How Structured Data Guides Feature Learning: A Case Study of the Parity Problem

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent works have shown that neural networks optimized by gradient-based methods can adapt to sparse or low-dimensional target functions through feature learning; an often studied target is classification of the sparse parity function on the unit hypercube. However, such isotropic data setting does not capture the anisotropy and low intrinsic dimensionality exhibited in realistic datasets. In this work, we address this shortcoming by studying how feature learning interacts with structured (anisotropic) input data: we consider the classification of sparse parity on high-dimensional orthotope where the feature coordinates have varying magnitudes. Specifically, we analyze the learning complexity of the mean-field Langevin dynamics (MFLD), which describes the noisy gradient descent update on two-layer neural network, and show that the statistical complexity (i.e. sample size) and computational complexity (i.e. network width) of MFLD can both be improved when prominent directions of the anisotropic input data aligns with the support of the target function. Moreover, we demonstrate the benefit of feature learning by establishing a kernel lower bound on the classification error, which applies to neural networks in the lazy regime.

1 Introduction

We consider the learning of a two-layer nonlinear neural network (NN) with N neurons:

$$f(z) = \frac{1}{N}\sum_{i=1}^{N} h_{x^{(i)}}(z), \quad z \in \mathbb{R}^d,\ h_{x^{(i)}}(z) : \mathbb{R}^d \to \mathbb{R}, \tag{1}$$

where $h_{x^{(i)}}(z)$ represents one neuron in the network with some trainable parameters $x^{(i)} \in \mathbb{R}^d$ and activation function $\sigma : \mathbb{R} \to \mathbb{R}$. One crucial benefit of the model (1) is the ability to learn representation that adapts to the learning problem, such as sparsity and low-dimensional structures. Indeed, recent works have shown that this *feature learning* ability enables NNs trained with gradient-based algorithms to outperform non-adaptive methods such as kernel models in learning various low-dimensional target functions (Abbe et al., 2022; Ba et al., 2022; Damian et al., 2022; Bietti et al., 2022; Mousavi-Hosseini et al., 2022; Abbe et al., 2023).

A noticeable example of low-dimensional problem is the classification of k-sparse parity, where the target label is defined as the sign of the product of $k \ll d$ coordinates: $f_*(z_i) = \mathrm{sign}\big(\prod_{i=1}^k z_i\big)$, where z_i denotes the i-th coordinate of vector z. Note that the XOR problem corresponds to the case where $k = 2$ and input on the unit hypercube. Efficiently learning this target function requires the first-layer parameters of the NN to identify the relevant k-dimensional subspace, which can be achieved via gradient-based feature learning (Daniely and Malach, 2020; Refinetti et al., 2021; Frei et al., 2022; Barak et al., 2022; Ben Arous et al., 2022).

Submitted to the Mathematics of Modern Machine Learning Workshop at NeurIPS 2023. Do not distribute.

data	result type	regime/method	sample size	width	iterations	authors
Isotropic	upper bound	NTK/SGD	d^2/ϵ	d^8	d^2/ϵ	Ji and Telgarsky (2019)
		two-phase SGD	d^{k+1}/ϵ^2	$\mathcal{O}(1)$	d/ϵ^2	Barak et al. (2022)
		mean-field/GF	d/ϵ	∞	∞	Wei et al. (2019)
		mean-field/GF	d/ϵ	$d^{d/2}$	∞	Telgarsky (2023)
		MFLD	d/ϵ	$\exp(d)$	$\exp(d)$	Suzuki et al. (2023b)
	lower bound	random features	–	d^k	–	Barak et al. (2022)
Anisotropic	upper bound	MFLD	$d^{\alpha'}/\epsilon$	$\exp(d^{\alpha'})$	$\exp(d^{\alpha'})$	Theorem 1
		MFLD (transformed)	$d^{\alpha' k} + 1/\epsilon$	d	$\mathcal{O}(1)$	Theorem 3
	lower bound	kernel	$d^{\alpha' k}$	–	–	Theorem 2

Table 1: Learning complexity for the k-sparse parity problem, omitting polylogarithmic terms. For the anisotropic bounds, we states the bounds for the spiked covariance model where the input magnitude in signal directions is d^α times larger than that in others, and define $\alpha' = 1 - 2\alpha \leq 1$. Wei et al. (2019); Telgarsky (2023) do not cover the general k-parity setting, so we state the complexity for the 2-parity (XOR). For the RF lower bound, we restate (Barak et al., 2022, Theorem 5) for bounded norm random features predictor.

One particularly relevant feature learning paradigm for the parity problem is the mean-field analysis, which lifts the optimization problem into the space of probability distribution of trainable parameters (Nitanda and Suzuki, 2017; Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018). In the setting of isotropic data ($z_i \in \{-1/\sqrt{d}, 1/\sqrt{d}\}$), it has been shown that mean-field NN can learn the parity function with *linear sample complexity*. Specifically, Wei et al. (2019); Chizat and Bach (2020); Telgarsky (2023) proved a $\mathcal{O}(d/n)$ classification error. Very recently, Suzuki et al. (2023b) considered a noisy variant of gradient descent termed the *mean-field Langevin dynamics* (MFLD), and showed that the $\mathcal{O}(d/n)$ error rate is achieved for the isotropic k-parity problem for dimension-free k. While the computational complexity is demanding due to the exponential network width required in the mean-field analysis, one remarkable feature is the statistical complexity *decouples* the degree k from the exponent in the dimension dependence; this contrasts the NTK analysis where a sample size of $n = \Omega(d^k)$ is typically needed to learn a degree-k polynomial (Ghorbani et al., 2019; Mei et al., 2022), and thus demonstrates the benefit of feature learning.

Feature learning under structured data. However, most existing analyses on the parity problem are restricted to the *isotropic* setting, where the input features do not provide any information of the support of the target function. On the other hand, realistic datasets are often structured, and different feature directions may have different magnitudes that guide the training algorithm towards more efficient learning. Recent works have indeed illustrated that in certain regression settings with low-dimensional target, structured data with a spiked covariance structure can improve the performance of both kernel methods and optimized NNs (Ghorbani et al., 2020; Ba et al., 2023; Mousavi-Hosseini et al., 2023). However, these regression analyses do not directly translate to the binary classification setting which the k-parity problem belongs to.

Therefore, our goal is to investigate the interplay between structured data and feature learning in the problem setting of classifying k-sparse parity function on *anisotropic* input data with mean-field NN.

1.1 Our Contributions

We study the statistical and computational complexity of the mean-field Langevin dynamics in learning a k-sparse parity target function on anisotropic input data. In particular, we show that

- When the feature directions of z with large magnitude align with the support of the target function I_k, then MFLD can achieve better statistical complexity (required sample size) and computational complexity (required network width) compared to the isotropic setting in Suzuki et al. (2023b). This highlights the role of structured data in the feature learning process. (Section 3 and Appendix C.1)
- If we apply a coordinate transform on the input data based on the gradient covariance matrix, then the required width can be made dimension-free, i.e., the problem can be learned by a constant width NN. This is equivalent to an anisotropic weight decay regularization, and we prove that the weighting matrix can be efficiently estimated from the first gradient descent step. (Appendix C.2)
- We prove the first lower bound on the classification error of kernel methods for general k-sparse parity problems, which is valid not only to the isotropic input but also to a spiked covariance model. The result shows that kernel methods requires larger sample size than the mean-field neural network, thus demonstrating the benefit of feature learning. (Section 4)

Due to space limitation, we defer the coordinate transform analysis to Appendix C.2.

In Table 1 we summarize and compare our results against prior works on learning sparse parity functions. To clearly illustrate the improved dimension dependence, we state our rates for a simple spiked covariance model analogous to the setting considered in Ghorbani et al. (2020); Ba et al. (2023): the data-label pairs (z, y) are generated as $= \mathrm{sign}\big(\prod_{i=1}^{k} z_i\big)$ for $k = \mathcal{O}_d(1)$, $z_i \in \{\pm d^{\alpha-\frac{1}{2}}\}$ $(i = 1, \cdots, k)$ for $0 \le \alpha \le \frac{1}{2}$, $z_i \in \{\pm d^{-\frac{1}{2}}\}$ $(i = k+1, \cdots, d)$. In this example setting, larger α corresponds to stronger anisotropy, which facilities feature learning due to the alignment between the low-dimensional structure and the target function. This benefit is evident in both the original MFLD algorithm and after the coordinate transform (or anisotropic weight decay regularization).

2 Problem Setting

k-sparse parity classification. The input random variable Z and the label Y are generated as

$$Z = \mathrm{diag}(s_1, \cdots, s_d)\tilde{Z}, \quad Y = \mathrm{sign}\big(\textstyle\prod_{i \in I_k} \tilde{Z}_i\big),$$

where $\tilde{Z}$ follows the uniform distribution on $\{\pm 1/\sqrt{d}\}^d$. We assume $s_i > 0$ and $\sum_{i=1}^{d} s_i^2 \lesssim 1$.

Mean-field two-layer network. Let $h_x(\cdot) : \mathbb{R}^d \to \mathbb{R}$ be one neuron associated with parameter $x = (x_1, x_2, x_3) \in \mathbb{R}^{d+1+1}$ in a two-layer neural network: given an input $z \in \mathbb{R}^d$,

$$h_x(z) = \bar{R}[\tanh(z^\top x_1 + x_2) + 2\tanh(x_3)]/3, \tag{2}$$

where $\bar{R} \in \mathbb{R}$ is an output scale of the network, and $\tanh$ for the bias $x_3 \in \mathbb{R}$ is placed to guarantee the boundedness following Suzuki et al. (2023b). Let $\mathcal{P}$ be the set of Borel probability measures on $\mathbb{R}^{\bar{d}}$ where $\bar{d} = d + 2$ and $\mathcal{P}_p$ be the subset of $\mathcal{P}$ with the finite p-th moment. The mean-field neural network is defined by integrating infinitely many neurons h_x over $\mathbb{R}^{\bar{d}}$ with the distribution $\mu \in \mathcal{P}$: $f_\mu(\cdot) = \int h_x(\cdot)\mu(\mathrm{d}x)$, We consider the logistic loss function $\ell(f, y) = \log(1 + \exp(-yf))$. We also denote $\ell(yf) = \ell(f, y)$. Then, the regularized empirical risk of f_μ are defined as

$$\mathcal{L}(\mu) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i f_\mu(z_i)) + +\lambda(\lambda_1 \mathbb{E}_{X\sim\mu}[\|X\|^2] + \mathrm{Ent}(\mu)), \tag{3}$$

with the regularization parameters $\lambda, \lambda_1 \ge 0$. $\mathbb{E}_{X\sim\mu}[\|X\|^2]$ is the L^2 regularization and $\mathrm{Ent}(\mu) = \int \log \mu \mathrm{d}\mu$ is the entropy regularization. A remarkable advantage of this setting is that the above objectives become convex functional with respect to the distribution μ since μ linearly acts on f_μ.

Mean-field Langevin dynamics. MFLD corresponds to the noisy gradient descent, where a Gaussian perturbation is added at each gradient step (Mei et al. (2018); Hu et al. (2019)). Let $\mathscr{X}_\tau = (X_\tau^i)_{i=1}^N \subset \mathbb{R}^{\bar{d}}$ be N neurons at the τ-th update, and define $\mu_\tau = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_\tau^i}$. Then, time- and space-discretized version of MFLD with step size η and N neurons is written as the following stochastic differential equation:

$$X_0^i \sim \mu_0 = N(0, I/(2\lambda_1)), \quad X_{\tau+1}^i = X_\tau^i - \eta\nabla\frac{\delta F(\mu_\tau)}{\delta\mu}(X_\tau^i) + \sqrt{2\lambda\eta}\xi_\tau^i, \tag{4}$$

where ξ_τ^i is an i.i.d. standard normal random variable $\xi_\tau^i \sim N(0, I)$, and $\frac{\delta F(\mu_t)}{\delta\mu}$ is the first variation of F, which, in our setting, is written as $\frac{\delta F(\mu)}{\delta\mu}(x) = \frac{1}{n}\sum_{i=1}^{n} \ell'(y_i f_\mu(z_i)) y_i h_x(z_i) + \lambda(\lambda_1\|x\|^2)$.

Logarithmic Sobolev Inequality. Convergence of MFLD crucially depends on the property of the *proximal Gibbs distribution* p_μ associated with $\mu \in \mathcal{P}$ Nitanda et al. (2022); Chizat (2022). The density of p_μ is given by $p_\mu(X) \propto \exp\left(-\frac{1}{\lambda}\frac{\delta F(\mu)}{\delta\mu}(X)\right)$. The key in our proof lies in controlling a constant in the following *logarithmic Sobolev inequality* (LSI) on the Gibbs measure by making use of anisotropy and extending Suzuki et al. (2023a,b). If we can find a good μ^* that achieves small loss and that $\mathrm{KL}(\mu_0||\mu^*)$ is small, then we can have a small LSI constant, which yields better convergence and generalization results. For more details of the analysis, please refer to the appendix.

Definition 1 (Logarithmic Sobolev inequality)**.** *Let μ be a Borel probability measure on $\mathbb{R}^d$. We say μ satisfies the LSI with a constant $\alpha > 0$ if for any smooth function $\phi : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_\mu[\phi^2] < \infty$,*

$$\mathbb{E}_\mu[\phi^2 \log(\phi^2)] - \mathbb{E}_\mu[\phi^2]\log(\mathbb{E}_\mu[\phi^2]) \le \frac{2}{\alpha}\mathbb{E}_\mu[\|\nabla\phi\|_2^2].$$

3 Statistical and Computational Complexity for Anisotropic Data

We have the following result on the anisotropic k-sparse parity setting.

Theorem 1 (k-sparse parity setting)**.** *Define* $S_{I_k}^2 := \sum_{j \in I_k} s_j^{-2}$*. We may take* $\bar{R} = k$ *and* $\lambda = O(1/(S_{I_k}^2 \log(k)^2))$ *so that the classification error is bounded by*

$$P(Y f_{\mu_{[\lambda]}} < 0) \le O\left(\frac{k S_{I_k}^2 \log(k)^2}{n} (\log(1/\delta) + \log\log(n)) \right),$$

with probability $1-\delta$*. Moreover, if* $n = \Omega(k^4 S_{I_k}^4 \log(k)^4)$*, then* $P(Y f_{\mu_{[\lambda]}} \le 0) = 0$ *with probability*

$$1 - \exp[-\Omega(n/(k^4 S_{I_k}^4 \log(k)^4))].$$

For the computational cost, it suffices to take the number of iterations T *and network width* N *as*

$$T = O(S_{I_k}^2 \log(k)^2 n \log(nd) \exp[O(k S_{I_k}^2 \log(k)^2)]), \quad N = O(n^2 \exp(O(k S_{I_k}^2 \log(k)^2)))),$$

respectively, to achieve the same statistical complexity as described above.

Notably, for sufficiently anisotropic data such that $S_{I_k}^2 = k^2$, the computational complexity becomes completely polynomial order with respect to the dimension d; this is in stark contrast to the isotropic setting, where the complexity has exponential order with respect to d.

We provide two examples of covariance structure that allows us to smoothly interpolate between the isotropic and anisotropic setting:

- *Power-law decay.* We set $I_k = \{1, \ldots, k\}$ and $s_i^2 = c_d i^{-\alpha}$ where $c_d = \Theta(d^{1-\alpha})$ for $\alpha \in [0, 1)$. Then, in this setting, we have that $S_{I_k}^2 = \mathcal{O}(d^{1-\alpha})$. This interpolates between the isotropic and the completely anisotropic setting $S_{I_k}^2 = k^2$ by adjusting α between $(0, 1)$.
- *Spiked covariance.* We set $s_i = d^{\alpha - 1/2}$ for $i \in I_k$, and $s_i = d^{-1/2}$ otherwise, for $\alpha \in [0, 1]$. In this case we have $S_{I_k}^2 = \mathcal{O}(d^{1-2\alpha})$, which becomes dimension-free when α approaches $\frac{1}{2}$. We verify Corollary 1 for this spiked covariance setting by conducting experiment in Appendix D.

4 Kernel lower bound for the anisotropic parity problem

To emphasize the benefit of feature learning, we prove a classification lower bound for kernel methods on the k-parity problem in the above spiked covariance setting. We remark that most existing kernel lower bounds are only valid for the regression setting, with the exception of Wei et al. (2019) which only handles the $k = 2$ case with the isotropic input.

Specifically, we consider an inner-product kernel, which is assumed to be expressed as

$$K(z, z') = \sum_{l=0}^{\infty} \alpha_l \left(z^\top z'\right)^l, \quad \{\alpha_0\}_{l=0}^{\infty}: \text{positive and bounded.}$$

Based on n i.i.d. training samples, we construct the kernel estimator $f_\beta(z)$ with $\beta \in \mathbb{R}^n$ chosen arbitrarily: $f_\beta(z) = \sum_{i=1}^n \beta_i K(z, z^i)$. For this f_β, we have the following lower bound.

Theorem 2. *Fix* $\delta > 0$ *arbitrarily. For sufficiently large d, draw* $n \lesssim d^{\lfloor (1-2\alpha)k \rfloor - \delta}$ *sample. Then, with probability at least* 0.99 *over the sample, for all choices of* $\beta \in \mathbb{R}^n$*,* $f_\beta = \sum_{i=1}^n \beta_i K(z, z^i)$ *will predict the sign of* y *wrong* $\Omega(1)$ *fraction of the time:*

$$\mathbb{P}_{z \sim P_Z}\left[f_\beta(z) y < 0\right] = \Omega(1).$$

The proof can be found in Appendix G. First, we lower bound the failure probability by the probability when $|f_\beta(z)|$ is large, by extending Wei et al. (2019) based on finer evaluation on the correlation $yK(z, z^i)$. Then, we reduce the problem into lower bounding the smallest eigenvalue of some kernel matrix, where we make use of the more refined characterization in Misiakiewicz (2022).

References

E. Abbe, E. B. Adsera, and T. Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

E. Abbe, E. Boix-Adsera, and T. Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. *arXiv preprint arXiv:2302.11055*, 2023.

J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=akddwRG6EGi.

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

D. Bakry and M. Émery. Diffusions hypercontractives. In J. Azéma and M. Yor, editors, *Séminaire de Probabilités XIX 1983/84*, pages 177–206, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg. ISBN 978-3-540-39397-9.

D. Bakry, I. Gentil, M. Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

B. Barak, B. L. Edelman, S. Goel, S. M. Kakade, eran malach, and C. Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=8XWP2ewX-im.

G. Ben Arous, R. Gheissari, and A. Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35:25349–25362, 2022.

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

F. Chen, Z. Ren, and S. Wang. Uniform-in-time propagation of chaos for mean field langevin dynamics. *arXiv preprint arXiv:2212.03050*, 2022.

L. Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=BDqzLH1gEm.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31*, pages 3040–3050, 2018.

L. Chizat and F. Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

A. Damian, J. Lee, and M. Soltanolkotabi. Neural networks can learn representations with gradient descent. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5413–5452. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/damian22a.html.

A. Daniely and E. Malach. Learning parities with neural networks. *arXiv preprint arXiv:2002.07400*, 2020.

S. Frei, N. S. Chatterji, and P. L. Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *arXiv preprint arXiv:2202.07626*, 2022.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. When do neural networks outperform kernel methods? *arXiv preprint arXiv:2006.13409*, 2020.

R. Holley and D. Stroock. Logarithmic sobolev inequalities and stochastic ising models. *Journal of statistical physics*, 46(5-6):1159–1194, 1987.

K. Hu, Z. Ren, D. Siska, and L. Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. *arXiv preprint arXiv:1905.07769*, 2019.

Z. Ji and M. Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

S. Mei, T. Misiakiewicz, and A. Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59: 3–84, 2022.

T. Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*, 2022.

A. Mousavi-Hosseini, S. Park, M. Girotti, I. Mitliagkas, and M. A. Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. In *The Eleventh International Conference on Learning Representations*, 2022.

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

A. Nitanda and T. Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

A. Nitanda, D. Wu, and T. Suzuki. Convex analysis of the mean field langevin dynamics. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9741–9757. PMLR, 28–30 Mar 2022.

R. O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

M. Refinetti, S. Goldt, F. Krzakala, and L. Zdeborova. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8936–8947. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/refinetti21b.html.

G. M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.

T. Suzuki, A. Nitanda, and D. Wu. Uniform-in-time propagation of chaos for the mean-field gradient langevin dynamics. In *The Eleventh International Conference on Learning Representations*, 2022.

T. Suzuki, D. Wu, and A. Nitanda. Mean-field langevin dynamics: Time-space discretization, stochastic gradient, and variance reduction. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023a.

T. Suzuki, D. Wu, K. Oko, and A. Nitanda. Feature learning via mean-field langevin dynamics: classifying sparse parities and beyond. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023b.

M. Telgarsky. Feature selection and low test error in shallow low-rotation ReLU networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=swEskiem99.

R. Vershynin. High-dimensional probability. *University of California, Irvine*, 2020.

C. Wei, J. D. Lee, Q. Liu, and T. Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9712–9724, 2019.

Appendix

The appendix is organized as follows. First, Appendices A and B complement the problem setting presented in the main text. Especially, Appendix B presents technical foundations of our analysis. Then, Appendix C discusses the upper bounds in detail: additional discussion on Section 3 can be found in Appendix C.1, and our second contribution on the significant improvement of the computational and statistical complexity utilizing the gradient covariance matrix is presented in Appendix C.2. We validate our theory by conducting a numerical experiment in Appendix D, which considers learning the 3-sparse parity in the spiked covariance setting. As for the proofs, Appendix E proves Section 3 and Appendix C.1, Appendix F provides the proof for Appendix C.2, and finally, Appendix G proves the kernel lower bound.

A Supplement for the Problem Setting

Here we recall the problem setting and provide additional explanations. We discuss the mean-field Langevin dynamics in the subsequent separate section.

k-sparse parity classification. We consider the binary classification problem where the labels are generated from a k-parity target function as follows. The following definition extends the one in the main text, that only referred to the axis-aligned case.

Definition 2 (k-sparse parity problem under linear transformation)**.** *The input random variable Z and the corresponding label Y are generated as*

$$Z = A\tilde{Z}, \quad Y = \mathrm{sign}\big(\textstyle\prod_{i \in I_k} \tilde{Z}_i\big),$$

where A is an invertible matrix and $\tilde{Z}$ is distributed from the uniform distribution on $\{\pm 1/\sqrt{d}\}^d$. We also assume $\|Z\| = \|A\tilde{Z}\| \lesssim 1$ almost surely.

Note that this definition includes the well-studied XOR problem (Wei et al., 2019; Telgarsky, 2023) as a special case.

Example 1 (Isotropic XOR)**.** *We take $A = I_d$ and $Y = \mathrm{sign}(\tilde{Z}_1\tilde{Z}_2)$ ($k = 2$).*

Similarly, the extension to k parity on isotropic data (Barak et al., 2022; Suzuki et al., 2023b) is also covered by our general definition.

The example that we considered in the main text is the following anisotropic and axis-aligned setting with $A = I_d$ and $I_k = [k]$. In this anisotropic setting the coordinates are independent but may have different magnitudes.

Example 2 (Axis-aligned anisotropic k parity)**.** *There exist positive reals $s_i > 0$ ($i = 1, \ldots, d$) such that the support of P_Z (the distribution of Z) is given by $\mathcal{S} := \{\pm s_1\} \times \{\pm s_2\} \times \cdots \times \{\pm s_d\}$, i.e., any element $z = (z_1, \ldots, z_d) \in \mathrm{supp}(P_Z)$ satisfies $z_i \in \{\pm s_i\} \quad (i = 1, \ldots, d)$. We also assume $(z_i)_{i=1}^d$ are mutually independent and $P(z_i = s_i) = P(z_i = -s_i) = 1/2$. The k-sparse parity label corresponds to the sign of the product of k-indices $I_k \subset \{1, \ldots, d\}$.*

Mean-field two-layer network. Let $h_x(\cdot) : \mathbb{R}^d \to \mathbb{R}$ be one neuron associated with parameter $x = (x_1, x_2, x_3) \in \mathbb{R}^{d+1+1}$ in a two-layer neural network: given an input $z \in \mathbb{R}^d$,

$$h_x(z) = \bar{R}[\tanh(z^\top x_1 + x_2) + 2\tanh(x_3)]/3, \tag{5}$$

where $\bar{R} \in \mathbb{R}$ is an output scale of the network and an extra $\tanh$ activation for the bias term $x_3 \in \mathbb{R}$ is placed to make the function bounded following Suzuki et al. (2023b). Let $\mathcal{P}$ be the set of Borel probability measures on $\mathbb{R}^{\bar{d}}$ where $\bar{d} = d + 2$ and $\mathcal{P}_p$ be the subset of $\mathcal{P}$ with finite p-th moment: $\mathbb{E}_\mu[\|X\|^p] < \infty$ ($\mu \in \mathcal{P}$). The mean-field neural network is defined by integrating infinitely many neurons h_x over $\mathbb{R}^{\bar{d}}$ with the distribution $\mu \in \mathcal{P}$,

$$f_\mu(\cdot) = \int h_x(\cdot)\mu(\mathrm{d}x),$$

Let $\ell(\cdot,\cdot):\mathbb{R}\times\mathbb{R}\to\mathbb{R}_{\geq 0}$ be a smooth and convex loss function for the binary classification. Typically, we consider the logistic loss function $\ell(f,y)=\log(1+\exp(-yf))$ where $f\in\mathbb{R},\ y\in\{\pm 1\}$. We also denote $\ell(yf)=\ell(f,y)$ Then, the empirical risk and the population risk of f_μ are defined as

$$L(\mu):=\frac{1}{n}\sum_{i=1}^{n}\ell(y_i f_\mu(z_i)),\ \ \bar{L}(\mu):=\mathbb{E}[\ell(Yf_\mu(Z))].$$

To avoid overfitting, we consider a regularized empirical risk $F(\mu):=L(\mu)+\lambda\mathbb{E}_{X\sim\mu}[\lambda_1\|X\|^2]$ with the regularization parameters $\lambda,\lambda_1\geq 0$. In addition, we introduce the entropy regularized risk:

$$\mathcal{L}(\mu)=F(\mu)+\lambda\mathrm{Ent}(\mu). \tag{6}$$

We can immediately see that $\mathcal{L}$ is equivalent to $L(\mu)+\lambda\mathrm{KL}(\nu,\mu)$ up to constant, where $\mathrm{KL}(\nu,\mu)=\int\log(\mu/\nu)\mathrm{d}\mu$ is the KL-divergence between ν and μ, and ν is the Gaussian distribution with mean 0 and variance $I/(2\lambda_1)$, i.e., $\nu=N(0,I/(2\lambda_1))$. A remarkable advantage of mean-field parameterization is that the above objectives become convex functional with respect to the distribution μ since μ linearly acts on f_μ.

B Mean-field Langevin dynamics

This section introduces tean-field Langevin dynamics in detail. In recent years, the theory of MFLD has been well established and it has been shown to optimize the functional $\mathcal{L}$. MFLD is defined by the following stochastic differential equation: $X_0\sim\mu_0$,

$$\mathrm{d}X_t=-\nabla\frac{\delta F(\mu_t)}{\delta\mu}(X_t)\mathrm{d}t+\sqrt{2\lambda}\mathrm{d}W_t,\quad \mu_t=\mathrm{Law}(X_t), \tag{7}$$

where $(W_t)_{t\geq 0}$ is the d-dimensional standard Brownian motion, and $\frac{\delta F(\mu_t)}{\delta\mu}$ is the first variation of F, which, in our setting, is written as $\frac{\delta F(\mu)}{\delta\mu}(x)=\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i f_\mu(z_i))y_i h_x(z_i)+\lambda(\lambda_1\|x\|^2)$. The Fokker-Planck equation of SDE (7) is given by[1]

$$\partial_t\mu_t=\lambda\Delta\mu_t+\nabla\cdot\left[\mu_t\nabla\tfrac{\delta F(\mu_t)}{\delta\nu}\right]=\nabla\cdot\left[\mu_t\nabla\left(\lambda\log(\mu_t)+\tfrac{\delta F(\mu_t)}{\delta\nu}\right)\right]. \tag{8}$$

Then, several studies (Mei et al., 2018; Hu et al., 2019; Nitanda et al., 2022; Chizat, 2022) showed the convergence $\mathcal{L}(\mu_t)\to\mathcal{L}(\mu_{[\lambda]})$, where $\mu_{[\lambda]}:=\mathrm{argmin}_{\mu\in\mathcal{P}}\mathcal{L}(\mu)$.

For a practical algorithm, we need to consider a space- and time-discretized version of the MFLD, that is, we approximate the solution μ_t by an empirical measure $\mu_{\mathscr{X}}=\frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}$ corresponding to a set of finite particles $\mathscr{X}=(X^i)_{i=1}^N\subset\mathbb{R}^{\bar{d}}$. Let $\mathscr{X}_\tau=(X_\tau^i)_{i=1}^N\subset\mathbb{R}^{\bar{d}}$ be N particles at the τ-th update ($\tau\in\{0,1,2,\dots\}$), and define $\mu_\tau=\mu_{\mathscr{X}_\tau}$ as a finite particle approximation of the population counterpart. Then, the discretized MFLD is defined as follows: $X_0^i\sim\mu_0$, and $\mathscr{X}_\tau$ is updated as

$$X_{\tau+1}^i=X_\tau^i-\eta\nabla\frac{\delta F(\mu_\tau)}{\delta\mu}(X_\tau^i)+\sqrt{2\lambda\eta}\xi_\tau^i, \tag{9}$$

where $\eta>0$ is the step size, ξ_τ^i is an i.i.d. standard normal random variable $\xi_\tau^i\sim N(0,I)$. Note that in the context of mean-field neural network (1), the discretized update (9) simply corresponds to the noisy gradient descent algorithm, where a Gaussian perturbation is added at each gradient step. We write $f_{\mathscr{X}}:=f_{\mu_{\mathscr{X}}}$ for simplicity of notation.

B.1 Logarithmic Sobolev inequality

Nitanda et al. (2022); Chizat (2022) have established the exponential convergence of MFLD by exploiting the *proximal Gibbs distribution* p_μ associated with $\mu\in\mathcal{P}$. The density of p_μ is given by

$$p_\mu(X)\propto\exp\left(-\frac{1}{\lambda}\frac{\delta F(\mu)}{\delta\mu}(X)\right).$$

[1] This should be interpreted in a weak sense, that is, for any continuously differentiable function ϕ with a compact support, $\int\phi\mathrm{d}\mu_t-\int\phi\mathrm{d}\mu_s=-\int_s^t\int\nabla\phi\cdot(\nabla\log(\mu_t)-\nabla\frac{\delta F(\mu_t)}{\delta\nu})\mathrm{d}\mu_\tau\mathrm{d}\tau$.

The smoothness of the loss function and the tanh activation guarantee the existence of the unique minimizer μ^* of $\mathcal{L}$, which also solves the equation: $\mu = p_\mu$ (see Proposition 2.5 of Hu et al. (2019)).

The key in their proofs is to show a *logarithmic Sobolev inequality* (LSI) on the Gibbs measure p_μ (see Definition 1). We can apply the classical Bakry-Emery and Holley-Stroock arguments (Bakry and Émery, 1985; Holley and Stroock, 1987) (Corollary 5.7.2 and 5.1.7 of Bakry et al. (2014)) to derive the LSI constant on the Gibbs distribution whose potential is the sum of the strongly convex function and bounded function. If $\|\frac{\delta L(\mu)}{\delta\mu}\|_\infty \leq B$, the proximal Gibbs distributions fall into this case and we can establish the LSI with $\alpha \geq \lambda_1 \exp\left(-4B/\lambda\right)$. In our case, since the logistic loss is employed and each neuron h_x is bounded by $\bar{R}$, we have $B = \bar{R}$ and therefore

$$\alpha \geq \lambda_1 \exp\left(-4\bar{R}/\lambda\right). \tag{10}$$

B.2 Quantitative Analysis of MFLD

Convergence guarantee. As shown in Chen et al. (2022); Suzuki et al. (2022), the LSI constant determines not only the rate of convergence, but also the number of particles (i.e., width of the neural network) to approximate the mean-field limit. Let us consider the linear functional of a distribution $\mu^{(N)}$ of N particles $\mathscr{X} = (X^i)_{i=1}^N \subset \mathbb{R}^{\bar{d}}$ defined by

$$\mathcal{L}^N(\mu^{(N)}) = N\mathbb{E}_{\mathscr{X}\sim\mu^{(N)}}[F(\mu_{\mathscr{X}})] + \lambda\mathrm{Ent}(\mu^{(N)}).$$

Let $\mu_\tau^{(N)}$ be the distribution of particles $\mathscr{X}_\tau = (X_\tau^i)_{i=1}^N$ at the τ-th iteration, and define $\Delta_\tau = \frac{1}{N}\mathcal{L}^N(\mu_\tau^{(N)}) - \mathcal{L}(\mu_{[\lambda]})$. Suzuki et al. (2023a) established the convergence rate of MFLD as follows.

Proposition 1. *Let $\bar{B}^2 := \mathbb{E}[\|X_0^i\|^2] + \frac{1}{\lambda\lambda_1}\left[\left(\frac{1}{4} + \frac{1}{\lambda\lambda_1}\right)\bar{R}^2 + \lambda d\right]$ and $\delta_\eta := C_1\bar{L}^2(\eta^2 + \lambda\eta)$, where $\bar{L} = 2\bar{R} + \lambda\lambda_1$ and $C_1 = 8(\bar{R}^2 + \lambda\lambda_1\bar{B}^2 + d) = O(d + \lambda^{-1})$. Then, if $\lambda\alpha\eta \leq 1/4$ and $\eta \leq 1/4$, then the neural network trained by MFLD converges to the optimal network $f_{[\lambda]}$ as*

$$\mathbb{E}_{\mathscr{X}_\tau\sim\mu_\tau^{(N)}}\left[\sup_{z\in\mathrm{supp}(P_Z)}(f_{\mathscr{X}_\tau}(z) - f_{\mu_{[\lambda]}}(z))^2\right] \leq \frac{4\bar{L}^2}{\lambda\alpha}\Delta_\tau + \frac{2}{N}\bar{R}^2,$$

where Δ_τ is further bounded by $\Delta_\tau \leq \exp\left(-\lambda\alpha\eta\tau/2\right)\Delta_0 + \frac{2}{\lambda\alpha}\bar{L}^2C_1\left(\lambda\eta + \eta^2\right) + \frac{4C_\lambda}{\lambda\alpha N}$.

In particular, for a given $\epsilon^* > 0$, the right hand side can be bounded by $\epsilon^* + \frac{2\bar{R}^2}{N}$ after $T = O\left(\frac{1}{\lambda\alpha\eta}\log(1/\epsilon^*)\right)$ iterations with the step size $\eta = O\left(\lambda\alpha^2\epsilon^*/C_1 + \lambda\alpha\sqrt{\epsilon^*/C_1}\right)$. In terms of generalization error (Proposition 2), the optimization error can be set as $\epsilon^* = O(1/(n\lambda)^2)$. Then, the required total number of iteration T and the number of particles N can be bounded by

$$T \leq O\left((d + \lambda^{-1})n^2\exp(16\bar{R}/\lambda)\log(n\lambda)\right), \quad N \leq O((\epsilon^*\lambda\alpha)^{-2}) = O\left(n^2\exp(8\bar{R}/\lambda)\right). \tag{11}$$

From this evaluation, it is crucial to carefully select the strength of regularization parameter λ to obtain a sufficiently small loss. In the following section, we evaluate λ and then investigate how structured data affects its value.

Generalization error bound. Now we state the classification error bound of the neural network optimized by MFLD. For this purpose, we introduce the following assumption which will be verified later on for the anisotropic parity setting.

Assumption 1. *There exists $c_0 > 0$ and $R > 0$ such that the following conditions are satisfied:*

- *There exists $\mu^* \in \mathcal{P}$ such that $\mathrm{KL}(\nu||\mu^*) \leq R$ and $L(\mu^*) \leq \ell(0) - c_0$.*
- *For any $\lambda < c_0/R$, the risk minimizer $\mu_{[\lambda]}$ of $\mathcal{L}(\mu)$ satisfies $Yf_{\mu_{[\lambda]}}(X) \geq c_0$ almost surely.*

Here c_0 plays a margin for a solution μ^* and R controls "difficulty" of the problem. Indeed, if larger R is required, the Bayes optimal solution should be far away from the prior ν, a Gaussian distribution. Hence, it is expected that obtaining a good classifier is more difficult. Let $\hat{\mu}$ be an approximately optimal solution of $\mathcal{L}$ with ϵ^* accuracy: $\mathcal{L}(\hat{\mu}) \leq \min_{\mu\in\mathcal{P}}\mathcal{L}(\mu) + \epsilon^*$; we have the following generalization error bounds.

Proposition 2 (Suzuki et al. (2023b)). *Let $M_0 = (\epsilon^* + 2(\bar{R}+1))/\lambda$ and suppose that $\lambda < c_0/R$.*

(i) If the sample size n satisfies

$$n > C\frac{\bar{R}^2}{c_0^2\lambda^2}\left[\lambda\left(\bar{R}+\frac{\lambda}{\bar{R}^2 n}\right)+\bar{R}^2(1+\log\log_2(n^2M_0\bar{R}))+n\lambda\epsilon^*\right] =: S,$$

with an absolute constant C, then $f_{\hat{\mu}}$ satisfies $P(Yf_{\hat{\mu}}(Z)\le 0)=0$ (the Bayes optimal classifier) with probability $1-\exp(-\frac{n\lambda^2}{32\bar{R}^4}(c_0^2 - S/n))$.

(ii) When the sample size does not satisfy the condition $n > S$, we still have that there exists an absolute constant $C > 0$ such that

$$P(Yf_{\hat{\mu}}(Z)\le 0) \le C\beta(c_0)\left[\frac{\bar{R}^2}{n\lambda}\left(1+t+\log\log_2(n^2M_0\bar{R})\right)+\frac{1}{n}\left(\bar{R}+\frac{\lambda}{\bar{R}^2 n}\right)+\epsilon^*\right],$$

with probability $1-\exp(-t)$, where $\beta(c_0) := 1/[\ell(0) - (\ell(c_0) - c_0\ell'(c_0))]$.

This result states that if we take the regularization parameter λ sufficiently small as $\lambda < \mathcal{O}(1/R)$, then for sufficiently large sample size such that $n > S = \Omega(1/\lambda^2)$, we have an exponential convergence of the expected classification error as $\mathbb{E}_{D^n}[P(Yf_{\hat{\mu}}(Z)\le 0)] \le \exp(-\Omega(n\lambda^2))$; otherwise, we sill have $\mathbb{E}_{D^n}[P(Yf_{\hat{\mu}}(Z)\le 0)] = \mathcal{O}(1/(n\lambda))$. Hence, the classification error and its convergence rate is almost completely characterized by R through the choice of $\lambda = \mathcal{O}(1/R)$: for a problem with large R, we need to pay greater sample complexity.

It is also worth noting that the value of R affects not only the statistical complexity but also the computational complexity. Remember that the number of iterations T and the network width N also depend on λ through Eq. (11). Indeed, by taking $\lambda = c_0/R$, we arrive at $T = \mathcal{O}(\exp(16\bar{R}R/c_0)\log(n))$ and $N = \mathcal{O}(\exp(8\bar{R}R/c_0))$, which has exponential dependence on R.

Therefore, the goal of the subsequent analysis is to answer the following question in the affirmative:

Can we utilize the anisotropy of input data to reduce the value of R,
hence improving the statistical and computational complexity of MFLD?

C Learning under Structured Data

C.1 Statistical and computational complexity for anisotropic data

This subsection explains how to obtain the result in Section 3. We analyze how the anisotropic property of the input affects the generalization error and the computational complexity through the aforementioned measure of problem difficulty R. We first present a framework for the general problem setting in Definition 2. Let $\tilde{\phi} = (\tilde{\phi}_1,\dots,\tilde{\phi}_d)^\top \in \mathbb{R}^d$ as

$$\tilde{\phi}_i = \begin{cases}\sqrt{d} & (i\in I_k),\\ 0 & (i\notin I_k).\end{cases} \tag{12}$$

Then, we have the following proposition that controls R in terms of the transformation matrix A.

Proposition 3. *Define $\phi := A^{-1}\tilde{\phi}$ where $\tilde{\phi}$ is defined by Eq. (12). For $\bar{R}=k$, there exists $\mu^*\in\mathcal{P}$ and R such that*

$$\mathrm{KL}(\nu||\mu^*) \le R = c_1(\|\phi\|^2 + k^2)\log(k)^2,$$

and $L(\mu^*) \le \ell(0) - c_2$, where $c_1, c_2 > 0$ are absolute constants.

Under this conditions in this proposition, we can show that the minimizer of the MFLD objective achieves the Bayes optimal classifier with a positive margin as follows.

Proposition 4. *Assume that there exists $\mu^*\in\mathcal{P}$ such that the conditions in Proposition 3 is satisfied with R and $\bar{R}$ in the statement. Then, if we choose the regulaization parameter λ as $\lambda < c_2/(2R)$, then the minimizer $\mu_{[\lambda]}$ of the MFLD objective satisfies*

$$\max\{\bar{L}(\mu_{[\lambda]}), L(\mu_{[\lambda]})\} < \ell(0) - \tfrac{c_2}{2},$$

and $f_{\mu_{[\lambda]}}$ is a perfect classifier with margin c_2, i.e., $Yf_{\mu_{[\lambda]}}(Z) \ge \frac{c_2}{2}$ almost surely.

The proofs of both propositions can be found in Appendix E in the appendix. These general results state that Assumption 1 is satisfied for the general problem setting in Definition 2. Now we consider special cases where concrete sample complexity and computational complexity can be derived. For example, we have the following evaluation for the k-sparse parity with anisotropic covariance.

Example: Anisotropic k-sparse parity. In the k-parity setting (Example 2), Assumption 1 is satisfied with constants specified in the following propositions, which follow from Proposition 4.

Corollary 1 (Anisotropic k-sparse parity). *Suppose that (Z, Y) is generated from the anisotropic k parity problem (Example 2). Then, for $\bar{R} = k$, there exists $\mu^* \in \mathcal{P}$ satisfying $\mathrm{KL}(\nu||\mu^*) \le R$ where*

$$R = c_1 \left(\sum_{i \in I_k} s_i^{-2} \right) \log(k)^2,$$

and $L(\mu^*) \le \ell(0) - c_2$, where $c_1, c_2 > 0$ are absolute constants.

This result highlights the benefit of structured data. Observe that isotropic covariance corresponds to $s_i = 1/\sqrt{d}$ $(i = 1, \ldots, d)$, where R needs to be $\tilde{\mathcal{O}}(kd)$, which then leads to exponential dimension dependency in the computational complexity, and also dimension-dependent sample complexity, as shown in Suzuki et al. (2023b). On the other hand, if the input covariance is anisotropic so that $s_j^2 > \Omega(1/k)$ for $j \in I_k$ (i.e., the input Z_j is large for the informative coordinates $j \in I_k$ and other coordinates are small), then the value of R becomes dimension-free: $R = \mathcal{O}(k^2 \log(k)^2)$.

Substituting the values of R and $\bar{R}$ to the generalization error and computational complexity bounds, we obtain the Corollary 1.

C.2 Utilizing Anisotropy via Coordinate Transform

This section explains our third contribution, i.e., a coordinate transform that enables learning even the isotropic k-sparse parity problem with a dimension-free constant width network.

From the previous analysis, we see that anisotropic data can indeed improve both the statistical and computational complexity. This being said, it is worth noting that unless the problem is sufficiently anisotropic such that R becomes cost, the computational cost would still be super-polynomial in terms of dimension dependence. The goal of this section is to show that the computational complexity can be further improved by exploiting the anisotropy of the learning problem. Specifically, we utilize the gradient covariance matrix to estimate the informative subspace, similar to the one-step gradient feature learning procedure studied in Ba et al. (2022); Damian et al. (2022); Barak et al. (2022).

Let $\sigma(w^\top z) = h_x(z)$ for $(x_1, x_2, x_3) = (w, b_1, b_2)$ for fixed b_1 and b_2. We initialize the particles $\mathscr{X}_0 = \{(w_l, b_1, b_2)\}_{l=1}^{N/2} \cup \{(-w_l, -b_1, -b_2)\}_{l=1}^{N/2}$ by generating w_l from the uniform distribution $\mathcal{U}(\mathcal{B}_{c_0})$ on the ball with sufficiently radius $c_0 > 0$. The gradient for each neuron is given as

$$g(w_l) = \frac{1}{n} \sum_{i=1}^{n} \ell'(y_i f_{\mathscr{X}_0}(z_i)) y_i z \sigma'(w_l^\top z).$$

Note that we have $f_{\mathscr{X}_0}(Z) = 0$ almost surely. We then calculate the covariance as

$$G = \tfrac{1}{N} \textstyle\sum_{l=1}^{N} g(w_l) g(w_l)^\top,$$

to estimate the informative subspace. Define the "regularized covariance" $\hat{G} = G + \hat{\lambda}_0 I$. For this choice of $\hat{G}$, we apply coordinate transform of the input Z as

$$\hat{Z} \leftarrow c_A \hat{G}^{1/2} Z,$$

where c_A is a scaling parameter so that $\|\hat{Z}\| \le 1$ almost surely. We denote by $\hat{z}_i = c_A \hat{G}^{1/2} z_i$ accordingly. After this coordinate transform, we train the neural network through MFLD; that is, we optimize the objective $\mu \mapsto \frac{1}{n} \sum_{i=1}^{n} \ell(f_\mu(\hat{z}_i) y_i) + \lambda(\lambda_1 \mathbb{E}_\mu[\|X\|^2] + \mathrm{Ent}(\mu))$. Intuitively, this coordinate transform tries to amplify the informative coordinates $(j \in I_k)$ and suppress the non-informative coordinates $(j \in I_k^c)$. More specifically, the covariance of the input becomes more well-specified to the target signal Y leading to a better LSI constant. We remark that such coordinate transformation is equivalent to employing an *anisotropic* weight decay regularization on the weight parameters $r(x) = \|x\|^2_{\hat{G}^{-1}}$.

Taken into account the sample complexity to estimate the gradient covariance, we obtain the following evaluation of the KL-divergence between the prior distribution ν and a Bayes optimal solution μ^*.

Theorem 3. *Suppose that c_0 is taken sufficiently small such that $\sum_{j=1}^d w_j^2 s_j^2 \leq 1$ almost surely for $w \sim \mathcal{U}(\mathcal{B}_{c_0})$ and $\mathbb{E}[w_j] = \Theta(1)$, and the regularization parameter $\hat{\lambda}_0$ is set to be $\hat{\lambda}_0 = \prod_{j' \in I_k} s_{j'}^2 \cdot \max_{j' \in I_k^c} s_{j'}^2$. We assume that the sample size n and the number of particles N satisfies*

$$n \geq C_k \frac{k^2 \bar{R}^2 \log(2N/\delta)^2}{\prod_{j' \in I_k} s_{j'}^2}, \quad N \geq C_k \frac{d \log(d/\delta)}{\max_{j' \notin I_k} s_{j'}^4}, \tag{13}$$

for given $\delta \in (0,1)$, where C_k is a constant depending on k. Then, for $\bar{R} = k$ and sufficiently small C_k, there exists $\mu^* \in \mathcal{P}$ such that $L(\mu^*) \leq \ell(0) - c_2$ and $\mathrm{KL}(\nu || \mu^*) \leq R$ where

$$R = c_1 \left(k \frac{\max_{j' \in [d]} s_{j'}^2}{\min_{j' \in I_k} s_{j'}^2} + k^2 \right) \log(k)^2,$$

for a constant c_1 independent of the dimensionality d, with probability $1 - \delta$. Here, the probability is with respect to the randomness of training data and generating the initial parameters $(w_l)_{l=1}^N$.

We make the following remarks on the theorem.

- This theorem implies a significant improvement on the LSI constant since R is independent of d as long as $\frac{\max_{j' \in [d]} s_{j'}^2}{\min_{j' \in I_k} s_{j'}^2} = \mathcal{O}(1)$, which is satisfied even for the isotropic setting. The dimension-free R then implies that no exponential dependence is present in the computational complexity.
- In order to accurately estimate the gradient matrix, there is an additional cost in the statistical complexity. For the isotropic setting, (13) implies a sample complexity of $n = \Omega(d^k)$, which matches the sample size to achieve nontrivial gradient concentration as in Barak et al. (2022).
- On the other hand, if the input is anisotropic so that $\prod_{j' \in I_k} s_{j'}^2 \gg d^{-k}$ (the most extreme case is $\prod_{j' \in I_k} s_{j'}^2 = \Omega(1)$), then the sample complexity to estimate the informative direction is also improved. Indeed, if the signal is well-specified by the principle components of the input (i.e., denominator is $\Omega(1)$), then the sample complexity is $\tilde{\mathcal{O}}(k^2)$, and hence we avoid the dimension dependence. This observation also demonstrates the benefit of structured data in feature learning.

Tradeoff between statistical and computational complexity. By comparing the complexity derived in Corollary 1 and Theorem 3, we observe a "tradeoff" between the statistical and computational complexity: estimating the gradient covariance matrix requires additional samples, but consequently the required width and iterations of the MFLD significantly decrease. An interesting question is whether such tradeoff naturally occurs in more general data settings and feature learning procedures.

D Experiment

We validate our theoretical analysis by numerical experiments. We considered an anisotropic d-dimensional 3-sparse parity problem: $y = z_1 z_2 z_3$, $s_1 = s_2 = s_3 = \alpha/\sqrt{d}$, and $s_4 = \cdots = s_d = 1/\sqrt{d}$ (note that α is not defined as an exponent of the signal-to-noise ratio, $s_1/s_4 = d^\alpha$, but is defined just as the ratio $s_1/s_4 = \alpha$). Here α controls the alignment of the distribution to the feature, or the signal-to-noise ratio. We fixed the dimension d to 300, and varied n and α. We trained the neural network (2) with $\bar{R} = 15$. Specifically, we employed the width $N = 2000$ as a finite neuron approximation, and initialized neurons so that each of them followed the standard normal distribution (and thus the network was rotation invariant at the initialization). By using the logistic loss, we updated the network by the discretized MFLD (4) by setting $\eta = 0.25$, $\lambda_1 = 0.1$, and $\lambda = 0.1\alpha^2/d$ (fixed during the training) by following Corollary 1, until $T = 10000$. We ran the experiment 5 times with different seeds and plotted the mean for each n and α.

In Figure 1 we plot the test accuracy as a function of the sample size n and α, which controls the level of anisotropy. As clearly seen, increasing α enables smaller the model to learn the problem with smaller sample complexity n, which demonstrates how anisotropy helps learning. Moreover, let us focus on the "phase transition" boundary between yellow and blue regions. According to Corollary 1, the classification error is bounded by $\sum_{j\in I_k} s_j^{-2}/n = \alpha^{-2}d/n$ up to a constant, which predicts that there would be a boundary around $\alpha^2 = \Theta(n)$, as indicated by the red line in the figure. We therefore conclude that the empirical findings match the theoretical result in Corollary 1.

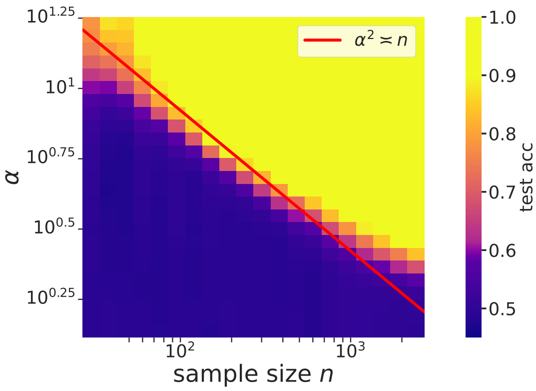

Figure 1: Test accuracy of NN trained by MFLD to learn the anisotropic d-dimensional 3-parity problem.

E Proofs of Propositions 3 and 4 and Corollary 1

Proof of Proposition 3. We follow the proof strategy from Suzuki et al. (2023b). Remember that

$$h_x(z) = \bar{R}[\tanh(z^\top x_1 + x_2) + 2\tanh(x_3)]/3.$$

Let $b_i = 2i - k$ for $i = 0, \ldots, k$, let $\zeta > 0$ be the positive real such that $\mathbb{E}_{u\sim N(0,1)}[2\tanh(\zeta + u)] = 1$ (note that, this also yields $\mathbb{E}_{u\sim N(0,1)}[2\tanh(-\zeta + u)] = -1$ by the symmetric property of tanh and the Gaussian distribution). Let

$$\Sigma := \begin{pmatrix} I/(2\lambda_1) & 0 & 0 \\ 0 & 1/(2\lambda_1) & 0 \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{(d+1+1)\times(d+1+1)},$$

and $\rho > 1$ be a constant which will be adjusted later on. Then, for $\xi_{2j} := [\log(\rho k)\phi^\top, -\log(\rho k)(b_j - 1), \zeta]^\top \in \mathbb{R}^{\bar{d}}$ and $\xi_{2j+1} := -[\log(\rho k)\phi^\top, -\log(\rho k)(b_j + 1), \zeta]^\top \in \mathbb{R}^{\bar{d}}$ for $j = 0, \ldots, k$, we define

$$\hat{\mu}_{2j} := N(\xi_{2j}, \Sigma), \quad \hat{\mu}_{2j+1} := N(\xi_{2j+1}, \Sigma).$$

Then, by noticing that for $z \in \mathrm{supp}(P_Z)$ there exists $\tilde{z} \in \{\pm 1/\sqrt{d}\}^d$ such that $z = A\tilde{z}$, we can see that

$$\mathbb{E}_{x\sim\hat{\mu}_{2j}}[h_x(z)] = \bar{R}\mathbb{E}_{u\sim N(0,1/\lambda_1)}\{\tanh[\log(\rho k)(\langle\tilde{\phi}, \tilde{z}\rangle - (b_j - 1)) + u] + 1\}/3$$

because we have

$$\begin{aligned}\langle x_1, z\rangle + x_2 &= \log(\rho k)(\langle\phi, z\rangle - (b_j - 1)) + \sum_{i=1}^d u_i z_i + u_{d+1} \\ &= \log(\rho k)(\langle A^{-1}\tilde{\phi}, A\tilde{z}\rangle - (b_j - 1)) + \sum_{i=1}^d u_i z_i + u_{d+1},\end{aligned}$$

for $x \sim N([\phi^\top, (b_j - 1)]^\top, I/(2\lambda_1))$ where $u_i \sim N(0, 1/(2\lambda_1))$ (i.i.d.) and $\sum_{i=1}^d u_i z_i + u_{d+1}$ obeys the Gaussian distribution with mean 0 and variance $\frac{1}{2\lambda_1}\|z\|^2 + \frac{1}{2\lambda_1} = \frac{1}{2\lambda_1}(1 + \|z\|^2) = \frac{1}{\lambda_1}$ for all $z \in \mathrm{supp}(P_Z)$, where we used the assumption on A. In the same vein, we also have

$$\mathbb{E}_{x\sim\hat{\mu}_{2j+1}}[h_x(z)] = -\bar{R}\mathbb{E}_{u\sim N(0,1/\lambda_1)}\{\tanh[\log(\rho k)(\langle\tilde{\phi}, \tilde{z}\rangle - (b_j + 1)) + u] + 1\}/3.$$

Here, define $|\tilde{z}| := |\{i \in I_k \mid \tilde{z}_i > 0\}|$ for $\tilde{z} \in \mathrm{supp}(P_{\tilde{Z}})$ which is the number of positive elements of z in the informative index set I_k. For a fixed number $j \in \{0, \ldots, k\}$, we let

$$f_1(z; u) = \{\tanh[\log(\rho k)(\langle\tilde{\phi}, \tilde{z}\rangle - (b_j - 1)) + u] + 1\}/3,$$
$$f_2(z; u) = \{\tanh[\log(\rho k)(\langle\tilde{\phi}, \tilde{z}\rangle - (b_j + 1)) + u] + 1\}/3,$$

then we can see that

$$f_1(z;0) = \begin{cases} O(1/(\rho k)) & (|\tilde{z}| < j), \\ 1 - O(1/(\rho k)) & (|\tilde{z}| \geq j), \end{cases}$$

and

$$f_2(z;0) = \begin{cases} O(1/(\rho k)) & (|\tilde{z}| < j+1), \\ 1 - O(1/(\rho k)) & (|\tilde{z}| \geq j+1), \end{cases}$$

because $\langle \tilde{\phi}, \tilde{z} \rangle - b_j = \sum_{j'=1}^{k} \mathrm{sign}(\tilde{z}_{j'}) - b_j = 2|\tilde{z}| - k - b_j = 2(|\tilde{z}| - j)$. Hence, we have that

$$f(z;u) := f_1(z;u) - f_2(z;u) = \begin{cases} \Omega(1) & (|\tilde{z}| = j), \\ O(1/(\rho k)) & (\text{otherwise}), \end{cases}$$

and $f(z;u) > 0$ for $|\tilde{z}| = j$. Then, since $\tanh(u) + 1 = \frac{e^u - e^{-u}}{e^u + e^{-u}} + 1 = \frac{2}{1+e^{-2u}}$, if $|\tilde{z}| = j$ and $|u| \leq 1/\lambda_1$,

$$f(z;u) \geq \Omega(1),$$

and if $|\tilde{z}| \neq j$ and $|u| \leq \log(\rho k)/2$,

$$f(z;u) \leq O(1/(\rho k)).$$

Therefore, when $|\tilde{z}| = j$,

$$\mathbb{E}_{u \sim N(0,1/\lambda_1)}[f(z;u)] \geq \int_{-1/\lambda_1}^{1/\lambda_1} f(z;u) g(u) \mathrm{d}u > \Omega(1).$$

where g is the density function of $N(0, 1/\lambda_1)$, and when $|\tilde{z}| \neq j$,

$$\begin{aligned} \mathbb{E}_{u \sim N(0,1/\lambda_1)}[f(z;u)] &\leq \int_{-\log(\rho k)/2}^{\log(\rho k)/2} f(z;u) g(u) \mathrm{d}u + \int_{|u| \geq \log(\rho k)/2} f(z;u) g(u) \mathrm{d}z \\ &\leq O(1/(\rho k)) + O\left(\frac{\exp(-\lambda_1 \log(\rho k)^2/2)}{\log(\rho k)} \right) \\ &= O(1/(\rho k)), \end{aligned}$$

where we used the upper-tail inequality of the Gaussian distribution in the second inequality. Hence, it holds that

$$\hat{f}_i(z) := \mathbb{E}_{x \sim \hat{\mu}_{2i}}[h_x(z)] + \mathbb{E}_{x \sim \hat{\mu}_{2i+1}}[h_x(z)] = \begin{cases} \Omega(k) & (|\tilde{z}| = j), \\ O(1/\rho) & (\text{otherwise}), \end{cases}$$

because $\bar{R} = k$. Therefore, by taking $\rho > 1$ sufficiently large, we also have

$$\hat{f}(z) := \frac{1}{2(k+1)} \sum_{i=0}^{k} (-1)^i \hat{f}_i(z) = \begin{cases} \Omega(1) & (|\tilde{z}| \text{ is even}), \\ -\Omega(1) & (|\tilde{z}| \text{ is odd}), \end{cases}$$

where the constant hidden in $\Omega(\cdot)$ is uniform over any $|\tilde{z}|$. Hence, there exists $c_2' > 0$ such that $Y\hat{f}(Z) > c_2'$ almost surely. Then, if we let $\mu_{\langle a \rangle}(B) := \mu(aB)$ for $a \in \mathbb{R}$, a probability measure μ and a measurable set B, then we can see that $\hat{f}$ is represented as

$$\hat{f}(\cdot) = \mathbb{E}_{x \sim \mu^*}[h_x(\cdot)],$$

where

$$\mu^* = \frac{1}{2(k+1)} \sum_{i=0}^{k} (\hat{\mu}_{2i,\langle (-1)^i \rangle} + \hat{\mu}_{2i+1,\langle (-1)^i \rangle}).$$

Then, by letting $c_2 = \ell(0) - \ell(c_2')$, we have

$$L(\mu^*) \leq \ell(0) - c_2.$$

Next, we bound the KL-divergence between ν and μ^*. Notice that the convexity of KL-divergence yields that

$$\mathrm{KL}(\nu, \mu^*) \leq \frac{1}{2(k+1)} \sum_{i=0}^{k} (\mathrm{KL}(\nu, \hat{\mu}_{2i}) + \mathrm{KL}(\nu, \hat{\mu}_{2i+1}))$$

$$\leq \lambda_1 \log(\rho k)^2[\|\phi\|^2 + (\max_j |b_j| + 1)^2] + \log(1/(2\lambda_1)) + \lambda_1(1+\zeta^2)$$
$$= O\left(\log(k)^2 \left(\|\phi\|^2 + k^2\right)\right).$$

This gives the assertion. □

Next, we prove Proposition 4.

Proof of Proposition 4. The proof of this statement resembles Proposition 4 of Suzuki et al. (2023b). The key step in their proof is to show that the optimal solution satisfies

$$|f_{\mu[\lambda]}(z)| = |f_{\mu[\lambda]}(z')|$$

for any $z, z' \in \mathrm{supp}(P_Z)$. We prove that this still holds in our general setting. Let $T_A : \mathbb{R}^{\bar{d}} \to \mathbb{R}$ be

$$T_A x = (Ax_1, x_2, x_3),$$

where $x = (x_1, x_2, x_3)$ for $x_1 \in \mathbb{R}^d$, $x_2 \in \mathbb{R}$ and $x_3 \in \mathbb{R}$. Then, we can see that

$$f_\mu(z) = f_{T_{A\#}\mu}(\tilde{z})$$

for $\mu \in \mathcal{P}$ and $T_{A\#}$ is the push-forward with respect to T_A, and $z = A\tilde{z}$. Based on this coordinate transform, we can reduce the problem to the standard parity setting where the input obeys the uniform distribution on $\{\pm 1/\sqrt{d}\}^d$. According to this coordinate transform, the prior distribution ν is transformed to $\nu_A := T_{A\#}\nu$, which is again a normal distribution with mean 0 and variance $AA^\top/(2\lambda_1)$. We also let T_j be the map which flips the sign of the i-th coordinate. Then, the key argument in the proof of Suzuki et al. (2023b) is to show that

$$\mathrm{KL}(\nu_A||\mu) = K(\nu_A||T_{j\#}\mu)$$

for a measure $\mu \in \mathcal{P}$ (which is supposed to be $T_{A\#}\hat{\mu}$ for a population risk minimizer $\hat{\mu}$). This equality is true because the normal distribution is point symmetric. Indeed, we have

$$\mathrm{KL}(\nu_A||\mu) = \mathrm{KL}(T_{j\#}\nu_A||T_{j\#}\mu) = \mathrm{KL}(\nu_A||T_{j\#}\mu),$$

where the first equality is by the invariance of the KL-divergence against any bijective coordinate transform and the second equality is by the point symmetricity of the normal distribution. Then, following the same argument to Suzuki et al. (2023b), we obtain the assertion. □

Then, Proposition 1 can be obtained as a corollary of Proposition 3 where we set $A = \mathrm{diag}\left(s_1\sqrt{d}, s_2\sqrt{d}, \dots, s_d\sqrt{d}\right)$. For this setting, we can easily see that

$$\|\phi\|^2 = \sum_{j \in I_k} s_j^{-2}.$$

Combining with this evaluation and the fact

$$k = \sum_{i \in I_k} 1 = \sum_{i \in I_k} s_i s_i^{-1} \leq \sqrt{\sum_{i \in I_k} s_i^2}\sqrt{\sum_{i \in I_k} s_i^{-2}} \leq \sqrt{\sum_{i \in I_k} s_i^{-2}}$$

we obtain the assertion.

F Estimating the information matrix

Without loss of generality, we may take $I_k = \{1, \dots, k\}$. Let $\sigma(w^\top z) = h_x(z)$ for $(x_1, x_2, x_3) = (w, b_1, b_2)$ for a fixed b_1 and b_2. Then,

$$\sigma(w^\top z) = \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \underbrace{\sigma^{(\ell)}(0)}_{=:c_\ell} (w^\top z)^\ell.$$

Note that the gradient of the loss with respect to w_j can be written as

$$g_j(w) = \frac{1}{n}\sum_{i=1}^n \ell'(y_i f_{\mu_0}(z_i)) y_i z_j \sigma'(w^\top z).$$

Suppose that $f_{\mu_0}(z_i) = 0$, then noticing that $Y = \prod_{j \in I_k}(s_j^{-1} Z_j)$, its expectation can be expressed as

$$\bar{g}_j(w) := \mathbb{E}\left[\prod_{j' \in I_k} (s_{j'}^{-1} Z_{j'}) Z_j \sigma'(w^\top Z)\right].$$

(1) If $j \in I_k$, then we have that

$$\bar{g}_j(w) := s_j \prod_{j' \in I_k \setminus j} s_{j'}^{-1} \mathbb{E}\left[\prod_{j' \in I_k \setminus j} Z_{j'} \sigma'(w^\top Z)\right].$$

Then, by the Taylor expansion of σ, it holds that

$$\begin{aligned}
\bar{g}_j(w) =& s_j \prod_{j' \in I_k \setminus j} s_{j'}^{-1} \left(\sum_{\ell=0}^{k-1} \frac{1}{\ell!} \partial_{\tilde{\theta}}^{(\ell)} \mathbb{E}\left[\prod_{j' \in I_k \setminus j} Z_{j'} \sigma'((\tilde{\theta} w)^\top Z)\right] \Bigg|_{\tilde{\theta}=0} \right. \\
& \left. + \sum_{\ell=k}^{\infty} \frac{1}{\ell!} \partial_{\tilde{\theta}}^{(\ell)} \mathbb{E}\left[\prod_{j' \in I_k \setminus j} Z_{j'} \sigma'((\tilde{\theta} w)^\top Z)\right] \Bigg|_{\tilde{\theta}=0} \right) \\
=& s_j \prod_{j' \in I_k \setminus j} s_{j'}^{-1} \left(\mathbb{E}\left[\prod_{j' \in I_k \setminus j} Z_{j'} \frac{c_k}{(k-1)!} (w^\top Z)^{k-1}\right] \right. \\
& \left. + \sum_{\ell=k}^{\infty} \mathbb{E}\left[\prod_{j' \in I_k \setminus j} Z_{j'} \cdot \frac{c_{\ell+1}}{\ell!} (w^\top Z)^{\ell}\right] \right) \\
=& s_j \prod_{j' \in I_k \setminus j} s_{j'}^{-1} \left(\prod_{j' \in I_k \setminus j} s_{j'}^2 \frac{c_k}{(k-1)!} (k-1)! \prod_{j' \in I_k \setminus j} w_{j'} + \underbrace{\text{(higher order term)}}_{=:(a)} \right) \\
=& c_k \cdot \prod_{j' \in I_k} s_{j'} \cdot \prod_{j' \in I_k \setminus j} w_{j'} + \text{(higher order term)}.
\end{aligned}$$

The higher order term (a) in the above expression can be evaluated as

$$\begin{aligned}
& \sum_{\ell=k}^{\infty} \mathbb{E}\left[\prod_{j' \in I_k \setminus j} Z_{j'} \cdot \frac{c_{\ell+1}}{\ell!} (w^\top Z)^{\ell}\right] \\
=& \sum_{\ell=k}^{\infty} \mathbb{E}\left[\prod_{j' \in I_k \setminus j} Z_{j'} \cdot \frac{c_{\ell+1}}{\ell!} \left(\frac{\ell!}{(k-1)!(\ell-k+1)!} (k-1)! \prod_{j' \in I_k \setminus j} w_{j'} \cdot \prod_{j' \in I_k \setminus j} Z_{j'} \cdot (w^\top Z)^{\ell-k+1} \right. \right. \\
& \left. \left. + \text{(the terms orthogonal to } \textstyle\prod_{j' \in I_k \setminus j} Z_{j'}) \right) \right] \\
=& \sum_{\ell=k}^{\infty} \frac{1}{(\ell-k+1)!} \mathbb{E}\left[\left(\prod_{j' \in I_k \setminus j} Z_{j'} \right)^2 \cdot c_{\ell+1} \prod_{j' \in I_k \setminus j} w_{j'} \cdot (w^\top Z)^{\ell-k+1} \right] \\
=& \prod_{j' \in I_k \setminus j} s_{j'}^2 \cdot \prod_{j' \in I_k \setminus j} w_{j'} \cdot \sum_{\ell=k}^{\infty} \frac{1}{(\ell-k+1)!} c_{\ell+1} \mathbb{E}\left[(w^\top Z)^{\ell-k+1}\right] \\
\leq& \prod_{j' \in I_k \setminus j} s_{j'}^2 \cdot \prod_{j' \in I_k \setminus j} w_{j'} \cdot \sum_{\ell=k}^{\infty} c_{\ell+1} (c \|w \odot s\|)^{\ell-k+1} K \frac{(\ell-k+1)^{(\ell-k+1)/2}}{(\ell-k+1)!},
\end{aligned}$$

where we used the moment bound of sub-Gaussian random variables in the last inequality by noting that $w^\top Z$ is a sub-Gaussian random variable with parameter $\|w \odot s\|^2$, that is, a sub-Gaussian random variable X with a parameter s satisfied $\mathbb{E}[|X|^\ell] \leq (cs)^\ell \ell^{\ell/2}$ with an absolute constant c (see Proposition 2.5.2 of Vershynin (2020), for example). Then, by the Stirling's formula, the absolute value of the right hand side can be bounded by

$$\begin{aligned}
&K \prod_{j' \in I_k \setminus j} s_{j'}^2 \cdot \prod_{j' \in I_k \setminus j} |w_{j'}| \cdot \sum_{\ell=k}^{\infty} c_{\ell+1} \|w \odot s\|^{\ell-k+1} \frac{(\ell-k+1)^{(\ell-k+1)/2}}{\sqrt{2\pi}(\ell-k+1)^{\ell-k+1+1/2} e^{-(\ell-k+1)}} \\
=&K \prod_{j' \in I_k \setminus j} s_{j'}^2 \cdot \prod_{j' \in I_k \setminus j} |w_{j'}| \cdot \sum_{\ell=k}^{\infty} c_{\ell+1} \|w \odot s\|^{\ell-k+1} \frac{1}{\sqrt{2\pi}} \left(\frac{e}{(\ell-k+1)^{1/2}} \right)^{\ell-k+1} \frac{1}{(\ell-k+1)^{1/2}} \\
\leq& \frac{c_k}{2} \prod_{j' \in I_k \setminus j} s_{j'}^2 \cdot \prod_{j' \in I_k \setminus j} |w_{j'}|,
\end{aligned}$$

where we used the assumption $\|w \odot s\|$ is sufficiently small such that $\sum_{\ell=k}^{\infty} c_{\ell+1} (c\|w \odot s\|)^{\ell-k+1} \frac{1}{\sqrt{2\pi}} \left(\frac{e}{(\ell-k+1)^{1/2}} \right)^{\ell-k+1} \frac{1}{(\ell-k+1)^{1/2}} \leq \frac{c_k}{2}$. Therefore, we can see that

$$\begin{aligned}
\bar{g}_j(w) &= c_k \cdot \prod_{j' \in I_k} s_{j'} \cdot \prod_{j' \in I_k \setminus j} w_{j'} + (\text{higher order term}), \\
|\bar{g}_j(w)| &\geq \frac{c_k}{2} \cdot \prod_{j' \in I_k} s_{j'} \cdot \prod_{j' \in I_k \setminus j} |w_{j'}|, \\
|\bar{g}_j(w)| &\leq \frac{3}{2} c_k \cdot \prod_{j' \in I_k} s_{j'} \cdot \prod_{j' \in I_k \setminus j} |w_{j'}|. \qquad (14)
\end{aligned}$$

(2) In the same vein, we also have for $j \notin I_k$, we have that

$$\begin{aligned}
\bar{g}_j(w) &= c_{k+2} \cdot \prod_{j' \in I_k \cup j} s_{j'} \cdot \prod_{j' \in I_k \cup j} w_{j'} + (\text{higher order term}), \\
|\bar{g}_j(w)| &\leq 2c_{k+2} \cdot \prod_{j' \in I_k \cup j} s_{j'} \cdot \prod_{j' \in I_k \cup j} |w_{j'}|. \qquad (15)
\end{aligned}$$

Next, we show the concentration of the empirical gradient $g_j(w)$ around its expectation. We observe that

$$\begin{aligned}
&\sup_{Y,Z} |\ell'(Y f_{\mu_0}(Z)) Y Z_j \sigma'(w^\top Z)| \leq \bar{R} s_j, \\
&\mathrm{Var}_{Y,Z}[\ell'(Y f_{\mu_0}(Z)) Y Z_j \sigma'(w^\top Z)] \leq \bar{R}^2 s_j^2.
\end{aligned}$$

Therefore, by the Bernstein's inequality, we obtain that

$$P\left(|g_j(w) - \bar{g}_j(w)| \geq \frac{4\bar{R} s_j}{\sqrt{n}} \log(2/\delta) \right) \leq \delta$$

for any $\delta \in (0, 1)$. Hence, if we let n

$$n \geq \frac{16k^2 \bar{R}^2 \log(2N/\delta)^2}{\left(C_0 c_k \cdot \prod_{j' \in I_k} s_{j'} \right)^2},$$

for a sufficiently small constant C_0, then we have that

$$|g_j(w_l) - \bar{g}_j(w_l)| \leq C_0 c_k \prod_{j' \in I_k} s_{j'} \cdot s_j, \qquad (16)$$

uniformly over $l = 1, \ldots, N$ with probability δ.

For that purpose, we evaluate the expectations of $g_{j_1}(w) g_{j_2}(w)$ carefully. Let $H(w) = \sum_{\ell=k}^{\infty} \frac{c_{\ell+1}}{(\ell-k+1)!} \mathbb{E}_Z \left[(w^\top Z)^{\ell-k+1} \right] = \frac{1}{2} \|w \odot s\|^2 + \sum_{\ell=0}^{\infty} \frac{c_{k+4+2\ell}}{(4+2\ell)!} \mathbb{E}_Z \left[(w^\top Z)^{4+2\ell} \right]$. We evaluate for each condition on j_1 and j_2.

468 (a) If $j_1 = j_2 \in I_k$, then it holds that

$$\mathbb{E}_W[g_{j_1}(W)g_{j_1}(W)] = c_k^2 \prod_{j' \in I_k} s_{j'}^2 \mathbb{E}_W\left[\prod_{j' \in I_k \setminus j_1} W_{j'}^2 (1 + H(W))^2\right]$$
$$= \Omega\left(\prod_{j' \in I_k} s_{j'}^2\right).$$

(b) If $j_1 \neq j_2$ and $j_1, j_2 \in I_k$, then it holds that

$$\mathbb{E}_W[g_{j_1}(W)g_{j_2}(W)] = c_k^2 \prod_{j' \in I_k} s_{j'}^2 \mathbb{E}\left[\prod_{j' \in I_k \setminus \{j_1, j_2\}} W_{j'}^2 \cdot W_{j_1} W_{j_2} (1 + H(W))^2\right] = 0,$$

469 where we used that the distribution of W is symmetric and $H(W)$ satisfies $H(W) = H(-W)$.

(c) If $j_1 \neq j_2$ and $j_1 \in I_k$ and $j_2 \notin I_k$, then

$$\mathbb{E}_W[g_{j_1}(W)g_{j_2}(W)] = c_k c_{k+2} \prod_{j' \in I_k} s_{j'}^2 s_{j_2} \mathbb{E}\left[\prod_{j' \in I_k \setminus j_1} W_{j'}^2 \cdot W_{j_2} (1 + H(W))^2\right] = 0.$$

470 (d) If $j_1 \notin I_k$ and $j_2 \notin I_k$, then

$$\mathbb{E}_W[g_{j_1}(W)g_{j_2}(W)] = c_{k+2}^2 \prod_{j' \in I_k} s_{j'}^2 s_{j_1} s_{j_2} \mathbb{E}\left[\prod_{j' \in I_k} W_{j'}^2 \cdot W_{j_1} W_{j_2} (1 + H(W))^2\right]$$
$$= \begin{cases} 0 & (j_1 \neq j_2), \\ \mathcal{O}(\prod_{j' \in I_k \cup j_1} s_{j'}^2) & (j_1 = j_2). \end{cases}$$

Summarizing these evaluations, we can see that $\bar{G} = (\bar{G}_{j_1,j_2})_{j_1=1,j_2=1}^{d,d} \in \mathbb{R}^{d \times d}$ defined by

$$\bar{G}_{j_1,j_2} = \mathbb{E}_W[g_{j_1}(W)g_{j_2}(W)]$$

is a diagonal matrix where $\bar{G}_{j_1,j_1}$ for $j_1 \in I_k$ has larger values than that for $j_1 \notin I_k$. We define its empirical average version $G = (G_{j_1,j_2})_{j_1=1,j_2=1}^{d,d} \in \mathbb{R}^{d \times d}$ as

$$G_{j_1,j_2} = \frac{1}{N} \sum_{l=1}^{N} g_i(w_l) g_j(w_l).$$

471 Now, we show the concentration of G around its population version $\bar{G}$. Note that

$$\frac{1}{N}\sum_{l=1}^{N} g_{j_1}(w_l)g_{j_2}(w_l) = \frac{1}{N}\sum_{l=1}^{N}(g_{j_1}(w_l) - \bar{g}_{j_1}(w_l) + \bar{g}_{j_1}(w_l))(g_{j_2}(w_l) - \bar{g}_{j_2}(w_l) + \bar{g}_{j_2}(w_l))$$
$$= \frac{1}{N}\sum_{l=1}^{N}(g_{j_1}(w_l) - \bar{g}_{j_1}(w_l))(g_{j_2}(w_l) - \bar{g}_{j_2}(w_l))$$
$$+ \frac{1}{N}\sum_{l=1}^{N}(g_{j_1}(w_l) - \bar{g}_{j_1}(w_l))\bar{g}_{j_2}(w_l)$$
$$+ \frac{1}{N}\sum_{l=1}^{N}(g_{j_2}(w_l) - \bar{g}_{j_2}(w_l))\bar{g}_{j_1}(w_l)$$
$$+ \frac{1}{N}\sum_{l=1}^{N}\bar{g}_{j_1}(w_l)\bar{g}_{j_2}(w_l).$$

Then, by the concentration bound (16) and the bounds (14) and (15) of $\bar{g}_j(w)$, $\Delta G_{j_1,j_2} = G_{j_1,j_2} - \bar{G}_{j_1,j_2}$ satisfies

$$\Delta G_{j_1,j_2} = \frac{1}{N}\sum_{l=1}^{N} \bar{g}_{j_1}(w_l)\bar{g}_{j_2}(w_l) - \bar{G}_{j_1,j_2} + \begin{cases} \mathcal{O}\left(C_0 \frac{1}{k}\prod_{j'\in I_k} s_{j'}^2 \cdot \max_{j'\in I_k} s_{j'}\right) & (j_1, j_2 \in I_k), \\ \mathcal{O}\left(C_0 \prod_{j'\in I_k} s_{j'}^2 \cdot s_{j_2}^2\right) & (j_1 \in I_k,\ j_2 \notin I_k), \\ \mathcal{O}\left(C_0 \prod_{j'\in I_k} s_{j'}^2 \cdot \max_{j'\in I_k^c} s_{j'} \max\{s_{j_1}, s_{j_2}\}\right) & (j_1, j_2 \notin I_k). \end{cases}$$

In addition to that, if we write $\hat{G}_{j_1,j_2} = \frac{1}{N}\sum_{l=1}^{N} \bar{g}_{j_1}(w_l)\bar{g}_{j_2}(w_l)$, then the matrix Bernstein's inequality yields that

$$P\left[\|\hat{G} - \bar{G}\|_{\mathrm{op}} \geq K\left(\sqrt{\frac{Q^2(t+\log(d))}{N}} + \frac{(t+\log(d))Q}{N}\right)\right] \leq \exp(-t),$$

where K is an absolute constant and $Q = d\prod_{j'\in I_k} s_{j'}^2$ because $\|\bar{g}(w_l)\bar{g}^\top(w_l)\|_{\mathrm{op}} \leq O(Q)$. Therefore, $N = \Omega(d\log(d/\delta)/(C_0 \max_{j'\notin I_k} s_{j'}^4))$ for sufficiently small C_0 yields that

$$\|G - \bar{G}\|_{\mathrm{op}} = \mathcal{O}\left(C_0 \prod_{j'\in I_k} s_{j'}^2 \cdot \max_{j'\notin I_k} s_{j'}^2\right),$$

with probability $1-\delta$.

Therefore, if we let $Q_1 = \prod_{j'\in I_k} s_{j'}^2$ and $Q_2 = \prod_{j'\in I_k} s_{j'}^2 \cdot \max_{j'\notin I_k} s_{j'}^2$, then it holds that

$$G_{j_1,j_1} = \begin{cases} \Theta(Q_1) & (j_i \in I_k), \\ \mathcal{O}(Q_2) & (j_1 \notin I_k). \end{cases}$$

If we let $\check{Q}_1 = \frac{1}{k}\prod_{j'\in I_k} s_{j'}^2 \cdot \max_{j'\in I_k^c} s_{j'}$, and $\check{Q}_2 = \frac{1}{k}\prod_{j'\in I_k} s_{j'}^2 \cdot \max_{j'\in I_k^c} s_{j'}^2$, then, for $j_1 \neq j_2$, it holds that

$$G_{j_1,j_2} = \begin{cases} \mathcal{O}(C_0\check{Q}_1) & (j_1 \in I_k \text{ and } j_2 \in I_k), \\ \mathcal{O}(C_0\check{Q}_2) & (\text{otherwise}). \end{cases}$$

Then, by modifying the objective as

$$L(\mu) + \lambda_1 \mathbb{E}_\mu[\|X\|^2_{(G+\hat{\lambda}_0 I)^{-1}}]$$

with a regularization parameter $\hat{\lambda}_0 = \check{Q}_2$. This is equivalent to the alternative objective $L(\mu) + \lambda_1\mathbb{E}_\mu[\|X\|^2]$ where the input is transformed as $Z \leftarrow A\tilde{Z}$ where $A = c_A\sqrt{G + \hat{\lambda}_0 I}B$ with $B = \mathrm{diag}\left(s_1\sqrt{d}, \ldots, s_d\sqrt{d}\right)$ and a constant $c_A = \mathcal{O}(\check{Q}_1^{-1/2}(\max_{j'} s_{j'})^{-1})$ such that $\|A\tilde{Z}\| \leq 1$. Then, we can see that

$$\|A^{-1}\tilde{\phi}\|^2 = c_A^{-2}\tilde{\phi}^\top B^{-1}(G + \hat{\lambda}_0 I)^{-1}B^{-1}\tilde{\phi} = c_A^{-2}\zeta_s^\top(G + \hat{\lambda}_0 I)^{-1}\zeta_s,$$

for $\zeta_s = (s_1^{-1}, \ldots, s_k^{-1}, 0, \ldots, 0)^\top$. Now, let

$$G + \hat{\lambda}_0 = \begin{pmatrix} G_{[1,1]} & G_{[1,2]} \\ G_{[2,1]} & G_{[2,2]} \end{pmatrix}.$$

Then, we can see that

$$(G + \hat{\lambda}_0)^{-1} = \begin{pmatrix} (G_{[1,1]} - G_{[1,2]}G_{[2,2]}^{-1}G_{[2,1]})^{-1} & * \\ * & * \end{pmatrix}.$$

We know that $\|G_{[2,2]}^{-1}\|_{\mathrm{op}} \leq \check{Q}_2^{-1}$ and $\|G_{[1,2]}\|_{\mathrm{op}} \leq C_0\sqrt{k\check{Q}_1^2 + (d-k)\check{Q}_2^2} \leq \sqrt{d}\max_{j'\in I_k^c} s_{j'}^2\check{Q}_1 = \sqrt{d}\check{Q}_2$. Hence, we can see that

$$G_{[1,1]} - G_{[1,2]}G_{[2,2]}^{-1}G_{[2,1]} \gtrsim \check{Q}_1 - \mathcal{O}(C_0 d\check{Q}_2).$$

Hence, by taking C_0 sufficiently small and under the assumption that $d \max_{j' \in I_k^c} s_{j'}^2 = O(1)$, we have that

$$(G_{[1,1]} - G_{[1,2]} G_{[2,2]}^{-1} G_{[2,1]})^{-1} \precsim \check{Q}_1^{-1} I.$$

Therefore, we finally arrive at

$$\|A^{-1}\tilde{\phi}\|^2 \le c_A^{-2} \|\zeta_s\|^2 \|(G_{[1,1]} - G_{[1,2]} G_{[2,2]}^{-1} G_{[2,1]})^{-1}\|_{\mathrm{op}}$$
$$\lesssim k \left(\min_{j' \in I_k} s_{j'}^2 \right)^{-1} \check{Q}_1^{-1} \check{Q}_1 \left(\max_{j'} s_{j'}^2 \right) = k \frac{\max_{j' \in [d]} s_{j'}^2}{\min_{j' \in I_k} s_{j'}^2}.$$

G Kernel lower bound

In this section, we derive the kernel lower bound for the k-parity classification problem (Example 2) with the spiked covariance setting. Before beginning the proof, we slightly change the notation. We assume $y = y(z) = \mathrm{sign}(\prod_{i=1}^k z_i)$, each z_i is independent, and $\mathbb{P}[z_i = \pm d^\alpha] = \frac{1}{2}$ $(i = 1, \cdots, k)$ or $\mathbb{P}[z_i = \pm 1] = \frac{1}{2}$ $(i = k+1, \cdots, d)$ for $0 \le \alpha < \frac{1}{2}$. This definition multiplies $\sqrt{d}$ to z compared to the original definition of the spiked covariance setting in the main text. This is because we intend to make the notation match to the previous literature on the kernel lower bounds like Wei et al. (2019) and Misiakiewicz (2022).

We consider the following inner-product Kernel, with positive and bounded coefficients $\{\alpha_0\}_{l=0}^\infty$.

$$K(z, z') = \sum_{l=0}^{\infty} \alpha_l \left(\frac{z^\top z'}{d} \right)^l$$

Based on the randomly drawn n sample, we construct the estimator $f_\beta(z)$ with $\beta \in \mathbb{R}^n$.

$$f_\beta(z) = \sum_{i=1}^{n} \beta_i K(z, z^i)$$

Then, the following lower bound on the accuracy of f_β can be obtained.

Theorem 2. *Fix $\delta > 0$ arbitrarily. For sufficiently large d, draw $n \lesssim d^{\lfloor (1-2\alpha)k \rfloor - \delta}$ sample. Then, with probability at least 0.99 over the sample, for all choices of $\beta \in \mathbb{R}^n$, $f_\beta = \sum_{i=1}^n \beta_i K(z, z^i)$ will predict the sign of y wrong $\Omega(1)$ fraction of the time:*

$$\mathbb{P}_{z \sim P_Z}\left[f_\beta(z) y < 0 \right] = \Omega(1).$$

The proof is divided into two steps. First, we translate the event when prediction fails into when the value of $|f_\beta(z)|$ is away from zero. We combine the proof for 2-parity (Wei et al., 2019) and an additional observation that $K(z, z^i)$ have d^{-k} correlation to y, to get the tighter bound for general higher order parities than (Wei et al., 2019). Then, we show that the probability of that event is evaluated by the the smallest eigenvalue of some other Kernel matrix defined in Lemma 3. Finally, we apply the lower bound of the smallest eigenvalue using (Misiakiewicz, 2022).

Note that, proving Theorem 2 for $\frac{1}{2} - \frac{2}{2k} < \alpha \le \frac{1}{2}$ means nothing. Thus in the following we assume $\frac{1}{2} - \alpha$ is not to small so that $d^{(1-2\alpha)} \gtrsim \log^{2k+1}(d)$.

Lemma 1. *For $n \le d^{(1-2\alpha)k}$, with probability $1 - \exp(-\Omega(d))$ over the random draws of the training sample, we have*

$$\mathbb{P}_{z \sim P_Z}\left[f_\beta(z) y < 0 \right] \gtrsim \mathbb{P}_{z \sim P_Z}\left[|f_\beta(z)| \ge \frac{c}{d^{(1-2\alpha)k}} \sum_{i=1}^{n} |\beta_i| \right] - 1/d,$$

where c is a constant depending on k and $\{\alpha_l\}_l$.

Proof. Randomly draw $z_{k+1:d}$, and fix it for the moment. Suppose $f_\beta(z) y(z) \ge 0$ for all choices of $z_{1:k}$ and $|f_\beta(z)| \gtrsim \frac{c}{d^{(1-2\alpha)k}} \sum_{i=1}^n |\beta_i|$ for some $z_{1:k}$ to show contradiction (with high probability). Then, consider the average of $K(z, z^i) y$ over the choices of $z_{1:k}$ as follows:

$$\mathbb{E}_{z_{1:k}}\left[K(z, z^i) y(z) \middle| z_{k+1:d} \right] = \mathbb{E}_{z_{1:k}}\left[\sum_{l=0}^{\infty} \alpha_l \left(\frac{z^\top z^i}{d} \right)^l y(z) \middle| z_{k+1:d} \right]$$

$$= \sum_{l=k}^{\infty} \alpha_l \mathbb{E}_{z_{1:k}} \left[\left(\frac{z^\top z^i}{d} \right)^l \prod_{j'=1}^{k} z_{j'} \middle| z_{k+1:d} \right] \tag{17}$$

Let us evaluate $\mathbb{E}_{z_{1:k}}[(\frac{z^\top z^i}{d})^l \prod_{j'=1}^k z_{j'} | z_{k+1:d}]$. For $k \le l \le 2k$, we expand $(\frac{z^\top z^i}{d})^l = (\sum_{i=j}^d \frac{z_j z_j^i}{d})^l$ to see

$$\mathbb{E}_{z_{1:k}} \left[\left(\frac{z^\top z^i}{d} \right)^l \prod_{j'=1}^{k} z_{j'} \middle| z_{k+1:d} \right] \le \underbrace{\sum_{l'=k}^{l} {}_l C_{l'} k^{l'} (d-k)^{l-l'}}_{\text{consider terms containing each } z_1, \cdots, z_k \text{ more than or equal to once}} \left(\frac{d^{2\alpha}}{d} \right)^{l'} \left(\frac{1}{d} \right)^{l-l'} \lesssim d^{-(1-2\alpha)k}.$$

For $l \ge 2k+1$, we have $|\frac{z^\top z^i}{d}| \lesssim d^{-(1-2\alpha)/2}\sqrt{\log d}$ with probability $1 - 1/d^{(1-2\alpha)k+1}$ over the choice of $z_{k+1:d}$, and therefore $\sum_{l=2k+1}^{\infty} \mathbb{E}_{z_{1:k}}[|\frac{z^\top z^i}{d}|^l | z_{k+1:d}] \lesssim d^{-(1-2\alpha)k}$. By using them for (17), we have

$$\mathbb{E}_{z_{1:k}} \left[K(z, z^i) y(z) \middle| z_{k+1:d} \right] = \mathbb{E}_{z_{1:k}} \left[\sum_{l=0}^{\infty} \alpha_l \left(\frac{z^\top z^i}{d} \right)^l y(z) \middle| z_{k+1:d} \right] \lesssim d^{-(1-2\alpha)k}$$

for randomly drawn $z_{k+1:d}$, with probability more than $1 - 1/d^{(1-2\alpha)k+1}$. Therefore,

$$\mathbb{E}_{z_{1:k}} \left[f_\beta(z) y(z) \middle| z_{k+1:d} \right] = \mathbb{E}_{z_{1:k}} \left[\sum_i \beta_i K(z, z^i) y(z) \middle| z_{k+1:d} \right] \lesssim \frac{1}{d^{(1-2\alpha)k}} \sum_{i=1}^{n} |\beta_i| \tag{18}$$

with probability more than $1 - 1/d$.

On the other hand, if $f_\beta(z) y(z) \ge 0$ for all $z_{1:k}$ and $|f_\beta(z)| \gtrsim \frac{c}{d^{1-2\alpha}} \sum_{i=1}^n |\beta_i|$ for some $z_{1:k}$, we have

$$\mathbb{E}_{z_{1:k}} \left[f_\beta(z) y(z) \middle| z_{k+1:d} \right] = \frac{1}{2^k} \sum_{z_{1:k}} f_\beta(z) y(z) \ge \frac{1}{2^k} \cdot \frac{c}{d^{(1-2\alpha)k}} \sum_{i=1}^{n} |\beta_i|. \tag{19}$$

By comparing (18) and (19), we have the contradiction for more than $1 - 1/d$ probability of the choice of $z_{k+1:d}$ by taking c sufficiently large. Therefore, if $|f_\beta(z)| \gtrsim \frac{c}{d^{(1-2\alpha)k}} \sum_{i=1}^n |\beta_i|$ for some $z_{1:k}$, there exists some $z_{1:k}$ that yields $f_\beta(z) y < 0$, for $z_{k+1:d}$ that is drawn with probability more than $1 - 1/d$, which yields the conclusion. □

From now, we evaluate the probability $\mathbb{P}_{z \sim P_Z}[|f_\beta(z)| \ge \frac{c}{d^{(1-2\alpha)k}} \sum_{i=1}^n |\beta_i|]$. However, $f_\beta(z)$ can have very high order term, so we approximate $f_\beta(z)$ as follows.

Lemma 2. *Let us define* $g_1 : [-1, 1] \to \mathbb{R}$ *as*

$$g_1(t) = \sum_{l=0}^{2k} \alpha_l t^l.$$

Suppose $n \le d^{(1-2\alpha)k}$. *Then,*

$$\mathbb{P}_{z \sim P_Z} \left[\exists i \in [n], \left| K(z, z^i) - g_1 \left(\frac{z^\top z^i}{d} \right) \right| \le d^{-(1-2\alpha)k} \right] \ge 1 - 1/d.$$

Proof. First, we note

$$\left| K(z, z^i) - g_1 \left(\frac{z^\top z^i}{d} \right) \right| = \left| \sum_{l=0}^{\infty} \alpha_l \left(\frac{z^\top z^i}{d} \right) - \sum_{l=0}^{2k} \alpha_l \left(\frac{z^\top z^i}{d} \right) \right| = \sum_{2k+1}^{\infty} \alpha_l \left| \frac{z^\top z^i}{d} \right|. \tag{20}$$

With probability $1 - 1/d^{(1-2\alpha)k+1}$, $\left| \frac{z^\top z^i}{d} \right| \lesssim d^{-(1-2\alpha)/2}\sqrt{\log d}$. This means that (20) is bounded by $\lesssim \left(\frac{\log d}{d^{1-2\alpha}} \right)^{(2k+1)/2} \le d^{-(1-2\alpha)k}$ for sufficiently large d. By taking the uniform bound over all i, we get the assertion. □

Because of this lemma, all we need is to bound $\mathbb{P}_{z\sim P_Z}[|\sum_{i=1}^n \beta_i g_1(\frac{z^\top z^i}{d})| \geq \frac{c}{d^{(1-2\alpha)k}} \sum_{i=1}^n |\beta_i|]$ by $\Omega(1)$, because

$$\mathbb{P}_{z\sim P_Z}\left[|f_\beta(z)| \geq \frac{c}{d^{(1-2\alpha)k}} \sum_{i=1}^n |\beta_i|\right] \geq \mathbb{P}_{z\sim P_Z}\left[\left|\sum_{i=1}^n \beta_i g_1\left(\frac{z^\top z^i}{d}\right)\right| \geq \frac{c+1}{d^{(1-2\alpha)k}} \sum_{i=1}^n |\beta_i|\right] - 1/d.$$

For this, we lower bound the second moment, which captures variation of f_β.

Lemma 3. *Suppose a_l are all positive and define $K_2 \in \mathbb{R}^{n\times n}$ as*

$$(K_2)_{i,j} = \sum_{l=0}^k \left(\frac{{z^i_{k+1:d}}^\top z^j_{k+1:d}}{d-k}\right)^l.$$

Then, for sufficiently large d, we have

$$\mathbb{E}_z\left[\left(\sum_{i=1}^n \beta_i g_1\left(\frac{z^\top z^i}{d}\right)\right)^2\right] \gtrsim d^{-\lfloor(1-2\alpha)k\rfloor} \beta^\top K_2 \beta.$$

The proof requires several auxiliary lemmas as follows. We defer the proofs of them after the proof of Lemma 3.

Lemma 4. *For any integers $p, g \geq 0$,*

$$\begin{aligned}
&\mathbb{E}_z\left[\left(\sum_{i=1}^n \beta_i (z^\top z^i)^p\right)\left(\sum_{i=1}^n \beta_i (z^\top z^i)^q\right)\right] \\
&\geq \mathbb{E}_{z_{k+1:d}}\left[\left(\sum_{i=1}^n \beta_i (z^\top_{k+1:d} z^i_{k+1:d})^p\right)\left(\sum_{i=1}^n \beta_i (z^\top_{k+1:d} z^i_{k+1:d})^q\right)\right] \geq 0
\end{aligned}$$

Lemma 5. *Let $z^i, z^j \in \{-1,1\}^d$, $z \in \{-1,1\}^d$ be a vector sampled uniformly from the hypercube, and let l be any integer. Then, we can expand the expectation as*

$$\mathbb{E}_z\left[\left(\frac{z^\top z^i}{d}\right)^l \left(\frac{z^\top z^j}{d}\right)^l\right] = \sum_{l'=0}^l d^{-l} c_{d,l,l'} \left(\frac{{z^i}^\top z^j}{d}\right)^{l'}.$$

Furthermore, for sufficiently large d, $c_{d,l,l'} \geq 0$ and especially $c_{d,l,l} = (l!)^2$.

Proof of Lemma 3. Let us first expand the target:

$$\begin{aligned}
&\mathbb{E}_z\left[\left(\sum_{i=1}^n \beta_i g_1\left(\frac{z^\top z^i}{d}\right)\right)^2\right] \\
&= \mathbb{E}_z\left[\left(\sum_{i=1}^n \beta_i \sum_{l=0}^{2k} \alpha_l \left(\frac{z^\top z^i}{d}\right)^l\right)^2\right] \\
&= \mathbb{E}_z\left[\left(\sum_{l=0}^{2k} \alpha_l \sum_{i=1}^n \beta_i \left(\frac{z^\top z^i}{d}\right)^l\right)^2\right] \\
&= \sum_{0\leq l_1,l_2\leq 2k} \alpha_{l_1}\alpha_{l_2} \mathbb{E}_z\left[\left(\sum_{i=1}^n \beta_i \left(\frac{z^\top z^i}{d}\right)^{l_1}\right)\left(\sum_{i=1}^n \beta_i \left(\frac{z^\top z^i}{d}\right)^{l_2}\right)\right] \qquad (21)
\end{aligned}$$

From Lemma 4 and $\alpha_{l_1}, \alpha_{l_2} > 0$, each term is non-negative and (21) is lower bounded by

$$\sum_{l=0}^{2k} \alpha_l^2 \mathbb{E}_{z_{k+1:d}}\left[\left(\sum_{i=1}^n \beta_i \left(\frac{z^\top_{k+1:d} z^i_{k+1:d}}{d}\right)^l\right)^2\right] \qquad (22)$$

$$\gtrsim \sum_{l=0}^{2k} \alpha_l^2 \mathbb{E}_{z_{k+1:d}} \left[\left(\sum_{i=1}^{n} \beta_i \left(\frac{z_{k+1:d}^\top z_{k+1:d}^i}{d-k} \right) \right)^{l\,2} \right]. \tag{23}$$

Let us define a matrix $K_1 \in \mathbb{R}^{n\times n}$ so that (23) is equal to $\beta^\top K_1 \beta$. For that, we define

$$(K_1)_{i,j} = \sum_{l=0}^{2k} a_l^2 \mathbb{E}_{z_{k+1:d}} \left[\left(\frac{z_{k+1:d}^\top z_{k+1:d}^i}{d-k} \right)^l \left(\frac{z_{k+1:d}^\top z_{k+1:d}^j}{d-k} \right)^l \right].$$

According to Lemma 5,

$$\begin{aligned}(K_1)_{i,j} &= \sum_{l=0}^{2k} a_l^2 \sum_{l'=0}^{l} (d-k)^{-l} c_{d-k,l,l'} \left(\frac{{z_{k+1:d}^i}^\top z_{k+1:d}^j}{d} \right)^{l'} \\ &= \sum_{l=0}^{2k} \left(\sum_{l''=l}^{2k} a_{l''}^2 (d-k)^{-l''} c_{d-k,l'',l} \right) \left(\frac{{z_{k+1:d}^i}^\top z_{k+1:d}^j}{d-k} \right)^{l}.\end{aligned}$$

Because $c_{d-k,l'',l} \geq 0$ and $c_{d-k,l,l} = (l!)^2$, $(d-k)^{-l} c_l := \left(\sum_{l''=l}^{2k} a_{l''}^2 (d-k)^{-l''} c_{d-k,l'',l} \right) \gtrsim d^{-l}$ holds. Thus, we have $(d-k)^{-l} c_l \geq d^{-\lfloor (1-2\alpha)k \rfloor} c$ for all $l \leq \lfloor (1-2\alpha)k \rfloor$ for sufficiently small c, and by defining $K_2, K_3 \in \mathbb{R}^{n\times n}$ as

$$\begin{aligned}(K_2)_{i,j} &= \sum_{l=0}^{\lfloor (1-2\alpha)k \rfloor} \left(\frac{{z_{k+1:d}^i}^\top z^j}{d-k} \right)^l \\ (K_3)_{i,j} &= \sum_{l=0}^{\lfloor (1-2\alpha)k \rfloor - 1} ((d-k)^{-l} c_l - d^{-(\lfloor (1-2\alpha)k \rfloor - 1)} c) \left(\frac{{z_{k+1:d}^i}^\top z_{k+1:d}^j}{d-k} \right)^l \\ &\quad + \sum_{l=\lfloor (1-2\alpha)k \rfloor + 1}^{2k} (d-k)^{-l} c_l \left(\frac{{z_{k+1:d}^i}^\top z_{k+1:d}^j}{d-k} \right)^l,\end{aligned}$$

we have $K_1 = c d^{-\lfloor (1-2\alpha)k \rfloor} K_2 + K_3$. Moreover, K_3 is positive semi-definite because K_3 is written as a sum of polynomial kernels with positive coefficients. Thus, we can lower bound $\beta^\top K_1 \beta$ by $d^{-\lfloor (1-2\alpha)k \rfloor} \beta^\top K_2 \beta$ (up to a constant factor). □

Proof of Lemma 4. The basic idea comes from Lemma B.9. of Wei et al. (2019). For a set $S \subseteq [k]$, we let $z^S = \prod_{i=1}^k z_i$, and for a set $T \subseteq [d] \setminus [k]$, we let $z^T = \prod_{i=1}^k z_i$. Expand $(z^\top z^i)^p$ as

$$(z^\top z^i)^p = \left(\sum_{j=1}^{d} z_j z_j^i \right)^p = \sum_{S,T} C_{|S|,|T|,p} z^S z^T (z^i)^S (z^i)^T,$$

where $c_{|S|,|T|,p} \geq 0$ depends only on $|S|$, T, and p considering the symmetry. Also, we let

$$(z_{k+1:d}^\top z_{k+1:d}^i)^p = \sum_T \bar{C}_{|T|,p} z_{k+1:d}^S z_{k+1:d}^T (z_{k+1:d}^i)^S (z_{k+1:d}^i)^T.$$

Note that $C_{0,|T|,p} \geq \bar{C}_{|T|,p} \geq 0$, because $C_{0,|T|,p}$ considers the case where $z_i (i \in [k])$ is multiplied even times.

As basic fact in the boolean function analysis, we have $\mathbb{E}_z[z^S z^T z^{S'} z^{T'}] = 0$ unless $S = S'$ and $T = T'$. Therefore,

$$\mathbb{E}_z \left[\left(\sum_{i=1}^{n} \beta_i (z^\top z^i)^p \right) \left(\sum_{i=1}^{n} \beta_i (z^\top z^i)^q \right) \right]$$

$$
\begin{aligned}
&= \mathbb{E}_z \left[\left(\sum_{i=1}^n \beta_i \sum_{S,T} C_{|S|,|T|,p} z^S z^T (z^i)^S (z^i)^T \right) \left(\sum_{i=1}^n \beta_i \sum_{S,T} C_{|S|,|T|,q} z^S z^T (z^i)^S (z^i)^T \right) \right] \\
&= \sum_{S,T} \mathbb{E}_z \left[\left(\sum_{i=1}^n \beta_i C_{|S|,|T|,p} (z^i)^S (z^i)^T \right) \left(\sum_{i=1}^n \beta_i C_{|S|,|T|,q} z^S z^T (z^i)^S (z^i)^T \right) \right] \\
&= \sum_{S,T} d^{2|S|\alpha} C_{|S|,|T|,p} C_{|S|,|T|,q} \left(\sum_{i=1}^n \beta_i \right)^2 \\
&\geq \sum_{T} C_{0,|T|,p} C_{0,|T|,q} \left(\sum_{i=1}^n \beta_i \right)^2
\end{aligned} \tag{24}
$$

Where we used $C_{|S|,|T|,p}, C_{|S|,|T|,q} \geq 0$. On the other hand,

$$
\mathbb{E}_{z_{k+1:d}} \left[\left(\sum_{i=1}^n \beta_i (z_{k+1:d}^\top z_{k+1:d}^i)^p \right) \left(\sum_{i=1}^n \beta_i (z_{k+1:d}^\top z_{k+1:d}^i)^q \right) \right] = \sum_T \bar{C}_{|T|,p} \bar{C}_{|T|,q} \left(\sum_{i=1}^n \beta_i \right)^2 \geq 0. \tag{25}
$$

Because $c_{|S|,|T|,p} \geq \bar{C}_{T,p}$ and $c_{|S|,|T|,q} \geq \bar{C}_{T,q}$, comparing (24) and (25) yields

$$
\begin{aligned}
&\mathbb{E}_z \left[\left(\sum_{i=1}^n \beta_i (z^\top z^i)^p \right) \left(\sum_{i=1}^n \beta_i (z^\top z^i)^q \right) \right] \\
&\geq \mathbb{E}_{z_{k+1:d}} \left[\left(\sum_{i=1}^n \beta_i (z_{k+1:d}^\top z_{k+1:d}^i)^p \right) \left(\sum_{i=1}^n \beta_i (z_{k+1:d}^\top z_{k+1:d}^i)^q \right) \right] \geq 0,
\end{aligned}
$$

which concludes the proof. □

Proof of Lemma 5. LHS is determined by how many coordinates are different between z^i and z^j, which is captured by ${z^i}^\top z^j$. Thus, LHS is the polynomial of ${z^i}^\top z^j$. Moreover, its degree is at most l because the degrees of $z^\top z^i$ and $z^\top z^j$ are at most l in LHS. Thus, we now find that LHS can be written as $\sum_{l'=0}^{l} c_{d,l,l'} (\frac{{z^i}^\top z^j}{d^2})^{l'}$. Note that, when l is even, LHS is invariant to the replacement $z^j \mapsto -z^j$, and therefore $c_{d,l,l'} = 0$ for odd l'. On the other hand, when l is odd, $c_{d,l,l'} = 0$ for even l'.

Let us evaluate $c_{d,l,l'}$. By multiplying d^l for both sides, we have

$$
\mathbb{E}_z \left[\left(\frac{z^\top z^i}{\sqrt{d}} \right)^l \left(\frac{z^\top z^j}{\sqrt{d}} \right)^l \right] = \sum_{l'=0}^{l} c_{d,l,l'} \left(\frac{{z^i}^\top z^j}{d} \right)^{l'}.
$$

By taking $d \to \infty$ (while fixing the angle $\frac{{z^i}^\top z^j}{d}$), LHS will converge into

$$
\mathbb{E}_g \left[\left(\frac{g^\top z^i}{\sqrt{d}} \right)^l \left(\frac{g^\top z^j}{\sqrt{d}} \right)^l \right], \tag{26}
$$

here g follows $\mathbb{S}^{d-1}(\sqrt{d})$.

Consider the Hermite expansion of $t^l = \sum_{l'=0}^{l} c_{l,l'} He_{l'}(t)$. If l is even, $c_{l,l'} = \frac{1}{2^{\frac{l-l'}{2}} (\frac{l-l'}{2})! l!} > 0$ for even l' and $c_{l,l'} = 0$ for odd l'. If l is odd, $c_{l,l'} = \frac{1}{2^{\frac{l-l'}{2}} (\frac{l-l'}{2})! l!} > 0$ for odd l' and $c_{l,l'} = 0$ for even l'. By using these Hermite coefficients, (26) is equal to

$$
\sum_{l'=0}^{l'} c_{l,l'}^2 \left(\frac{{z^i}^\top z^j}{d} \right)^{l'}.
$$

Note that, as a function of the angle $\frac{z^{i\top} z^j}{d} \in [-1, 1]$, the convergence is uniform. Therefore, we get

$$d^{-l} c_{d,l,l'} \to c_{l,l'}^2 \quad (d \to \infty)$$

for all l and l'. When $c_{l,l'}^2 = 0$, $c_{d,l,l'} = 0$ for all d as we saw above. When $c_{l,l'}^2 > 0$, there exists d such that $c_{d',l,l'} > 0$ for all $d' \geq d$. Therefore, for sufficiently large d, we have $c_{d,l,l'} \geq 0$. Moreover, by direct calculation, $c_{d,l,l} = (l!)^2$. □

After obtained Lemma 3 we would like to bound $d^{-\lfloor (1-2\alpha)k \rfloor} \beta^\top K_2 \beta$. For this, we use the lower bound the smallest eigenvalue of K_2.

Let $K_{(d)}$ $(d = 1, 2, \cdots)$ be a sequence of inner-product kernels with $K_{(d)}(z, z') = h_{(d)}(\frac{z^\top z'}{d})$. Consider the case when each $K_{(d)}$ is associated with the same Kernel function $h\colon [-1, 1] \to \mathbb{R}$, so that $h_{(d)} = h$ holds for all $z, z' \in \{-1, 1\}^d$. Suppose that h is a degree-k polynomial and its coefficients are positive for all degrees. Note that K_2 satisfies these conditions. Then, we have the following.

Lemma 6 (Misiakiewicz (2022)). *Assume the following conditions hold:*

(a) $h^{(k')}(0) > 0$ *for* $k' = 0, \cdots, k-1$

(b) $h^{(k)}(0) > 0$

(c) $h(t)$ *is* k*-times differentiable*

They are Assumption 1 of Misiakiewicz (2022) for the case of $h_d = h$ *at* $l = k - 1$*, but are trivially true for a degree-*k *polynomial with positive coefficients. Also, fix* $\delta > 0$ *arbitrarily, and assume that* $d \gg 1$ *and* $n \lesssim d^k e^{-a_d \sqrt{\log d}}$ *for some* $\{a_d\}$ *with* $a_d \to \infty (d \to \infty)$*.*

Draw n *i.i.d. sample* $\{z^i\}_{i=1}^n$ *from* P_Z *to construct a Kernel matrix* $K \in \mathbb{R}^{n \times n}$ *as* $(K_{(d)})_{i,j} = h(\frac{z^{i\top} z^j}{d})$*. Then, for the Kernel matrix* $K_{(d)}$ *is decomposed into two positive semi-definite Kernel* $K_{>k-1}$ *and* $K_{\leq k-1}$*, and the spectrum of* $K_{>k-1}$ *is bounded by*

$$\mathbb{E}_{\{z^i\}_{i=1}^n} \left[\| K_{>k-1} - h^{(k)}(0) I \|_{\text{op}}^2 \right] \to 0 \quad (d \to \infty).$$

Proof. See Section 3.2 of Misiakiewicz (2022), where we take $\kappa = k - \delta$. □

Therefore, by fixing $\delta > 0$ arbitrarily, for $d \gg 1$ and $n \lesssim d^{\lfloor (1-2\alpha)k \rfloor - \delta}$, all the assumptions are satisfied for K_2 with $k = \lfloor (1 - 2\alpha)k \rfloor$ (if we regard K_2 as a kernel in $\mathbb{R}^{d-k} \times \mathbb{R}^{d-k}$). Note that we can take $a_d = (\log d)^{\frac{1}{4}}$ so that and $d^k e^{-a_d \sqrt{\log d}} \gtrsim d^{k-\delta}$. Then, the smallest eigenvalue of $K_{>k-1}$ is lower bounded by $\Omega(1)$ with probability at least 0.99 over the randomly drawn sample, for sufficiently large d. This immediately implies that the smallest eigenvalue of K_2 is bounded by $\Omega(1)$ with probability at least 0.99.

Now we finalize the proof of Theorem 2.

Proof of Theorem 2. According to Lemmas 3 and 6, for all choices of β, with probability at least 0.99 over the randomly drawn sample, we have

$$\mathbb{E}_z \left[\left(\sum_{i=1}^n \beta_i g_1 \left(\frac{z^\top z^i}{d} \right) \right)^2 \right] \gtrsim d^{-\lfloor (1-2\alpha)k \rfloor} \sum_{i=1}^n \beta_i^2 \tag{27}$$

$$\geq \frac{1}{d^{\lfloor (1-2\alpha)k \rfloor} n} \left(\sum_{i=1}^n |\beta_i| \right)^2 \tag{28}$$

$$\gtrsim \frac{1}{d^{2\lfloor (1-2\alpha)k \rfloor - \delta}} \left(\sum_{i=1}^n |\beta_i| \right)^2. \tag{29}$$

Because g_1 is the degree-$2k$ polynomial, Bonami's Lemma (e.g., Theorem 9.21 of (O'Donnell, 2014)) yields

$$\mathbb{E}_z\left[\left(\sum_{i=1}^{n}\beta_i g_1\left(\frac{z^\top z^i}{d}\right)\right)^4\right] \geq \frac{1}{(2k-1)^{4k}}\mathbb{E}_z\left[\left(\sum_{i=1}^{n}\beta_i g_1\left(\frac{z^\top z^i}{d}\right)\right)^2\right]^2$$

As a result, the Paley–Zygmund inequality (see Theorem 9.4 of (O'Donnell, 2014)) yields

$$\mathbb{P}_z\left[\left|\sum_{i=1}^{n}\beta_i g_1\left(\frac{z^\top z^i}{d}\right)\right| \geq t\mathbb{E}_z\left[\left(\sum_{i=1}^{n}\beta_i g_1\left(\frac{z^\top z^i}{d}\right)\right)^2\right]^{\frac{1}{2}}\right] \geq \frac{(1-t^2)^2}{(2k-1)^{4k}} \tag{30}$$

for all $0 \leq t \leq 1$.

Combining (27) and (30), with probabilty 0.99 over the sample, we have

$$\left|\sum_{i=1}^{n}\beta_i g_1\left(\frac{z^\top z^i}{d}\right)\right| \gtrsim \frac{1}{d^{\lfloor(1-2\alpha)k\rfloor-\delta/2}}\sum_{i=1}^{n}|\beta_i|.$$

with probability larger than $\Omega(1)$ over the choice of z. By taking sufficiently large d, $\frac{1}{d^{\lfloor(1-2\alpha)k\rfloor-\delta/2}}$ is larger than $\frac{1+c}{d^{\lfloor(1-2\alpha)k\rfloor}}$ (c is a constant from Lemma 1). Thus, using Lemma 2, we get

$$\mathbb{P}_{z\sim P_Z}\left[|f_\beta(z)| \geq \frac{c}{d^{\lfloor(1-2\alpha)k\rfloor}}\sum_{i=1}^{n}|\beta_i|\right] \gtrsim 1 - 1/d.$$

Now we apply Lemma 1 and finally get

$$\mathbb{P}_{z\sim P_Z}\left[f_\beta(z)y < 0\right] \gtrsim 1 - 2/d,$$

which concludes the proof. □