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

Behnam Neyshabur, Russ R Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Eshaan Nichani, Alex Damian, and Jason D. Lee. Provable guarantees for nonlinear feature learning in three-layer neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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 *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4009–4081. PMLR, 2024a.

Kazusato Oko, Denny Wu, Jason D. Lee, and Taiji Suzuki. Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024b.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18420–18432. Curran Associates, Inc., 2021.

Dan Qiao and Yu-Xiang Wang. Does flatness imply generalization for logistic loss in univariate two-layer relu network?, 2025.

Dan Qiao, Kaiqi Zhang, Esha Singh, Daniel Soudry, and Yu-Xiang Wang. Stable minima cannot overfit in univariate reLU networks: Generalization by large step sizes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1 – 9, 2013. doi: 10.1214/ECP.v18-2865.

Holger Sambale. Some notes on concentration for α-subexponential random variables. In *High Dimensional Probability IX*, pp. 167–192, Cham, 2023. Springer International Publishing. ISBN 978-3-031-26979-0.

Matan Schliserman, Shira Vansover-Hager, and Tomer Koren. Flat minima and generalization: Insights from stochastic convex optimization, 2025.

Jonathan W. Siegel and Jinchao Xu. High-order approximation rates for shallow neural networks with cosine and reluk activation functions. *Applied and Computational Harmonic Analysis*, 58: 1–26, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.005.

Berfin Simsek, Amire Bendjeddou, and Daniel Hsu. Learning gaussian multi-index models with gradient flow: Time complexity and directional convergence. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4204–4212. PMLR, 03–05 May 2025.

Elias M. Stein and Rami Shakarchi. *Functional Analysis: Introduction to Further Topics in Analysis*, volume 4 of *Princeton Lectures in Analysis*. Princeton University Press, Princeton, NJ, 2011. ISBN 978-0691113876.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9636–9647. PMLR, 13–18 Jul 2020.

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), January 2020. ISSN 2049-1573. doi: 10.1002/sta4.318.

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

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How sharpness-aware minimization minimizes sharpness? In *The Eleventh International Conference on Learning Representations*, 2023.

Jingfeng Wu, Vladimir Braverman, and Jason D. Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Lei Wu and Weijie J Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37656–37684. PMLR, 23–29 Jul 2023.

Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Lei Wu, Mingze Wang, and Weijie Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35:4680–4693, December 2022.

Yunfei Yang and Ding-Xuan Zhou. Optimal rates of approximation by shallow relu^k neural networks and applications to nonparametric regression. *Constructive Approximation*, 62(2):329–360, February 2024. ISSN 1432-0940. doi: 10.1007/s00365-024-09679-z.

Ping yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, February 2021. ISSN 0001-0782. doi: 10.1145/3446776.

Huiming Zhang and Songxi Chen. Concentration inequalities for statistical inference. *Communications in Mathematical Research*, 37(1):1–85, June 2021. ISSN 2707-8523.

Huiming Zhang and Haoyu Wei. Sharper sub-weibull concentrations. *Mathematics*, 10(13), 2022. ISSN 2227-7390. doi: 10.3390/math10132252.

Bo Zhao, Robin Walters, and Rose Yu. Symmetry in neural network parameter spaces. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

CONTENTS

1 Introduction **1**

1.1 Main Contributions 2

1.2 Comparison to Related Works 3

2 Setup **4**

2.1 Flatness under Rescaling Symmetry 5

3 Bad Interpolators That are Not Flattest **5**

4 Flattest Interpolators Generalize **7**

5 Proof Sketches **8**

6 Conclusion **9**

A Related Works **16**

B Experiments **17**

B.1 Setup 17

B.2 Results 18

C Preliminaries **20**

C.1 Population Loss 20

C.2 Flatness and Rescaling Symmetry 21

C.3 Properties of Piece-wise Polynomial Activations (Assump. 1) 22

D Main Proofs **22**

D.1 Derivation of $\Upsilon(\mathbf{w})$: Proof of Lem. 1 23

D.2 Lower bounds on $\Upsilon(\mathbf{w})$: Proof of Thm. 1 24

D.3 Flatness of Bad Interpolators: Proof of Thm. 2 28

D.4 Population loss of Flattest Interpolators : Proof of Thm. 3 32

E Technical Tools and Missing Proofs **38**

E.1 Sub-Weibull Distributions 38

E.2 Proofs for App. C 42

E.2.1 Proof of Prop. 3 42

E.2.2 Proof of Lem. 2 44

E.3 Proof of Lem. 4 45

E.4 Proofs for App. D.3 46

E.4.1 Proof of Prop. 4 46

E.4.2 Proof of Example 1 46
E.4.3 Proof of Lem. 8 47
E.4.4 Proof of Lem. 9 47
E.4.5 Proof of Lem. 10 48
E.4.6 Proof of Lem. 11 48
E.5 Proofs for App. D.4 49
E.5.1 Proof of Thm. 4 49
E.5.2 Proof of Prop. 2 49
E.6 Proof of Lem. 13 49

A Related Works

Existing works focus on flatness of even local minima of empirical loss, not just interpolators. Throughout this section, we will discuss flatness of an arbitrary minima of the empirical loss. Interpolators are global minima of the empirical loss, with 0 empirical loss.

Experimental Evidence Linking Flatness and Generalization There have been several attempts to create algorithms that seek flatter minima in non-convex neural networks. Hochreiter & Schmidhuber (1994) were the first to use variants of gradient descent to explicitly search for flat minima. For a large class of algorithms to find flat minima, some form of noise is added to the updates of gradient descent, either independent noise like SGLD (Chaudhari et al., 2019) or dependent noise from stochastic gradients (Keskar et al., 2017). In practice, these algorithms also converge to good minima. Foret et al. (2021) propose the Sharpness-Aware Minimization (SAM) as an alternative to SGD. This algorithm intuitively minimizes sharpness and in practice converges to good minima. yeh Chiang et al. (2023) implement non-gradient based algorithms that only depend on flatness of the empirical landscape also converge to good minima. Jiang et al. (2020) and Andriushchenko et al. (2023) benchmark the correlation between flatness and generalization for a large class of real-world NN architectures and datasets. Jiang et al. (2020) compute flatness by measuring the sensitivity of the loss to perturbations, while Andriushchenko et al. (2023) use a version of this flatness metric that is invariant under rescaling symmetry. Jiang et al. (2020) show strong correlation between flatness and generalization for most cases, while Andriushchenko et al. (2023) show that this correlation is high only for specific architecture and dataset combinations.

Theoretical Flatness-based Generalization bounds Most flatness-based generalization bounds can be classified into 2 categories – minima stability bounds and PAC-Bayesian bounds. Minima stability (Wu et al., 2018; 2022; Mulayoff & Michaeli, 2024; Nacson et al., 2023; Chemnitz & Engel, 2024) shows that (S)GD with a large step size can stably converge to a flat minima whose flatness is proportional to the step size. The goal of minima stability-based generalization is to bound the generalization of the set of flat minima. For specific examples, the set of flat minima has weights with small norm. Then, they use generalization bounds for models with small weight norm (Neyshabur et al., 2015). These bounds are available for a few specific settings:classification and regression on 2-layer ReLU networks (Mulayoff et al., 2021; Qiao et al., 2024; Liang et al., 2025b; Qiao & Wang, 2025; Liang et al., 2025a) and diagonal neural networks (Wu & Su, 2023). Some of these bounds become vacuous under rescaling symmetry (Wu & Su, 2023), while others do not apply to our setting as they need $d = 1$ (Mulayoff et al., 2021; Qiao et al., 2024; Liang et al., 2025b; Qiao & Wang, 2025) or $d \geqslant 1$, but $\mathbf{x}$ is a finite mixture of low-dimensional subspaces (Liang et al., 2025a). The only bounds that are applicable for $d \geqslant 1$ are from Liang et al. (2025b), however these have excess population risk $\mathcal{O}(d^{-\frac{1}{d}}) = \mathcal{O}(1)$ for our setting with optimal sample complexity $n = \tilde{\Theta}(d)$. The core idea of PAC-Bayesian approaches is to separate a term in the upper bound on excess population loss that is related to the flatness, $\nabla^2 F_S(\mathbf{w})$. Dziugaite & Roy (2018) show that this happens for Entropy-SGD algorithm in (Chaudhari et al., 2019), while Haddouche et al. (2025) show this more generally. However, these bounds are still sensitive to rescaling symmetry, and are not tight for arbitrary neural networks with our data distribution. In fact, note that any flatness-based generalization bound that measures the population loss of all empirical minima that are Υ-flat for some $\Upsilon > 0$,

i.e., $\mathrm{Tr}(\nabla^2 F_S(\mathbf{w})) \leqslant \Upsilon$, is vacuous for $\Upsilon \gtrsim (dn)^{\frac{c}{c+1}} m^{-\frac{c-1}{c+1}}$. This is due to the existence of bad interpolators in Definition 2, which can achieve this flatness using rescaling symmetry (Theorem 2).

Flattest Minima under Rescaling Dinh et al. (2017) show that good minima can be made sharp for most reasonable definitions of sharpness using this rescaling symmetry. Several attempts have been made to handle this rescaling symmetry. For simpler models like matrix factorization, 2-layer NN with quadratic activation (Ding et al., 2024), and deep linear NN (Mulayoff & Michaeli, 2020), the flattest minima under rescaling can be computed. Our settings cover most of their results, apart from the case of deep linear networks, where they establish a connection between depth and flatness. Tsuzuku et al. (2020) and Adilova et al. (2023) use generalization bounds to obtain algorithms that find flat minima invariant to rescaling symmetry. To the best of our knowledge, only Petzka et al. (2021) provide a generalization bound invariant under scaling for classification with a general loss function, model architecture and data distribution. Their generalization bound encodes flatness using a form of robustness to features. However, their assumptions are very restrictive and not applicable to our setting. In particular, they cannot be extended to regression, or non-zero label noise, or if the labels are not locally constant. Additionally, their generalization bound for $n \asymp d$, which is the optimal sample complexity in our setting, is $\mathcal{O}(d^{-1/d}) = \mathcal{O}(1)$, making their bounds vacuous. As for deep linear networks (Gatmiry et al., 2023), under regularity conditions like the restricted isometry property (RIP) (Wainwright, 2019, Chapter 7), the flattest interpolator of deep linear networks for low-rank matrix factorization recovers the ground truth and thus perfectly generalizes. However, the regularity conditions, like RIP, are not satisfied in our settings, as we have $m \geqslant n$ (Gatmiry et al., 2023, Corollary 1). This result crucially uses uniform convergence bounds (Gatmiry et al., 2023, Theorem 6) for generalization, which are vacuous in our overparametrized settings.

Multi-Index Data Distributions. Multi-index data distributions have been extensively studied in traditional statistics (Li, 1991); however, more recently, these have become an appropriate test-bed to study learning in 2-layer neural networks. For 2-layer neural networks, learning guarantees for gradient-based algorithms, for instance, Online-SGD or SGD with batch-reuse, exist for single-index (Arous et al., 2021; Oko et al., 2024b) and sum of single-index models (Oko et al., 2024a). For deeper networks, there has been significant work on multi-index models with a hierarchical structure (Damian et al., 2026; Abbe et al., 2023) or 3-layer networks (Nichani et al., 2023). All these works require much smaller sample complexity than the NTK regime (Jacot et al., 2018), often dependent on Hermite coefficients of the link function $\sigma^\star$, for instance, the information (Arous et al., 2021), generative (Damian et al., 2023) and leap exponents (Abbe et al., 2023; Damian et al., 2026). For more general multi-index models, existing works can guarantee that gradient-based algorithms can recover the true direction $\Theta^\star$; however, they cannot show generalization without additional assumptions Bietti et al. (2025); Simsek et al. (2025); Defilippis et al. (2026). Note that all these works characterize the population loss of the output of gradient-based methods; however, they don't consider the flatness of the loss landscape. A detailed survey of existing works in multi-index models is provided in (Bruna & Hsu, 2025).

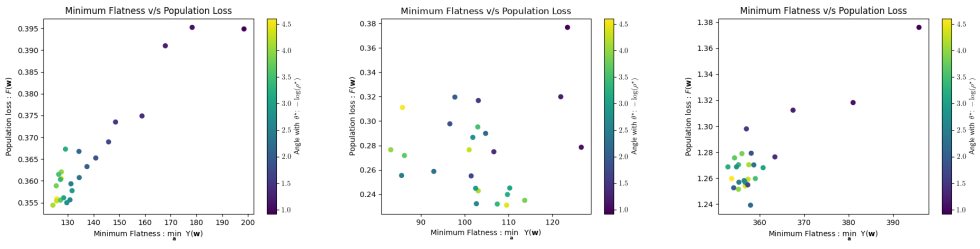
B Experiments

In this section, we empirically verify our theoretical results for several single-index and multi-index problems. We first describe our experimental setup.

B.1 Setup

For all our experiments, we set $d = 100, n = 500$ and $m = 5000$. We use 3 different activation functions. For $c = 1$, we use ReLU and LeakyReLU with coefficient $c'' = -0.1$. For $c = 2$, we use quadratic activation.

For the data distribution, we consider single-index and sum of single-index models as described in Section 2. In each of these distributions, the link functions $\sigma^\star$ and $\tilde{\sigma}^\star$ are obtained by adding a scalar sampled from $\mathrm{Unif}([-\Delta,\Delta])$ to a sum of activations, thus satisfying Assump. 3. For all cases other than ReLU activation, we set $\Delta = m^{-\frac{1}{2}} n^{-\frac{1}{2}-\epsilon_1}$ and $\zeta = m^{-\frac{1}{2}} n^{-\frac{1}{2}-\epsilon_2}$ with $\epsilon_1, \epsilon_2 \sim \mathrm{Unif}([0,0.1])$. For ReLU activation, we consider $\zeta = \Delta = 0$, so the activation matches the link function. For the sum of single-index models, we select the coefficients $a_j^\star \sim \mathrm{Unif}([-1,1])$ and set $m^\star = 5$.



(a) Activation: Leaky ReLU (b) Activation: ReLU (c) Activation: Quadratic

Figure 2: Flatness and population loss for learning single-index link functions close to activations. In all cases except ReLU, the flattest solution generalizes.

We use 3 additional data distributions to cover a varied set of examples. First, we use a single-index model with a linear link function $\sigma^\star(b) = b, \forall b \in \mathbb{R}$ and a ReLU activation. Note that for this model, $\mathrm{ReLU}(b) - \mathrm{ReLU}(-b) = \sigma^\star(b), \forall b \in \mathbb{R}$. Therefore, using a ReLU network should be able to learn this data model. Second, we use a single-index model with linear link function and linear activation. Note that linear activation has $c' = 1, c'' = -1$, so it doesn't satisfy our requirements for Piecewise-Polynomial activations in Assump. 1 and our analysis should not cover it. This example corresponds to the case of matrix factorization with $r = 1$, matching the example of Gatmiry et al. (2023), however, as $m > n$, RIP property should not hold. Finally, we use an example that is not multi-index with $m^\star = \mathcal{O}(1)$, where the link function is quadratic, but $m^\star = \frac{d}{2}$. As $m^\star = \Theta(d)$, this problem is no longer a multi-index problem, so our results should not hold. We use a ReLU activation for this link. Additionally, while the link function $\sigma^\star$ satisfies Assump. 2 for $c = 2$, learning it with ReLU activation implies a mismatch in homogeneity of the activation and link as $c = 1$ for ReLU activation.

For all our data distributions and activations, we first sample a set of random inner-layer weights $\{\theta_j\}_{j\in[m]}$ each of unit norm such that the correlation of each weight to the subspace spanned by the vectors in $\Theta^\star$, is exactly $\rho^\star$. Note that this is similar to our definition of bad minima in Def. 2. We then find the vector $\mathbf{a} \in \mathbb{R}^m$ that is the solution of the following optimization problem,

$$\min_{\mathbf{a}\in\mathbb{R}^m} \Upsilon(\mathbf{w}) \coloneqq \sum_{j=1}^{m} B(\theta_j)|a_j|^{\frac{2c}{c+1}}, \text{ such that } \sum_{j=1}^{m} a_j \sigma(\langle \theta_j, \mathbf{x}_i \rangle) = y_i, \forall i \in [n] \tag{2}$$

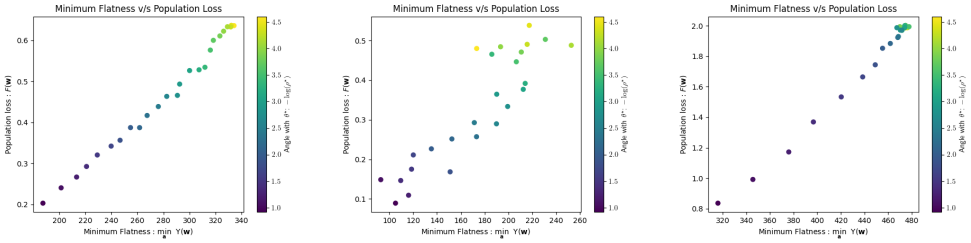
The solution to this problem should give us the outer-layer weights that correspond to the flattest interpolator after rescaling according to Lemma 1. For fixed inner-layer weights, the above problem is a convex program with linear constraints. Therefore, we use the convex optimization solver CVXPY (Diamond & Boyd, 2016; Agrawal et al., 2018) with MOSEK (MOSEK ApS, 2026), to solve it.

We choose 25 different logarithmically equally spaced values for $\rho^\star$ in the range $[d^{-2}, \frac{1}{2}]$. For each value of $\rho^\star$, we solve Eq. (2) to compute the outer-layer weights $\mathbf{a}$ and minimum flatness $\Upsilon(\mathbf{w})$ corresponding to this $\mathbf{a}$. Then, we compute the population loss for this choice of inner-layer and outer-layer weights for the 2-layer network by measuring the squared error on a fresh sample of 1500 datapoints. We provide a scatter plot of the flatness $\Upsilon(\mathbf{w})$ and the population loss $F(\mathbf{w})$ for each $\rho^\star$, where the color denotes the value of $-\log(\rho^\star)$. The plots are averaged over 5 random seeds, and the experiments took 5 hours on a single CPU with 15 cores and $20GB$ RAM.

We plot the results for single-index, sum of single-index, and the 3 special data distributions in Figures 2, 3 and 4 respectively.

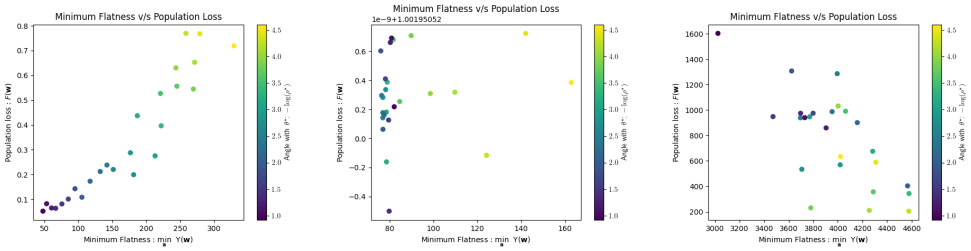
B.2 Results

Note that our theoretical results establish that bad interpolators are not flattest (Thm. 2) and that flattest interpolators generalize (Thm. 3). Therefore, to verify them in the flatness v/s population loss plots, we should ideally check if large population loss interpolators are not flattest, i.e., the top-right corner of the plot, and if flattest interpolators generalize, i.e., the bottom-left corner of the plot.



(a) Activation: Leaky ReLU (b) Activation: ReLU (c) Activation: Quadratic

Figure 3: Flatness and population loss for learning sum of single-index link functions, each close to activations.In all cases, the flattest solution generalizes.



(a) Link: Linear, Activation: ReLU (b) Link: Linear, Activation: Linear (c) Link: Quadratic, Activation: ReLU, $m^\star = \frac{d}{2}$

Figure 4: Flatness and population loss for special data distributions. For learning linear link with ReLU activation, flattest interpolator generalizes. For linear activation with linear link all presented interpolators, achieve small flatness and generalize, however for the last case, flatttest interpolator doesn't generalize.

Single-Index Models. For single-index models, there is a link between flattest interpolators and their population loss. From Fig 2, we can see that for LeakyReLU and Quadratic activation, if the minimum flatness after rescaling, $\min_{\mathbf{a}} \Upsilon(\mathbf{w})$, is low, then population loss is also low. However, extremely high correlation with the true direction, i.e., large $\rho^\star$ forces large $\|\mathbf{a}\|_\infty$ to ensure interpolation. Thus, large $\rho^\star$ does not necessarily achieve the minimum flatness and, in turn, has a high population loss. This phenomenon results in a poor connection between flatness and generalization for ReLU activation, where very small $\rho^\star$ can easily interpolate with small $\|\mathbf{a}\|_\infty$, but as $\rho^\star$ is small, its population loss is large.

Sum of Single-Index Models. For the sum of single-index models, the connection between flatness and generalization is much stronger. Further, there is also a connection between the correlation $\rho^\star$ and flatness. From Fig 3, this connection holds for all activations, even ReLU, where it seemed to be absent for single-index models. If we increase $\rho^\star$, the correlation to the true direction, the population loss decreases, and the minimum flatness $\min_{\mathbf{a}} \Upsilon(\mathbf{w})$ also decreases. This stronger connection is precisely because multi-index models with diverse features can interpolate much better for a large $\rho^\star$ than single-index models, as their activation matrices Φ are appropriately conditioned.

Special Cases. For the special case of learning a linear function using ReLU in Fig 4a, although this is a single-index model, it behaves similarly to our results for multi-index models, with a strong connection between the correlation, population loss and flatness. For a linear link function with linear activation in Fig 4b, it seems that most interpolators obtain the same flatness. While their population losses can be different, all these population losses are very small, even for different $\rho^\star$. This is because for linear activation, the outer-layer $\mathbf{a}$ weights can be selected so that $\sum_{j=1}^m a_j \theta_j$ is parallel to $\theta^\star$. Therefore, for linear link function and linear activation, empirically flattest interpolators seem to generalize, however the theoretical justification in (Gatmiry et al., 2023) is still loose. For multi-index model in Fig 4c, note that it is no longer a multi-index model, hence flattest interpolators need not generalize.

ReLU Single-Index and Sum of Single-Index Models (Fig 2b and 3b) For a ReLU link function, in addition to the issues discussed above, note that a constant fraction of the responses y_i are exactly 0. This further makes interpolation hard. However, as we set $\Delta = 0, \zeta = 0$ for this case, we can exactly compute the flattest interpolators for this case using Thm. 4. This flattest interpolator repeats the true link function ,$\sigma^\star(\Theta^\star \mathbf{x})$, $\frac{m}{m^\star}$ times with $\|\mathbf{a}\|_\infty = \frac{m^\star}{m}$. For the single-index case, Fig 2b, this interpolator achieves flatness 8.471 on this data and has population loss exactly 0. For sum of single-index case, Fig 3b, this interpolator achieves flatness 7.484, and als has population loss exactly 0. Therefore, adding these points to the corresponding plots in Fig 2b and Fig 3b, we find that these plots are indeed at the bottom-left and thus show that flattest interpolators generalize. Therefore, although the original Fig 2b and 3b might appear to show no connection between flatness and generalization, our theoretical results of flattest interpolators generalizing, and bad interpolators not being flattest are still true for these cases.

C PRELIMINARIES

In this section, we describe core technical results required for our results. Most of their proofs are deferred to App. E.

Additional Notation Recall that $\mathbf{w} \in \mathbb{R}^{m(d+1)}$ is used to denote the weights of an interpolator with unit-norm inner-layer weights $\{\theta_j\}_{j\in[m]}$, succintly represented by the matrix $\Theta \in \mathbb{R}^{m\times d}$, and outer layer weights $\mathbf{a} \in \mathbb{R}^m$. Further, $\mathbf{y} \in \mathbb{R}^n$ denotes the vector of labels, with $\mathbf{y} = \mathbf{y}^\star + \mathbf{n}$, where $\mathbf{n}$ is the vector of label noise $\{\xi_i\}_{i\in[n]}$ and $\mathbf{y}^\star$ is the vector of true responses, $\{\sigma^\star(\Theta^\star \mathbf{x}_i)\}_{i\in[n]}$. We use $\widetilde{\mathbf{U}} \coloneqq (\mathbf{U}\mathbf{U}^\top)^{-1}$. We use $\mathbf{1}_k, \mathbf{0}_k \in \mathbb{R}^k$ to represent the vectors with all its coordinates 1 or 0.

C.1 POPULATION LOSS

We first provide the expression of population in terms of easier to handle quantities. We define the functions $\widetilde{\psi}: [-1,1]^{m^\star} \to \mathbb{R}$ and $\phi: [-1,1] \to \mathbb{R}$ as the following,

$$\forall \Xi \in [-1,1]^{m^\star}, \quad \widetilde{\psi}(\Xi) = \mathbb{E}[\sigma^\star(\mathbf{q}_2)\sigma(q_1)], \text{ where } \begin{bmatrix} q_1 \\ \mathbf{q}_2 \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \Xi^\top \\ \Xi & \mathbb{I}_{m^\star} \end{bmatrix}\right)$$
$$\forall \Xi \in [-1,1], \quad \phi(\Xi) = \mathbb{E}[\sigma(q_2)\sigma(q_1)], \text{ where } \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \Xi \\ \Xi & 1 \end{bmatrix}\right).$$

Note that $\phi(1) = \phi^\star = \mathop{\mathbb{E}}_{q\sim\mathcal{N}(0,1)}\left[\sigma^2(q)\right]$. Using these functions, we can compute the population loss of any weight $\mathbf{w} \in \mathbb{R}^d$ with inner layer weights $\{\theta_j\}_{j\in[m]}$, $\theta_j \in \mathbb{S}^{d-1}, \forall j \in [m]$, and outer-layer weights $\mathbf{a} \in \mathbb{R}^m$.

$$\begin{aligned} 2F(\mathbf{w}) &= \mathbb{E}\left[\left(y - \sum_{j=1}^m a_j \sigma(\langle \theta_j, \mathbf{x}\rangle)\right)^2\right] = \mathop{\mathbb{E}}_{\mathbf{x},\xi}\left[\left(\sigma^\star(\Theta^\star\mathbf{x}) + \xi - \sum_{j=1}^m a_j \sigma(\langle \theta_j, \mathbf{x}\rangle)\right)^2\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\sigma^\star(\Theta^\star\mathbf{x}) - \sum_{j=1}^m a_j \sigma(\langle \theta_j, \mathbf{x}\rangle)\right)^2\right] + \mathop{\mathbb{E}}_{\xi}\left[\xi^2\right] + 2\mathop{\mathbb{E}}_{\xi}[\xi]\mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\sigma^\star(\Theta^\star\mathbf{x}) - \sum_{j=1}^m a_j \sigma(\langle \theta_j, \mathbf{x}\rangle)\right)\right] \end{aligned}$$

We first expand the terms corresponding to the noise ξ and plug in its mean and variance.

Now, we expand the terms of activation.

$$
\begin{aligned}
2F(\mathbf{w}) &= \mathop{\mathbb{E}}_{\mathbf{x}}\left[(\sigma^\star(\Theta^\star\mathbf{x}) - \sum_{j=1}^{m} a_j\sigma(\langle\theta_j,\mathbf{x}\rangle))^2\right] + \zeta^2 \\
&= \mathop{\mathbb{E}}_{\mathbf{x}}\left[(\sigma^\star(\Theta^\star\mathbf{x}))^2\right] + \mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\sum_{j=1}^{m} a_j\sigma(\langle\theta_j,\mathbf{x}\rangle)\right)^2\right] - 2\sum_{j\in[m]} a_j \mathop{\mathbb{E}}_{\mathbf{x}}[\sigma^\star(\Theta^\star\mathbf{x})\sigma(\langle\theta_j,\mathbf{x}\rangle)] + \zeta^2 \\
&= \psi^\star + \zeta^2 + \sum_{j,j'\in[m]}\sum a_j a_{j'} \mathop{\mathbb{E}}_{\mathbf{x}}[\sigma(\langle\theta_j,\mathbf{x}\rangle))\sigma(\langle\theta_{j'},\mathbf{x}\rangle)] - 2\sum_{j\in[m]} a_j\widetilde{\psi}(\Theta^\star\theta_j) \\
&= \psi^\star + \zeta^2 + \sum_{j,j'\in[m]}\sum a_j a_{j'}\phi(\langle\theta_j,\theta_{j'}\rangle) - 2\left\langle \mathbf{a},\widetilde{\Psi}\right\rangle \\
2F(\mathbf{w}) &= \psi^\star + \zeta^2 + \mathbf{a}^\top\Phi\mathbf{a} - 2\left\langle \mathbf{a},\widetilde{\Psi}\right\rangle
\end{aligned}
$$

Here, $\widetilde{\Psi}\in\mathbb{R}^m$ is a vector with its jth coordinate being $\widetilde{\psi}(\Theta^\star\theta_j), \forall j\in[m]$ and $\Phi\in\mathbb{R}^{m\times m}$ is a matrix with its j,j^{th} entry being $\Phi_{j,j'} = \phi(\langle\theta_j,\theta_{j'}\rangle, \forall j,j'\in[m]$. Next, we find the optimal outer-layer weights that minimize the population loss for single-index data distributions.

C.2 Flatness and Rescaling Symmetry

In this section, we discuss the impact of symmetry on flatness. First, we state the full expression of flatness for any $\mathbf{w}\in\mathbb{R}^{m(d+1)}$.

$$
\begin{aligned}
\nabla F_S(\mathbf{w}) &= -\frac{1}{n}\sum_{i=1}^{n}(y_i - h(\mathbf{w},\mathbf{x}_i))\nabla_{\mathbf{w}}h(\mathbf{w},\mathbf{x}_i) \\
\nabla^2 F_S(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^{n}\big(\nabla_{\mathbf{w}}h(\mathbf{w},\mathbf{x}_i)(\nabla_{\mathbf{w}}h(\mathbf{w},\mathbf{x}_i))^\top - (y_i - h(\mathbf{w},\mathbf{x}_i))\nabla^2 h(\mathbf{w},\mathbf{x}_i)\big).
\end{aligned}
$$

If $\mathbf{w}$ is an interpolator, $F_S(\mathbf{w}) = 0$, so $h(\mathbf{w},\mathbf{x}_i) = y_i, \forall i\in[n]$. This removes the second term in $\nabla^2 F_S(\mathbf{w})$. Therefore, for any interpolator, flatness is,

$$
\mathrm{Tr}(\nabla^2 F_S(\mathbf{w})) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}\big(\nabla_{\mathbf{w}}h(\mathbf{w},\mathbf{x}_i)(\nabla_{\mathbf{w}}h(\mathbf{w},\mathbf{x}_i))^\top\big) = \frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\mathbf{w}}h(\mathbf{w},\mathbf{x}_i)\|_2^2.
$$

Now, we compute the value of $\nabla_{\mathbf{w}}h(\mathbf{w},\mathbf{x})$. Note that $h(\mathbf{w},\mathbf{x}) = \sum_{j=1}^{m} a_j\sigma(\langle\theta_j,\mathbf{x}\rangle)$. Therefore,

$$
\nabla_{a_j}h(\mathbf{w},\mathbf{x}) = \sigma(\langle\theta_j,\mathbf{x}\rangle), \quad \nabla_{\theta_j}h(\mathbf{w},\mathbf{x}) = a_j\sigma'(\langle\theta_j,\mathbf{x}\rangle)\mathbf{x}.
$$

Plugging these values into $\mathrm{Tr}(\nabla^2 F_S(\mathbf{w}))$, we obtain the flatness of any interpolator as follows,

$$
\mathrm{Tr}(\nabla^2 F_S(\mathbf{w})) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big(\sigma^2(\langle\theta_j,\mathbf{x}_i\rangle) + a_j^2(\sigma')^2(\langle\theta_j,\mathbf{x}_i\rangle)\|\mathbf{x}_i\|_2^2\Big).
$$

Now, applying rescaling symmetry from Def. 1, for rescaling coefficients $\{\alpha_j\}_{j\in[m]}$, the flatness of the new interpolator is,

$$
\mathrm{Tr}(\nabla^2 F_S(\widetilde{\mathbf{w}})) = \frac{1}{n}\sum_{i\in[n],j\in[m]}\sum(\alpha_j^{2c}\sigma^2(\langle\theta_j,\mathbf{x}_i\rangle) + \alpha_j^{-2}a_j^2(\sigma')^2(\langle\theta_j,\mathbf{x}_i\rangle)\|\mathbf{x}_i\|_2^2)
$$

We use the fact that σ and σ' are c and $(c-1)$ homogeneous functions from Assump. 1. Here, $\widetilde{\mathbf{w}}$ is the interpolator obtained after applying the rescaling coefficients. Therefore, $\widetilde{\mathbf{w}} = \big[\alpha_1^{-c}a_1,...,\alpha_m^{-c}a_m,\alpha_1\theta_1^\top,...,\alpha_m\theta_m^\top.\big]$

To show that $\tilde{\mathbf{w}}$ is still an interpolator, note that,

$$
h(\widetilde{\mathbf{w}},\mathbf{x}) = \sum_{j=1}^{m}\alpha_j^{-c}a_j\sigma(\langle\alpha_j\theta_j,\mathbf{x}\rangle) = \sum_{j=1}^{m}\alpha_j^{-c}a_j\alpha_j^{c}\sigma(\langle\theta_j,\mathbf{x}\rangle) = \sum_{j=1}^{m}a_j\sigma(\langle\theta_j,\mathbf{x}\rangle) = h(\mathbf{w},\mathbf{x}), \quad \forall\mathbf{x}\in\mathbb{R}^d.
$$

Therefore, $\widetilde{\mathbf{w}}$ is also an interpolator and has the same population loss as $\mathbf{w}$.

Note that changing the values of $\{\alpha_j\}_{j\in[m]}$ changes the value of $\mathrm{Tr}(\nabla^2 F_S(\widetilde{\mathbf{w}})$. The only case when this remains constant for any α_j is if $\sigma(\langle\theta_j,\mathbf{x}_i\rangle)=0,a_j,\forall j\in[m],i\in[n]$. Since $F_S(\mathbf{w})=0$, this implies, $h(\mathbf{w},\mathbf{x}_i)=y_i=0,\forall y_i\in[n]$. If $\zeta=0$, $y_i=\sigma(\langle\theta^\star,\mathbf{x}_i\rangle)$. For piecewise polynomial activations, $\sigma(b)=0$ for either $b\geqslant 0$ or $b\leqslant 0$. Since $\langle\theta^\star,\mathbf{x}_i\rangle\sim\mathcal{N}(0,1)$, this event can happen with probability atmost $\frac{1}{2}$ for each $i\in[n]$. Since we would need $y_i=0,\forall i\in[n]$, the probability of such an event is bounded by 2^{-n} by the independence of $\mathbf{x}_i$. If $\zeta\neq 0$, then $y_i=0,\forall i\in[n]$ if $\xi_i=-\sigma(\langle\theta^\star,\mathbf{x}_i\rangle),\forall i\in[n]$. Since ξ_i is a gaussian independent of $\mathbf{x}_i$, this event happens with probability 0 for each $i\in[n]$.

Note that by setting any $\alpha_j\to\infty$ or $\alpha_j\to 0$, we can increase $\mathrm{Tr}(\nabla^2 F_S(\widetilde{\mathbf{w}})\to\infty$. Therefore, we can arbitrarily make any interpolator sharper. However, it is a convex function in each α_j, hence, we cannot make any interpolator flatter.

C.3 Properties of Piece-wise Polynomial Activations (Assump. 1)

Proposition 3 (Activations and Sum of Activations are Valid Link Functions)**.** *Any Piece-wise Polynomial activation function σ (Assump. 1) satisfies Assump 2 with $m^\star=1$. Further, $\hat{\sigma}:\mathbb{R}^{m^\star}\to\mathbb{R}$ defined as the following also satisfies Assump. 2.*

$$\hat{\sigma}(\mathbf{b})=\sum_{j=1}^{m^\star}\hat{a}_j\sigma(b_j),\quad \forall\mathbf{b}\in\mathbb{R}^{m^\star}\text{ with }a_j\neq 0,|\hat{a}_j|=\mathcal{O}(1),\mathrm{sign}(\hat{a}_j)=\mathrm{sign}(\hat{a}_{j'}),\forall j,j'\in[m^\star],j'\neq j.$$

The proofs of this Prop. are provided in App. E.2.1. This proposition ensures that the sum of activations and activations themselves are valid link functions, and thus have Sub-Weibull tails.

This provides us with the following Lem. on the behaviour of their eigenvalues, which we crucially use to establish a gap between bad and good interpolators.

Lemma 2 (Eigenvalues of Activation matrix)**.** *With probability $1-\frac{\delta}{2}$, for every $b\in[0,1]$, $\exists k\in\mathbb{N}$ with $k=\mathcal{O}(m^{1-b})$ such that,*

$$\sigma_{\max}(\mathbf{U}\mathbf{U}^\top)\asymp mn,\quad \sigma_i(\mathbf{U}\mathbf{U}^\top)\lesssim m^{1+b},\forall i\in\{k,k+1,...,n\}. \tag{3}$$

Further, if $\lambda_{\min}(\bar{\Phi})=\Omega(1)$, and $\sigma_{\max}(\Theta)=\mathcal{O}(1)$, then with probability $1-\frac{\delta}{2}$,

$$\sigma_n(\mathbf{U}\mathbf{U}^\top)\gtrsim m. \tag{4}$$

Here, $\sigma_{\max}$ is the largest singular value, σ_i is the ith largest singular value of a matrix.

The proof of this Lem. is provided in App. E.2.2.

D Main Proofs

In this section, we describe our main proofs for all our theoretical results. Based on the proof sketch from Section 5, we break down our core proof into 4 important parts.

1. Derivation of $\Upsilon(\mathbf{w})$ (App. D.1).
2. Lower Bounds on $\Upsilon(\mathbf{w})$ for All Interpolators (App. D.2).
3. Lower Bounds on $\Upsilon(\mathbf{w})$ for Bad Interpolators (App. D.3).
4. Connection of $\Upsilon^\star$ to population loss (App. D.4).

We will prove Lem. 1 in the first part to derive $\Upsilon(\mathbf{w})$. Then, we obtain a lower bound on $\Upsilon(\mathbf{w})$ in terms of only the inner-layer weights, $\{\theta_j\}_{j\in[m]}$, in Lem. 5. This allows us to compute $\Upsilon^\star$, proving Thm. 1. Then, we use Lem. 5 along with the definition of bad interpolators from Def. 2 to obtain the lower bound on flatness for bad interpolators, proving Thm. 2. Finally, we utilize the value of $\Upsilon^\star$ to characterize the population loss of the set of flattest interpolators, thus proving Thm. 3. Proofs of any intermediate Lems. that are not provided in Appendices D.1 - D.4 are provided in App. E.

D.1 Derivation of $\Upsilon(\mathbf{w})$: Proof of Lem. 1

Our proof heavily relies on the expression of $\Upsilon(\mathbf{w})$ from Lem. 1. We first provide a proof for this Lem. by using the following two Lems.

Lemma 3. *For any interpolator $\mathbf{w}\in\mathbb{R}^{m(d+1)}$ with unit norm inner-layer weights,*

$$\Upsilon(\mathbf{w}) = \inf_{\hat{\mathbf{w}}\in\mathcal{W}_{\text{rescale}(\mathbf{w})}} \operatorname{Tr}(\nabla^2 F_S(\hat{\mathbf{w}})) = \sum_{j=1}^{m} |a_j|^{\frac{2c}{c+1}} B(\theta_j),$$

$$\text{where } B(\theta_j) = (c^{-\frac{c}{c+1}} + c^{\frac{1}{c+1}}) B_1^{\frac{1}{c+1}}(\theta_j) B_2^{\frac{c}{c+1}}(\theta_j).$$

Here, $B_1, B_2 : \mathbb{S}^{d-1} \to \mathbb{R}_+$ are functions given by $B_1(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^n \sigma^2(\langle\theta,\mathbf{x}_i\rangle)$ and $B_2(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^n (\sigma')^2(\langle\theta,\mathbf{x}_i\rangle)\|\mathbf{x}_i\|_2^2$.

Lemma 4 (Bounds on $B(\theta)$)**.** *For Piece-wise Polynomial Activation σ (Assump. 1), from Prop. 3, with probability $1-6\delta$, for a set $\{\theta_j\}_{j\in[m]}$, with $\theta_j\in\mathbb{S}^{d-1}, \forall j\in[m]$,*

$$\max_{j\in[m]} B(\theta_j) \lesssim d^{\frac{c}{c+1}}, \quad \min_{j\in[m]} B(\theta_j) \gtrsim d^{\frac{c}{c+1}}, \implies B(\theta_j) \asymp d^{\frac{c}{c+1}}, \forall j\in[m]$$

Using Lem. 3 and 4, we can show that with probability $1-6\delta$, each $B(\theta_j)\asymp d^{\frac{c}{c+1}}$, therefore,

$$\Upsilon(\mathbf{w}) \asymp d^{\frac{c}{c+1}} \sum_{j=1}^{m} |a_j|^{\frac{2c}{c+1}} = d^{\frac{c}{c+1}} \|\mathbf{a}\|_{\frac{2c}{c+1}}^{\frac{2c}{c+1}}.$$

This completes the proof of Lem. 1. The proof of the intermediate Lem. 4 is deferred to App. E.3. We provide a proof for Lem. 3 as it extends the flatness beyond ReLU ($c=1$ in Wen et al. (2023)) and quadratic activations ($c=2$ in Ding et al. (2024)) to any piece-wise Polynomial activations.

Proof of Lem. 3. To find the flattest empirical minimizer in $\mathcal{W}_{\text{rescale}}(\mathbf{w})$ for an empirical minimizer $\mathbf{w}\in\mathbb{R}^d$, we need to solve the optimization problem,

$$\begin{aligned}
&\min_{\{\alpha_j\}_{j\in[m]},\alpha_j>0} \frac{1}{n} \sum_{i\in[n],j\in[m]} (\alpha_j^{2c}\sigma^2(\langle\theta_j,\mathbf{x}_i\rangle) + \alpha_j^{-2} a_j^2 (\sigma')^2(\langle\theta_j,\mathbf{x}_i\rangle)\|\mathbf{x}_i\|_2^2) \\
&= \min_{\{\alpha_j\}_{j\in[m]},\alpha_j>0} \sum_{j\in[m]} \alpha_j^{2c} B_1(\theta_j) + \alpha_j^{-2} a_j^2 B_2(\theta_j).
\end{aligned}$$

Here, $B_1(\theta) \coloneqq \frac{1}{n}\sum_{i\in[n]} \sigma^2(\langle\theta,\mathbf{x}_i\rangle)$, and $B_2(\theta) \coloneqq \frac{1}{n}\sum_{i\in[n]} (\sigma')^2(\langle\theta,\mathbf{x}_i\rangle)\|\mathbf{x}_i\|_2^2$, for any $\theta\in\mathbb{S}^{d-1}$. Note that $B_1, B_2 \geqslant 0$, $\forall\theta\in\mathbb{S}^{d-1}$. As the objective is separable in α_j, we can minimize each function of α_j individually.

Consider the objective function $Q:\mathbb{R}_+\setminus\{0\}\to\mathbb{R}_+$. Then,

$$\begin{aligned}
Q(\alpha_j) &= B_1(\theta_j)\alpha_j^{2c} + a_j^2\alpha_j^{-2} B_2(\theta_j) \\
\nabla Q(\alpha_j) &= 2cB_1(\theta_j)\alpha_j^{2c-1} - 2a_j^2 B_2(\theta_j)\alpha_j^{-3} \\
\nabla^2 Q(\alpha_j) &= 2c(2c-1)B_1(\theta_j)\alpha_j^{2c-2} + 6a_j^2 B_2(\theta_j)\alpha_j^{-4}.
\end{aligned}$$

Note that $\nabla^2 Q(\alpha_j) > 0, \forall\alpha_j > 0$, so the function is minimized at α_j where $\nabla Q(\alpha_j) = 0$. This value corresponds to

$$\alpha_j = \left(\frac{a_j^2 B_2(\theta_j)}{cB_1(\theta_j)}\right)^{\frac{1}{2c+2}}.$$

Therefore, the flatness of the flattest minima obtained by purely rescaling $\mathbf{a}$ and θ_j is given by,

$$\sum_{j\in[m]} |a_j|^{\frac{2c}{c+1}} B(\theta_j), \text{ where } B(\theta_j) \coloneqq (c^{-\frac{c}{c+1}} + c^{\frac{1}{c+1}}) B_1^{\frac{1}{c+1}}(\theta_j) B_2^{\frac{c}{c+1}}(\theta_j). \tag{5}$$

Note that if any $a_j = 0$ for some $j\in[d]$, $\alpha_j = 0$. So, for any neuron that does not contribute to the output, we can completely remove its flatness by rescaling as long as we allow $\alpha_j \geqslant 0$ instead of $\alpha_j > 0$. If a vector $\mathbf{a}$ has some coordinates that are 0, we will assume $\alpha_j \geqslant 0$ for those coordinates. Note that the final expression in Lem. 3 doesn't change as $|a_j|^{\frac{2c}{c+1}}$ is still 0 for all the coordinates that are 0. □

D.2 Lower bounds on $\Upsilon(\mathbf{w})$: Proof of Thm. 1

Lemma 5 (Lower bound on $\Upsilon(\mathbf{w})$)**.** *If the conditions of Lem. 2 hold, for an interpolator $\mathbf{w} \in \mathbb{R}^{m(d+1)}$, with unit-norm inner-layer weights, $\{\theta_j\}_{j\in[m]}$, with probability $1-6\delta$,*

$$\Upsilon(\mathbf{w}) \gtrsim d^{\frac{c}{c+1}} \left(\frac{\|\mathbf{a}_{\min,\ell_2}\|_2^2}{\|\mathbf{a}_{\min,\ell_2}\|_{\frac{2c}{c-1}}} \right)^{\frac{2c}{c+1}},$$

$$\textit{where, } \mathbf{a}_{\min,\ell_2} = \mathbf{U}^\top \widetilde{\mathbf{U}}\mathbf{y} = \min_{\mathbf{a}\in\mathbb{R}^m, \mathbf{U}\mathbf{a}=\mathbf{y}} \|\mathbf{a}\|_2.$$

If $\frac{\|\mathbf{a}_{\min,\ell_2}\|_\infty}{\|\mathbf{a}_{\min,\ell_2}\|_2} = \mathcal{O}(\frac{1}{\sqrt{m}})$,

$$\Upsilon(\mathbf{w}) \gtrsim d^{\frac{c}{c+1}} m^{\frac{1}{c+1}} \|\mathbf{a}_{\min,\ell_2}\|_2^{\frac{2c}{c+1}}.$$

Lemma 6 (Lower bound on $\|\mathbf{a}_{\min,\ell_2}\|_2$)**.** *If $\zeta^2 = o(1)$, and the conditions of Lem. 2 hold, then with probability $1-4\delta$,*

$$\|\mathbf{y}\|_2 \gtrsim \sqrt{n}, \quad \|\mathbf{a}_{\min,\ell_2}\|_2 \gtrsim \sigma_{\max}^{-1}(\mathbf{U})\|\mathbf{y}\|_2 \gtrsim \frac{1}{\sqrt{m}}.$$

Further, if $\mathbf{a}_{\min,\ell_2}$ is aligned to the singular vector corresponding to the singular values $\sigma_i(\mathbf{U}) \asymp \sqrt{mn}$, then $\frac{\|\mathbf{a}_{\min,\ell_2}\|_\infty}{\|\mathbf{a}_{\min,\ell_2}\|_2} = \mathcal{O}(\frac{1}{\sqrt{m}})$.

To prove Thm. 1, consider an interpolator $\mathbf{a} \in \mathbb{R}^m$ with activation matrix $\mathbf{U} \in \mathbb{R}^{n\times m}$. Due to interpolation, we have,

$$\|\mathbf{U}\mathbf{a}\|_2 = \|\mathbf{y}\|_2.$$

From Lem. 6, with probability $1-4\delta$, we have,

$$\|\mathbf{y}\|_2 \gtrsim \sqrt{n}.$$

Further, using Holder's inequality,

$$\|\mathbf{U}\mathbf{a}\|_2 \leqslant \|\mathbf{a}\|_1 \max_{j\in[m]} \sqrt{\sum_{j=1}^n \sigma^2(\langle\theta_j, \mathbf{x}_i\rangle)}.$$

From Eq (23), we have, $\frac{1}{n}\sum_{i=1}^n \sigma^2(\langle\theta_j,\mathbf{x}_i\rangle) = B_1(\theta_j) = \Theta(1), \forall j \in [m]$. Therefore, for interpolation.

$$\|\mathbf{a}\|_1 \gtrsim 1.$$

Now, consider the flatness objective $\|\mathbf{a}\|_{\frac{2c}{c+1}}^{\frac{2c}{c+1}}$ in Lem. 3. Let

$$\hat{\mathbf{a}} \in \operatorname*{argmin}_{\|\mathbf{a}\|_1 \gtrsim 1} \|\mathbf{a}\|_{\frac{2c}{c+1}}^{\frac{2c}{c+1}}.$$

Then, using Holder's inequality, we have,

$$\|\mathbf{a}\|_{\frac{2c}{c+1}} \geqslant m^{-\frac{c-1}{2c}} \|\mathbf{a}\|_1 \gtrsim m^{-\frac{c-1}{2c}}.$$

Therefore,

$$\|\mathbf{a}\|_{\frac{2c}{c+1}}^{\frac{2c}{c+1}} \gtrsim m^{-\frac{c-1}{c+1}}.$$

This proves Theorem 1.

From Lems. 5 and 6, plugging in $\mathbf{a}_{\min,\ell_2}$ aligned to the singular vector along $\sigma_{\max}(\mathbf{U})$, we have,

$$\Upsilon(\mathbf{w}) \gtrsim d^{\frac{c}{c+1}} m^{-\frac{1}{c+1}} \|\mathbf{a}_{\min,\ell_2}\|_2^{\frac{2c}{c+1}} \gtrsim d^{\frac{c}{c+1}} m^{\frac{1}{c+1}} \left(\frac{1}{\sqrt{m}}\right)^{\frac{2c}{c+1}} = d^{\frac{c}{c+1}} m^{-\frac{c-1}{c+1}} = \Upsilon^\star.$$

Thus, the lower bound of $\Upsilon^\star$ is obtained when $\mathbf{a}_{\min,\ell_2}$ is aligned along the singular vectors of $\mathbf{U}$ that ensure $\sigma_i(\mathbf{U}) \asymp \sqrt{mn}$. This provides an important connection to the flattest interpolator.

In the remainder of this Section, we provide the proofs for Lems. 5 and 6.

Proof of Lem. 5. We find a lower bound on $\Upsilon(\mathbf{w})$ for a fixed set of inner-layer weights $\{\theta_j\}_{j\in[m]}$. Note that $\mathbf{w}$ is an interpolator, so,

$$\begin{aligned} & h(\mathbf{w},\mathbf{x}_i)=y_i, \forall i\in[n] \\ & \sum_{i=1}^{n} a_j\sigma(\langle\theta_j,\mathbf{x}_i\rangle)=y_i \\ & (\mathbf{U}\mathbf{a})_i=y_i \\ \implies & \mathbf{U}\mathbf{a}=\mathbf{y}. \end{aligned}$$

From Lem. 1, a lower bound on flatness in terms of only inner-layer weights requires minimizing $\Upsilon(\mathbf{w})$ in terms of $\mathbf{a}$. Note that this is a constrained minimization, as interpolation forces the linear constraint, $\mathbf{U}\mathbf{a}=\mathbf{y}$. The following statement holds with probability $1-9\delta$,

$$\Upsilon(\mathbf{w}) \geqslant \min_{\mathbf{a}\in\mathbb{R}^m, \mathbf{U}\mathbf{a}=\mathbf{y}} \gtrsim d^{\frac{c}{c+1}} \min_{\mathbf{a}\in\mathbb{R}^m, \mathbf{U}\mathbf{a}=\mathbf{y}} \|\mathbf{a}\|^{\frac{2c}{c+1}}_{\frac{2c}{c+1}}.$$

Note that the above optimization in $\mathbf{a}$ minimizes a convex ℓ_p norm, with a linear equality constraint. Therefore, strong duality holds (Boyd & Vandenberghe, 2004, Chapter 5), and we can use duality to find a lower bound on this objective. Using dual variables $\mathbf{r}\in\mathbb{R}^n$ for the interpolation constraint $\mathbf{U}\mathbf{a}=\mathbf{y}$, the primal problem has the following form.

$$\min_{\mathbf{a}\in\mathbb{R}^m}\max_{\mathbf{r}\in\mathbb{R}^n} \|\mathbf{a}\|^{\frac{2c}{c+1}}_{\frac{2c}{c+1}} + \langle \mathbf{r}, \mathbf{U}\mathbf{a}-\mathbf{y}\rangle. \tag{6}$$

As strong duality holds, the above optimal primal objective is equal to its optimal dual objective. Its dual problem is given by the following.

$$\max_{\mathbf{r}\in\mathbb{R}^n}\min_{\mathbf{a}\in\mathbb{R}^m} \|\mathbf{a}\|^{\frac{2c}{c+1}}_{\frac{2c}{c+1}} + \langle \mathbf{r}, \mathbf{U}\mathbf{a}-\mathbf{y}\rangle \tag{7}$$

We first solve the inner unconstrained minimization in terms of $\mathbf{a}$. Note that the Lagrangian is a convex function, therefore, we can minimize it by setting its gradient to 0. We consider 2 cases, when $c=1$ and $c>1$, to handle the ℓ_1 norm and arbitrary ℓ_p norms for $p>1$ separately. We first find the lower bound for $c>1$.

Case I : $c>1$ **.** By setting the first derivative of the Lagrangian in Eq (7) to 0 for $c>1$, we obtain, $\forall j\in[m]$,

$$\begin{aligned} & \operatorname{sign}(a_j)\frac{2c}{c+1}|a_j|^{\frac{c-1}{c+1}} + \sum_{i\in[n]} r_i\mathbf{U}_{i,j}=0, \\ & |a_j| = \left(-\operatorname{sign}(a_j)\frac{(c+1)\sum_{i\in[n]} r_i\mathbf{U}_{i,j}}{2c}\right)^{\frac{c+1}{c-1}}. \end{aligned}$$

Plugging this value into Eq (7), the Lagrangian becomes,

$$\|\mathbf{a}\|^{\frac{2c}{c+1}}_{\frac{2c}{c+1}} + \langle \mathbf{r}, \mathbf{U}\mathbf{a}-\mathbf{y}\rangle = -\widetilde{c}\sum_{j\in[m]}\left|\sum_{i\in[n]} r_i\mathbf{U}_{i,j}\right|^{\frac{2c}{c-1}} - \langle \mathbf{r},\mathbf{y}\rangle = -\widetilde{c}\|\mathbf{U}^\top\mathbf{r}\|^{\frac{2c}{c-1}}_{\frac{2c}{c-1}} - \langle \mathbf{r},\mathbf{y}\rangle. \tag{8}$$

Here $\widetilde{c}=(c-1)(c+1)^{\frac{c+1}{c-1}}(2c)^{-\frac{2c}{c-1}}$. The above function is a concave objective in $\mathbf{r}$ as it is negative of an ℓ_p norm of $\mathbf{r}$ with $p=\frac{2c}{c-1}$. As $p=\frac{2c}{c-1}\geqslant 2$, the p-norm is convex, and the negative of a convex function is concave.

To maximize Eq (8) with respect to $\mathbf{r}$, we can set the first derivative with respect to $\mathbf{r}$ to 0. However, computing a closed-form solution to $\mathbf{r}$ is not feasible for any arbitrary $c>1$. Instead, we find a lower bound to Eq (8). As the original dual problem, Eq (7) maximizes the objective in Eq (8) in terms of $\mathbf{r}$, if we plug in any value of $\mathbf{r}\in\mathbb{R}^n$, it should be a lower bound Eq (7). We plug in $\mathbf{r}=-\beta(\mathbf{U}\mathbf{U})^{-1}\mathbf{y}$

in Eq (8) for some $\beta > 0$.

$$\max_{\mathbf{r}\in\mathbb{R}^n} -\widetilde{c}\left\|\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y}\right\|_{\frac{2c}{c-1}}^{\frac{2c}{c-1}} - \langle \mathbf{r},\mathbf{y}\rangle$$
$$\geqslant \max_{\beta\in\mathbb{R}} -\widetilde{c}\beta^{\frac{2c}{c-1}}\left\|\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y}\right\|_{\frac{2c}{c-1}}^{\frac{2c}{c-1}} + \beta\left\|\widetilde{\mathbf{U}}^{\frac{1}{2}}\mathbf{y}\right\|_2^2$$

To obtain a tight lower bound, we can now maximize this lower bound in terms of β by setting the first derivative with respect to β to 0, as this is a concave function β. This gives us the optimal value of $\beta^\star$ as the following.

$$\beta^\star = \left(\frac{(c-1)\left\|\widetilde{\mathbf{U}}^{\frac{1}{2}}\mathbf{y}\right\|_2^2}{2c\widetilde{c}\left\|\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y}\right\|_{\frac{2c}{c-1}}^{\frac{2c}{c-1}}}\right)^{\frac{c-1}{c+1}}.$$

This is a concave function as its hessian with respect to β is $-\widetilde{c}\frac{2c(c+1)}{(c-1)}\beta^{\frac{2}{c-1}}\left\|\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y}\right\|_2$, hence it is maximized at this value of $\beta^\star$. Plugging in this value of $\beta^\star$ and ignoring the constant terms of c, we obtain the following lower bound on flatness for inner layer weights $\{\theta_j\}_{j\in[m]}$.

$$\min_{\mathbf{a}\in\mathbb{R}^m,\mathbf{U}\mathbf{a}=\mathbf{y}} \Upsilon(\mathbf{w}) \gtrsim d^{\frac{c}{c+1}}\left(\frac{\left\|\widetilde{\mathbf{U}}^{\frac{1}{2}}\mathbf{y}\right\|_2^2}{\left\|\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y}\right\|_{\frac{2c}{c-1}}}\right)^{\frac{2c}{c+1}} \tag{9}$$

Case II : $c = 1$**.** We need to minimize ℓ_1 norm of weights subject to the constraint $\mathbf{U}\mathbf{a} = \mathbf{y}$. The dual problem takes the following form,

$$\max_{\mathbf{r}\in\mathbb{R}^n}\min_{\mathbf{a}\in\mathbb{R}^d} \sum_{j\in[m]} |a_j|(1-\text{sign}(a_j)(\mathbf{U}^\top\mathbf{r})_j) - \langle \mathbf{r},\mathbf{y}\rangle.$$

If $\exists j \in [m]$ such that $\left|(\mathbf{U}^\top\mathbf{r})_j\right| > 1$, then each $|a_j|$ is multiplied by a negative term, thus it can be minimized by setting $|a_j| \to \infty$, and the objective would be $-\infty$ making the problem infeasible. If $\left\|\mathbf{U}^\top\mathbf{r}\right\|_\infty \leqslant 1, \forall j \in [m]$, then, the coefficient of each $|a|_j$ is non-negative. This is minimized for $\mathbf{a} = \mathbf{0}$. Therefore, solving the inner minimization in terms of $\mathbf{a}$ gives us the following bound.

$$\max_{\mathbf{r}\in\mathbb{R}^n} -\langle \mathbf{r},\mathbf{y}\rangle, \quad \text{s.t.} \quad \left\|\mathbf{U}^\top\mathbf{r}\right\|_\infty \leqslant 1$$

We can again convert this to a 1-dimensional optimization problem by setting $\mathbf{r} = -\beta\widetilde{\mathbf{U}}\mathbf{y}$ for some $\beta > 0$. From the constraint, we obtain $\beta \leqslant \left\|\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y}\right\|_\infty^{-1}$. Therefore, a lower bound on this function is $\frac{\left\|\widetilde{\mathbf{U}}^{\frac{1}{2}}\mathbf{y}\right\|_2^2}{\left\|\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y}\right\|_\infty}$, which can also be obtained by taking the limit of $c \to 1$ for Eq (9).

Final Bound. We now try to find the lower bound on RHS of Eq (9). To do this, we consider the minimum ℓ_2 norm interpolator of $\mathbf{a}$ defined as,

$$\mathbf{a}_{\min,\ell_2} = \operatorname*{argmin}_{\mathbf{a}\in\mathbb{R}^m,\mathbf{U}\mathbf{a}=\mathbf{y}} \|\mathbf{a}\|_2.$$

If we fix the matrix of activations $\mathbf{U}$, the constraint $\mathbf{U}\mathbf{a} = \mathbf{y}$ corresponds to an interpolator of overparametrized linear regression for $\mathbf{a}$ as $m \geqslant n$. The minimum ℓ_2-norm interpolator Bartlett et al. (2020) for overparmetrized linear regression is obtained by solving the above problem. Further, from Lem. 2, the $\sigma_{\min}(\mathbf{U}) > 0$, so $\mathbf{U}$ is full-rank, and therefore, $\mathbf{U}\mathbf{U}^\top \in \mathbb{R}^{n\times n}$ is invertible. In this case, the minimum ℓ_2-norm interpolator for $\mathbf{a}$, is given by Bartlett et al. (2020),

$$\mathbf{a}_{\min,\ell_2} = \mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y}.$$

Additionally, the ℓ_2 norm of this interpolator is given by,

$$\begin{aligned}\|\mathbf{a}_{\min,\ell_2}\|_2^2 &= \left\|\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y}\right\|_2^2 \\ &= \mathbf{y}^\top\widetilde{\mathbf{U}}\mathbf{U}\mathbf{U}^\top\widetilde{\mathbf{U}} \\ &= \mathbf{y}^\top\widetilde{\mathbf{U}}\mathbf{y} \\ &= \left\|\widetilde{\mathbf{U}}^{\frac{1}{2}}\mathbf{y}\right\|_2^2.\end{aligned}$$

Therefore, the RHS of Eq (9) takes the following form,

$$d^{\frac{c}{c+1}}\left(\frac{\left\|\widetilde{\mathbf{U}}^{\frac{1}{2}}\mathbf{y}\right\|_2^2}{\left\|\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y}\right\|_{\frac{2c}{c-1}}}\right)^{\frac{2c}{c+1}} = d^{\frac{c}{c+1}}\left(\frac{\|\mathbf{a}_{\min,\ell_2}\|_2^2}{\|\mathbf{a}_{\min,\ell_2}\|_{\frac{2c}{c-1}}}\right)^{\frac{2c}{c+1}}.$$

If $\frac{\|\mathbf{a}_{\min,\ell_2}\|_\infty}{\|\mathbf{a}_{\min,\ell_2}\|_2} = \mathcal{O}(\frac{1}{\sqrt{m}})$, then, by using Riesz-Thorin Interpolation Theorem (Stein & Shakarchi, 2011), we have,

$$\|\mathbf{a}_{\min,\ell_2}\|_{\frac{2c}{c-1}} \leqslant \|\mathbf{a}_{\min,\ell_2}\|_2^{\frac{c-1}{c}}\|\mathbf{a}_{\min,\ell_2}\|_\infty^{\frac{1}{c}} \leqslant m^{-\frac{1}{2c}}\|\mathbf{a}_{\min,\ell_2}\|_2.$$

Plugging this in, we obtain the required final expression.

$$\min_{\mathbf{a}\in\mathbb{R}^m,\mathbf{U}\mathbf{a}=\mathbf{y}}\Upsilon(\mathbf{w}) = d^{\frac{c}{c+1}}\left(\frac{\|\mathbf{a}_{\min,\ell_2}\|_2^2}{\|\mathbf{a}_{\min,\ell_2}\|_{\frac{2c}{c-1}}}\right)^{\frac{2c}{c+1}} \asymp d^{\frac{c}{c+1}}m^{\frac{1}{c+1}}\|\mathbf{a}_{\min,\ell_2}\|_2^{\frac{2c}{c+1}}.$$

This completes the proof. □

Proof of Lem. 6. Note that, by Cauchy-Schwarz,

$$\|\mathbf{a}_{\min,\ell_2}\|_2^2 = \mathbf{y}^\top\widetilde{\mathbf{U}}\mathbf{y} \geqslant \lambda_{\min}(\widetilde{\mathbf{U}})\|\mathbf{y}\|_2^2 = \sigma_{\max}^{-2}(\mathbf{U})\|\mathbf{y}\|^2$$

From Lem. 2, with probability, $1-\delta$, $\sigma_{\max}^{-2}(\mathbf{U}) \gtrsim \frac{1}{mn}$.

Therefore, to bound $\|\mathbf{a}_{\min,\ell_2}\|$, we need to bound $\|\mathbf{y}\|_2^2$. Decomposing it into the signal and noise terms $\mathbf{y}^\star$ and $\mathbf{n}$ respectively, we obtain,

$$\|\mathbf{y}\|_2^2 = \|\mathbf{y}^\star+\mathbf{n}\|_2^2 = \|\mathbf{y}^\star\|_2^2+\|\mathbf{n}\|_2^2+2\langle\mathbf{n},\mathbf{y}^\star\rangle. \tag{10}$$

We will bound each of the 3 terms in the above equation. For the first term $\|\mathbf{y}^\star\|_2^2$. Note that it is the ℓ_2 norm of an n-dimensional each. Each coordinate of $\mathbf{y}^\star$ is an independent random variable $\sigma^\star(\Theta^\star\mathbf{x}_i)$. As $\mathbf{x}_i$ is Gaussian and $\Theta^\star$ is an orthonormal matrix, $\Theta^\star\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\mathbb{I}_{m^\star})$. Further, from Assump. 2, $\sigma^\star(\Theta^\star\mathbf{x}_i)$ is a $(\frac{c}{2},\mathcal{O}(1))$- Sub-Weibull random variable, as $\|\mathbb{I}_{m^\star}\|_F = \sqrt{m^\star} = \mathcal{O}(1)$, and $\mathbb{E}\big[(\sigma^\star)^2(\Theta^\star\mathbf{x}_i)\big] = \psi^\star = \Theta(1)$. Further, $\|\mathbf{y}^\star\|_2^2$ is a quadratic form of Sub-Weibull random variables, so from Lem. 16, with probability $1-\delta$,

$$\|\mathbf{y}^\star\|_2^2 \asymp n.$$

As for the second term which is an inner product, using Sub-Gaussian concentration (Wainwright, 2019, Chapter 2) for $\mathbf{n}$ with probability $1-\delta$,

$$2\langle\mathbf{n},\mathbf{y}^\star\rangle \geqslant -2\zeta\|\mathbf{y}^\star\|_2\sqrt{2\log(1/\delta)} \gtrsim -\zeta\sqrt{n}.$$

For the third term $\|\mathbf{n}\|_2^2$, which is a quadratic form of the gaussian random variables $\{\xi_i\}_{i\in[n]}$, by Hanson-Wright (Rudelson & Vershynin, 2013, Thm. 1), with probability $1-\delta$, we have,

$$\|\mathbf{n}\|_2^2 \geqslant \zeta^2 n - \zeta^2\nu_1\max\{\sqrt{n\log(2/\delta)},\log(2/\delta)\} \gtrsim n\zeta^2.$$

Plugging these bounds into Eq (10), along with the bounds on $\sigma_{\max}(\mathbf{U})$, with probability, $1-4\delta$, we have,

$$\|\mathbf{a}_{\min,\ell_2}\|_2^2 = \mathbf{y}^\top \widetilde{\mathbf{U}}\mathbf{y} \gtrsim \frac{1+\zeta^2}{m} - \frac{\zeta}{m\sqrt{n}} \gtrsim \frac{1}{m}.$$

Note that $\zeta = o(1)$.

To prove the condition between, ℓ_2 and ℓ_∞ norms of $\mathbf{a}_{\min,\ell_2}$. Note that $\sigma_{\max}(\mathbf{U})$ corresponds to the all-ones singular vector. Therefore, to achieve the lower bound on $\|\mathbf{a}_{\min,\ell_2}\|_2 \asymp \frac{1}{\sqrt{m}}$, we need $\mathbf{a}_{\min,\ell_2}$ to be almost parallel to $\mathbf{1}_m$. Since, $\|\mathbf{1}_m\|_\infty = \frac{\|\mathbf{1}_m\|_2}{\sqrt{m}}$, any vector almost parallel to $\mathbf{1}_m$ also satisfies the required condition. Further, if $\sigma_i(\mathbf{U}) \asymp \sqrt{mn}$, from Lem. 2, there are only $k = \mathcal{O}(1)$ such singular values. If $\mathbf{u}_i \in \mathbb{R}^m$ are the singular vectors corresponding to these singular values, then,

$$\|\mathbf{u}_i\|_\infty \leqslant \max_{j\in[m]} \frac{\sqrt{nB(\theta_j)}}{\sigma_i(\mathbf{U})} \lesssim \frac{1}{\sqrt{m}}.$$

Therefore, all vectors along $\sigma_i(\mathbf{U}) \asymp \sqrt{mn}$ satisfy the incoherence condition in Lem. 5. This completes the proof. □

In the next section, we obtain a tighter bound than Lem. 6 for the case of bad interpolators.

D.3 Flatness of Bad Interpolators: Proof of Thm. 2

In this section, we provide lower bounds on flatness after rescaling, $\Upsilon(\mathbf{w})$, for the class of bad interpolators in Def. 2. Before we describe our lower bounds, we verify that bad interpolators defined in Def. 2 do exist.

Existence of Bad Interpolators. In Def. 2, we provide conditions on the inner-layer weights $\{\theta_j\}_{j\in[m]}$ and the population loss. However, we do not verify that interpolators verifying these conditions actually exist. In this section, we first show that satisfying the first condition in Def. 2 ensures that such an interpolator always exists. Further, we provide sufficient conditions on $\rho^\star$ in Def. 2, such that any interpolator satisfying the third condition satisfies the second condition, i.e., its population loss is large.

Note that an interpolator does not exist for any choice of inner-layer weights. An easy counter-example to this is setting all inner layer weights $\theta_j = \theta_\perp^\star$, where $\theta_\perp^\star \in \mathbb{S}^{d-1}, \theta_\perp^\star \perp \theta^\star$. For piece-wise polynomial activations, these weights will a.s. not interpolate in the presence of label noise ($\zeta > 0$). However, from condition 3. of Def. 2 and Lem. 2, we have, $\sigma_n(\mathbf{U}\mathbf{U}^\top) \neq 0$. Therefore, $\mathbf{U}$ has rank n and thus, it is $\mathbf{a}_{\min,\ell_2} = \mathbf{U}^\top(\mathbf{U}\mathbf{U}^\top)^{-1}\mathbf{y}$ is always a valid interpolator for given inner-layer weights.

Note that we have only shown that inner-layer weights corresponding to those in Def. 2 lead to an interpolator. To ensure that this interpolator is actually bad, i.e., its population loss is large, we first prove a sufficient condition.

Proposition 4 (Sufficient Conditions for Large Population Loss)**.** *For an interpolator* $\mathbf{w}$*, if the first condition in Def. 2 is satisfied, and if* $\Phi > 0$*,* ψ *is coordinate-wise non-decreasing function,* ϕ *is non-decreasing functions, and*

$$2m\psi^2(\rho_1^\star \mathbf{1}_{m^\star}) \leqslant (1-2\kappa)\lambda_{\min}(\phi)(\zeta^2 + \psi^\star)$$

, then second condition of Def. 2 is also satisfied, i.e, $F(\mathbf{w}) - F^\star \geqslant \kappa(\zeta^2 + \psi^\star)$*, where* $\kappa \leqslant \frac{1}{2}$*.*

We assume that $F^\star = o(1)$, so that the above condition can hold. The proof of this Prop. is provided in App. E.4.1. The above sufficient condition provides an upper bound on $\rho^\star$ in terms of the functions ψ and ϕ and the width. Crucially, a small $\rho^\star$ will ensure that the population loss will always be large. This corresponds to the inner-layer weights have very low alignment with the true direction.

Prop. 4 allows us to prove that for small ρ, these interpolators are also bad. Combining our argument about $\mathbf{U}$ being full-rank with Prop. 4, we have shown that bad interpolators defined in Def. 2 do exist for specific values of $\rho^\star, \kappa, \kappa_1, \kappa_2$. We corroborate the existence result with an actual example of a bad interpolator according to Def. 2.

Example 1 (Bad Interpolator for Def. 2)**.** *Consider the subspace perpendicular to* $\operatorname{span}(\{\theta_j^\star\}_{j\in[m^\star]})$ *of dimension* $d-m^\star$*. If we sample* m *unit norm vectors* $\{\theta_j\}_{j\in[m]}$ *iid from this subspace. Then, we can find* $\mathbf{a}\in\mathbb{R}^m$ *such that* $\mathbf{w}$ *satisfies Def. 2 with* $\kappa \coloneqq \frac{1}{2}-\frac{1}{2}\left(\mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{N}(0,\mathbb{I}_{m^\star})}[\sigma^\star(\mathbf{b})]\right)^2(\psi^\star)^{-1}$.

Proof for this example is provided in App. E.4.2.

Proof of Thm. 2. The core contribution of this proof is the following lower bound on $\|\mathbf{a}_{\min,\ell_2}\|_2$ that is tighter than Lem. 6.

Lemma 7 (Lower bound on $\|\mathbf{a}_{\min,\ell_2}\|_2$ for Bad Interpolators(Def. 2))**.** *If* $\zeta^2=o(1)$*,* $\rho_1=o((m^\star)^{-\frac{1}{4}}$*, and the conditions of Lem. 2 holds, then with probability* $1-7\delta$*,*

$$\|\mathbf{a}_{\min,\ell_2}\|_2^2 \gtrsim \sum_{i=1}^{n}\sigma_i^{-2}(\mathbf{U}) \gtrsim \frac{n}{m}.$$

Further, $\frac{\|\mathbf{a}_{\min,\ell_2}\|_\infty}{\|\mathbf{a}_{\min,\ell_2}\|_2}=\mathcal{O}(\frac{1}{\sqrt{m}})$.

Combining Lem. 7 with Lem. 5, with probability $1-12\delta$, for a bad interpolator $\mathbf{w}$ defined in Def. 2, we have,

$$\Upsilon(\mathbf{w}) \gtrsim d^{\frac{c}{c+1}}m^{\frac{1}{c+1}}\left(\sqrt{\frac{n}{m}}\right)^{\frac{2c}{c+1}} = (dn)^{\frac{c}{c+1}}m^{-\frac{c-1}{c+1}}.$$

This completes the proof. □

In the remainder of this section, we provide the proof for Lem. 7.

Proof of Lem. 7. Note that,

$$\|\mathbf{a}_{\min,\ell_2}\|_2^2 = \mathbf{y}^\top\widetilde{\mathbf{U}}\mathbf{U}\mathbf{U}^\top\widetilde{\mathbf{U}}\mathbf{y} = \mathbf{y}^\top\widetilde{\mathbf{U}}\mathbf{y}.$$

To bound this sum, we first remove the contribution of the noise terms $\mathbf{n}$. We use the following Lem. for this task.

Lemma 8 (Removing contribution of label noise)**.** *With probability* $1-2\delta$*,*

$$\|\mathbf{a}_{\min,\ell_2}\|_2^2 \geqslant \frac{1}{2}(\mathbf{y}^\star)^\top\widetilde{\mathbf{U}}\mathbf{y}^\star - \zeta^2\frac{n}{m}$$

The proof of this Lem. is provided in App. E.4.3.

The remaining term in Lem. 8 depends on the true signal $\mathbf{y}^\star$ and the activation matrix $\widetilde{\mathbf{U}}$. A lower bound for this term can be obtained from Lem. 6. However, this lower bound is loose for the set bad interpolators in Def. 2.

To obtain a tight lower bound for Def. 2, we utilize the observation that the new labels $\mathbf{y}^\star$ and the new activation $\mathbf{U}$ are obtained from random variables that have very low correlation from 3rd of Def. 2. Therefore, there is a large component of $\mathbf{y}^\star$ that is independent of the activation matrix $\mathbf{U}$.

To quantify this bound, we first derive the conditional distribution of the labels $\mathbf{y}^\star$ conditioned on the activations $\mathbf{U}$. Let $\mathbf{U}_i\in\mathbb{R}^m$ be the ith row of $\mathbf{U}$, $\forall i\in[n]$. Then, due to independence of samples, only the ith coordinate of $\mathbf{y}^\star$, $\mathbf{y}_i^\star$, is not independent of $\mathbf{U}_i$. Therefore, to characterize the distribution of $\mathbf{y}^\star|\hat{\mathbf{U}}$, we only need to characterize the distribution of $\mathbf{y}_i^\star|\hat{\mathbf{U}}_i$. The following Lem. provides this distribution. We will characterize this distribution conditioned on $\hat{\mathbf{U}}_i=\Theta\mathbf{x}_i$ where $\mathbf{U}_i$ is obtained by applying the activation σ element-wise to this term.

Lemma 9 (Distribution of $\mathbf{y}_i^\star|\hat{\mathbf{U}}_i$). *If Assump. 2 holds, for all $i\in[n]$,*

$$\textbf{\textit{Mean}}: \left\|\mathbb{E}\left[\mathbf{y}_i^\star|\hat{\mathbf{U}}_i\right]\right\|_2 = \Theta\left(\|\widetilde{\mu}_i\|_2^c + \|\varsigma\|_F^{\frac{c}{2}}\right), \qquad \textbf{\textit{Variance}}: Var\left[\mathbf{y}_1^\star|\hat{\mathbf{U}}_1\right] = \Omega\left(\|\varsigma\|_F\|\widetilde{\mu}_i\|_2^{2c-2} + \|\varsigma\|_F^c\right),$$

$$\textbf{\textit{Tails}}: \mathbf{y}_i^\star|\hat{\mathbf{U}}_i \text{ is a } (\frac{c}{2}, \mathcal{O}(\|\widetilde{\mu}_i\|_2^c + \|\varsigma\|_F^{\frac{c}{2}}))\text{-Sub-Weibull random vector.}$$

where $\tilde{\mu}_i \coloneqq \Theta^\star\Theta^\top(\Theta\Theta^\top)^\dagger\Theta\mathbf{x}_i \sim \mathcal{N}(0, \mathbb{I}_{m^\star} - \varsigma), \varsigma = \mathbb{I}_{m^\star} - \Theta^\star\Theta^\top(\Theta\Theta^\top)^\dagger\Theta(\Theta^\star)^\top$ and $A^\dagger$ is the pseudo-inverse of A for a square matrix A with real entries.

The proof of this Lem. is provided in App. E.4.4. We use Assump. 2 to obtain all these bounds as $\mathbf{y}^\star$ is obtained from the link function $\sigma^\star$.

Let $\mathbf{y}^\star_{\hat{\mathbf{U}}} \coloneqq \mathbb{E}\left[\mathbf{y}^\star|\hat{\mathbf{U}}\right]$, where $\hat{\mathbf{U}} \in \mathbb{R}^{n\times m}$ is the matrix of $\hat{\mathbf{U}}_i$'s. Then,

$$(\mathbf{y}^\star)^\top\widetilde{\mathbf{U}}\mathbf{y}^\star = (\mathbf{y}^\star - \mathbf{y}^\star_{\hat{\mathbf{U}}})^\top\widetilde{\mathbf{U}}(\mathbf{y}^\star - \mathbf{y}^\star_{\hat{\mathbf{U}}}) + (\mathbf{y}^\star_{\hat{\mathbf{U}}})^\top\widetilde{\mathbf{U}}\mathbf{y}^\star_{\hat{\mathbf{U}}} + 2(\mathbf{y}^\star - \mathbf{y}^\star_{\hat{\mathbf{U}}})^\top\widetilde{\mathbf{U}}\mathbf{y}^\star_{\hat{\mathbf{U}}}.$$

For the last term, note that its mean is 0, conditioned on $\hat{\mathbf{U}}$. Therefore, using the (Zhang & Wei, 2022, Corollary 2), with probability $1-\delta$, we have,

$$(\mathbf{y}^\star - \mathbf{y}^\star_{\hat{\mathbf{U}}})^\top\widetilde{\mathbf{U}}\mathbf{y}^\star_{\hat{\mathbf{U}}} \gtrsim -\bar{\nu}'\max\left\{\left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{y}_{\hat{\mathbf{U}}}\right\|_2\sqrt{\log(1/\delta)}, \left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{y}_{\hat{\mathbf{U}}}\right\|_\infty\log(1/\delta)\right\}$$

Then, conditioned on $\hat{\mathbf{U}}$, using Hanson-Wright for Sub-Weibull random variables from Lem. 16 for the first term, with probability $1-\delta$, we have,

$$\begin{aligned}(\mathbf{y}^\star)^\top\widetilde{\mathbf{U}}\mathbf{y}^\star \geqslant &\sum_{i\in[n]}\widetilde{\mathbf{U}}_{i,i}\mathrm{Var}\left[\mathbf{y}_i^\star|\hat{\mathbf{U}}_i\right] + (\mathbf{y}^\star_{\hat{\mathbf{U}}})^\top\widetilde{\mathbf{U}}\mathbf{y}^\star_{\hat{\mathbf{U}}} \\ &-\bar{\nu}'\max\left\{\left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{y}_{\hat{\mathbf{U}}}\right\|_2\sqrt{\log(1/\delta)}, \left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{y}_{\hat{\mathbf{U}}}\right\|_\infty\log(1/\delta)\right\} \\ &-\bar{\nu}_{\frac{c}{2}}\max\{\left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{K}\right\|_F\sqrt{\log(2/\delta)}, \left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{K}\right\|_2(\log(2/\delta))^c\}\end{aligned} \tag{11}$$

where the matrix $\mathbf{K}\in\mathbb{R}^{n\times n}$ satisfies, $\mathbf{K} \lesssim \mathbf{K}' + \|\varsigma\|_F^{\frac{c}{2}}\mathbb{I}_n$ with the matrix $\mathbf{K}'\in\mathbb{R}^{n\times n}$ being the diagonal matrix, $\mathbf{K}' \coloneqq \mathrm{diag}(\|\widetilde{\mu}_1\|_2^c, \|\widetilde{\mu}_2\|_2^c, \cdots, \|\widetilde{\mu}_n\|_2^c)$. Further, we can apply the bounds on $\mathrm{Var}[\mathbf{y}_i^\star|\mathbf{U}_i]$ and $\|\mathbf{y}^\star_{\mathbf{U}}\|$ from Lem. 9 to simplify each of the terms above. We use the following Lem. to obtain appropriate bounds for each of the terms in the above equation. Note that to remove the conditioning on $\mathbf{U}$, we compute an upper bound on the expected value of probability of error with the expectation over $\mathbf{U}$. Since it is a constant δ, its expectation over $\mathbf{U}$ is also δ, and thus the above bound holds with probability $1-\delta$ even without the conditioning.

Lemma 10 (Simplification). *Note that,*

$$\begin{aligned}&\sum_{i\in[n]}\widetilde{\mathbf{U}}_{i,i}Var\left[\mathbf{y}_i^\star|\hat{\mathbf{U}}_i\right] \gtrsim \|\varsigma\|_F\mathrm{Tr}((\mathbf{K}')^{\frac{2c-2}{c}})\sigma_{\max}^{-2}(\mathbf{U}) + \|\varsigma\|_F^c\mathrm{Tr}(\widetilde{\mathbf{U}}), \\ &\mathbf{y}^\star_{\hat{\mathbf{U}}}\widetilde{\mathbf{U}}\mathbf{y}^\star_{\hat{\mathbf{U}}} \gtrsim (\mathrm{Tr}((\mathbf{K}')^2) + n\|\varsigma\|_F)\sigma_{\max}^{-2}(\mathbf{U}), \\ &\left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{K}\right\|_F \leqslant (\sqrt{\mathrm{Tr}((\mathbf{K}')^4)} + \|\varsigma\|_F^c)\sqrt{\sum_{i=1}^n\sigma_i^{-4}(\mathbf{U})} \\ &\left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{y}_{\hat{\mathbf{U}}}\right\|_2 \lesssim \sigma_n^{-2}(\mathbf{U})\left(\mathrm{Tr}((\mathbf{K}')^2) + \sqrt{\mathrm{Tr}((\mathbf{K}')^2)}\|\varsigma\|_F^{\frac{c}{2}}\right).\end{aligned}$$

The proof of this Lem. is provided in App. E.4.5. In addition to the above simplification, we use the fact that $\|A\|_2 \leqslant \|A\|_F$ for any PSD matrix A. Applying the above simplification to Eq. (11), with probability $1-\delta$ conditioned on $\mathbf{U}$, we obtain,

$$\begin{aligned}\mathbf{y}^\star\widetilde{\mathbf{U}}\mathbf{y}^\star \gtrsim &\|\varsigma\|_F^c(\mathrm{Tr}(\widetilde{\mathbf{U}}) + n\sigma_{\max}^{-2}(\mathbf{U})) + \sigma_{\max}^{-2}(\mathbf{U})(\mathrm{Tr}((\mathbf{K}')^{\frac{2c-2}{c}})\|\varsigma\|_F + \mathrm{Tr}((\mathbf{K}')^4)) \\ &-\bar{\nu}'\sigma_n^{-2}(\mathbf{U})\left(\mathrm{Tr}((\mathbf{K}')^2) + \sqrt{\mathrm{Tr}((\mathbf{K}')^2)}\|\varsigma\|_F^{\frac{c}{2}}\right)\sqrt{\log(1/\delta)} \\ &-\bar{\nu}_{\frac{c}{2}}\sqrt{\sum_{i=1}^n\sigma_i^{-4}(\mathbf{U})}(\sqrt{\mathrm{Tr}((\mathbf{K}')^4)} + \|\varsigma\|_F^c)(\log(2/\delta))^c.\end{aligned}$$

We need to bound the terms of $\mathbf{K}$ and the terms of activation matrices, $\mathbf{U}$ and $\widetilde{\mathbf{U}}$. From Lem. 2, with probability $1-\delta$, we have, $\mathrm{Tr}(\widetilde{\mathbf{U}}) = \sum_{i=1}^{n} \sigma_i^{-2}(\mathbf{U}) \asymp \frac{n}{m}$, $\sqrt{\sum_{i=1}^{n} \sigma_i^{-4}(\mathbf{U})} \asymp \frac{\sqrt{n}}{m}$, and $\sigma_{\max}^{-2}(\mathbf{U}) \asymp \frac{1}{mn}$. Plugging these bounds into the above equation, with probability $1-2\delta$, conditioned on $\mathbf{U}$, we obtain,

$$\begin{aligned}\mathbf{y}^\star \widetilde{\mathbf{U}} \mathbf{y}^\star \gtrsim & \|\varsigma\|_F^c \frac{n}{m} + \frac{1}{mn}(\mathrm{Tr}((\mathbf{K}')^{\frac{2c-2}{c}})\|\varsigma\|_F + \mathrm{Tr}((\mathbf{K}')^2)) \\ & - \bar{\nu}' \frac{1}{m}\left(\mathrm{Tr}((\mathbf{K}')^2) + \sqrt{\mathrm{Tr}((\mathbf{K}')^2)}\|\varsigma\|_F^{\frac{c}{2}}\right)\sqrt{\log(1/\delta)} \\ & - \bar{\nu}_{\frac{c}{2}} \frac{\sqrt{n}}{m}(\sqrt{\mathrm{Tr}((\mathbf{K}')^4)} + \|\varsigma\|_F^c)(\log(2/\delta))^c.\end{aligned} \tag{12}$$

Note that the only random variable in the lower bound is obtained from the matrix $\mathbf{K}'$. The following lemma provides a bound on this.

Lemma 11 (Bounds on $\mathbf{K}'$). *For any $\varphi \geqslant 0$, with probability $1-\delta$,*

$$\mathrm{Tr}((\mathbf{K}')^\varphi) \asymp n\|\mathbb{I}_{m^\star} - \varsigma\|_F^{\frac{c\varphi}{2}}$$

The proof of this Lem. is provided in App. E.4.6. We can apply the above Lem. for $\varphi = \frac{2(c-1)}{c}$,2 and 4 to in Eq. (12), with probability $1-5\delta$, we have,

$$\begin{aligned}\mathbf{y}^\star \widetilde{\mathbf{U}} \mathbf{y}^\star \gtrsim & \|\varsigma\|_F^c \frac{n}{m} + \frac{\|\mathbb{I}_{m^\star} - \varsigma\|_F^{c-1}}{mn}(\|\varsigma\|_F + \|\mathbb{I}_{m^\star} - \varsigma\|_F) \\ & - \bar{\nu}' \frac{\sqrt{n}}{m}\left(\sqrt{n}\|\mathbb{I}_{m^\star} - \varsigma\|_F^c + \|\mathbb{I}_{m^\star} - \varsigma\|_F^{\frac{c}{2}}\|\varsigma\|_F^{\frac{c}{2}}\right)\sqrt{\log(1/\delta)} \\ & - \bar{\nu}_{\frac{c}{2}} \frac{\sqrt{n}}{m}(\sqrt{n}\|\mathbb{I}_{m^\star} - \varsigma\|_F^c + \|\varsigma\|_F^c)(\log(2/\delta))^c.\end{aligned} \tag{13}$$

We can now see sufficient conditions for $\mathbf{y}^\star \widetilde{\mathbf{U}} \mathbf{y}^\star \gtrsim \frac{n}{m}$. This is $\|\mathbb{I}_{m^\star} - \varsigma\|_F = o(1)$.

We first find an upper bound on $\|\mathbb{I}_{m^\star} - \varsigma\|_F$.

Upper bound on $\|\mathbb{I}_{m^\star} - \varsigma\|_F$ Note that,

$$\|\mathbb{I}_{m^\star} - \varsigma\|_F = \left\|\Theta^\star \Theta^\top (\Theta\Theta^\top)^\dagger \Theta (\Theta^\star)^\top\right\|_F.$$

Let $\Theta^\top(\Theta\Theta^\top)^\dagger\Theta \coloneqq P_\Theta$. Note that $P_\Theta \in \mathbb{R}^{d\times d}$ projects any d-dimensional vector to the subspace of $\mathrm{span}(\{\theta_j\}_{j\in[m]})$. Therefore, the second term can be bounded in the following way,

$$\begin{aligned}\left\|\Theta^\star \Theta^\top (\Theta\Theta^\top)^\dagger \Theta (\Theta^\star)^\top\right\|_F = \left\|\Theta^\star P_\Theta (\Theta^\star)^\top\right\|_F &= \sqrt{\sum_{j,j'\in[m^\star]} \left\langle \theta_j^\star, P_\Theta \theta_{j'}^\star \right\rangle^2} \\ &= \sqrt{\sum_{j\in[m^\star]} \left\langle \theta_j^\star, P_\Theta \theta_j^\star \right\rangle^2}.\end{aligned}$$

We use the fact that $\left\langle \theta_j^\star, \theta_{j'}^\star \right\rangle = 0$ if $j \neq j'$. Note that $P_\Theta \theta_j^\star$ is a vector in $\mathrm{span}(\{\theta_j\}_{j\in[m]})$. Therefore, by Def. 2,

$$\left\langle \theta_j^\star, P_\Theta \theta_j^\star \right\rangle^2 \leqslant (\rho^\star)^2 \left\|P_\Theta \theta_j^\star\right\|^2 = \rho^4$$

We use the fact that $\left\|P_\Theta \theta_j^\star\right\| = \left|\left\langle \mathbf{v}, \theta_j^\star \right\rangle\right| \leqslant \rho^\star$ for some $\mathbf{v} \in \mathbb{S}^{d-1}$ and $\mathbf{v} \in \mathrm{span}(\{\theta_j\}_{j\in[m]})$

Using these bounds, we find that,

$$\left\|\Theta^\star \Theta^\top (\Theta\Theta^\top)^\dagger \Theta (\Theta^\star)^\top\right\|_F \leqslant (\rho^\star)^2 \sqrt{m^\star}.$$

From the condition in Thm. 2, $\rho^\star = o((m^\star)^{-\frac{1}{4}})$, so $\|\mathbb{I}_{m^\star} - \varsigma\|_F = o(1)$. Therefore, plugging this bound into Eq. (13), with probability $1-5\delta$, we have,

$$(\mathbf{y}^\star)^\top \widetilde{\mathbf{U}} \mathbf{y}^\star \gtrsim \frac{n}{m}.$$

From Lem. 8, with probability $1-6\delta$, we have,

$$\|\mathbf{a}_{\min,\ell_2}\|_2^2 \gtrsim \frac{n}{m} - \zeta^2\frac{n}{m} \gtrsim \frac{n}{m}.$$

Using $\zeta^2 = o(1)$ in the last step provides our required bound.

We now prove the bound on $\frac{\|\mathbf{a}_{\min,\ell_2}\|_\infty}{\|\mathbf{a}_{\min,\ell_2}\|_2}$. If $\mathbf{e}_j \in \mathbb{R}^m$ is the jth coordinate vector, then, we can bound,

$$\|\mathbf{a}_{\min,\ell_2}\|_\infty = \max_{j\in[m]} \left|\left\langle \mathbf{e}_j, \mathbf{U}^\top \widetilde{\mathbf{U}}\mathbf{y}\right\rangle\right|$$

First, conditioned on $\mathbf{U}$, using Lem. 15, with probability $1-\delta$, by taking a union bound over all m coordinates, we obtain,

$$\max_{j\in[m]} \left|\left\langle \mathbf{e}_j, \mathbf{U}^\top \widetilde{\mathbf{U}}\mathbf{y}\right\rangle\right| \lesssim \max_{j\in[m]} \left(\left|\left\langle \mathbf{e}_j, \mathbf{U}^\top \widetilde{\mathbf{U}}\mathbf{y}^\star_{\mathbf{U}}\right\rangle\right| + \left(\left\|\widetilde{\mathbf{U}}\mathbf{U}\mathbf{e}_j\right\|_\infty + \zeta^2\left\|\widetilde{\mathbf{U}}\mathbf{U}\mathbf{e}_j\right\|_2\right)\log(\frac{m}{\delta})\right)$$

We use the fact that $\rho^\star = \mathcal{O}((m^\star)^{-\frac{1}{4}}$ and $\zeta^2 = o(1)$ to eliminate the first and the third term. Then, we bound the dominating second term with ℓ_∞ norm by ℓ_2 norm. We also absorb the $\log(m/\delta)$ term inside $\lesssim$ notation.

Finally, we require the following bound,

$$\left\|\widetilde{\mathbf{U}}\mathbf{U}\mathbf{e}_j\right\|_\infty \leqslant \left\|\widetilde{\mathbf{U}}\mathbf{U}\mathbf{e}_j\right\|_2 \leqslant \sigma_{\min}^{-1}(\mathbf{U}) \leqslant \frac{1}{\sqrt{m}}.$$

Using Lem. 2, we obtain, $\|\mathbf{a}_{\min,\ell_2}\|_\infty \lesssim \frac{1}{\sqrt{m}}$. From our previous analysis, $\|\mathbf{a}_{\min,\ell_2}\|_2 \gtrsim \sqrt{\frac{n}{m}}$, therefore,

$$\frac{\|\mathbf{a}_{\min,\ell_2}\|_\infty}{\|\mathbf{a}_{\min,\ell_2}\|_2} \lesssim \frac{1}{\sqrt{n}} = \tilde{\mathcal{O}}(\frac{1}{\sqrt{m}}).$$

We finally use $n = \Theta(m)$. This completes the proof.

□

In the next section, we find conditions under which an interpolator $\mathbf{w}$ achieves the minimum flatness $\Upsilon^\star$ and characterize its population loss.

D.4 Population loss of Flattest Interpolators : Proof of Thm. 3

Note that Assump. 3 quantifies sufficient conditions to achieve the minimum flatness. There is a simple instance of our problem setting that achieves this condition, which serves as a motivation for both Assump. 3 and Prop. 2. We state it here for completeness.

Theorem 4 (Flattest Good Minima for Learning Activation Without Noise)**.** *For $\zeta = 0$, the following interpolator $\mathbf{w} \in \mathbb{R}^{m(d+1)}$ has $F(\mathbf{w}) = 0$, and with high probability, $\Upsilon(\mathbf{w}) \asymp \Upsilon^\star$.*

- *Single-Index : $\sigma^\star = \sigma$, $a_j = \frac{1}{m}, \theta_j = \theta^\star, \forall j \in [m]$,*
- *Sum of Single-Index : $\widetilde{\sigma}^\star_{j'} = \sigma$, $a_j = \frac{a^\star_{j'} m^\star}{m}$ and $\theta_j = \theta^\star_{j'}, \forall j \in \left[\frac{(j'-1)m}{m^\star}+1, \frac{jm}{m^\star}\right], \forall j' \in [m^\star]$, where m is divisible by $m^\star$.*

The proof of this Thm. is provided in App. E.5.1. Now, we prove Thm. 3. Note that this Thm. contains an existence and generalization result. We restate these two results separately below.

Lemma 12 (Existence of a Flattest Interpolator)**.** *Suppose Assump. 1, 2 and 3 hold, and $n \asymp d$. Then, with high probability, $\exists \hat{\mathbf{w}}^\star = \left[(\hat{\mathbf{a}}^\star)^\top, (\hat{\theta}^\star_1)^\top, ..., (\hat{\theta}^\star_m)^\top\right]$ such that $\hat{\theta}^\star_j \in \mathbb{S}^{d-1}, \forall j \in [m]$, $F_S(\hat{\mathbf{w}}^\star) = 0$, $\|\hat{\mathbf{a}}^\star\|_\infty \lesssim m^{-1}$ and $\Upsilon(\hat{\mathbf{w}}^\star) \asymp \Upsilon^\star$.*

Theorem 5 (Population Loss of Flattest Interpolator)**.** *Suppose Assump. 1, 2 and 3 hold. Then, for all interpolators $\mathbf{w} \in \mathbb{R}^{m(d+1)}$ with unit norm inner-layer weights and $\|\mathbf{a}\|_1 = \mathcal{O}(1)$, with probability $1-3\delta$,*

$$F(\mathbf{w}) - F(\mathbf{w}^\star) \lesssim \sqrt{\frac{d+\log(1/\delta)}{n}} + \frac{(d+\log(1/\delta))^{\frac{c}{2}}}{n^{\min\{1,\frac{c}{2}\}}}.$$

Essentially, interpolation with approximation error and label noise increases $\|\mathbf{a}\|_\infty$ and thus flatness $\Upsilon(\mathbf{w})$ for any interpolator $\mathbf{w}$ by a term proportional to $nm(\Delta^2+\zeta^2)$. Hence, we require these terms to be small to achieve a small approximation error.

We now provide the proof of Lem. 12 and Thm. 5, and defer the proof of Prop. 2 to App. E.5.2.

Proof of Lem. 12. We divide the m layer weights into 2 parts of sizes $m-Cn$ and Cn neurons for some constant $C>1$. The first part is used to learn the signal and the rest is used to learn noise and approximation error.

We assume that $m-Cn$ is divisble by $\breve{m}$. For the first $m-n$ layer-weights, we set it to exactly the optimal approximation in Assump. 3.

$$\hat{\theta}_j = \breve{\theta}_{j'}, \quad \forall j \in \left[\frac{(m-Cn)(j'-1)}{\breve{m}}+1, \frac{(m-Cn)j'}{\breve{m}}\right], \forall j' \in [\breve{m}].$$

Further, we set the first $m-Cn$ outer-layer weights to also match those in Assump. 3 with an appropriate scaling by m. Define $\hat{\mathbf{a}} \in \mathbb{R}^m$ such that,

$$\hat{a}_j = \frac{\breve{m}\breve{a}_{j'}}{m-Cn}, \quad \forall j \in \left[\frac{(m-Cn)(j'-1)}{\breve{m}}+1, \frac{(m-Cn)j'}{\breve{m}}\right], \quad \forall j' \in [\breve{m}].$$

Note that this choice matches the optimal weights corresponding to the sum of single-index case in Thm. 4. Summing up the first $m-Cn$ layer weights, we obtain,

$$\sum_{j=1}^{m-Cn} \frac{\breve{m}\breve{a}_j}{m-Cn}\sigma(\left\langle \hat{\theta}_j, \mathbf{x}_i \right\rangle) = \frac{m-Cn}{\breve{m}}\frac{\breve{m}}{m-Cn}\sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\left\langle \breve{\theta}_j, \mathbf{x}_i \right\rangle).$$

Define the vector $\mathbf{v} \in \mathbb{R}^n$, as

$$(\mathbf{v})_i = \sigma^\star(\Theta^\star \mathbf{x}_i) - \sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\left\langle \breve{\theta}_j, \mathbf{x}_i \right\rangle).$$

Note that $\Theta^\star \mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(0, \mathbb{I}_{m^\star})$ and $\left\langle \breve{\theta}_j, \mathbf{x}_i \right\rangle \overset{iid}{\sim} \mathcal{N}(0,1)$, $\forall j \in [\breve{m}]$ and $\forall i \in [n]$. Using Maximal inequality, with probability $1-\delta$

$$\forall i \in [n], \qquad \max\left\{\|\Theta^\star \mathbf{x}_i\|_\infty, \max_{j\in[\breve{m}]} \left|\left\langle \breve{\theta}_j, \mathbf{x}_i \right\rangle\right|\right\} \lesssim C'\sqrt{\log(n(m^\star+\breve{m}))\log(1/\delta)}.$$

As $n \asymp d$, from Assumption 3, with high probability this bound is satisfied and each coordinate of $\mathbf{v}$ has absolute value at most Δ.

To ensure interpolation, we need $\mathbf{U}\mathbf{a} = \mathbf{y}$. If $\breve{\mathbf{U}} \in \mathbb{R}^{n \times Cn}$ corresponds to the activations of the last Cn weights, and $\tilde{\mathbf{a}} \in \mathbb{R}^{Cn}$ is their corresponding outer-layer weights, we can ensure interpolation as long as,

$$\breve{\mathbf{U}}\tilde{\mathbf{a}} = \mathbf{v} + \mathbf{n}$$

Note that $Cn \geqslant n$, so we have enough overparametrization to fit any n-dimensional target. We select the last Cn inner-layer weights such that we satisfy the condition 3 required for a bad interpolator (Def 2). These are $\lambda_{\min}(\bar{\Phi}) = \Omega(1)$ and $\sigma_{\max}(\Theta) = \mathcal{O}(1)$ and their pairwise inner-products are small. From Lemma 2, we can ensure that these inner-layer weights will always interpolate any target, and that the minimum singular value of their activations, $\sigma_n(\breve{\mathbf{U}})$ will be $\gtrsim \sqrt{n}$. We replace m by Cn

We will set $\tilde{\mathbf{a}}$ as the minimum ℓ_2 norm interpolator. Thus,

$$\tilde{\mathbf{a}} = \breve{\mathbf{U}}^\top (\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1}(\mathbf{v}+\mathbf{n}).$$

Now, we characterize the $\|\mathbf{a}\|_\infty$ of this interpolator. For the first $m-Cn$ outer-layer weights, from Assump. 3, their ℓ_∞ norm is bounded by $\tilde{\mathcal{O}}\left(\frac{\breve{m}}{m-Cn}\frac{1}{\breve{m}}\right) = \tilde{\mathcal{O}}(1)$.

For the last Cn layer weights, we have,

$$\|\tilde{\mathbf{a}}\|_\infty \leqslant \|\tilde{\mathbf{a}}\|_2 \leqslant \sqrt{(\mathbf{v}+\mathbf{n})^\top (\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1}(\mathbf{v}+\mathbf{n})}.$$

As $\mathbf{n}$ is independent of $\mathbf{v}$, conditioned on $\mathbf{v}$, $\mathbf{n}+\mathbf{v}\sim\mathcal{N}(\mathbf{v},\mathbb{I}_n)$. By applying Hanson-Wright (Rudelson & Vershynin, 2013, Thm. 1) and gaussian concentration for the gaussian random vector $\mathbf{n}$, with probability $1-\delta$,

$$\begin{aligned}(\mathbf{n}+\mathbf{v})^\top(\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1}(\mathbf{n}+\mathbf{v}) &\leqslant \zeta^2\mathrm{Tr}((\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1})+\mathbf{v}(\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1}\mathbf{v}+\zeta\nu'\left\|(\breve{\mathbf{U}}\breve{\mathbf{U}})^{-1}\mathbf{v}\right\|_2\sqrt{\log(2/\delta)}\\ &\quad+\zeta^2\nu\max\left\{\left\|(\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1}\right\|_F\sqrt{\log(2/\delta)},\left\|(\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1}\right\|_2\log(2/\delta)\right\}\\ &\lesssim \zeta^2\frac{n}{n}+\|\mathbf{v}\|_2^2\frac{1}{n}+\zeta\|\mathbf{v}\|_2\frac{1}{n}+\zeta^2\frac{\sqrt{n}}{n}\lesssim\zeta^2+\Delta^2.\end{aligned}$$

From Lem. 2 and the bounds computed in App. D.3, we know that with probability $1-\delta$, $\mathrm{Tr}((\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1})\asymp\frac{n}{n},\sigma_{\min}^{-2}(\mathbf{U})\asymp\frac{1}{n},\left\|(\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1}\right\|_2\leqslant\left\|(\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1}\right\|_F$ and $\left\|(\breve{\mathbf{U}}\breve{\mathbf{U}}^\top)^{-1}\right\|_F\asymp\frac{\sqrt{n}}{n}$. These bounds are obtained after replacing m by Cn. We plug all these bounds into the equation above.

Now, we plug in the values of Δ and ζ from Assump. 3. Therefore, with probability $1-2\delta$,

$$\|\tilde{\mathbf{a}}\|_2^2\lesssim\frac{n(\zeta^2+\Delta^2)}{m}\lesssim\frac{n^{-2\min\{\epsilon_1,\epsilon_2\}}}{m^2}.$$

We use the fact that $m=\Theta(n)$. Plugging in the bounds on Δ^2 and ζ^2 from Assump. 3, we obtain,

$$\|\tilde{\mathbf{a}}\|_\infty\lesssim\frac{n^{-\min\{\epsilon_1,\epsilon_2\}}}{m}$$

Therefore, with high probability,

$$\|\mathbf{a}\|_\infty\lesssim m^{-1}\Longrightarrow\|\mathbf{a}\|_1=\tilde{\mathcal{O}}(1).$$

Plugging this expression in Lem. 1, with high probability, we obtain,

$$\Upsilon(\mathbf{w})\lesssim d^{\frac{c}{c+1}}m^{-\frac{c-1}{c+1}}.$$

This completes the proof. □

Proof of Thm. 5. Let $\mathcal{W}_{flat}\coloneqq\{\mathbf{w}\in\mathbb{R}^{m(d+1)}:F_S(\mathbf{w})=0,\|\mathbf{a}\|_1\leqslant R,\|\theta_j\|_2=1,\forall j\in[m]\}$, be a superset of the flattest interpolators.

Then, we will bound the quantity $\sup_{\mathbf{w}\in\mathcal{W}_{flat}}F(\mathbf{w})-F^\star$.

First, using Assumption 3, we have,

$$F(\mathbf{w})=\frac{1}{2}\zeta^2+\frac{1}{2}\mathbb{E}\left[(y-\sum_{j=1}^m a_j\sigma(\langle\theta_j,\mathbf{x}\rangle))^2\right]\leqslant\frac{1}{2}\zeta^2+\Delta^2+\mathbb{E}\left[\left(\sum_{j=1}^{\breve{m}}\breve{a}_j\sigma(\langle\breve{\theta}_j,\mathbf{x}\rangle)-\sum_{j=1}^m a_j\sigma(\langle\theta_j,\mathbf{x}\rangle)\right)^2\right].$$

Now, consider the last term in the RHS.

Truncating the square loss. We first consider a truncated version of this term at level B to handle the unbounded square loss. Define $\tilde{f}_B(t)=\min\{t^2,B^2\}$ for some $B>0$. Then, note that the loss function $\tilde{f}_B(t)$ is bounded by B^2.

$$\begin{aligned}&\mathbb{E}\left[\left(\sum_{j=1}^{\breve{m}}\breve{a}_j\sigma(\langle\breve{\theta}_j,\mathbf{x}\rangle)-\sum_{j=1}^m a_j\sigma(\langle\theta_j,\mathbf{x}\rangle)\right)^2\right]-\mathbb{E}\left[\tilde{f}_B\left(\sum_{j=1}^{\breve{m}}\breve{a}_j\sigma(\langle\breve{\theta}_j,\mathbf{x}\rangle)-\sum_{j=1}^m a_j\sigma(\langle\theta_j,\mathbf{x}\rangle)\right)\right]\\ &=\mathbb{E}\left[\max\left\{\left(\sum_{j=1}^{\breve{m}}\breve{a}_j\sigma(\langle\breve{\theta}_j,\mathbf{x}\rangle)-\sum_{j=1}^m a_j\sigma(\langle\theta_j,\mathbf{x}\rangle)\right)^2-B^2,0\right\}\right]\end{aligned}$$

Note that each $\sigma(\langle\theta_j,\mathbf{x}\rangle)$ is a $(\frac{c}{2},\mathcal{O}(1))$-SubWeibull random variable as the activation is piece-wise polynomial. Therefore, as $\|\mathbf{a}\|_1=\mathcal{O}(1)$ from $\mathcal{W}_{flat}$ and $\|\breve{\mathbf{a}}\|_1=\mathcal{O}(1)$ from Assumption 3, the term

$\mathfrak{z} := \sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x} \rangle) - \sum_{j=1}^{m} a_j \sigma(\langle \theta_j, \mathbf{x} \rangle)$ is also a $(\frac{c}{2}, \mathcal{O}(1))$- SubWeibull random variable. Further, $\mathfrak{z}^2$ a $(c, \mathcal{O}(1))$ random variable. Additionally, $\mathbb{E}\left[\mathfrak{z}^2\right] = \mathcal{O}(1)$. For a large enough B, we have,

$$\mathbb{E}\left[\max\left\{\left(\sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x} \rangle) - \sum_{j=1}^{m} a_j \sigma(\langle \theta_j, \mathbf{x} \rangle)\right)^2 - B^2, 0\right\}\right] \leqslant \int_{B^2}^{\infty} \mathbb{P}\left[\mathfrak{z}^2 \geqslant t\right] dt \lesssim \exp(-B^{\frac{2}{c}}).$$

Choosing $B \asymp (\log(n))^{\frac{c}{2}}$, this deviation is bounded by $\lesssim \frac{1}{n}$. Therefore,

$$F(\mathbf{w}) \lesssim \frac{1}{2}\zeta^2 + \Delta^2 + \frac{1}{n} + \mathbb{E}\left[\tilde{f}_B\left(\sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x} \rangle) - \sum_{j=1}^{m} a_j \sigma(\langle \theta_j, \mathbf{x} \rangle)\right)\right] \tag{14}$$

Talagrand's contraction for bounded, Lipschitz losses Consider the last term in Eq (14). Note that, $\tilde{f}_B$ is a B^2-bounded loss function. Further, it is also $2B$-Lipschitz as the subderivative of $\tilde{f}_B$ is bounded by $2B$. For a bounded loss function, using symmetrization ((Wainwright, 2019, Theorem 4.10)), with probability $1-\delta$, we have,

$$\sup_{\mathbf{w} \in \mathcal{W}_{flat}} \mathbb{E}\left[\tilde{f}_B\left(\sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x} \rangle) - \sum_{j=1}^{m} a_j \sigma(\langle \theta_j, \mathbf{x} \rangle)\right)\right] - \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_B\left(\sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x}_i \rangle) - \sum_{j=1}^{m} a_j \sigma(\langle \theta_j, \mathbf{x}_i \rangle)\right)$$
$$\leqslant 2\hat{\mathfrak{R}}_n(\tilde{f}_B) + B^2\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Using $B^2 \asymp (\log(n))^c$, we can remove the last term in RHS. The first term in RHS is the empirical Rademacher complexity (Wainwright, 2019, Chapter 4) given by,

$$\hat{\mathfrak{R}}_n(\tilde{f}_B) = \mathop{\mathbb{E}}_{\{\varepsilon_i \sim \mathrm{Unif}(\{\pm 1\})\}}\left[\sup_{\mathbf{w} \in \mathcal{W}_{flat}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \tilde{f}_B\left(\sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x}_i \rangle) - \sum_{j=1}^{m} a_j \sigma(\langle \theta_j, \mathbf{x}_i \rangle)\right)\right|\right].$$

Here, ε_i are the Rademacher random variables independent of $\mathbf{x}_i$. First, note that the Rademacher complexity of a $\tilde{f}_B$ and its centered version, which can be obtained by subtracting its mean is the same order-wise from (Wainwright, 2019, Proposition 4.11) if it is Lipschitz. Further, from Talagrand's contraction Lemma ((Wainwright, 2019, Proposition 5.28)), for a centered $2B$-Lipschitz function, its rademacher complexity is $2B$ times the rademacher complexity of its argument. Therefore,

$$\hat{\mathfrak{R}}_n(\tilde{f}_B) \lesssim B\hat{\mathfrak{R}}_n(\mathcal{W}_{flat}), \quad \text{where,}$$
$$\hat{\mathfrak{R}}_n(\mathcal{W}_{flat}) = \mathop{\mathbb{E}}_{\{\varepsilon_i \sim \mathrm{Unif}(\{\pm 1\})\}}\left[\sup_{\mathbf{w} \in \mathcal{W}_{flat}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left(\sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x}_i \rangle) - \sum_{j=1}^{m} a_j \sigma(\langle \theta_j, \mathbf{x}_i \rangle)\right)\right|\right]. \tag{15}$$

We will bound Rademacher complexity $\hat{\mathfrak{R}}_n \breve{\mathcal{W}}_{flat}$ in the next paragraph, but, before that, we need to bound the empirical loss. Note that,

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{f}_B\left(\sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x}_i \rangle) - \sum_{j=1}^{m} a_j \sigma(\langle \theta_j, \mathbf{x}_i \rangle)\right) \leqslant \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x}_i \rangle) - \sum_{j=1}^{m} a_j \sigma(\langle \theta_j, \mathbf{x}_i \rangle)\right)^2$$
$$\leqslant \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{\breve{m}} \breve{a}_j \sigma(\langle \breve{\theta}_j, \mathbf{x}_i \rangle)\right)^2$$
$$\leqslant \frac{2}{n}\sum_{i=1}^{n} \xi_i^2 + 2\Delta^2.$$

We use interpolation and Assumption 3, for the intermediate inequalities. Note that, $\frac{\xi_i^2}{\zeta^2}$ are χ_1^2 random variables, so using χ^2 concentration, with probability $1-\delta$, we have,

$$\frac{2}{n}\sum_{i=1}^{n} \xi_i^2 \lesssim 2\zeta^2 + \zeta^2\sqrt{\frac{\log(1/\delta)}{n}}.$$

Therefore, with probability $1-2\delta$, we have,

$$\sup_{\mathbf{w}\in\mathcal{W}_{flat}} F(\mathbf{w}) \lesssim \zeta^2+\Delta^2+\sqrt{\frac{\log(1/\delta)}{n}}+(\log(n))^c\hat{\mathfrak{R}}_n(\mathcal{W}_{flat}). \tag{16}$$

Rademacher complexity of $\mathcal{W}_{flat}$ To obtain $\hat{\mathfrak{R}}_n(\mathcal{W}_{flat})$, we first change the weights from $\mathbf{w}$ to $\tilde{\mathbf{w}}$, where $\tilde{\mathbf{w}}\in\mathbb{R}^{m+\breve{m}}$. From Assumption 3, $\|\breve{\mathbf{a}}\|_1=\mathcal{O}(1)$ and $\left\|\breve{\theta}_j\right\|_2=1,\forall j\in[\breve{m}]$. Therefore, the last term can be represented as the square loss applied on a function $\tilde{h}(\tilde{\mathbf{w}})=\sum_{j=1}^{m+\breve{m}}\tilde{a}_j\sigma(\left\langle\tilde{\theta}_j,\mathbf{x}\right\rangle)$, where $\tilde{a}_j=a_j,\tilde{\theta}_j=\theta_j,\forall j\in[m]$ and $\tilde{a}_j=-\breve{a}_j,\tilde{\theta}_j=\breve{\theta}_j,\forall j\in\{m+1,m+2,\ldots,m+\breve{m}\}$. Here, $\tilde{\mathbf{w}}\in\mathbb{R}^{(m+\breve{m})(d+1)}$. We define the class,

$$\tilde{\mathcal{W}}_{flat}=\{\tilde{\mathbf{w}}\in\mathbb{R}^{m+\breve{m}}:\|\tilde{\mathbf{a}}\|_1\leqslant\tilde{R},\left\|\tilde{\theta}_j\right\|_2=1,\forall j\in[m+\breve{m}]\}.$$

Note that $\tilde{\mathcal{W}}_{flat}$ contains $\mathcal{W}_{flat}$ for a constant $\tilde{R}>R+\|\breve{\mathbf{a}}\|_1$, therefore, its rademacher complexity is an upper bound on $\hat{\mathfrak{R}}_n(\mathcal{W}_{flat})$, i.e.

$$\hat{\mathfrak{R}}_n(\mathcal{W}_{flat})\leqslant\hat{\mathfrak{R}}_n(\tilde{\mathcal{W}}_{flat}),\quad \text{where,}$$
$$\hat{\mathfrak{R}}_n\left(\tilde{\mathcal{W}}_{flat}\right)=\mathop{\mathbb{E}}_{\{\varepsilon_i\sim\mathrm{Unif}(\{\pm1\})\}}\left[\sup_{\tilde{\mathbf{w}}\in\tilde{\mathcal{W}}_{flat}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\left(\sum_{j=1}^{m+\breve{m}}\tilde{a}_j\sigma(\left\langle\tilde{\theta}_j,\mathbf{x}_i\right\rangle)\right)\right|\right]$$

Now, to bound $\hat{\mathfrak{R}}_n(\tilde{\mathcal{W}}_{flat})$, we use the fact that its outer-layer weights have $\mathcal{O}(1)$ ℓ_1 norm. Using Cauchy-Schwarz inequality, we have,

$$\begin{aligned}\sup_{\tilde{\mathbf{w}}\in\tilde{\mathcal{W}}_{flat}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\left(\sum_{j=1}^{m+\breve{m}}\tilde{a}_j\sigma(\left\langle\tilde{\theta}_j,\mathbf{x}_i\right\rangle)\right)\right|&=\sup_{\tilde{\mathbf{w}}\in\tilde{\mathcal{W}}_{flat}}\left|\sum_{j=1}^{m+\breve{m}}\tilde{a}_j\left(\frac{1}{n}\sum_{i=1}^n\varepsilon_i\sigma(\left\langle\tilde{\theta}_j,\mathbf{x}_i\right\rangle)\right)\right|\\&\leqslant\|\tilde{\mathbf{a}}\|_1\sup_{\|\theta\|_2=1}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\sigma(\langle\theta,\mathbf{x}_i\rangle)\right|.\end{aligned}$$

As $\|\tilde{\mathbf{a}}\|_1\leqslant\tilde{R}$, we can bound $\mathfrak{R}_n(\tilde{\mathcal{W}}_{flat})$ as follows,

$$\hat{\mathfrak{R}}_n(\tilde{\mathcal{W}}_{flat})\leqslant\tilde{R}\hat{\mathfrak{R}}_n(\sigma),\text{ where,}$$
$$\hat{\mathfrak{R}}_n(\sigma)=\mathop{\mathbb{E}}_{\{\varepsilon_i\sim\mathrm{Unif}(\{\pm1\})\}}\left[\sup_{\|\theta\|_2=1}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\sigma(\langle\theta,\mathbf{x}_i\rangle)\right|\right]$$

Note that $\hat{\mathfrak{R}}_n(\sigma)$ is simply the Rademacher complexity of our activation restricted to θ lying on the unit sphere.

Rademacher complexity of activation Note that our activation is piece-wise polynomial. If $c=1$, then $|\sigma'(u)|\lesssim1$, and the function σ is Lipschitz. Using Talagrand contraction Lemma (Wainwright, 2019, Propositon 5.28) for this Lipschitz function and the Rademacher complexity of Linear functions on unit ball in d-dimensions, we have,

$$\begin{aligned}\hat{\mathfrak{R}}_n(\sigma)&\lesssim\mathop{\mathbb{E}}_{\varepsilon_i\sim\mathrm{Unif}(\{\pm1\})}\left[\sup_{\|\theta\|_2=1}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\langle\theta,\mathbf{x}_i\rangle\right|\right]=\frac{\mathbb{E}_{\varepsilon_i}\left[\|\sum_{i=1}^n\varepsilon_i\mathbf{x}_i\|_2\right]}{n}\leqslant\frac{1}{n}\sqrt{\mathbb{E}_{\varepsilon_i}\left[\left\|\sum_{i=1}^n\varepsilon_i\mathbf{x}_i\right\|_2^2\right]}\\&\lesssim\frac{1}{n}\sqrt{\sum_{i=1}^n\|\mathbf{x}_i\|_2^2}.\end{aligned}$$

Note that $\sum_{i=1}^n\|\mathbf{x}_i\|_2^2\asymp nd$ with probability $1-\delta$, using χ^2_d concentration. Therefore, for $c=1$, $\hat{\mathfrak{R}}_n(\sigma)\lesssim\sqrt{\frac{d}{n}}$.

For $c>1$, we need to use a sharper concentration that follows the steps of (Dirksen, 2015, Theorem 3.5). Let $\mathfrak{G}(\theta)=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\sigma(\langle\theta,\mathbf{x}_i\rangle)$. Then, $\mathbb{E}[\mathfrak{G}(\theta)]=0$.

Consider $\theta_1\neq\theta_2$ Note that $\varepsilon_i(\sigma(\langle\theta_1,\mathbf{x}_i\rangle)-\sigma(\langle\theta_2,\mathbf{x}_i\rangle))$ is a $(\frac{c}{2},\mathcal{O}(\|\theta_1-\theta_2\|_2))$-SubWeibull random variable, as ε_i only changes the sign. To check that this variable is indeed $\mathcal{O}(\|\theta_1-\theta_2\|_2)$-SubWeibull, we compute its kth moment.

Note that, for a piece-wise polynomial activation σ,

$$|\sigma(\langle\theta_1,\mathbf{x}\rangle)-\sigma(\langle\theta_2,\mathbf{x}\rangle)|\lesssim\|\theta_1-\theta_2\|_2\|\mathbf{x}\|_2\left(|\langle\theta_1,\mathbf{x}\rangle|^{c-1}+|\langle\theta_2,\mathbf{x}\rangle|^{c-1}\right).$$

Since $\langle\theta_1,\mathbf{x}\rangle\sim\mathcal{N}(\|\theta_1\|_2^2),\langle\theta_2,\mathbf{x}\rangle\sim\mathcal{N}(0,\|\theta_2\|_2^2)$, therefore, its kth moment scales as $\|\theta_1-\theta_2\|_2$, as long as $\|\theta_1\|_2,\|\theta_2\|_2\leqslant 1$.

Now, using the precise form of Lemma 15 from (Zhang & Wei, 2022, Theorem 1) for mean of Sub-Weibull random variables, we have, for all $t\geqslant 0$,

$$\mathbb{P}\left[|\mathfrak{G}(\theta_1)-\mathfrak{G}(\theta_2)|\geqslant C\|\theta_1-\theta_2\|_2\left(\sqrt{\frac{t}{n}}+\frac{t^{\frac{c}{2}}}{n^b}\right)\right]\leqslant\exp(-t). \tag{17}$$

Here, $b=1$ if $c\geqslant 2$ and $b=\frac{c}{2}$ if $c\in[1,2]$ and $C>0$ is some constant. We use the precise version as (Dirksen, 2015, Theorem 3.5) handles the tails carefully to get a good sample complexity.

Now, we will use the following Lemma to obtain bounds on this random variable.

Lemma 13 (Chaining for Sub-Weibull Random Variables)**.** *For a continuous function $\mathfrak{G}:\mathbb{R}^d\to\mathbb{R}$ that satisfies Eq* (17) *for $\theta_1\neq\theta_2$ such that $\|\theta_1\|_2,\|\theta_2\|_2\leqslant 1$, with $\mathfrak{G}(0)=0$, we have, for all $t\geqslant 0$,*

$$\mathbb{P}\left[\sup_{\|\theta\|_2=1}\mathfrak{G}(\theta)\geqslant C'\left(\sqrt{\frac{d}{n}}+\sqrt{\frac{t}{n}}+\frac{d^{\frac{c}{2}}}{n^b}+\frac{t^{\frac{c}{2}}}{n^b}\right)\right]\lesssim\exp(-t).$$

The proof of this Lemma is provided in Appendix E.6.

Now, we compute $\mathbb{E}\left[\sup_{\|\theta\|_2=1}\mathfrak{G}(\theta)\right]$. Define $\tilde{\mathfrak{G}}=\max\left\{\sup_{\|\theta\|_2=1}\mathfrak{G}(\theta)-C'\left(\sqrt{\frac{d}{n}}+\frac{d^{\frac{c}{2}}}{n^b}\right),0\right\}$.

Then, using Lem. 13, for any $p\geqslant 1$, we have,

$$\mathbb{E}\left[\tilde{\mathfrak{G}}^p\right]\leqslant p\int_0^\infty u^{p-1}\mathbb{P}\left[\tilde{\mathfrak{G}}\geqslant u\right]du\lesssim p\int_0^\infty C''\left(\exp\left(-\frac{nu^2}{(C')^2}\right)+\exp\left(-\left(\frac{n^bu}{C'}\right)^{\frac{2}{c}}\right)\right)du.$$

Note that this term is a sum of gaussian tail and a Sub-Weibull tail. Plugging in the value of Gaussian and Sub-Weibull moments from Def. 3, we obtain,

$$\left\|\tilde{\mathfrak{G}}\right\|_{L_p}\lesssim\sqrt{\frac{p}{n}}+\frac{p^{\frac{c}{2}}}{n^b}.$$

Now, using Jensen's inequality, we have,

$$\mathop{\mathbb{E}}_{\mathbf{x}_i}\left[\max\left\{\hat{\mathfrak{R}}_n(\sigma)-C'\left(\sqrt{\frac{d}{n}}+\frac{d^{\frac{c}{2}}}{n^b}\right),0\right\}^p\right]\leqslant\mathop{\mathbb{E}}_{\mathbf{x}_i,\varepsilon_i}\left[\tilde{\mathfrak{G}}^p\right]$$

Using Markov's inequality, and setting $p=\max\{1,\log(C'''/\delta)\}$ for some constant $C'''>0$ and $\delta\in(0,1)$, we obtain,

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathbf{x}_i}\left[\hat{\mathfrak{R}}_n(\sigma)\geqslant C'\left(\sqrt{\frac{d}{n}}+\frac{d^{\frac{c}{2}}n^b}{+}\sqrt{\frac{\log(C'''/\delta)}{n}}+\frac{(\log(C'''/\delta))^{\frac{c}{2}}}{n^b}\right)\right]\\
&\leqslant\frac{\mathop{\mathbb{E}}_{\mathbf{x}_i}\left[\max\left\{\hat{\mathfrak{R}}_n(\sigma)-C'\left(\sqrt{\frac{d}{n}}+\frac{d^{\frac{c}{2}}}{n^b}\right),0\right\}^p\right]}{\exp(\log(C'''/\delta))(C')^p\left(\sqrt{\frac{\log(C'''/\delta)}{n}}+\frac{(\log(C'''/\delta))^{\frac{c}{2}}}{n^b}\right)^p}\\
&\leqslant\delta.
\end{aligned}$$

Therefore, with probability $1-\delta$, we have,

$$\hat{\mathfrak{R}}_n(\sigma) = \mathop{\mathbb{E}}_{\varepsilon_i}\left[\sup_{\|\theta\|_2=1} |\mathfrak{G}(\theta)|\right] \lesssim \sqrt{\frac{d}{n}} + \frac{d^{\frac{c}{2}} n^b}{+}\sqrt{\frac{\log(1/\delta)}{n}} + \frac{(\log(1/\delta))^{\frac{c}{2}}}{n^b}.$$

Combining this with Eq (16), with probability $1-3\delta$, we have,

$$\sup_{\mathbf{w}\in\mathcal{W}_{flat}} F(\mathbf{w}) \lesssim \zeta^2 + \Delta^2 + \sqrt{\frac{d}{n}} + \frac{d^{\frac{c}{2}}}{n^b}.$$

From Assumption 3, the terms of $\zeta^2, \Delta^2 \lesssim \sqrt{\frac{d}{n}}$. This completes the proof.

□

E Technical Tools and Missing Proofs

E.1 Sub-Weibull Distributions

A comprehensive treatment of Sub-Weibull random variables is provided in (Vladimirova et al., 2020; Sambale, 2023; Zhang & Wei, 2022; Zhang & Chen, 2021). We restate the main results that we use for Sub-Weibull random variables.

We first provide definition and properties of Sub-Weibull random variables, mostly adapted from (Zhang & Chen, 2021, Corollary 6.1).

Definition 3 ((ϱ, K)-Sub-Weibull Random Variable)**.** *For a random variable v the following statements are equivalent.*

- *v is a (ϱ,K)-Sub-Weibull random variable for $\varrho \geqslant \frac{1}{2}$ and $K>0$.*
- *The tails of v satisfy $\mathbb{P}[|v-\mathbb{E}[v]| \geqslant t] \leqslant \exp\left(-\left(\frac{t}{K_1}\right)^{\frac{1}{\varrho}}\right)$ for all $t>0$ where $K_1 = \Theta(K)$.*
- *The moments of v satisfy $\|v\|_{L_k} \coloneqq (\mathbb{E}\left[|v-\mathbb{E}[v]|^k\right])^{\frac{1}{k}} \leqslant K_2 k^{\varrho}$ for all $k \geqslant \min\{1,\varrho^{-1}\}$, where $K_2 = \Theta(K)$.*
- *The MGF of $|v-\mathbb{E}[v]|^{\varrho}$ satisfies, $\mathbb{E}[\exp((\lambda|v-\mathbb{E}[v]|)^{\varrho})] \leqslant \exp((\lambda K_3)^{\varrho}), \forall|\lambda| \leqslant K_3^{-1}$ where $K_3 = \Theta(K)$.*
- *Orlicz Norm: $K = \inf\left\{K_4 \in (0,\infty) : \mathbb{E}\left[\exp\left(\frac{|v-\mathbb{E}[v]|}{K_4}\right)^{\varrho}\right] \leqslant 2\right\}$.*

The following Lem. provides guarantees on sum and products of Sub-Weibull random variables. This has been adapted from (Zhang & Wei, 2022).

Lemma 14 (Sum and Product of Sub-Weibull Random Variables)**.** *If v_i is a (ϱ_i, K_i)-Sub-Weibull random variable $\forall i \in [r]$ then,*

- *$\sum_{i=1}^r v_i$ is a (ϱ,K')-Sub-Weibull random variable if $\varrho = \varrho_i, \forall i \in [r]$ and $K' = \Theta(\sum_{i=1}^r K_i)$.*
- *$\prod_{i=1}^r v_i$ is a (ϱ,K'')-Sub-Weibull random variable where $\varrho = \sum_{i=1}^r \varrho_i$ and $K'' = \prod_{i=1}^r K_i$.*

The following Lemma provides concentration for sum of independent Sub-Weibull random variables. This has been adapted from (Zhang & Wei, 2022, Prop. 3).

Lemma 15 (Concentration of sum of Sub-Weibull Random Variables)**.** *If $\{v_i\}_{i\in[r]}$ are n iid (ϱ,K)-Sub-Weibull random variables for $\varrho \geqslant \frac{1}{2}$, then with probability $1-\delta$,*

$$\left|\frac{1}{n}\sum_{i=1}^n v_i - \mathbb{E}[v_1]\right| \leqslant \mathcal{O}\left(K \max\left\{\sqrt{\frac{\log(1/\delta)}{n}}, \frac{(\log(1/\delta))^{\frac{c}{2}}}{n^b}\right\}\right).$$

Here, $b=1$ for $c>2$ and $b=\frac{c}{2}$ for $c\in[1,2]$.

This Lem. provides bounds on quadratic forms for Sub-Weibull random variables and has been adapted from Sambale (2023).

Lemma 16 (Hanson-Wright for Sub-Weibull distributions)**.** *If $\mathbf{v}\in\mathbb{R}^n$ is a random vector with $\mathbf{v}_i$ being an independent (ϱ,K_i)-Sub-Weibull random variable. Then, for any deterministic symmetric matrix $\bar{\mathbf{U}}\in\mathbb{R}^{n\times n}$, with probability $1-\delta$,*

$$\left|\mathbf{v}^\top\bar{\mathbf{U}}\mathbf{v}-\mathbb{E}\left[\mathbf{v}^\top\bar{\mathbf{U}}\mathbf{v}\right]\right|\leqslant\bar{\nu}_\varrho\max\{\left\|\mathbf{K}\bar{\mathbf{U}}\mathbf{K}\right\|_F\sqrt{\log(2/\delta)},\left\|\mathbf{K}\bar{\mathbf{U}}\mathbf{K}\right\|_2(\log(2/\delta))^{2\varrho}\} \tag{18}$$

for some absolute constant $\bar{\nu}_\varrho$ depending on ϱ and $\mathbf{K}\in\mathbb{R}^{n\times n}$ is given by $\mathbf{K}=\operatorname{diag}(K_1,K_2,...,K_n)$.

Proof. The proof simply replaces $\bar{\mathbf{U}}$ by $\mathbf{K}\bar{\mathbf{U}}\mathbf{K}$ which makes $\tilde{\mathbf{v}}=\mathbf{K}^{-1}\mathbf{v}$ a vector of indepedent $(\varrho,1)$-Sub-Weibull random variables. Further, the term α in (Sambale, 2023, Thm. 2.1) is equal to $\frac{1}{\varrho}$ in our notation. □

Lemma 17 (Bai-Yin for Sub-Weibull Random Matrices with Independent Columns)**.** *For a matrix $\bar{\mathbf{U}}\in\mathbb{R}^{n\times m}$ with independent rows, each element $\bar{\mathbf{U}}_{i,j}$ being a 0-mean $(\varphi,\mathcal{O}(1))$-Sub-Weibull random variable, with probability $1-\frac{\delta}{2}$,*

$$\sigma_{\max}(\bar{\mathbf{U}})\lesssim\sqrt{\lambda_{\max}(\bar{\Phi})}(\sqrt{m}+\sqrt{n}).$$

Further, if $\lambda_{\min}(\bar{\Phi})=\Omega(1)$ and $\sigma_{\max}(\Theta)=\mathcal{O}(1)$, then, with probability $1-\frac{\delta}{2}$,

$$\sigma_n(\bar{\mathbf{U}})\gtrsim\sqrt{m}.$$

Here, $\bar{\Phi}=\frac{1}{n}\mathbb{E}\left[\bar{\mathbf{U}}^\top\bar{\mathbf{U}}\right]$ is the covariance of the matrix $\bar{\mathbf{U}}$, and each entry $\bar{\mathbf{U}}_{i,j}=\sigma(\langle\theta_j,\mathbf{x}_i\rangle)-\mathbb{E}\left[\sigma(\langle\theta_j,\mathbf{x}_i\rangle)\right]$, where $\Theta\in\mathbb{R}^{m\times d}$ is the matrix of the vectors $\{\theta_j\}_{j\in[m]}$ and σ is a Piece-wise polynomial activation.

Proof. **Maximum Singular Value** For $\sigma_{\max}$, define two $\frac{1}{2}$-covers, $\mathcal{C}_{\frac{1}{2}}(\mathbb{S}^{n-1})$ and $\mathcal{C}_{\frac{1}{2}}(\mathbb{S}^{m-1})$. Note that,

$$\sigma_{\max}(\bar{\mathbf{U}})=\max_{\mathbf{v}_1\in\mathbb{S}^{n-1},\mathbf{v}_2\in\mathbb{S}^{m-1}}\mathbf{v}_1^\top\bar{\mathbf{U}}\mathbf{v}_2^\top\leqslant 4\max_{\mathbf{v}_1\in\mathcal{C}_{\frac{1}{2}}(\mathbb{S}^{n-1}),\mathbf{v}_2\in\mathcal{C}_{\frac{1}{2}}(\mathbb{S}^{m-1})}\mathbf{v}_1^\top\bar{\mathbf{U}}\mathbf{v}_2^\top.$$

For a fixed $\mathbf{v}_1,\mathbf{v}_2$, we have $\mathbb{E}\left[\mathbf{v}_1^\top\bar{\mathbf{U}}\mathbf{v}_2\right]=0$. Further,

$$\mathbf{v}_1^\top\bar{\mathbf{U}}\mathbf{v}_2=\sum_{i=1}^n(\mathbf{v}_1)_i\left\langle\bar{\mathbf{U}}_i,\mathbf{v}_2\right\rangle.$$

Note that $\left\langle\bar{\mathbf{U}}_i,\mathbf{v}_2\right\rangle$ is a sum of dependent Sub-Weibull variables, hence, it is $(\varphi,\mathcal{O}(\mathbf{v}_2^\top\bar{\Phi}\mathbf{v}_2))$-Sub-Weibull. Note that $\mathbf{v}_2^\top\bar{\Phi}\mathbf{v}_2\leqslant\lambda_{\max}(\bar{\Phi})\leqslant\operatorname{Tr}(\bar{\Phi})=\mathcal{O}(m)$. Further, the term $\mathbf{v}_1^\top\bar{\mathbf{U}}\mathbf{v}_2$ is a sum of n iid $(\varphi,\mathcal{O}(1))$-Sub-Weibull random variables. Hence, by Lem. 15, and taking a union bound over the size of the covers, from (Wainwright, 2019, Chapter 5), we have, with probability $1-\delta$,

$$\mathbf{v}_1^\top\bar{\mathbf{U}}\mathbf{v}_2\leqslant\widetilde{C}_1\sqrt{\lambda_{\max}(\bar{\Phi})}(\sqrt{m}+\sqrt{n})+\widetilde{C}_2(\log(1/\delta))^c$$

for some constants $\widetilde{C}_1,\widetilde{C}_2>0$. By using the bounds from the cover, $\sigma_{\max}(\bar{\mathbf{U}})\lesssim\sqrt{\lambda_{\max}}(\bar{\Phi})(\sqrt{m}+\sqrt{n})$.

Minimum Singular Value For $\sigma_{\min}$, we will use the thin-shell property from (Bartl & Mendelson, 2022, Definition 1.4). This requires us to show that that $\mathbf{V}=\bar{\mathbf{U}}\bar{\Phi}^{-\frac{1}{2}}$ and its component vector $\mathbf{V}_i=\bar{\Phi}^{-\frac{1}{2}}\bar{\mathbf{U}}_i\in\mathbb{R}^m$ are 0-mean, independent, symmetric, and satisfy the following equations.

$$\mathbb{P}\left[\max_{1\leqslant i\leqslant n}\left|\frac{\|\mathbf{V}_i\|^2}{m}-1\right|\geqslant\upsilon\right]\leqslant\delta' \tag{19}$$

$$\|\langle\mathbf{V}_i,\mathbf{v}\rangle\|_{L_p}\lesssim p^{\frac{1}{\upsilon'}}\|\langle\mathbf{V}_i,\mathbf{v}\rangle\|_{L_2},\qquad\forall 2\leqslant p\lesssim\log(n),\forall i\in[n]. \tag{20}$$

Here, $\upsilon,\delta'\in(0,1),\upsilon'\in(0,2]$, and $\|A\|_{L_p}=(\mathbb{E}[|A|^p])^{\frac{1}{p}}$ is the L_p-norm of a random variable for some $p\geqslant 0$.

Note that $\mathbf{V}_i$ are already 0-mean and independent. To ensure that these are symmetric, we can multiply each $\mathbf{V}_i$ by an independent Rademacher random variable $\varepsilon_i \sim \text{Unif}\{\pm 1\}$. This ensures that $\varepsilon_i \mathbf{V}_i$ is symmetric. Further, this transformation doesn't change Eq (19) or Eq (20) as $|\varepsilon_i| = 1$. Further, the eigenvalues of $\{\varepsilon \mathbf{V}_i\}_{i\in[n]}$ are also the same, as $\left\|\mathbf{V}^\top \mathbf{v}\right\|_2^2 = \mathbf{v}^\top \mathbf{V}\mathbf{V}\mathbf{D}\mathbf{V}^\top \mathbf{v}$, where $\mathbf{D} = \text{diag}\{\varepsilon_i^2\} = \mathbb{I}_d$.

If $\mathbf{V}_i$ satisfies the above bounds, then applying (Bartl & Mendelson, 2022, Theorem 1.5) to $\mathbf{V}$ for the set $\mathbb{S}^{n-1}$, we obtain, with probability $1-\frac{\delta}{2}$

$$\sup_{\mathbf{v}\in\mathbb{S}^{n-1}} \frac{1}{m}\left\|\left\|\mathbf{V}^\top \mathbf{v}\right\|_2^2 - 1\right\|_2^2 \leqslant C\sqrt{\frac{n}{m}}$$

for some small constant $C > 0$. We use the fact that the gaussian width of the set $\mathbb{S}^{n-1}$ is $\sqrt{n}$ (Wainwright, 2019, Chapter 5). Choosing a small enough constant C and as $m \geqslant n$, we obtain,

$$\|\mathbf{V}\mathbf{v}\|_2^2 \asymp \sqrt{m}, \quad \forall \mathbf{v} \in \mathbb{S}^{n-1}.$$

Note that,

$$\sigma_n(\bar{\mathbf{U}}) \geqslant \sqrt{\lambda_{\min}(\bar{\Phi})}\sigma_n(\mathbf{V}) \gtrsim \inf_{\mathbf{v}\in\mathbb{S}^{n-1}} \|\mathbf{V}\mathbf{v}\|_2^2 \gtrsim \sqrt{m}.$$

We use the fact that $\lambda_{\min}(\bar{\Phi}) = \Omega(1)$.

Therefore, we only need to show that Eq (19) and Eq (20) hold for our piece-wise polynomial activations.

We will first show that Eq (19) holds. For a fixed $i \in [n]$, with probability, $1 - m^{-C_1}$, for some $C_1 > 0$, we have,

$$\max_{i\in[n], j\in[m]} |\langle \theta_j, \mathbf{x}_i\rangle| \leqslant T$$

Note that each $\langle \theta_j, \mathbf{x}_i\rangle \sim \mathcal{N}(0,1)$, so by gaussian maximal inequality, $T \asymp \sqrt{\log(m)}$, as $m = \Theta(n)$.

We consider a vector $\mathbf{V}_{T,i} \in \mathbb{R}^m$ where we replace the jth coordinate of $\bar{\mathbf{U}}_i$ by the truncated $\sigma(\max\{-T, \min\{\langle \theta_j, \mathbf{x}_i\rangle, T\}\} \leqslant T\})$ and still remove the mean of σ and the apply $\bar{\Phi}^{-\frac{1}{2}}$. We will bound the probability of this truncated version of $\mathbf{V}_i$ satisfying Eq (19), as with probability $1 - m^{-C_1}$, $\mathbf{V}_{T,i} = \mathbf{V}_i$.

If $\|\mathbf{V}_{T,i}\|_2^2 = \tilde{L}(\mathbf{x}_i)$, then the function $\tilde{L}$ is Lipschitz, with its Lipschitz constant given by $\lesssim \lambda_{\min}^{-1}(\bar{\Phi})\sqrt{m}T^{2c-1}\sigma_{\max}^{2c}(\Theta)$, as σ is a c-homogeneous function. Plugging in $\lambda_{\min}(\bar{\Phi}) = \Omega(1)$ and $T \asymp \sqrt{\log(m)}$, and $\sigma_{\max}(\Theta) = \mathcal{O}(1)$, the Lipschitz constant of the function is $\lesssim \sqrt{m}(\log(m))^{\frac{2c-1}{2}}$. Note that $\tilde{L}$ is a Lipschitz function of a Gaussian, therefore, we can bound its deviation from its mean. We only need to establish that the mean of this function is close m. Note that, $\mathbb{E}\left[\|\mathbf{V}_i\|_2^2\right] = m, \forall i \in [n]$.

Now, we will bound the error between, $\mathbb{E}\left[\|\mathbf{V}_{T,i} - \mathbf{V}_i\|_2^2\right]$.

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{V}_{T,i} - \mathbf{V}_i\|_2^2\right] &\leqslant \lambda_{\min}^{-1}(\bar{\Phi}) m \mathbb{E}\left[(\sigma(\langle \theta_j, \mathbf{x}_i\rangle) - T)^2 \mathbf{1}\{|\langle \theta_j, \mathbf{x}_i\rangle| > T\}\right] \\
&\lesssim m\mathbb{E}\left[|\langle \theta_j, \mathbf{x}_i\rangle|^{2c} \mathbf{1}\{|\langle \theta_j, \mathbf{x}_i\rangle| > T\}\right] \lesssim m\frac{2^c}{\sqrt{\pi}}\Gamma\left(\frac{2c+1}{2}, \frac{T^2}{2}\right) \\
&\lesssim mT^{2c-1}\exp\left(-\frac{T^2}{2}\right) \lesssim m(\log(m))^{\frac{2c-1}{2}} m^{-2} \lesssim m^{-1}.
\end{aligned}$$

We first use Cauchy-Schwarz, then use the bound on $\lambda_{\min}(\bar{\Phi})$, followed by the fact that σ is a piece-wise polynomial activation. Then, we compute the exact value of the expectation which is given by a Γ function. We then substitute a bound on the Γ function and finally the value of T. We can keep $T \geqslant 2\sqrt{\log(m)}$), so that we get the given bound of m^{-1}. By triangle inequality, we have,

$$\left|\mathbb{E}\left[\|\mathbf{V}_{T,i}\|_2^2\right] - m\right| \lesssim m^{-\frac{1}{2}}.$$

Now, note that $\tilde{L}$ is a Lipschitz function of the gaussian $\mathbf{x}_i$, so using concentration of Lipschitz functions of Gaussians from (Wainwright, 2019, Chapter 2) and with a union bound over all $i \in [m]$, we obtain,

$$\mathbb{P}\left[\max_{i\in[n]}\left|\frac{\mathbf{V}_{T,i}^2}{m} - 1\right| \geqslant \upsilon\right] \lesssim n\exp\left(-\frac{m^2\nu^2}{m(\log(m))^{\frac{2c-1}{2}}}\right).$$

Adding the probability of error of m^{-C_1} from thresholding and using the fact that $m \geqslant n$, we prove Eq (19).

Now, to prove Eq (20), as $\mathbb{E}\big[\mathbf{V}_i^\top \mathbf{V}_i\big] = \mathbb{I}_m$, we have,

$$\|\langle \mathbf{V}_i, \mathbf{v}\rangle\|_{L_2}^2 = \mathbb{E}\Big[\langle \mathbf{V}_i, \mathbf{v}\rangle^2\Big] = \|\mathbf{v}\|_2^2.$$

Therefore, we only need to compare the L_p norm of $\langle \mathbf{V}_i, \mathbf{v}\rangle$ to $\|\mathbf{v}\|_2$.

Using $\lambda_{\min}(\Phi) = \Omega(1)$, we have,

$$\begin{aligned}\|\langle \mathbf{V}_i, \mathbf{v}\rangle\|_{L_p}^p = \mathbb{E}[|\langle \mathbf{V}_i, \mathbf{v}\rangle|^p] &\leqslant \lambda_{\min}^{\frac{p}{2}}(\bar{\Phi})\mathbb{E}\big[\big|\big\langle \mathbf{U}_i - \mathbb{E}\big[\bar{\mathbf{U}}_i\big], \mathbf{v}\big\rangle\big|^p\big] \\ &\lesssim \mathbb{E}\big[\big|\big\langle \mathbf{U}_i - \mathbb{E}\big[\bar{\mathbf{U}}_i\big], \mathbf{v}\big\rangle\big|^p\big] = \big\|\big\langle \mathbf{U}_i - \mathbb{E}\big[\bar{\mathbf{U}}_i\big], \mathbf{v}\big\rangle\big\|_{L_p}^p.\end{aligned}$$

Further, using Gaussian-Poincare L_p inequality (Adamczak & Wolff, 2014), for any function L, we have, $\|f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]\|_{L_p} \lesssim \sqrt{p}\big\|\|\nabla f(\mathbf{x})\|_2\big\|_{L_p}$. Here, $\langle \mathbf{U}_i, \mathbf{v}\rangle = \sum_{j=1}^m \mathbf{v}_j \sigma(\langle \theta_j, \mathbf{x}_i\rangle)$. Therefore, its gradient with respect to $\mathbf{x}_i$ is given by the vector with j^{th} coordinate $\mathbf{v}_j \sigma'(\langle \theta_j, \mathbf{x}_i\rangle)\theta_j$. Using Gaussian-Poincare L_p inequality, we have,

$$\begin{aligned}\|\langle \mathbf{V}_i, \mathbf{v}\rangle\|_{L_p} \lesssim \sqrt{p}\left\|\sqrt{\sum_{j=1}^m \mathbf{v}_j^2 (\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2 \|\theta_j\|_2^2}\right\|_{L_p} &\lesssim \sigma_{\max}(\Theta)\sqrt{p}\left\|\sqrt{\sum_{j=1}^m \mathbf{v}_j^2 (\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2}\right\|_{L_p}, \\ &\lesssim \sqrt{p}\left\|\sqrt{\sum_{j=1}^m \mathbf{v}_j^2 (\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2}\right\|_{L_p}.\end{aligned}$$

We use the fact that $\sigma_{\max}(\Theta) = \mathcal{O}(1)$. Now, note that,

$$\begin{aligned}\left\|\sqrt{\sum_{j=1}^m \mathbf{v}_j^2 (\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2}\right\|_{L_p}^p &= \mathbb{E}\left[\left(\sum_{j=1}^m \mathbf{v}_j^2 (\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2\right)^{\frac{p}{2}}\right], \\ &= \left\|\sum_{j=1}^m \mathbf{v}_j^2 (\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2\right\|_{L_{\frac{p}{2}}}^{\frac{p}{2}}.\end{aligned}$$

We use the definition of the L_p norm for the above equation.

Now, consider the following term, and use triangle inequality for the $L_{\frac{p}{2}}$ norm,

$$\left\|\sum_{j=1}^m \mathbf{v}_j^2 (\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2\right\|_{L_{\frac{p}{2}}} \leqslant \sum_{j=1}^m \mathbf{v}_j^2 \big\|(\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2\big\|_{L_{\frac{p}{2}}}.$$

For each term on the RHS, note that as σ' is a $(c-1)$ homogeneous piece-wise polynomial function, we have,

$$\begin{aligned}\big\|(\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2\big\|_{L_{\frac{p}{2}}}^{\frac{p}{2}} = \mathbb{E}\big[(\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^p\big] &\leqslant C_2 + C_3 \mathbb{E}\Big[|\langle \theta_j, \mathbf{x}_i\rangle|^{p(c-1)}\Big] \\ &\leqslant C_2 + C_3 (p(c-1))^{\frac{p(c-1)}{2}} \lesssim p^{\frac{p(c-1)}{2}}.\end{aligned}$$

We use the fact that $\langle \theta_j, \mathbf{x}_i\rangle \sim \mathcal{N}(0,1)$ and for a unit normal L_p norm is bounded by $\sqrt{p}$.

Substituting this value into all the intermediate bounds, we obtain,

$$\left\|\sum_{j=1}^m \mathbf{v}_j^2 (\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2\right\|_{L_{\frac{p}{2}}} \leqslant \sum_{j=1}^m \mathbf{v}_j^2 \big\|(\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2\big\|_{L_{\frac{p}{2}}} \lesssim \|\mathbf{v}\|_2^2 p^{c-1}$$

$$\|\langle \mathbf{V}_i, \mathbf{v}\rangle\|_{L_p} \lesssim \sqrt{p}\left\|\sqrt{\sum_{j=1}^m \mathbf{v}_j^2 (\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2}\right\|_{L_p} \lesssim \sqrt{p}\sqrt{\left\|\sum_{j=1}^m \mathbf{v}_j^2 (\sigma'(\langle \theta_j, \mathbf{x}_i\rangle))^2\right\|_{L_{\frac{p}{2}}}}$$

$$\implies \|\langle \mathbf{V}_i, \mathbf{v}\rangle\|_{L_p} \lesssim \sqrt{p} p^{\frac{c-1}{2}} \|\mathbf{v}\|_2 = p^{\frac{c}{2}} \|\mathbf{v}\|_2.$$

As $\nu' = \frac{2}{c} \leqslant 2$ for $c \geqslant 1$, Eq (20) is also satisfied.

This completes the proof.

□

E.2 Proofs for App. C

E.2.1 Proof of Prop. 3

Properties of σ First, for a piece-wise polynomial σ defined in Assump. 1, we compute the expectation, variance and tail behavior, for $\sigma(b)$ where $\mathbf{b} \sim \mathcal{N}(\mu, \varsigma^2)$. We first compute this for the case of $m^\star = 1$

Mean of σ We use the transformation $b' = \frac{b-\mu}{\varsigma}$

$$\mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}[\sigma(b)] = c' \int_{-\frac{\mu}{\varsigma}}^{\infty} \left|\mu+\varsigma b'\right|^c \phi(b')db' + c'' \int_{-\infty}^{-\frac{\mu}{\varsigma}} \left|\mu+\varsigma b'\right|^c \phi(b')db'$$

Note that $|b_1+b_2|^c \leqslant \nu_c(|b_1|^c + |b_2|^c)$ for the constant $\widetilde{\nu}_c \leqslant 2^{c-1}$ for any $c \geqslant 1$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}[\sigma(b)] \leqslant& \widetilde{\nu}_c|\mu|^c \left(\left|c'\right| \int_{-\frac{\mu}{\varsigma}}^{\infty} \phi(b')db' + \left|c''\right| \int_{-\infty}^{-\frac{\mu}{\varsigma}} \phi(b')db' \right) \\
&+ \widetilde{\nu}_c\varsigma^c \left(\left|c'\right| \int_{-\frac{\mu}{\varsigma}}^{\infty} \left|b'\right|^c \phi(b')db' + \left|c''\right| \int_{-\infty}^{-\frac{\mu}{\varsigma}} \left|b'\right|^c \phi(b')db' \right) \\
=& \widetilde{\nu}_c|\mu|^c \left(\left|c'\right| + (\left|c''\right| - \left|c'\right|)\Phi\left(-\frac{\mu}{\varsigma}\right) \right) \\
&+ \widetilde{\nu}_c\varsigma^c 2^{\frac{c}{2}-1}\pi^{-\frac{1}{2}} \left((\left|c'\right| + \left|c''\right|)\Xi_1\left(\frac{c+1}{2}, \frac{2\mu^2}{\varsigma^2}\right) + 2\left|c'\right|\Xi_2\left(\frac{c+1}{2}, \frac{2\mu^2}{\varsigma^2}\right) \right)
\end{aligned}$$

Here, $\Xi_1(b_1, b_2) = \int_{b_2}^{\infty} t^{b_1-1}dt$ and $\Xi_2(b_1, b_2) = \int_0^{b_2} t^{b_1-1}dt$ represent the upper and the lower incomplete Gamma functions for $b_2 > 0, b_1 > 0$. We further use inequalities to bound these gamma functions. For any $b_1, b_2 > 0$, the following hold.

$$\Xi_1(b_1,b_2) \leqslant \Xi_1(b_1) \leqslant \lceil b_1 \rceil ! = \mathcal{O}(b_1^{b_1} e^{-b_1}), \quad \Xi_2(b_1,b_2) \quad \leqslant \frac{b_2^{b_1}}{b_1} \tag{21}$$

Plugging in these upper bounds, and setting $c, c', c'' = \Theta(1)$, we obtain,

$$\mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}[\sigma(b)] \leqslant \mathcal{O}(|\mu|^c + \varsigma^c) \tag{22}$$

The lower bound for the expectation is also of the order of $\Omega(|\mu|^c + |\sigma|^c)$. For the case of $\mu = 0$, the lower bound is exactly of the order of $\Omega(\varsigma^c)$

$$\left| \mathop{\mathbb{E}}_{b\sim\mathcal{N}(0,\varsigma^2)}[\sigma(b)] \right| = \varsigma^c \frac{c'+c''}{2} \mathop{\mathbb{E}}_{b'\sim\mathcal{N}(0,1)}\left[\left|b'\right|^c\right] = \varsigma^c \frac{(|c'+c''|)2^{\frac{c}{2}}}{2\sqrt{\pi}} \Xi_1\left(\frac{c+1}{2}\right) = \Omega(\varsigma^c)$$

Here, $\Xi_1(b_1) = \int_0^{\infty} t^{b_1-1}e^{-t}dt$ is the complete Gamma function.

The only case when $\mathop{\mathbb{E}}_{b\sim\mathcal{N}(0,\sigma)}[\sigma(b)] = 0$ is when σ is an odd-degree monomial, i.e., $c' + c'' = 0$.

For a general mean μ, using Jensen's inequality for the convex function $|b|^c$ with $c \geqslant 1$,

$$\left| \mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}[\sigma(b)] \right| \geqslant \min\{|c'|, |c''|\} \left| \mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}[|b|^c] \right| \geqslant \min\{|c'|, |c''|\} \left| \mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}[b] \right|^c = \Omega(|\mu|^c)$$

Variance of σ

For a 0-mean differentiable function $\widetilde{\sigma}:\mathbb{R}\to\mathbb{R}$, using Cauchy-Schwartz, we have,

$$\Big(\mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}\big[\widetilde{\sigma}(b)(b-\mu)\big]\Big)^2 \leqslant \mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}\big[(b-\mu)^2\big]\cdot\mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}\big[\widetilde{\sigma}^2(b)\big] = \varsigma^2\mathop{\mathrm{Var}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}\big[\widetilde{\sigma}(b)\big]$$

$$\varsigma^4\Big(\mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}\big[\widetilde{\sigma}'(b)\big]\Big)^2 \leqslant \varsigma^2\mathop{\mathrm{Var}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}\big[\widetilde{\sigma}(b)\big]$$

$$\varsigma^4\Big(\mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}\big[\widetilde{\sigma}'(b)\big]\Big)^2 \leqslant \mathop{\mathrm{Var}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}\big[\widetilde{\sigma}(b)\big]$$

We use Stein's identity for the second inequality.

For a differentiable function σ, we can set $\widetilde{\sigma}$ to be $\sigma(b)-\mathop{\mathbb{E}}_{b'\sim\mathcal{N}(\mu,\varsigma}[\sigma(b)]$ and $\sigma^\star(b)-\mathop{\mathbb{E}}_{b'\sim\mathcal{N}(\mu,\varsigma}[\sigma^\star(b)]$. Note that adding or subtracting a constant to σ doesn't change its first derivative.

Using the above inequality, we can lower bound the variance of σ.

$$\mathop{\mathrm{Var}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}\big[\sigma(b)\big] \geqslant \varsigma^2\left|\mathop{\mathbb{E}}_{b\sim\mathcal{N}(\mu,\varsigma^2)}\big[\sigma'(b)\big]\right|^2 = \Omega\Big(\varsigma^2|\mu|^{2c-2}+\varsigma^{2c}\}\Big)$$

Since $\sigma'(b)$ is also a piece-wise polynomial, but of degree $c-1$, we can use the previous bounds on the mean of piece-wise polynomial.

Tails of σ We will now compute a concentration inequality for the tail of $\sigma(b)$ for $b\sim\mathcal{N}(\mu,\varsigma^2)$.

Since σ is a piece-wise polynomial function of degree c, $\sigma(b)$ for $b\sim\mathcal{N}(\mu,\varsigma^2)$ is a $(\frac{c}{2},\mathcal{O}(|\mu|^c+\varsigma^c))$-Sub-Weibull (Vladimirova et al., 2020). Note that the kth moment of (ϱ,K)-Sub-Weibull distribution is upper bounded by $K(k)^\varrho$. As piece-wise polynomial activations are c-degree monomials of gaussian $\mathcal{N}(\mu,\varsigma^2)$, their k^{th} moment is the $(ck)^{th}$ moment of a gaussian which grows as $k^{\frac{c}{2}}$, with the constant $K=\mathcal{O}((|\mu|+\varsigma)^c)=\mathcal{O}(|\mu|^c+|\varsigma|^c)$.

Mean, variance and tails of $\hat{\sigma}$: From Assump. 2, the mean, variance and tails of the unknown link function $\sigma^\star$ match those of the activation function, if $m^\star=1$. This directly corresponds to the single-index link function in Section 2. To extend this notion to Sum of Single-Index link functions in Section 2, we need to find compute the mean, variance and tails of for the random variable $q=\sum_{j=1}^{m^\star}a_j^\star\sigma(b_j)$ for a random vector $\mathbf{b}\sim\mathcal{N}(\mu,\varsigma)$ with $\mu\in\mathbb{R}^{m^\star}$ and $\varsigma\in\mathbb{R}^{m^\star\times m^\star}$ being a Positive Semi-Definite matrx.

Mean.

$$\mathbb{E}[q]=\sum_{j=1}^{m^\star}a_j^\star\mathbb{E}[\sigma(b_j)]$$

As each $b_j\sim\mathcal{N}(\mu_j,\varsigma_{j,j})\ \forall j\in[m^\star]$, using the bounds on mean of σ from previous section, we have,

$$|\mathbb{E}[\sigma(b_j)]|=\Theta(|\mu|^c,(\varsigma_{j,j})^{\frac{c}{2}})$$

Therefore,

$$|\mathbb{E}[q]|=\Theta(\sum_{j=1}^{m^\star}|\mu_j|^c+\sum_{j=1}^{m^\star}(\varsigma_{j,j})^{\frac{c}{2}})$$

As $m^\star=\mathcal{O}(1)$, and all $a_j^\star\neq 0$ and $\left|a_j^\star\right|=\mathcal{O}(1)$, we have the above bound. Additionally, since $1\leqslant m^\star=\mathcal{O}(1)$, by the equivalence of ℓ_p-norms, we have,

$$\sum_{j=1}^{m^\star}|\mu_j|^c=\Theta(\|\mu\|_2^c),\quad \sum_{j=1}^{m^\star}(\varsigma_{j,j})^{\frac{c}{2}}=\|\varsigma\|_F^{\frac{c}{2}}$$

Variance. For variance, we use a similar bound based on Stein's method. Define $\mathbf{q}'\in\mathbb{R}^{m^\star}$ as the vector with its jth coordinate being $\sigma'(b_j)$, $\forall j\in[m^\star]$.

$$\|\mathbb{E}[(q-\mathbb{E}[q])(\mathbf{b}-\mu)]\|_2^2 \leqslant \mathrm{Var}[\mathbf{b}]\mathrm{Var}[q]$$

$$\|\varsigma\mathbf{q}'\|_2^2 \leqslant \mathrm{Tr}(\varsigma)\mathrm{Var}[q] \leqslant \sqrt{m^\star}\|\varsigma\|_F\mathrm{Var}[q]$$

As σ' is also a piece-wise polynomial, and from previous bounds on the mean of piece-wise polynomials, we have,

$$\mathrm{Var}[q] \geqslant \|\varsigma\|_F \Omega(\|\mu\|_2^{c-1} + \|\varsigma\|_F^{\frac{c-1}{2}})$$

Tails. Here, each $\sigma(b_j)$ is a $(\frac{c}{2}, \mathcal{O}(|\mu_j|^c + (\varsigma_{j,j})^{\frac{c}{2}})$-Sub-Weibull random variable. From Def. 14, q is a $(\frac{c}{2}, \mathcal{O}(\max_{j\in[m^\star]}|\mu_j|^c + \max_{j\in[m^\star]}(\varsigma_{j,j})^{\frac{c}{2}}))$-Sub-Weibull random variable. Since $m^\star = \mathcal{O}(1)$, we have,

$$\max_{j\in[m^\star]} |\mu_j|^c = \Theta(\|\mu\|_2^c), \quad \max_{j\in[m^\star]} (\varsigma_{j,j})^{\frac{c}{2}} = \Theta(\|\varsigma\|_F^{\frac{c}{2}})$$

Therefore, q is a $(\frac{c}{2}, \mathcal{O}(\frac{c}{2}, \mathcal{O}(\|\mu\|_2^c + \|\varsigma\|_F^{\frac{c}{2}})$ random variable.

Through the above calculations, we have shown that the sum of activations satisfies Assump. 2.

E.2.2 Proof of Lem. 2

Note that $\mathbf{U}$ is a tall random matrix with iid rows and dependent columns. Let $\bar{\phi} \coloneqq \mathbb{E}[\mathbf{U}_{i,j}]$. If we set $\bar{\mathbf{U}}_{i,j} \coloneqq \mathbf{U}_{i,j} - \bar{\phi}$, then $\mathbb{E}\big[\bar{\mathbf{U}}_{i,j}\big] = 0, \mathbb{E}\big[\bar{\mathbf{U}}^\top \bar{\mathbf{U}}^\top\big] = \bar{\Phi}$. This decomposition can be written as,

$$\mathbf{U} = \bar{\mathbf{U}} + \bar{\phi}\mathbf{1}_n^\top \mathbf{1}_m$$

From Prop. 3, $\big|\bar{\phi}\big| = \Theta(1)$ and each $\mathbf{U}_{i,j}$ is a $(\frac{c}{2}, \mathcal{O}(1))$-Sub-Weibull random variable. Therefore, each $\bar{\mathbf{U}}_{i,j}$ is a 0-mean, $(\frac{c}{2}, \mathcal{O}(1))$-Sub-Weibull random variable. Note that $\mathrm{Tr}(\bar{\Phi})) = \Theta(m)$ as there are m diagonal elements each of constant variance using Assumption 2. Note that $\lambda_{\max}(\bar{\Phi}) \leqslant \mathrm{Tr}(\bar{\Phi})$, therefore, $\lambda_{\max}(\bar{\Phi}) \lesssim m$.

Using Lem. 17 for the whole $\bar{\mathbf{U}}$, with probability $1-\delta$, $\sigma_i(\bar{\mathbf{U}}) \lesssim \sqrt{\lambda_{\max}(\bar{\Phi})}(\sqrt{m} + \sqrt{n})$. Since $m \geqslant n$, $\sigma_n(\bar{\mathbf{U}}) > 0$. For $\mathbf{U}$, note that it has a non-zero mean component $\phi\mathbf{1}_n^\top \mathbf{1}_m^\top$. Since $\big|\bar{\phi}\big| = \Theta(1)$, the largest singular vector of $\mathbf{U}$ is along the all-ones vector. The singular value corresponding to this vector obtained from the mean component is $\sqrt{mn}$.

$$\sigma_{\max}(\mathbf{U}) \leqslant \mathcal{O}(\sigma_{\max}(\bar{\mathbf{U}}) + \sigma_{\max}(\mathbf{1}_n^\top \mathbf{1}_m)) = \mathcal{O}(\sqrt{\lambda_{\max}(\bar{\Phi})}(\sqrt{m} + \sqrt{n}) + \sqrt{mn})$$

Therefore, $\sigma_{\max}(\mathbf{U}) = \mathcal{O}(\sqrt{mn})$.

Note that all the other singular values of $\mathbf{U}$ are obtained from singular values of $\bar{\mathbf{U}}$ as the mean corresponds to a rank-1 matrix.

To show that atleast $\Theta(n)$ eigenvalues of $\mathbf{U}$ are $\lesssim \sqrt{m}$, let $k \in [m]$ be the smallest eigenvalue of $\bar{\Phi}$ that is atleast $\Omega(m^b)$ for some $b > 0$. Then, note that,

$$\Theta(m) = \mathrm{Tr}(\bar{\Phi}) \geqslant \sum_{j=1}^{k} \lambda_i(\bar{\Phi}) \geqslant k m^b.$$

So, $k = \mathcal{O}(m^{1-b})$.

Now, consider the decomposition of matrix $\bar{\Phi}$ into 2 PSD matrices $\bar{\Phi} = \bar{\Phi}_k + \bar{\Phi}_{k^c}$, where $\bar{\Phi}_k$ is of rank k and contains the top-k eigenvalues and $\bar{\Phi}_{k^c}$ contains all the remaining eigenvalues that are atmost $\mathcal{O}(1)$. Further, consider the decomposition of $\bar{\mathbf{U}}$ into $\bar{\mathbf{U}}_k$ and $\bar{\mathbf{U}}_{k^c}$ according to these eigenvalues. Applying Lemma 17 to $\bar{\mathbf{U}}_{k^c}$, we have $\sigma_{\max}(\bar{\mathbf{U}}_{k^c}) \lesssim \sqrt{\lambda_{\max}(\bar{\Phi}_{k^c})}(\sqrt{m} + \sqrt{n}) \lesssim \sqrt{m^{1+b}}$. Since these correspond to the $n-k$ singular values of $\mathbf{U}$, the last $n-k$ singular values of $\mathbf{U}$ are $\lesssim \sqrt{m^{1+b}}$.

Under the additional conditions of $\lambda_{\min}(\bar{\Phi}) = \Omega(1)$ and $\sigma_{\max}(\Theta) = \mathcal{O}(1)$, we can use Lemma 17 again, to show that,

$$\sigma_n(\mathbf{U}) \gtrsim \sqrt{m}.$$

This completes the proof.

E.3 Proof of Lem. 4

First, consider the expression for $B(\theta)$ for any $\theta \in \mathbb{S}^{d-1}$,

$$B(\theta) = 2B_1^{\frac{1}{c+1}}(\theta) B_2^{\frac{c}{c+1}}(\theta)$$
$$B_1(\theta) = \frac{1}{n}\sum_{i=1}^{n} \sigma^2(\langle \theta, \mathbf{x}_i \rangle), \quad B_2(\theta) = \frac{1}{n}\sum_{i=1}^{n} (\sigma')^2(\langle \theta, \mathbf{x}_i \rangle) \|\mathbf{x}_i\|_2^2$$

Bound on B_1 Note that $\sigma^2(\langle \theta, \mathbf{x}_i \rangle)$ is $(c, \mathcal{O}(1))$-Sub-Weibull(Def. 3) as $\langle \theta, \mathbf{x}_i \rangle \sim \mathcal{N}(0,1)$. Further, $\mathbb{E}\left[\sigma^2(\langle \theta, \mathbf{x}_i \rangle\right] = \phi^\star = \Theta(1)$ (App. E.2.1). Therefore, the mean of n independent Sub-Weibull random variables can be bounded by using concentration. For a fixed $\theta \in \mathbb{S}^{d-1}$, using Lem. 15, with probability $1-\delta$,

$$|B_1(\theta) - \phi^\star| \lesssim \sqrt{\frac{\log(2/\delta)}{n}} \tag{23}$$

Bound on B_2

$$B_2(\theta) = \frac{1}{n}(\sigma')^2(\langle \theta, \mathbf{x}_i \rangle)\|\mathbf{x}_i\|_2^2 = \frac{1}{n}(\sigma')^2(\langle \theta, \mathbf{x}_i \rangle)\left(\langle \theta, \mathbf{x}_i \rangle^2 + \left\|(\mathbb{I}_d - \theta\theta^\top)\mathbf{x}_i\right\|_2^2\right)$$
$$= \underbrace{\frac{1}{n}(\sigma')^2(\langle \theta, \mathbf{x}_i \rangle)\langle \theta, \mathbf{x}_i \rangle^2}_{B_{2,1}(\theta)} + \underbrace{\frac{1}{n}(\sigma')^2(\langle \theta, \mathbf{x}_i \rangle)\left\|(\mathbb{I}_d - \theta\theta^\top)\mathbf{x}_i\right\|_2^2}_{B_{2,2}(\theta)}$$

We bound the two terms separately. Note that we have decomposed $\mathbf{x}_i$ in the direction along θ and the direction perpendicular to θ.

For the first term, since σ' is a $(c-1)$-degree piecewise-polynomial function of $\langle \theta, \mathbf{x}_i \rangle$, $(\sigma')^2(\langle \theta, \mathbf{x}_i \rangle)\langle \theta, \mathbf{x}_i \rangle^2$ is a $2c$-degree piece-wise polynomial function of $\langle \theta, \mathbf{x}_i \rangle$. This random variable has mean $\mu_{B_{2,1}} = \Theta(1)$(App. E.2.1), and is $(c, \mathcal{O}(1))$-Sub-Weibull (Def. 3). For a fixed $\theta \in \mathbb{S}^{d-1}$, using Lem. 15, with probability $1-\delta$,

$$\left|B_{2,1}(\theta) - \mu_{B_{2,1}}\right| \lesssim \sqrt{\frac{\log(2/\delta)}{n}}$$

For the second term, note that $(\mathbb{I}_d - \theta\theta^T)\mathbf{x} \sim \mathcal{N}(0, \mathbb{I}_d - \theta\theta^\top)$ and this random variable is independent of $\langle \theta, \mathbf{x} \rangle$. Note that the norm of this variable is Sub-Exponential with mean $d-1$, and Sub-Exponential parameter proportional to d. Note that this Sub-Exponential variable is $(1,d)$-Sub-Weibull(Def. 3). Since $(\sigma')^2$ itself is $(c-1, \mathcal{O}(1))$-Sub-Weibull (Def. 3), and independent of this Sub-Exponential variable, the product of these two variables is $(c, \mathcal{O}(d))$-Sub-Weibull (Lem. 14). Let $\mathop{\mathbb{E}}_{b \sim \mathcal{N}(0,1)}\left[(\sigma')^2(b)\right] \coloneqq \mu_{B_{2,2}} = \Theta(1)$. Using Lem. 15, with probability $1-\delta$, for a fixed $\theta \in \mathbb{S}^{d-1}$, we have,

$$\left|B_{2,2}(\theta) - (d-1)\mu_{B_{2,1}}\right| \lesssim (d-1)\sqrt{\frac{\log(2/\delta)}{n}}$$

Lower Bound on $\min_{j\in[m]} B(\theta_j)$ We use the bounds for B_1 and B_2 for each $\theta = \theta_j \ \forall j \in [m]$, and take a union bound over the probability of error. We set the probability of error for each θ_j to be $\frac{\delta}{m}$ instead of δ. Therefore, with probability $1-3\delta$,

$$\min_{j\in[m]} B(\theta_j) \gtrsim \left(\phi^\star - \frac{\sqrt{\log(m)}}{\sqrt{n}}\right)^{\frac{1}{c+1}} \left(\mu_{B_{2,1}} + (d-1)\mu_{B_{2,2}} - \frac{d\sqrt{\log(m)}}{\sqrt{n}}\right)^{\frac{c}{c+1}} \gtrsim d^{\frac{c}{c+1}}$$

Upper Bound on $\max_{j\in[m]} B(\theta_j)$ Using the same union bound as before, with probability $1-3\delta$,

$$\max_{j\in[m]} B(\theta_j) \lesssim \left(\phi^\star + \frac{\sqrt{\log(m)}}{\sqrt{n}}\right)^{\frac{1}{c+1}} \left(\mu_{B_{2,1}} + (d-1)\mu_{B_{2,2}} + \frac{d\sqrt{\log(m)}}{\sqrt{n}}\right)^{\frac{c}{c+1}} \lesssim d^{\frac{c}{c+1}}$$

E.4 Proofs for App. D.3

E.4.1 Proof of Prop. 4

From App. C, the population loss for any weight vector $\mathbf{w}$ is given by,

$$2F(\mathbf{w}) = \zeta^2 + \psi^\star + \mathbf{a}^\top \Phi \mathbf{a} - 2\left\langle \widetilde{\Psi}, \mathbf{a} \right\rangle.$$

Note that this is a quadratic in $\mathbf{a}$. If $\Phi > 0$, this quadratic is minimized in $\mathbf{a}$ at $\bar{\mathbf{a}} = \Phi^{-1}\widetilde{\Psi}$. Therefore, a lower bound on the population loss independent of $\mathbf{a}$ is given by,

$$2F(\mathbf{w}) \geqslant \zeta^2 + \psi^\star - \widetilde{\Psi}\Phi^{-1}\widetilde{\Psi} \geqslant \zeta^2 \geqslant \zeta^2 + \psi^\star - \lambda_{\min}^{-1}(\Phi)\left\|\widetilde{\Psi}\right\|_2^2 \geqslant \zeta^2 + \psi^\star - m\lambda_{\min}^{-1}(\Phi)\max_{j\in[m]} \widetilde{\psi}^2(\Theta^\star\theta_j)$$

Each coordinate of $\widetilde{\Psi}$ is $\widetilde{\psi}(\Theta^\star\theta_j)$. If $\widetilde{\psi}$ is coordinate-wise an increasing function, and each coordinate of the vector $\Theta^\star\theta_j$ lies in $[-\rho^\star, \rho^\star]$, $\widetilde{\psi}(\Theta^\star\theta_j) \leqslant \widetilde{\psi}(\rho^\star \mathbf{1}_{m^\star}), \forall j \in [m]$. Plugging in the sufficient condition from Prop. 4 completes the proof.

E.4.2 Proof of Example 1

If we select $\theta_j \perp \theta_{j'}^\star, \forall j \in [m], j' \in [m^\star]$, then $\langle \theta_j, \mathbf{x}\rangle$ is independent of $\langle \theta_{j'}, \mathbf{x}\rangle, \forall j \in [m], j' \in [m^\star]$. This implies that $h(\mathbf{w},\mathbf{x})$ is independent of $\sigma^\star(\Theta^\star\mathbf{x})$. Therefore, the population loss for such an interpolator is given by the following.

$$\begin{aligned} 2F(\mathbf{w}) &= \zeta^2 + \mathbb{E}\left[(y - h(\mathbf{w},\mathbf{x}))^2\right] = \zeta^2 + \psi^\star + \mathbb{E}\left[h^2(\mathbf{w},\mathbf{x})\right] - 2\mathbb{E}[h(\mathbf{w},\mathbf{x})]\mathbb{E}[\sigma^\star(\Theta^\star\mathbf{x})] \\ &= \zeta^2 + \psi^\star + \mathbb{E}\left[h^2(\mathbf{w},\mathbf{x})\right] - (\mathbb{E}[h(\mathbf{w},\mathbf{x})])^2 - (\mathbb{E}[\sigma^\star(\Theta^\star\mathbf{x})])^2 + (\mathbb{E}[h(\mathbf{w},\mathbf{x})] - \mathbb{E}[\sigma^\star(\Theta^\star\mathbf{x})])^2 \end{aligned}$$

The second equation is obtained by completing the squares while the first equation is obtained by independence of $h(\mathbf{w},\mathbf{x})$ and $\sigma^\star(\Theta^\star\mathbf{x})$. Note that $\mathrm{Var}[h(\mathbf{w},\mathbf{x})] = \mathbb{E}\left[h^2(\mathbf{w},\mathbf{x})\right] - (\mathbb{E}[h(\mathbf{w},\mathbf{x}])^2 \geqslant 0$. Additionally, the term inside the square is non-negative. This provides the required lower bound on population risk.

$$2F(\mathbf{w}) = \zeta^2 + \psi^\star - (\mathbb{E}[\sigma^\star(\Theta^\star\mathbf{x})])^2 = \zeta^2 + \left(1 - \frac{(\mathbb{E}[\sigma^\star(\Theta^\star(\mathbf{x}))])^2}{\psi^\star}\right)\psi^\star$$

If we set $\kappa \coloneqq \frac{1}{2}\left(1 - \frac{(\mathbb{E}[\sigma^\star(\Theta^\star\mathbf{x})])^2}{\psi^\star}\right)$, we get the required result. Note that $\Theta^\star\mathbf{x} \sim \mathcal{N}(0, \mathbb{I}_{m^\star})$.

To show that this interpolator satisfies Condition 3 of Def. 2, we will first replace θ_j by $\tilde{\theta}_{j'}$ which is sampled iid from the gaussian instead of unit sphere. This gives us a matrix $\tilde{\Theta} \in \mathbb{R}^{m\times d}$. If we multiply this matrix by the diagonal matrix of entries $\left\|\tilde{\theta}_j\right\|_2^{-1}$, then we get our original matrix $\Theta \in \mathbb{R}^{m\times d}$, as sampling vectors from the unit sphere is same as sampling vectors from a gaussian and then normalizing them.

Now, for a gaussian matrix, is largest singular value is given by, $\sigma_{\max}(\tilde{\Theta}) \lesssim \sqrt{m} + \sqrt{d}$ with high probability using (Wainwright, 2019, Chapter 5). Further, each $\left\|\tilde{\theta}_j\right\|_2^2 \asymp d$ with high probability using χ^2 concentration and maximal inequality. Therefore, $\sigma_{\max}(\Theta) \leqslant \frac{\sigma_{\max}(\tilde{\Theta}_j)}{\min_{j\in[m]}\left\|\tilde{\theta}_j\right\|_2} \lesssim \sqrt{\frac{m}{d}} + 1$. Since $m = \Theta(d)$, we have $\sigma_{\max}(\Theta) = \mathcal{O}(1)$.

For the lower bound, note that,

$$\bar{\Phi} \geq \tilde{c}_2 \mathfrak{F}(\Theta)$$

where $\tilde{c}_2$ is the second hermite coefficient of the activation, and $(j,j)^{th}$ element of $\mathfrak{F}(\theta)$ is $\langle \theta_j, \theta_{j'}\rangle^2$.

For piece-wise polynomials in Assump. 1, this is a constant. We need to show minimum eigenvalue of the second order tensor product is atleast constant.

We can replace Θ by $\tilde{\Theta}$. Note that,

$$\lambda_{\min}(\mathfrak{F}(\Theta)) \gtrsim \frac{1}{d^2}\lambda_{\min}(\mathfrak{F}(\tilde{\Theta})).$$

Using (El Karoui, 2010, Theorem 1) to the matrix $\mathfrak{F}(\frac{\tilde{\Theta}}{\sqrt{d}})$, we obtain,

$$\lambda_{\min}\left(\mathfrak{F}(\frac{\tilde{\Theta}}{\sqrt{d}})\right)=\Omega(1).$$

Therefore, $\lambda_{\min}(\bar{\Phi})=\Omega(1)$.

E.4.3 Proof of Lem. 8

Note that, $\mathbf{y}=\mathbf{y}^\star+\mathbf{n}$. By triangle inequality for ℓ_2 norm, for any two vectors $\mathbf{b}_1,\mathbf{b}_2\in\mathbb{R}^m$, $\|\mathbf{b}_1\|_2^2\geqslant\frac{1}{2}\|\mathbf{b}_1-\mathbf{b}\|_2^2-\|\mathbf{b}_2\|_2^2$.

Therefore,

$$\|\mathbf{a}_{\min,\ell_2}\|_2^2\geqslant\frac{1}{2}\left\|\widetilde{\mathbf{U}}^{\frac{1}{2}}\mathbf{y}^\star\right\|_2^2-\left\|\widetilde{\mathbf{U}}^{\frac{1}{2}}\mathbf{n}\right\|_2^2$$

We can bound the term of noise separately. Note that,

$$\left\|\widetilde{\mathbf{U}}^{\frac{1}{2}}\mathbf{n}\right\|_2^2\leqslant\lambda_{\max}(\widetilde{\mathbf{U}})\|\mathbf{n}\|_2^2=\sigma_{\min}^{-2}(\mathbf{U})\|\mathbf{n}\|_2^2$$

From Lem. 2, with probability $1-\delta$, $\sigma_{\min}(\mathbf{U})\gtrsim\frac{1}{\sqrt{m}}$. Further, note that $\zeta^{-2}\|\mathbf{n}\|_2^2$ is a χ_n^2-random variable. By χ^2 concentration from (Laurent & Massart, 2000, Lem. 1), with probability $1-\delta$,

$$\|\mathbf{n}\|_2^2\leqslant\zeta^2(n+2\sqrt{n\log(1/\delta)}+2\log(1/\delta))\lesssim\zeta^2 n$$

After eliminating the contribution of noise from $\|\mathbf{a}_{\min,\ell_2}\|_2^2$, with probability $1-2\delta$, we obtain,

$$\|\mathbf{a}_{\min,\ell_2}\|_2^2\gtrsim\frac{1}{2}\left\|\widetilde{\mathbf{U}}^{\frac{1}{2}}\mathbf{y}^\star\right\|_2^2-\zeta^2\frac{n}{m}$$

This completes the proof.

E.4.4 Proof of Lem. 9

Note that the matrix $\mathbf{U}$ is obtained by applying a deterministic function $\sigma(\cdot)$ on the matrix $\hat{\mathbf{U}}$ element-wise. Since $\widetilde{\mathbf{U}}$ is obtained from $\mathbf{U}$, it is also a deterministic function of the matrix $\hat{\mathbf{U}}\in\mathbb{R}^{n\times m}$. Let $\hat{\mathbf{U}}_i=\Theta\mathbf{x}_i\in\mathbb{R}^n$ denote the ith row of $\hat{\mathbf{U}}$. To compute the distribution of $\mathbf{y}_i^\star|v\hat{U}_i$, it is enough for us to compute the distribution of $\mathbf{y}^\star|\hat{\mathbf{U}}_i$. Since $\mathbf{y}_i^\star=\sigma^\star(\Theta^\star\mathbf{x}_i)$, we first need to specify the distribution of $\Theta^\star\mathbf{x}_i|\hat{\mathbf{U}}_i$.

Distribution of $\Theta^\star\mathbf{x}_1|\hat{\mathbf{U}}_1$ Note that,

$$\begin{bmatrix}\Theta^\star\mathbf{x}_1\\ \hat{\mathbf{U}}_1\end{bmatrix}\sim\mathcal{N}\left(\mathbf{0}_{m+m^\star},\begin{bmatrix}\mathbb{I}_{m^\star} & \Theta^\star\Theta^\top\\ \Theta(\Theta^\star)^\top & \Theta\Theta^\top\end{bmatrix}\right).$$

Therefore, $\Theta^\star\mathbf{x}_1|\hat{\mathbf{U}}_1$ is also a gaussian distribution with mean and variance determined by the Schur's complement of the joint distribution. Note that here $\Theta\Theta^\top$ is not invertible as its rank is atmost d, while it is a $m\times m$ matrix. Therefore, we compute its pseudo-inverse in the Schur's complement.

$$\Theta^\star\mathbf{x}_1|\hat{\mathbf{U}}_1\sim\mathcal{N}(\Theta^\star\Theta^\top(\Theta\Theta^\top)^\dagger\hat{\mathbf{U}}_1,\mathbb{I}_{m^\star}-\Theta^\star\Theta^\top(\Theta\Theta^\top)^\dagger\Theta(\Theta^\star)^\top)$$

Here, $A^\dagger$ is the pseudo-inverse of A for a square matrix with real entries A. In this case, $\mathbf{y}_i^\star|\mathbf{U}=\sigma^\star(\Theta^\star\mathbf{x}_i)|\hat{\mathbf{U}}$ where $\Theta^\star\mathbf{x}_i|\hat{\mathbf{U}}\sim\mathcal{N}(\widetilde{\mu}_i,\varsigma)$ with $\varsigma=\mathbb{I}_{m^\star}-\Theta^\star\Theta^\top(\Theta\Theta^\top)^\dagger\Theta(\Theta^\star)^\top\in\mathbb{R}^{m^\star\times m^\star}$ and $\widetilde{\mu}_i=\Theta^\star\Theta^\top(\Theta\Theta^\top)^\dagger\hat{\mathbf{U}}_i,\forall i\in[n]$. Therefore, $\mathbf{y}_i^\star|\hat{\mathbf{U}}$ is a $(\frac{c}{2},\mathcal{O}(\|\widetilde{\mu}_i\|_2^c+\|\varsigma\|_F^{\frac{c}{2}}))$-Sub-Weibull random variable, using Assump. 2. The other conditions on mean and variance are obtained by plugging in the corresponding expressions from Assump. 2. This completes the proof.

E.4.5 PROOF OF LEM. 10

Lower Bound on $(\mathbf{y}^\star_{\hat{\mathbf{U}}})^\top \widetilde{\mathbf{U}} \mathbf{y}^\star_{\hat{\mathbf{U}}}$

$$(\mathbf{y}^\star_{\hat{\mathbf{U}}})^\top \widetilde{\mathbf{U}} \mathbf{y}^\star_{\hat{\mathbf{U}}} \geqslant \left\|\mathbf{y}^\star_{\hat{\mathbf{U}}}\right\|_2^2 \lambda_{\min}(\widetilde{\mathbf{U}}) \gtrsim (\mathrm{Tr}((\mathbf{K}')^2) + n\|\varsigma\|_F^c)\sigma^{-2}_{\max}(\mathbf{U})$$

We use the fact that $\left\|\mathbf{y}^\star_{\hat{\mathbf{U}}}\right\|^2 = \sum_{i=1}^n \|\widetilde{\mu}_i\|_2^{2c} = \mathrm{Tr}((\mathbf{K}')^2)$.

Lower Bound on $\sum_{i\in[n]} \widetilde{\mathbf{U}}_{i,i} \mathbf{Var}\left[\mathbf{y}^\star_i | \hat{\mathbf{U}}_i\right]$

$$\begin{aligned}\sum_{i\in[n]} \widetilde{\mathbf{U}}_{i,i} \mathrm{Var}\left[\mathbf{y}^\star_i | \hat{\mathbf{U}}\right] &\gtrsim \|\varsigma\|_F \mathrm{Tr}((\mathbf{K}')^{\frac{2(c-1)}{c}} \widetilde{\mathbf{U}}) + \|\varsigma\|_F^c \mathrm{Tr}(\widetilde{\mathbf{U}}) \\ &\gtrsim \|\varsigma\|_F (\mathrm{Tr}((\mathbf{K}')^{\frac{2(c-1)}{c}}) \lambda_{\min}(\widetilde{\mathbf{U}}) + \|\varsigma\|_F^c \mathrm{Tr}(\widetilde{\mathbf{U}})) \\ &\gtrsim \|\varsigma\|_F (\mathrm{Tr}((\mathbf{K}')^{\frac{2(c-1)}{c}}) \sigma^{-2}_{\max}(\mathbf{U}) + \|\varsigma\|_F^c \mathrm{Tr}(\widetilde{\mathbf{U}}))\end{aligned}$$

Upper Bound on $\left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{K}\right\|_F$ We use triangle inequality for the norm,

$$\left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{K}\right\|_F \lesssim \left\|\mathbf{K}'\widetilde{\mathbf{U}}\mathbf{K}'\right\|_F + \|\varsigma\|_F^c \left\|\widetilde{\mathbf{U}}\right\|_F$$

For the first term,

$$\begin{aligned}\left\|\mathbf{K}'\widetilde{\mathbf{U}}\mathbf{K}'\right\|_F &= \sqrt{\mathrm{Tr}(\mathbf{K}'\widetilde{\mathbf{U}}(\mathbf{K}')^2\widetilde{\mathbf{U}}\mathbf{K}')} = \sqrt{\mathrm{Tr}((\mathbf{K}')^4\widetilde{\mathbf{U}}^2)} \\ &\leqslant \sqrt{\mathrm{Tr}((\mathbf{K}')^4)}\sqrt{\mathrm{Tr}(\widetilde{\mathbf{U}}^2)} \leqslant \sqrt{\mathrm{Tr}((\mathbf{K}')^4)}\left\|\widetilde{\mathbf{U}}\right\|_F\end{aligned}$$

We use Cauchy-Schwartz for Positive Semi-Definite matrices $\mathbf{K}'$ and $\widetilde{\mathbf{U}}$ for the inequality and commute the terms inside the trace. Finally, we use an upper bound on $\left\|\widetilde{\mathbf{U}}\right\|_F$.

$$\left\|\widetilde{\mathbf{U}}\right\|_F = \sqrt{\mathrm{Tr}(\widetilde{\mathbf{U}}^\top\widetilde{\mathbf{U}})} = \sqrt{\mathrm{Tr}(\widetilde{\mathbf{U}}^2)} = \sqrt{\sum_{i=1}^n \lambda_i^2(\widetilde{\mathbf{U}})} = \sqrt{\sum_{i=1}^n \sigma_i^{-4}(\mathbf{U})}$$

We use the definition of frobenius norm in the first step. Then, we use the fact that $\widetilde{\mathbf{U}}$ is symmetric. We then use the connection between trace and eigenvalues of a PD symmetric matrix $\widetilde{\mathbf{U}}$ and the fact that $\lambda_i(\widetilde{\mathbf{U}}) = \sigma_i^{-2}(\mathbf{U})$.

Upper Bound on $\left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{y}_{\hat{\mathbf{U}}^\star}\right\|_2$ Note that, using Cauchy-Schwarz, we have,

$$\left\|\mathbf{K}\widetilde{\mathbf{U}}\mathbf{y}_{\hat{\mathbf{U}}^\star}\right\|_2 \lesssim \left\|\mathbf{K}'\widetilde{\mathbf{U}}\mathbf{y}_{\hat{\mathbf{U}}^\star}\right\|_2 + \|\varsigma\|_F^c \left\|\widetilde{\mathbf{U}}\mathbf{y}^\star_{\hat{\mathbf{U}}}\right\|_2 \leqslant \lambda_{\max}(\widetilde{\mathbf{U}})\left(\mathrm{Tr}((\mathbf{K}')^2) + \sqrt{\mathrm{Tr}((\mathbf{K}')^2)}\|\varsigma\|_F^c\right).$$

E.4.6 PROOF OF LEM. 11

From Lem. 9, $\|\widetilde{\mu}_i\|_2^2$ is a sum of gaussian random variables, hence it is subexponetial and in turn $(1, \mathcal{O}(\|\mathbb{I}_{m^\star} - \varsigma\|_F)$-Sub-Weibull. Raising a Sub-Weibull to the exponent $\frac{\varphi}{2}$ for some $\varphi > 0$, we obtain another Sub-Weibull random variable. Therefore, $\|\widetilde{\mu}_i\|_2^\varphi$ is also a $(\frac{\varphi}{2}, \mathcal{O}(\|\mathbb{I}_{m^\star} - \varsigma\|_F^{\frac{\varphi}{2}})$-Sub-Weibull random variable (Def. 3). In addition to this, the Sub-Weibull random variable is a polynomial of sum of sub-Weibull random variables, hence its mean follows the same properties as that of the link function in App. E.2.1, with its mean bounded by $\mathcal{O}(\|\mathbb{I}_{m^\star} - \varsigma\|_F^{\frac{\varphi}{2}})$.

Since we want to bound the sum of these iid random variables, we can use concentration inequalties for sum of Sub-Weibull random variables. Here, $\mathrm{Tr}((\mathbf{K}')^\varphi)$ for any $\varphi > 0$ can be represented as

$$\mathrm{Tr}((\mathbf{K}')^\varphi) = \sum_{i=1}^n \|\widetilde{\mu}_i\|_2^{c\varphi}$$

Therefore, we have a sum of $\left(\frac{c\varphi}{2}, \|\mathbb{I}_{m^\star} - \varsigma\|_F^{\frac{c\varphi}{2}}\right)$-Sub-Weibull independent random variables. This sum is a $\left(\frac{c\varphi}{2}, n\|\mathbb{I}_{m^\star} - \varsigma\|_F^{\frac{c\varphi}{2}}\right)$-Sub-Weibull random variable by triangle inequality of the Orlicz norm, as these are independent random variables. Using Lem. 15, with probability $1-\delta$,

$$\left|\mathrm{Tr}((\mathbf{K}')^{\varphi}) - n\mathbb{E}[\|\widetilde{\mu}_1\|_2^{c\varphi}]\right| \lesssim n\|\mathbb{I}_{m^\star} - \varsigma\|_F^{\frac{c\varphi}{2}} \left(\log(2/\delta)\right)^{\frac{c\varphi}{2}}$$
$$\mathrm{Tr}((\mathbf{K}')^{\varphi}) \asymp n\|\mathbb{I}_{m^\star} - \varsigma\|_F^{\frac{c\varphi}{2}} \asymp n\|\mathbb{I}_{m^\star} - \varsigma\|^{\frac{c\varphi}{2}}.$$

For $\varphi = 0$, $\mathrm{Tr}((\mathbf{K}')^0) = \mathrm{Tr}(\mathbb{I}_n) = n$.

E.5 Proofs for App. D.4

E.5.1 Proof of Thm. 4

For **Single-Index** models defined in Section 2, we consider a good interpolator given by $a_j = \frac{1}{m}, \theta_j = \theta^\star, \forall j \in [m]$. For this interpolator, $\forall \mathbf{x} \in \mathbb{R}^d$, we have,

$$h(\mathbf{w}, \mathbf{x}) = \sum_{j=1}^{m} \frac{1}{m} \sigma(\langle \theta^\star, \mathbf{x} \rangle) = y.$$

Therefore, this is indeed an interpolator, and it is good, as $F_S(\mathbf{w}) = F(\mathbf{w}) = 0$. The flatness of this good interpolator is obtained by plugging in the value of $\mathbf{a}$ into Lem. 1.

$$\|\mathbf{a}\|_{\frac{2c}{c+1}}^{\frac{2c}{c+1}} = \sum_{j=1}^{m} m^{-\frac{2c}{c+1}} = m^{-\frac{c-1}{c+1}}.$$

Therefore, $\Upsilon(\mathbf{w}) \asymp d^{\frac{c}{c+1}} m^{-\frac{c-1}{c+1}} \asymp \Upsilon^\star$.

Similarly for the **Sum of Single-Index** case in Section 2, for the given good interpolator., $\forall \mathbf{x} \in \mathbb{R}^d$, we have,

$$h(\mathbf{w}, \mathbf{x}) = \frac{m}{m^\star} \sum_{j \in [m^\star]} \frac{a_j^\star m^\star}{m} \sigma(\langle \theta_j^\star, \mathbf{x} \rangle) = \sum_{j \in [m^\star]} a_j^\star \sigma(\langle \theta_j^\star, \mathbf{x} \rangle) = y.$$

Therefore, $F_S(\mathbf{w}) = F(\mathbf{w}) = 0$ and it is indeed a good interpolator. From Lem. 1, the flatness can be computed from $\mathbf{a}$ as,

$$\|\mathbf{a}\|_{\frac{2c}{c+1}}^{\frac{2c}{c+1}} = \left(\frac{m}{m^\star}\right)^{-\frac{c-1}{c+1}} \sum_{j \in [m^\star]} \left|a_j^\star\right|^{\frac{2c}{c+1}} \lesssim \left(\frac{m}{m^\star}\right)^{-\frac{c-1}{c+1}} m^\star \left(\max_{j \in [m^\star]} \left|a_j^\star\right|\right)^{\frac{2c}{c+1}} \lesssim m^{-\frac{c-1}{c+1}}$$

We use the fact that $m^\star = \mathcal{O}(1)$, $\left|a_j^\star\right| = \mathcal{O}(1)$, $\forall j \in [m^\star]$. This bound implies that $\Upsilon(\mathbf{w}) \asymp d^{\frac{c}{c+1}} m^{-\frac{c-1}{c+1}} = \Upsilon^\star$.

E.5.2 Proof of Prop. 2

For the necessary condition, from Lem. 1, with high probability for a single interpolator $\mathbf{w}$,

$$\Upsilon(\mathbf{w}) \asymp \Upsilon^\star \Longrightarrow \|\mathbf{a}\|_{\frac{2c}{c+1}}^{\frac{2c}{c+1}} \asymp m^{-\frac{c-1}{c+1}}.$$

Using Holder's inequality for ℓ_p norms, we have,

$$\|\mathbf{a}\|_1 \leqslant m^{1-\frac{c+1}{2c}} \|\mathbf{a}\|_{\frac{2c}{c+1}} \lesssim m^{\frac{c-1}{2c}} \|\mathbf{a}\|_{\frac{2c}{c+1}} \lesssim m^{\frac{c-1}{2c}} m^{-\frac{c-1}{c+1} \cdot \frac{c+1}{2c}} = \mathcal{O}(1).$$

E.6 Proof of Lem. 13

This proof heavily builds on that of (Dirksen, 2015, Theorem 3.5) and applies the generic chaining proof techniques from (Wainwright, 2019, Section 5.3.3). We first consider the ℓ_2 ball given by $\|\theta\|_2 \leqslant 1$. As the unit sphere is a subset of this, $\sup_{\|\theta\|_2 = 1} \mathfrak{G}(\theta) \leqslant \sup_{\|\theta\|_2 \leqslant 1} \mathfrak{G}(\theta)$.

Let $\mathcal{C}_{2^{-k}}$ be a 2^{-k} cover of the unit ball in d-dimensions for $k \geqslant 1$. Then, size of this cover is $|\mathcal{C}_{2^{-k}}| \lesssim \exp(C'd(k+2))$. for some constant $C' > 0$.

Now, fix a θ on the unit ball. Let $\theta_k = \operatorname{argmin}_{\theta' \in \mathcal{C}_{2^{-k}}} \|\theta - \theta'\|_2$, be its closest point on the kth-level cover.

We will write $\mathfrak{G}(\theta)$ as,

$$\mathfrak{G}(\theta) = \sum_{k=1}^{\infty} \mathfrak{G}(\theta_k) - \mathfrak{G}(\theta_{k-1}).$$

Note that by convention, we set $\theta_0 = 0$, the 0 vector, and $\lim_{k\to\infty} \theta_k = \theta$. Then, by definition of the covers,

$$\|\theta_k - \theta_{k-1}\| \leqslant 3 \cdot 2^{-k}.$$

Now, we will apply Eq (17) at each level separately. Apply Eq (17) for $\theta_1 = \theta_k$ and $\theta_2 = \theta_{k-1}$. Note that there are $|\mathcal{C}_{2^{-k}}||\mathcal{C}_{2^{-(k-1)}}| \lesssim \exp(C'2dk)$ such pairs, so we need to take a union bound over all of them. Applying Eq (17) for at least one pair with $t = t_k$ at level k, we obtain, $\forall k \geqslant 1$

$$\mathbb{P}\left[\forall k \geqslant 1, |\mathfrak{G}(\theta_k) - \mathfrak{G}(\theta_{k-1})| \geqslant 3C2^{-k}\left(\sqrt{\frac{t_k}{n}} + \frac{t_k^{\frac{c}{2}}}{n^b}\right)\right] \leqslant \sum_{k=1}^{\infty} \exp(-t_k + 2C'dk).$$

We set $t_k = t + 2C'dk + 2k$. Then, the union bound becomes,

$$\sum_{k=1}^{\infty} \exp(-t_k + 2C'dk) = \exp(-t) \sum_{k=1}^{\infty} \exp(-2k) \leqslant \exp(-t) \frac{1}{\exp(2) - 1} \leqslant \exp(-t).$$

To compute a lower bound on $\mathfrak{G}(\theta)$, note that,

$$\mathbb{P}\left[\mathfrak{G}(\theta) \geqslant \sum_{k=1}^{\infty} 3C2^{-k}\left(\sqrt{\frac{t_k}{n}} + \frac{t_k^{\frac{c}{2}}}{n^b}\right)\right] \leqslant \mathbb{P}\left[\forall k \geqslant 1, |\mathfrak{G}(\theta_k) - \mathfrak{G}(\theta_{k-1})| \geqslant 3C2^{-k}\left(\sqrt{\frac{t_k}{n}} + \frac{t_k^{\frac{c}{2}}}{n^b}\right)\right].$$

We now compute the summation that bounds $\mathfrak{G}(\theta)$.

$$\sum_{k=1}^{\infty} 3C2^{-k}\left(\sqrt{\frac{t_k}{n}} + \frac{t_k^{\frac{c}{2}}}{n^b}\right) \lesssim \sum_{k=1}^{\infty} 2^{-k}\left(\sqrt{\frac{dk+t}{n}} + \frac{(dk+t)^{\frac{c}{2}}}{n^b}\right) \lesssim \sum_{k=1}^{\infty} 2^{-k}\left(\sqrt{\frac{d}{n}}\sqrt{k} + \sqrt{\frac{t}{n}} + \frac{d^{\frac{c}{2}}}{n^b}k^{\frac{c}{2}} + \frac{t^{\frac{c}{2}}}{n^b}\right).$$

We decompose $\sqrt{\cdot}$ and the $\frac{c}{2}$-th powers in terms of their components. The additional constants depend on c so they are subsumed inside the $\lesssim$ notation. Note that the coefficients of the terms of t in the RHS above are $\sum_{k=1}^{\infty} 2^{-k} \leqslant 2$. For the other terms, note that the series $\sum_{k=1}^{\infty} 2^{-k}\sqrt{k}$ and $\sum_{k=1}^{\infty} 2^{-k}k^{\frac{c}{2}}$ converge to constants, as the polynomial growth of $\sqrt{k}$ or $k^{\frac{c}{2}}$ is overpowered by the exponential decay of 2^{-k}. Therefore, the above term is bounded by,

$$\sqrt{\frac{d}{n}} + \sqrt{\frac{t}{n}} + \frac{d^{\frac{c}{2}}}{n^b} + \frac{t^{\frac{c}{2}}}{n^b}.$$

Using $\mathfrak{G}(0) = 0$, and repeating the same steps for $-G(\theta)$, we complete the proof.