First Steps Toward Understanding the Extrapolation of Nonlinear Models to Unseen Domains

Kefan Dong
Stanford University
kefandong@stanford.edu

Tengyu Ma
Stanford University
tengyuma@stanford.edu

Abstract

Real-world machine learning applications often involve deploying neural networks to domains that are not seen in the training time. Hence, we need to understand the extrapolation of *nonlinear* models—under what conditions on the distributions and function class, models can be guaranteed to extrapolate to new test distributions. The question is very challenging because even two-layer neural networks cannot be guaranteed to extrapolate outside the support of the training distribution without further assumptions on the domain shift. This paper makes some initial steps towards analyzing the extrapolation of nonlinear models for structured domain shift. We primarily consider settings where the *marginal* distribution of each coordinate of the data (or subset of coordinates) does not shift significantly across the training and test distributions, but the joint distribution may have a much bigger shift. We prove that the family of nonlinear models of the form $f(x) = \sum f_i(x_i)$, where f_i is an *arbitrary* function on the subset of features x_i, can extrapolate to unseen distributions, if the covariance of the features is well-conditioned. To the best of our knowledge, this is the first result that goes beyond linear models and the bounded density ratio assumption, even though the assumptions on the distribution shift and function class are stylized.

1 Introduction

In real-world applications, machine learning models are often deployed on domains that are not seen in the training time. For example, we may train machine learning models for medical diagnosis on data from hospitals in Europe and then deploy them to hospitals in Asia.

Thus, we need to understand the extrapolation of models to new test distributions — how the model trained on one distribution behaves on another unseen distribution. This extrapolation of neural networks is central to various robustness questions such as domain generalization (Gulrajani and Lopez-Paz [2020], Ganin et al. [2016], Peters et al. [2016] and references therein) and adversarial robustness [Goodfellow et al., 2014, Kurakin et al., 2018], and also plays a critical role in nonlinear bandits and reinforcement learning where the distribution is constantly changing during training [Dong et al., 2021, Agarwal et al., 2019, Lattimore and Szepesvári, 2020, Sutton and Barto, 2018].

This paper focuses on the following mathematical abstraction of this extrapolation question:

Under what conditions on the source distribution P, target distribution Q, and function class $\mathcal{F}$ do we have that any functions $f, g \in \mathcal{F}$ that agree on P are also guaranteed to agree on Q?

Here we can measure the agreement of two functions on P by the ℓ_2 distance between f and g under distribution P, that is, $\|f - g\|_P \triangleq \mathbb{E}_{x\sim P}[(f(x) - g(x))^2]^{1/2}$. The function f can be thought of as the learned model, g as the ground-truth function, and thus $\|f - g\|_P$ as the error on the source distribution P.

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

This question is well-understood for linear function class $\mathcal{F}$. Essentially, if the covariance of Q can be bounded from above by the covariance of P (in any direction), then the error on Q is guaranteed to be bounded above from the error on P. We refer the reader to Lei et al. [2021], Mousavi Kalan et al. [2020] and references therein for more recent advances along this line.

By contrast, theoretical results for extrapolation of *nonlinear* models is rather limited. Classical results have long settled the case where P and Q have bounded density ratios [Ben-David and Urner, 2014, Sugiyama et al., 2007]. Bounded density ratio implies that the support of Q must be a subset of the support of P, and thus arguably these results do not capture the extrapolation behavior of models *outside* the training domain.

Without the bounded density ratio assumption, there was limited prior *positive* result for characterizing the extrapolation power of neural networks. Ben-David et al. [2010] show that the model can extrapolate when the $\mathcal{H}\Delta\mathcal{H}$-distance between training and test distribution is small. However, it remains unclear for what distributions and function class, the $\mathcal{H}\Delta\mathcal{H}$-distance can be bounded.[1] In general, the question is challenging partly because of the existence of such a strong impossibility result. As soon as the support of Q is not contained in the support of P (and they satisfy some non-degeneracy condition), it turns out that even two-layer neural networks cannot extrapolate—there are two-layer neural networks f and g that agree on P perfectly but behave very differently on Q (See Proposition 6 for a formal statement.)

The impossibility result suggests that any positive results on the extrapolation of nonlinear models require more fine-grained structures on the relationship between P and Q (which are common in practice Koh et al. [2021], Sagawa et al. [2022]) as well as the function class $\mathcal{F}$. The structure in the domain shift between P and Q may also need to be compatible with the assumption on the function class $\mathcal{F}$. This paper makes some first steps towards proving certain family of nonlinear models can extrapolate to a new test domain with structured shift.

We consider a setting where the joint distribution of the data can does not have much overlap across P and Q (and thus bounded density ratio assumption does not hold), whereas the marginal distributions for each coordinate of the data does overlap. Such a scenario may practically happen when the features (coordinates of the data) exhibit different correlations on the source and target distribution. For example, consider the task of predicting the probability of a lightning storm from basic meteorological information such as precipitation, temperature, etc. We learn models from some cities on the west coast of United States and deploy them to the east coast. In this case, the joint test distribution of the features may not necessarily have much overlap with the training distribution—correlation between precipitation and temperature could be vastly different across regions, e.g., the rainy season coincides with the winter's low temperature on the west coast, but not so much on the east coast. However, the individual feature's marginal distributions are much more likely to overlap between the source and target—the possible ranges of temperature on east and west coasts are similar.

Concretely, we assume that the features $x \in \mathbb{R}^{s_1+s_2}$ have Gaussian distributions and can be divided into two subsets $x_1 \in \mathbb{R}^{s_1}$ and $x_2 \in R^{s_2}$ such that each set of feature x_i ($i \in \{1, 2\}$) has the same marginal distributions on P and Q. Moreover, we assume that x_1 and x_2 are not exactly correlated on P—the covariance of features x on distribution P has a strictly positive minimum eigenvalue.

As argued before, restricted assumptions on the function class $\mathcal{F}$ are still necessary (for almost any P and Q without the bounded density ratio property). Here, we assume that $\mathcal{F}$ consists of all functions of the form $f(x) = f_1(x_1) + f_2(x_2)$ for *arbitrary* functions $f_1 : \mathcal{R}^{s_1} \to \mathbb{R}$ and $f_2 : \mathcal{R}^{s_2} \to \mathbb{R}$. The function class $\mathcal{F}$ does not contain all two-layer neural networks (so that the impossibility result does not apply), but still consists of a rich set of functions where each subset of features independently contribute to the prediction with arbitrary nonlinear transformations. We show that under these assumptions, if any two models approximately agree on P, they must also approximately agree on Q — formally speaking, $\forall f, g \in \mathcal{F},\ \|f - g\|_Q \lesssim \|f - g\|_P$ (Theorem 5).

We also prove a variant of the result above where we divide features vector $x \in \mathbb{R}^d$ into d coordinate, denoted by $x = (x_1, \ldots, x_d)$ where $x_i \in \mathbb{R}$. The function class consists of all combinations of *nonlinear* transformations of x_i's, that is, $\mathcal{F} = \{\sum_{i=1}^d f_i(x_i)\}$. Assuming coordinates of x

[1] In fact, the $\mathcal{H}\Delta\mathcal{H}$-distance likely cannot be bounded when the function class contains two-layer neural networks, and the supports of the training and test distributions do not overlap —when there exists a function that can distinguish the source and target domain, the $\mathcal{H}\Delta\mathcal{H}$ divergence will be large.

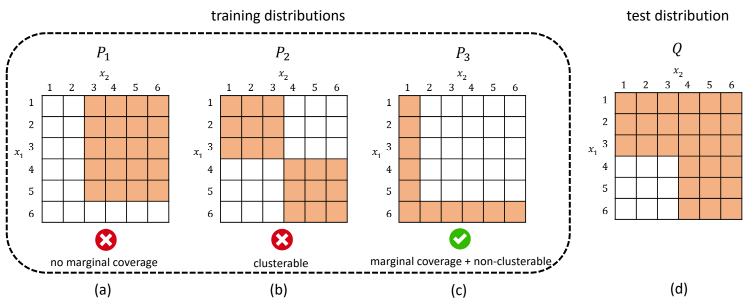

Figure 1: Visualization of three different training distributions and a test distribution, where the orange color blocks marks the support. **(a)** and **(b):** distributions that do not satisfy our conditions and cannot extrapolate. **(c):** a distribution that satisfies our conditions even with a sparse support.

are pairwise Gaussian and a non-degenerate covariance matrix, the nonlinear model $f \in \mathcal{F}$ can extrapolate to any distribution Q that has the same marginals as P (Theorem 4).

These results can be viewed as first steps for analyzing extrapolation beyond linear models. Compared the works of Lei et al. [2021] on linear models, our assumptions on the covariance of P are qualitatively similar. We additionally require P and Q have overlapping marginal distributions because it is even necessary for the extrapolation of one-dimensional functions on a single coordinate. Our results work for a more expressive family of nonlinear functions, that is, $\mathcal{F} = \{\sum_{i=1}^d f_i(x_i)\}$, than the linear functions.

We also present a result on the case where x_i's are discrete variables, which demonstrates the key intuition and also may be of its own interest. Suppose we have two discrete random variable x_1 and x_2. In this case, the joint distribution of P and Q can be both presented by a matrix (as visualized in Figure 1), and the marginal distributions are the column and row sums of this joint probability matrix. We prove that extrapolation occurs when (1) the support of the *marginals* of Q is covered by P, and (2) the density matrix of P is non-clusterable — we cannot find a subset of rows and columns such the support of P in these rows and columns lies within their intersections.

In Figure 1, we visualize a few interesting cases. First, distributions P_1, P_2 visualized in Figures 1a and 1b, respectively, do not satisfy our conditions. Fundamentally, there are two models that agreed on the support of distribution P_1 (or P_2), but still differ much on the non-support. In contrast, our result predict that models trained on distribution P_3 in Figure 1c can extrapolate to the distribution Q in Figure 1d, despite that the support of P_3 is sparse and their support have very little overlap.

We also note that the failure of P_1 and P_2 demonstrate the non-triviality of our results. The overlapping marginal assumption by itself does not guarantees extrapolation, and condition (2), or analogously the minimal eigenvalue condition for the Gaussian cases, is critical for extrapolation.

Our proof techniques for the theorems above is generally viewing $\|h\|_P^2$ as $K_P(h,h)$ for some kernel $K_P : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$. Here h is a shorthand for the error function $f - g$. Note that the kernel takes in two functions in $\mathcal{F}$ as inputs and captures relationship between the functions. Hence, the extrapolation of $\mathcal{F}$ (i.e., proving $\|h\|_Q \lesssim \|h\|_P$ for all $h \in \mathcal{F}$) reduces to the relationship of the kernels (i.e., whether $K_Q(h,h) \lesssim K_P(h,h)$ for all $h \in \mathcal{F}$), which is then governed by properties of the eigenspaces of kernel K_P, K_Q. Thanks to the special structure of our model class $\mathcal{F} = \sum f_i(x_i)$, we can analytically relate the eigenspace of the kernels K_P, K_Q to more explicitly and interpretable property of the data distribution of P and Q.

2 Problem Setup and Preliminaries

We use P and Q to denote the source and target distribution over the space of features $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$ respectively. We measure the extrapolation of a model class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ from P to Q by the the following quantity:[2]

$$\tau(P, Q, \mathcal{F}) \triangleq \sup_{f,g\in\mathcal{F}} \frac{\mathbb{E}_Q[(f(x)-g(x))^2]}{\mathbb{E}_P[(f(x)-g(x))^2]}. \tag{1}$$

When $\tau(P, Q, \mathcal{F})$ is small, if two models $f, g \in \mathcal{F}$ approximately agree on P (meaning that $\mathbb{E}_P[(f(x)-g(x))^2]$ is small), they must approximately agree on Q because $\mathbb{E}_Q[(f(x)-g(x))^2] \le \tau\mathbb{E}_P[(f(x)-g(x))^2]$.

If the model class $\mathcal{F}$ is expressive enough to include the ground-truth labeling function, τ becomes an upper bound of the ratio between the loss on distribution Q and the loss on distribution P (formally stated in Proposition 1), which provides the robustness guarantee of the trained model. This is because when g corresponds to the ground-truth label, $\mathbb{E}_P[(f(x)-g(x))^2]$ becomes the ℓ_2 loss of model f.

Proposition 1. *Let τ the quantity defined in Eq.* (1). *For any distribution P, Q and model class $\mathcal{F}$, if there exists a model in $\mathcal{F}$ that can represent the true labeling $y : \mathcal{X} \to \mathbb{R}$ on both P and Q:*

$$\exists f^\star \in \mathcal{F} \text{ such that } \mathbb{E}_{\frac{1}{2}(P+Q)}[(y(x)-f^\star(x))^2] \le \epsilon_{\mathcal{F}}, \tag{2}$$

then we have

$$\forall y \in \mathcal{F}, \mathbb{E}_Q[(y(x)-f(x))^2] \le (8\tau+4)\epsilon_{\mathcal{F}} + 4\tau\mathbb{E}_P[(y(x)-f(x))^2]. \tag{3}$$

Proof of this proposition is deferred to Appendix F.1

Relationship to the $\mathcal{H}\Delta\mathcal{H}$-distance. The quantity τ is closely related to the $\mathcal{H}\Delta\mathcal{H}$-distance [Ben-David et al., 2010]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) = 2\sup_{f,g\in\mathcal{F}} |\mathrm{Pr}_{x\sim P}[f(x) \ne g(x)] - \mathrm{Pr}_{x\sim Q}[f(x) \ne g(x)]|, \tag{4}$$

with the differences that (1) we consider ℓ_2 loss instead of classification loss, and (2) τ focuses on the ratio of losses whereas $d_{\mathcal{H}\Delta\mathcal{H}}$ focuses on the absolute difference. As we will see later, these differences bring the mathematical simplicity to prove concrete conditions for model extrapolation.

In this paper, we focus on the model class $\mathcal{F} = \{\sum_{i=1}^k f_i(x_i) : \mathbb{E}_P[f_i(x_i)^2] < \infty, \forall i \in [d]\}$ where $f_i : \mathcal{X}_i \to \mathbb{R}$ is an *arbitrary* function. Since $\mathcal{F}$ is closed under addition, we can simplify Eq. (1) to $\tau(P, Q, \mathcal{F}) = \sup_{f\in\mathcal{F}} \frac{\mathbb{E}_Q[f(x)^2]}{\mathbb{E}_P[f(x)^2]}$. For simplicity, we omit the dependency on $P, Q, \mathcal{F}$ when the context is clear.

3 Summary of Main Results

3.1 Features with Discrete Values

In this section, we relate the extrapolation power of a training distribution P to the clusterablity of its density matrix. Without loss of generality, we assume that x_i takes the value in $\{1, 2, \cdots, r_i\}$.

We measure the (approximate) clusterablity by eigenvalues of the Laplacian matrix of a bipartite graph associated with the density matrix $P \in \mathbb{R}^{r_1\times r_2}$. Concretely, let G_P be the bipartite graph where $P(x_1 = i, x_2 = j)$ corresponds to the weight of the edge between the i-th vertex on the left part of the graph and the j-th vertex on the right part. The Laplacian of G_P is computed as follows.

Let $d_1 \in \mathbb{R}^{r_1}$ and $d_2 \in \mathbb{R}^{r_2}$ be the row and column sums of the weight matrix P (in other words, degree of the vertices). Define diagonal matrices $D_1 = \mathrm{diag}(d_1) \in \mathbb{R}^{r_1\times r_1}$ and $D_2 = \mathrm{diag}(d_2) \in \mathbb{R}^{r_2\times r_2}$. Then the (signless) Laplacian K_P and normalized Laplacian of $\bar{K}_P$ is:

$$K_P = \begin{pmatrix} D_1 & P \\ P^\top & D_2 \end{pmatrix}, \quad \bar{K}_P = \mathrm{diag}(K_P)^{-1/2} K_P \mathrm{diag}(K_P)^{-1/2}. \tag{5}$$

And our main theorem in this section is stated as follows.

Theorem 2. *For any two dimensional distribution P, Q over discrete random variables x_1, x_2, and the model class $\mathcal{F} = \{f_1(x_1) + f_2(x_2)\}$ where $f_i : \mathcal{X}_i \to \mathbb{R}$ is an arbitrary function, we have*

$$\tau(P, Q, \mathcal{F}) \le 2\lambda_2(\bar{K}_P)^{-1} \max_{i\in[2], t\in[r_i]} \frac{P(x_i = t)}{Q(x_i = t)}. \tag{6}$$

[2]For simplicity, we set $0/0 = 0$.

Remarks. Cheeger's inequality provides upper and lower bounds of $\lambda_2(\bar{K}_P)$ by the sparest cut $\phi(G_P)$ of the bipartite graph G_P [Chung, 1996, Alon, 1986]:

$$\lambda_2(\bar{K}_P)/2 \leq \phi(G_P) \leq \sqrt{2\lambda_2(\bar{K}_P)}, \tag{7}$$

When the bipartite graph is not connected (e.g., the distribution in Figure 1b), we get $\phi(G_P) = 0$ and therefore $\lambda_2(\bar{K}_P) = 0$. In this case, Theorem 2 is vacuous because the upper bound in Eq. (6) becomes infinity, and therefore we have no extrapolation guarantees. In fact, we can even find two models f, g and a input $x^\dagger$ such that $\|f - g\|_P = 0$ but $f(x^\dagger) \neq g(x^\dagger)$ (see Proposition 7).

The following corollary proves sufficient conditions for $\tau(P, Q, \mathcal{F}) < \infty$, which are visualized in Figure 1.[3]

Corollary 3. *In the setting of Theorem 2, $\tau(P, Q, \mathcal{F}) < \infty$ if*

1. *the marginals of Q has a smaller support than that of P, and*
2. *we cannot find a subset of rows and columns such that the support of P in these rows and columns lies within their intersections.*

The proof of Corollary 3 is deferred to Appendix F.4

Compared with prior works that assumes a bounded density ratio on the entire distribution (e.g., Ben-David and Urner [2014], Sugiyama et al. [2007]), we only require a bounded density ratio of the marginal distributions. This is because the model class $f(x) = f_1(x_1) + f_2(x_2)$ can extrapolate to distributions with a larger support (see Figure 1c). In contrast, for an unstructured model considered in prior works (in this case, $f(x)$ is an arbitrary function of the entire input x) we do not have any control of the model's behavior on data points outside the support of P.

The proof sketch of Theorem 2 is deferred to Appendix B.1.

3.2 Features with Real Values

We also consider the case when $x_1, x_2, \cdots, x_k$ are real-valued random variables. Let $\Sigma_P = \mathbb{E}_P[xx^\top], \Sigma_Q = \mathbb{E}_Q[xx^\top]$ be the covariance matrix of x on distributions P, Q. Recall that our model has the structure $f(x) = \sum_{i=1}^k f_i(x_i)$ where f_i is an arbitrary one-dimensional function. Without loss of generality, we assume $\operatorname{diag}(\Sigma_P) = \operatorname{diag}(\Sigma_Q) = I$ (a detailed discussion is deferred to Appendix B.2), then we have the following theorem.

Theorem 4. *For any distributions P, Q over variables $x = (x_1, \cdots, x_d)$ with matching marginals, if (x_i, x_j) has the distribution of a two-dimensional Gaussian random variable for every $i, j \in [d]$ on both P and Q, and $\operatorname{diag}(\Sigma_P) = \operatorname{diag}(\Sigma_Q) = I$, then*

$$\tau \leq d(\lambda_{\min}(\Sigma_P))^{-1}. \tag{8}$$

Compared with linear models $\mathcal{F}_{\text{linear}} \triangleq \{v^\top x : v \in \mathbb{R}^d\}$, Theorem 4 proves comparable conditions for the structured nonlinear model class $f(x) = \sum_{i=1}^d f_i(x_i)$ to extrapolate because linear model requires $\tau(P, Q, \mathcal{F}_{\text{linear}}) = \sup_{v \in \mathbb{R}^d} \frac{\|v^\top x\|_Q^2}{\|v^\top x\|_P^2} = \sup_{v \in \mathbb{R}^d} \frac{v^\top \Sigma_Q v}{v^\top \Sigma_P v} \lesssim \lambda_{\min}(\Sigma_P)^{-1}$ to be small, which is the requirement of Eq. (8) (up to factors that is independent of the distributions P, Q).

We can extend to the case where $x_1 \in \mathbb{R}^{d_1}, x_2 \in \mathbb{R}^{d_2}$ are two subsets of the input x, and the input $x = (x_1, x_2)$ has Gaussian distribution (Theorem 5).

Acknowledgment

The authors would like to thank Yuanhao Wang, Yuhao Zhou, Hong Liu, Ananya Kumar, Jason D. Lee for helpful discussions. The authors would also like to thank the support from NSF CIF 2212263.

[3] In fact, we can prove Corollary 3 directly without referring to Theorem 2. We start with the fact that $\|f\|_P = 0$ implies $f(x) = 0$ for all $x \in \operatorname{supp}(P)$. Then we can iteratively expand the set of points that we know $f(x) = 0$, using the fact that for any x_1, x_2, x_1', x_2', the three equations $f_1(x_1) + f_2(x_2) = 0$, $f_1(x_1') + f_2(x_2) = 0$, and $f_1(x_1) + f_2(x_2') = 0$ together implies $f_1(x_1') + f_2(x_2') = 0$.

References

Alekh Agarwal, Nan Jiang, and Sham M Kakade. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Noga Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, 1986.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Shai Ben-David and Ruth Urner. Domain adaptation–can quantity compensate for quality? *Annals of Mathematics and Artificial Intelligence*, 70(3):185–202, 2014.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Tianle Cai, Ruiqi Gao, Jason Lee, and Qi Lei. A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning*, pages 1170–1182. PMLR, 2021.

Enrico Celeghini, Manuel Gadella, and Mariano A del Olmo. Hermite functions and fourier series. *Symmetry*, 13(5):853, 2021.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. *arXiv preprint arXiv:2006.10032*, 2020.

Fan RK Chung. Laplacians of graphs and cheeger's inequalities. *Combinatorics, Paul Erdos is Eighty*, 2(157-172):13–2, 1996.

Kefan Dong, Jiaqi Yang, and Tengyu Ma. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *arXiv preprint arXiv:2102.04168*, 2021.

Gregory E Fasshauer. Positive definite kernels: past, present and future. *Dolomites Research Notes on Approximation*, 4:21–63, 2011.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*, pages 597–613. Springer, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Jeff Z HaoChen, Colin Wei, Ananya Kumar, and Tengyu Ma. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *arXiv preprint arXiv:2204.02683*, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Qi Lei, Wei Hu, and Jason Lee. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, pages 6164–6174. PMLR, 2021.

Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3918–3930, 2020a.

Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13944, 2020b.

Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Advances in Neural Information Processing Systems*, 33:1959–1969, 2020.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78:947–1012, 2016.

Alexander D Poularikas. *Handbook of formulas and tables for signal processing*. CRC press, 2018.

Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Kendrick Shen Sang Michael Xie, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Sara Beery Henrik Marklund, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Tatsunori Hashimoto Kate Saenko, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the wilds benchmark for unsupervised adaptation. In *International Conference on Learning Representations*, 2022.

Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 19847–19878. PMLR, 2022.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data, 2020. URL `https://openreview.net/forum?id=rC8sJ4i6kaH`.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael I Jordan. Bridging theory and algorithm for domain adaptation. *arXiv preprint arXiv:1904.05801*, pages 7404–7413, 2019.

Huaiyu Zhu, Christopher KI Williams, Richard Rohwer, and Michal Morciniec. Gaussian regression and optimal finite dimensional linear models. 1997.

Contents

A Useful Notations

In this section, we introduce some useful notations.

Let $\mathbb{I}[E]$ be the indicator function that equals 1 if the condition E is true, and 0 otherwise. For an integer n, let $[n]$ be the set $\{1, 2, \cdots, n\}$. For a vector $x \in \mathbb{R}^d$, we use $[x]_i$ to denote its i-th coordinate. Similarly, $[M]_{i,j}$ denotes the (i, j)-th element of a matrix M. We use $M^{\odot n}$ to represent the element-wise n-th power of the matrix M (i.e., $[M^{\odot n}]_{i,j} = ([M]_{i,j})^n$). Let $I_d \in \mathbb{R}^{d\times d}$ be the identity matrix, $\mathbf{1}_d \in \mathbb{R}^d$ the all-1 vector and $e_{i,d}$ the i-th base vector. We omit the subscript d when the context is clear. For a square matrix $P \in \mathbb{R}^{d\times d}$, we use $\mathrm{diag}(P) \in \mathbb{R}^{d\times d}$ to denote the matrix generated by masking out all non-diagonal terms of P. For list $\sigma_1, \cdots, \sigma_d$, let $\mathrm{diag}(\{\sigma_1, \cdots, \sigma_d\}) \in \mathbb{R}^{d\times d}$ be the diagonal matrix whose diagonal terms are $\sigma_1, \cdots, \sigma_d$.

For a symmetric matrix $M \in \mathbb{R}^{d\times d}$, let $\lambda_1(M) \le \lambda_2(M) \le \cdots \le \lambda_d(M)$ be its eigenvalues in ascending order, and $\lambda_{\max}(M), \lambda_{\min}(M)$ the maximum and minimum eigenvalue, respectively. Similarly, we use $\sigma_1(M) \le \cdots \le \sigma_{\min(d_1,d_2)}(M)$ to denote the singular values of $M \in \mathbb{R}^{d_1\times d_2}$.

B Main Results

In this section, we present our additional main results.

B.1 Proof Sketch of Theorem 2 Values

Proof sketch of Theorem 2. In the following we present a proof sketch of Theorem 2, and defer the complete proof to Appendix F.3. We start with a high-level proof strategy and then instantiate the proof on the setting of Theorem 2.

Suppose we can find a set of (not necessarily orthogonal) basis $\{b_1, \cdots, b_r\}$ where $b_i : \mathcal{X} \to \mathbb{R}$, such that any model $f \in \mathcal{F}$ can be represented as a linear combination of basis $\sum_{i=1}^r v_i b_i$. Since the model family $\mathcal{F}$ is closed under subtraction, we have

$$\tau = \sup_{f\in\mathcal{F}} \frac{\|f\|_Q^2}{\|f\|_P^2} = \sup_{v\in\mathbb{R}^r} \frac{\mathbb{E}_Q[\big(\sum_{i=1}^k v_i b_i(x)\big)^2]}{\mathbb{E}_P[\big(\sum_{i=1}^k v_i b_i(x)\big)^2]} = \sup_{v\in\mathbb{R}^r} \frac{\sum_{i,j=1}^r [v]_i[v]_j \mathbb{E}_Q[b_i(x)b_j(x)]}{\sum_{i,j=1}^r [v]_i[v]_j \mathbb{E}_P[b_i(x)b_j(x)]}. \tag{9}$$

If we define the kernel matrix $[K_P]_{i,j} = \mathbb{E}_P[b_i(x)b_j(x)]$, it follows that

$$\sup_{v\in\mathbb{R}^r} \frac{\sum_{i,j=1}^k v_i v_j \mathbb{E}_Q[b_i(x)b_j(x)]}{\sum_{i,j=1}^k v_i v_j \mathbb{E}_P[b_i(x)b_j(x)]} = \sup_{v\in\mathbb{R}^r} \frac{v^\top K_Q v}{v^\top K_P v}. \tag{10}$$

Hence, upper bounding τ reduces to bounding the eigenvalues of kernel matrices K_P, K_Q.

Since the model has the structure $f(x) = f_1(x_1) + f_2(x_2)$, we can construct the basis $\{b_t\}_{t=1}^r$ explicitly. For any $i \in [2], t \in [r_i]$, with little abuse of notation, let

$$b_{i,t}(x) = \mathbb{I}[x_i = t]. \tag{11}$$

We can verify that the set $\{b_{i,t}\}_{i\in[r],t\in[r_i]}$ is indeed a complete set of basis. As a result, the kernel matrices K_P can be computed directly using its definition:

$$\mathbb{E}_P[b_{i,t}(x)b_{j,s}(x)] = \begin{cases} P(x_i = t, x_j = s), & \text{when } i \neq j, \\ P(x_i = t)\mathbb{I}[s = t], & \text{when } i = j, \end{cases} \tag{12}$$

which is exactly the Laplacian matrix defined in Eq. (5).

Let $D_P = \mathrm{diag}(K_P)$ and $D_Q = \mathrm{diag}(K_Q)$. To prove Eq. (6), the high level intuition is that

$$\sup_{v\in\mathbb{R}^r} \frac{v^\top K_Q v}{v^\top K_P v} = \sup_{v\in\mathbb{R}^r} \frac{v^\top D_Q^{1/2} \bar{K}_Q D_Q^{1/2} v}{v^\top D_P^{1/2} \bar{K}_P D_P^{1/2} v} \tag{13}$$

$$\le \frac{\lambda_{\max}(\bar{K}_Q)}{\lambda_{\min}(\bar{K}_P)} \sup_{v\in\mathbb{R}^r} \frac{\|D_Q^{1/2} v\|_2^2}{\|D_P^{1/2} v\|_2^2} \le \frac{\lambda_{\max}(\bar{K}_Q)}{\lambda_{\min}(\bar{K}_P)} \max_{i\in[2],t\in[r_i]} \frac{P(x_i = t)}{Q(x_i = t)}. \tag{14}$$

However, this naive bound is vacuous because for any P we have $\lambda_{\min}(\bar{K}_P) = 0$. In fact, K_P and K_Q share the eigenvalue 0 and the corresponding eigenvector. Therefore we can ignore this directions first, and then the desired result directly follows.

Theorem 2 can be easily extend to the case when $k > 2$. We choose to present the current version because of simplicity, and because the kernel matrix K_P coincides with the Laplacian of a bipartite graph when $k = 2$, which provides further insights and better intuitions.

B.2 Features with Real Values

In this section we extend our analysis to the case where $x_1, x_2, \cdots, x_d$ are real-valued random variables. Let $\Sigma_P = \mathbb{E}_P[xx^\top], \Sigma_Q = \mathbb{E}_Q[xx^\top]$ be the covariance matrix of x on distributions P, Q. Recall that our model has the structure $f(x) = \sum_{i=1}^d f_i(x_i)$ where f_i is an arbitrary one-dimensional function. We first focus on the case where $\operatorname{diag}(\Sigma_P) = \operatorname{diag}(\Sigma_Q) = I$ for simplicity, and then extend to general cases.

When $d = 2$, this setting is almost the same as the one in Section 3.1, and the only difference is that the Laplacian "matrix" will be infinite dimensional (indexed by a pair of real numbers) because the feature x_i is a real number. However, we can still compute the eigenvalues of the kernel with some Gaussianity assumptions on the distribution of features, as stated in the following theorem (restate of Theorem 4).

Theorem 4. *For any distributions P, Q over variables $x = (x_1, \cdots, x_d)$ with matching marginals, if (x_i, x_j) has the distribution of a two-dimensional Gaussian random variable for every $i, j \in [d]$ on both P and Q, and $\operatorname{diag}(\Sigma_P) = \operatorname{diag}(\Sigma_Q) = I$, then*

$$\tau \leq d(\lambda_{\min}(\Sigma_P))^{-1}. \tag{8}$$

Remarks. For the general case where $\operatorname{diag}(\Sigma_P) \neq I$, we can normalize the input x_i to $t_i \triangleq \operatorname{var}_P(x_i)^{-1/2}(x_i - \mathbb{E}_P[x_i])$ first. Let P', Σ'_P be the density and covariance of t when $x \sim P$ (and similarly Q'). Then we have $\operatorname{diag}(\Sigma'_P) = \operatorname{diag}(\Sigma'_Q) = I$ when P, Q have matching marginals. Hence, we can prove that $\tau(P, Q, \mathcal{F}) = \tau(P', Q', \mathcal{F}) \leq \frac{d}{\lambda_{\min}(\Sigma'_P)}$ (Lemma 8).

Compared with linear models, Theorem 4 proves comparable results for the structured nonlinear model class $f(x) = \sum_{i=1}^d f_i(x_i)$ to extrapolate. For linear models $\mathcal{F}_{\text{linear}} \triangleq \{v^\top x : v \in \mathbb{R}^d\}$ we have

$$\tau(P, Q, \mathcal{F}_{\text{linear}}) = \sup_{v \in \mathbb{R}^d} \frac{\|v^\top x\|_Q^2}{\|v^\top x\|_P^2} = \sup_{v \in \mathbb{R}^d} \frac{v^\top \Sigma_Q v}{v^\top \Sigma_P v} \lesssim \lambda_{\min}(\Sigma_P)^{-1},$$

which is the RHS of Eq. (8) (up to factors that is independent of the distributions P, Q).

We emphasize that we only assume the marginals on every *pair* of features x_i, x_j is Gaussian, which does not imply the Gaussianity of the joint distribution of x. In fact, there exists a non-Gaussian distribution that satisfies our assumption.

Proof sketch of Theorem 4. On a high level, the proof intuition is to pretend the real numbers are discrete, and follow the same proof strategy as Theorem 2. We present this non-rigorous proof sketch below for its simplicity, and the rigorous proof is deferred to Appendix F.5.

First we consider a simplified case when $d = 2$. Following the same argument as Theorem 2, we can compute the normalized kernel $\bar{K}_P$ explicitly. Because x_1, x_2 are continuous random variables, the kernel $\bar{K}_P$ will be infinite dimensional, and have the form $\bar{K}_P = \begin{pmatrix} I & A \\ A^\top & I \end{pmatrix}$, where A is an infinite dimensional "matrix" indexed by real numbers $x_1, x_2 \in \mathbb{R}$, with values $[A]_{x_1, x_2} = P(x_1, x_2)/\sqrt{P(x_1)P(x_2)}$, and I is the "identity matrix" (rigorously speaking they are kernel functions that map $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$.)

Similar to Theorem 2, we need to lower bound the second smallest eigenvalue of $\bar{K}_P$. To this end, we first decompose A using singular value decomposition $A = U\Lambda V^\top$, where $UU^\top = I, VV^\top = I$ and $\Lambda = \operatorname{diag}(\{\sigma_n\}_{n \geq 0})$ with $\sigma_0 \geq \sigma_1 \geq \cdots$. Then we get

$$\bar{K}_P = \begin{pmatrix} I & A \\ A^\top & I \end{pmatrix} = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} I & \Lambda \\ \Lambda^\top & I \end{pmatrix} \begin{pmatrix} U^\top & 0 \\ 0 & V^\top \end{pmatrix}. \tag{15}$$

Since the matrix $\hat{K}_P \triangleq \begin{pmatrix} I & \Lambda \\ \Lambda^\top & I \end{pmatrix}$ consists of four diagonal sub-matrices, we can shuffle the rows and columns of $\hat{K}_P$ to form a block-diagonal matrix with blocks $\left\{ \begin{pmatrix} 1 & \sigma_n \\ \sigma_n & 1 \end{pmatrix} \right\}_{n=0,1,2,\cdots}$.

As a result, $\lambda_2(\bar{K}_P) = \lambda_2(\hat{K}_P) = 1 - \sigma_1$. By the assumption that (x_1, x_2) follows from Gaussian distribution, the "matrix" A is actually a Gaussian kernel, whose eigenvalues and eigenfunctions can be computed analytically — Theorem 11 proves that $\sigma_1 = |\mathbb{E}_P[x_1 x_2]|$. Consequently, $\lambda_2(\bar{K}_P) = 1 - \sigma_1 = 1 - |\mathbb{E}_P[x_1 x_2]| = \lambda_{\min}(\Sigma_P)$.

Now we briefly discuss the case when $d = 3$, and the most general cases (i.e., $d > 3$) are proved similarly. When $d = 3$, the normalized kernel will have the following form

$$\bar{K}_P = \begin{pmatrix} I & A & B \\ A^\top & I & C \\ B^\top & C^\top & I \end{pmatrix}. \tag{16}$$

By the assumption that $x_1, x_2, \cdots, x_d$ have zero mean and unit variance, matrices A, B, C are symmetric. In addition, Theorem 11 shows that the eigenfunctions of the Gaussian kernel is *independent* of the value $\mathbb{E}_P[x_i x_j]$. Hence, the matrices A, B, C shares the same eigenspace and can be diagonalized simultaneously:

$$\bar{K}_P = \begin{pmatrix} I & A & B \\ A^\top & I & C \\ B^\top & C^\top & I \end{pmatrix} = \begin{pmatrix} U & 0 & 0 \\ 0 & U & 0 \\ 0 & 0 & U \end{pmatrix} \begin{pmatrix} I & \Lambda_A & \Lambda_B \\ \Lambda_A^\top & I & \Lambda_C \\ \Lambda_B^\top & \Lambda_C^\top & I \end{pmatrix} \begin{pmatrix} U^\top & 0 & 0 \\ 0 & U^\top & 0 \\ 0 & 0 & U^\top \end{pmatrix}. \tag{17}$$

By reshuffling the columns and rows, the eigenvalues of $\bar{K}_P$ are the union of the eigenvalues of following matrices

$$\{\bar{K}_P^{(n)}\}_{n=0,1,2,\cdots} \triangleq \left\{ \begin{pmatrix} 1 & \sigma_n(A) & \sigma_n(B) \\ \sigma_n(A) & 1 & \sigma_n(C) \\ \sigma_n(B) & \sigma_n(C) & 1 \end{pmatrix} \right\}_{n=0,1,2,\cdots}. \tag{18}$$

Theorem 11 implies that $\sigma_n(A) = ([\Sigma_P]_{1,2})^n, \sigma_n(B) = ([\Sigma_P]_{1,3})^n$ and $\sigma_n(C) = ([\Sigma_P]_{2,3})^n$. Consequently we get $\bar{K}_P^{(n)} = \Sigma_P^{\odot n}$. Finally, this theorem follows directly from Lemma 13 which shows $\lambda_{\min}(\Sigma_P^{\odot n}) \ge \lambda_{\min}(\Sigma_P)$ for $n \ge 1$.

B.3 Two Features with Multi-dimensional Gaussian Distribution

Now we extend Theorem 4 to the case where $x_1 \in \mathbb{R}^{d_1}, x_2 \in \mathbb{R}^{d_2}$ are two subsets of the input x, and the input $x = (x_1, x_2)$ has Gaussian distribution.[4] Since f_1, f_2 are arbitrary functions and P, Q have the same marginal distribution, we can assume without loss of generality that $\mathbb{E}_P[x_i x_i^\top] = \mathbb{E}_Q[x_i x_i^\top] = I_{d_i}, \forall i \in [2]$.

The main theorem for this case is given below, whose proof is deferred to Appendix F.6.

Theorem 5. *For any distributions P, Q over variables $x = (x_1, x_2)$ where $x_1 \in \mathbb{R}^{d_1}, x_2 \in \mathbb{R}^{d_2}$, let $\Sigma_P = \mathbb{E}_P[xx^\top]$. If (x_1, x_2) follows from Gaussian distribution on P and Q with matching marginals and $E_P[x_1 x_1^\top] = I_{d_1}, E_P[x_2 x_2^\top] = I_{d_2}$, then*

$$\tau \le \frac{2}{\lambda_{\min}(\Sigma_P)}. \tag{19}$$

Compared with Theorem 4 where the features are not grouped, our condition for the covariance is almost the same — that is, $\lambda_{\min}(\Sigma_P)$ is bounded away from zero. However, the model class considered Theorem 5 is more powerful because it captures nonlinear interactions between features within the same group. As a compromise, the assumption on the marginals of P and Q is stronger because Theorem 5 requires matching marginals on each group of the features, whereas Theorem 4 only requires matching marginals on each individual feature.

In the following, we show the proof sketch of Theorem 5.

[4] Ideally, we want to handle the general version where the inputs are divided into multiple subsets, but due to technical reasons we can only use our proof techniques to two subsets.

Proof sketch of Theorem 5. We start by considering a simpler case when $d_1 = d_2$ and $\Sigma_{12} \triangleq \mathbb{E}_P[x_1 x_2^\top] \in \mathbb{R}^{d_1 \times d_2}$ has only diagonal terms. In other words, $([x_1]_i, [x_2]_i)$ are independent with every other coordinates in the input. In this case, we can decompose the multi-dimensional Gaussian kernel into products of one-dimensional Gaussian kernels in the following way

$$[A]_{x_1,x_2} = P(x_1, x_2)/\sqrt{P(x_1)P(x_2)} = \prod_{i=1}^{d_1} P([x_1]_i, [x_2]_i)/\sqrt{P([x_1]_i)P([x_2]_i)}. \tag{20}$$

Consequently, the singular values of A will be the products of singular values of these one-dimensional Gaussian kernels: $\{\prod_{i=1}^{d_1} |[\Sigma_{12}]_{i,i}|^{k_i}\}_{k_1,k_2,\cdots,k_{d_1} \geq 0}$. Therefore, the second largest singular value of A will be $\max_{i\in[d_1]} |[\Sigma_{12}]_{i,i}|$. Following the same reasoning as Theorem 4 we get

$$\tau \leq \frac{2}{1 - \max_{i\in[d_1]} |[\Sigma_{12}]_{i,i}|} = \frac{2}{1 - \sigma_{\max}(\Sigma_{12})}$$

where $\sigma_{\max}(\Sigma_{12})$ is the largest singular value of Σ_{12}. Combining with Lemma 15 we get Eq. (19).

Now we turn to the general cases where Σ_{12} is not diagonal. Recall that our model is $f(x) = f_1(x_1) + f_2(x_2)$ for some arbitrary functions f_1, f_2. Hence, we can rotate the inputs x_1, x_2 without affecting the model class. Formally speaking, for any orthogonal matrices $U \in \mathbb{R}^{d_1 \times d_1}$ and $V \in \mathbb{R}^{d_2 \times d_2}$,

$$\sup_{f\in\mathcal{F}} \frac{\mathbb{E}_Q[(f_1(x_1) + f_2(x_2))^2]}{\mathbb{E}_P[(f_1(x_1) + f_2(x_2))^2]} = \sup_{f\in\mathcal{F}} \frac{\mathbb{E}_Q[(f_1(Ux_1) + f_2(Vx_2))^2]}{\mathbb{E}_P[(f_1(Ux_1) + f_2(Vx_2))^2]}. \tag{21}$$

If U, V are the orthonormal matrices in the singular value decomposition $\Sigma_{12} = U^\top \Lambda_{12} V$ where Λ_{12} is a diagonal matrix, we get

$$\mathbb{E}_P[(Ux_1)(Vx_2)^\top] = \Lambda_{12}, \quad \mathbb{E}_P[(Ux_1)^\top (Ux_1)] = I, \quad \mathbb{E}_P[(Vx_2)^\top (Vx_2)] = I. \tag{22}$$

As a result, reusing the result from the previous case on inputs Ux_1, Vx_2 proves the desired result.

Remarks. Our current techniques can only handle the case when the input is divided into $k = 2$ subsets. This is because for $k \geq 3$ we must diagonalize multiple multi-dimensional Gaussian kernels simultaneously using the same set of eigenfunctions, as required in the proof of Theorem 4. However, these multi-dimensional Gaussian kernels do not share the same eigenfunctions because the rotation matrix U, V depends on the covariance $E_P[x_i x_j^\top]$. Hence, the proof strategy for Theorem 4 fails.

C Lower Bound

In this section, we prove a lower bound as a motivation to consider structured distributions shifts. On the high level, the following proposition shows that models learned on P cannot extrapolate to Q when the support of distribution Q is not contained in the support of P.

Proposition 6. *Let the model class $\mathcal{F}$ be the family of two-layer neural networks with ReLU activation: $\mathcal{F} = \left\{\sum_i a_i \mathrm{ReLU}(w_i^\top x + b_i) : w_i \in \mathbb{R}^d, a_i, b_i \in \mathbb{R}\right\}$. Suppose for simplicity that all the inputs have unit norm (i.e., $\|x\|_2 = 1$). If Q has non-zero probability mass on the set of points well-separated from the support of P in the sense that*

$$\exists \epsilon > 0, \quad Q(\{x : \|x\|_2 = 1, \mathrm{dist}(x, \mathrm{supp}(P)) \geq \epsilon\}) > 0, \tag{23}$$

we can construct a model $f \in \mathcal{F}$ such that $\|f\|_P = 0$ but $\|f\|_Q$ can be arbitrarily big.

A complete proof of this proposition is deferred to Appendix F.7. On a high level, we prove this proposition by construct a two-layer neural network g_t that represents a bump function around any given input $t \in S^{d-1}$. As a result, when t is a point in $\mathrm{supp}(Q) \setminus \mathrm{supp}(P)$, the model $g_t(x)$ will have zero ℓ_2 norm on P but have a positive ℓ_2 norm on Q. This construction is inspired by Dong et al. [2021, Theorem 5.1].

D Related Works

The most related work is Ben-David et al. [2010], where they use the $\mathcal{H}\Delta\mathcal{H}$-distance to measure the maximum discrepancy of two models $f, g \in \mathcal{F}$ on any distributions P, Q. However, it remains an open question to determine when $\mathcal{H}\Delta\mathcal{H}$-distance is small for *concrete* nonlinear model classes and distributions. On the technical side, the quantity τ is an analog of the $\mathcal{H}\Delta\mathcal{H}$-distance for regression problems, and we provide concrete examples where τ is upper bounded even if the distributions P, Q have significantly different support.

Another closely related settings are domain adaptation [Ganin and Lempitsky, 2015, Ghifary et al., 2016, Ganin et al., 2016] and domain generalization [Gulrajani and Lopez-Paz, 2020, Peters et al., 2016], where the algorithm actively improve the extrapolation of learned model either by using unlabeled data from the test domain [Sun and Saenko, 2016, Li et al., 2020a,b, Zhang et al., 2019], or learn an invariant model across different domains [Arjovsky et al., 2019, Peters et al., 2016, Gulrajani and Lopez-Paz, 2020]. In comparison, this paper studies a more basic question: whether a model trained on one distribution (without any implicit bias and unlabeled data from test domain) extrapolates to new distributions. There are also prior works that theoretically analyze algorithms that use additional (unlabeled) data from the test distribution, such as self-training [Wei et al., 2020, Chen et al., 2020], contrastive learning [Shen et al., 2022, HaoChen et al., 2022], label propagation [Cai et al., 2021], etc.

E Conclusions

In this paper, we propose to study domain shifts between P and Q with the structure that each feature's marginal distribution has good overlap between source and target domain but the joint distribution of the features may have a much bigger shift. As a first step toward understanding the extrapolation of nonlinear models, we prove sufficient conditions for the model $f(x) = \sum_{i=1}^{k} f_i(x_i)$ to extrapolate where f_i is an arbitrary function of a single feature. Even though the assumptions on the shift and function class is stylized, to the best of our knowledge, this is the first analysis of how nonlinear models extrapolate when source and target distribution *do not* have shared support in concrete settings.

There still remain many interesting open questions, which we leave as future works:

1. Our current proof can only deal with a restricted nonlinear model family of the special form $f(x) = \sum_{i=1}^{k} f_i(x_i)$ due to technical reasons. Can we extend to a more general model class?
2. In this paper, we focus on regression tasks with ℓ_2 loss for mathematical simplicity, whereas majority of the prior works focus on the classification problems. Do similar results also hold for classification problem?

F Missing Proofs

In the following, we present the missing proofs.

F.1 Proof of Proposition 1

In this section, we prove Proposition 1.

Proof of Proposition 1. By the definition of τ, for any $f \in \mathcal{F}$ we get

$$\frac{\mathbb{E}_Q[(f(x) - f^\star(x))^2]}{\mathbb{E}_P[(f(x) - f^\star(x))^2]} \leq \tau. \tag{24}$$

Or equivalently,

$$\mathbb{E}_Q[(f(x) - f^\star(x))^2] \leq \tau \mathbb{E}_P[(f(x) - f^\star(x))^2]. \tag{25}$$

As a result,

$$\mathbb{E}_Q[(y - f(x))^2] \leq 2\mathbb{E}_Q[(y - f^\star(x))^2] + 2\mathbb{E}_Q[(f(x) - f^\star(x))^2] \tag{26}$$

$$\leq 4\epsilon_{\mathcal{F}} + 2\tau \mathbb{E}_P[(f(x) - f^\star(x))^2] \leq 4\epsilon_{\mathcal{F}} + 4\tau \left(\mathbb{E}_P[(y - f(x))^2] + \mathbb{E}_P[(y - f^\star(x))^2]\right) \tag{27}$$

$$\leq (8\tau + 4)\epsilon_{\mathcal{F}} + 4\tau \mathbb{E}_P[(y - f(x))^2]. \tag{28}$$

□

F.2 Proof of Proposition 7

In this section, we state and prove Proposition 7.

Proposition 7. *Let $P(x_1, x_2)$ be a distribution over discrete random variables $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$, and recall that $\mathcal{F} = \{f_1(x_1) + f_2(x_2)\}$. If there exists subsets $U_1 \subseteq \mathcal{X}_1, U_2 \subseteq \mathcal{X}_2$ such that (1) $0 < |U_1| < |\mathcal{X}_1|$ and $0 < |U_2| < |\mathcal{X}_2|$, and (2) $P(x_1, x_2) = 0$ when $\mathbb{I}[x_1 \in U_1] \neq \mathbb{I}[x_2 \in U_2]$ (e.g., the distribution visualized in Figure 1b), then we can construct models $f, g \in \mathcal{F}$ and $x^\dagger \in \mathcal{X}_1 \times \mathcal{X}_2$ where $\|f - g\|_P = 0$ and $f(x^\dagger) \neq g(x^\dagger)$.*

Proof. Since the model class $\mathcal{F}$ is closed under addition, we only need to construct $h \in \mathcal{F}$ such that $\|h\|_P = 0$ but $h(x^\dagger) \neq 0$.

Let $h_1(x_1) = (-1)^{\mathbb{I}[x_1 \in U_1]}$ and $h_2(x_2) = -(-1)^{\mathbb{I}[x_2 \in U_2]}$. Consider the model $h(x) = h_1(x_1) + h_2(x_2)$. For any x_1, x_2 where $\mathbb{I}[x_1 \in U_1] = \mathbb{I}[x_2 \in U_2]$, we get $h(x) = 0$. Combining with condition (2) we get $\|h\|_P = 0$.

On the other hand, let $x^\dagger = (x_1^\dagger, x_2^\dagger)$ with $x_1^\dagger \in U_1$ and $x_2^\dagger \in \mathcal{X}_2 \setminus U_2$. Such $x^\dagger$ exists because of condition (1). Then $h(x^\dagger) = -2$. □

F.3 Proof of Theorem 2

In this section, we prove Theorem 2.

Proof of Theorem 2. Let

$$b_t(x_1, x_2) = \begin{cases} \mathbb{I}[x_1 = t], & \text{when } t \le r_1, \\ \mathbb{I}[x_2 = t - r_1], & \text{when } r_1 < t \le r_1 + r_2. \end{cases} \tag{29}$$

Then for any $f, f' \in \mathcal{F}$, we can always find $v \in \mathbb{R}^{r_1+r_2}$ such that $f(x_1, x_2) - f'(x_1, x_2) = \sum_{i=1}^{r_1+r_2} v_i b_i(x_1, x_2)$ for all x_1, x_2. Indeed, for any $f, f' \in \mathcal{F}$ the architecture of our model implies $(f - f')(x_1, x_2) = (f_1 - f_1')(x_1) + (f_2 - f_2')(x_2)$. Therefore, we can simply set $v_t = (f_1 - f_1')(t)$ for $1 \le t \le r_1$ and $v_t = (f_2 - f_2')(t - r_1)$ for $r_1 < t \le r_2$.

Consequently,

$$\tau = \sup_{f,f' \in \mathcal{F}} \frac{\mathbb{E}_Q[(f(x) - f'(x))^2]}{\mathbb{E}_P[(f(x) - f'(x))^2]} = \sup_{v \in \mathbb{R}^{r_1+r_2}} \frac{\mathbb{E}_Q[(\sum_{i=1}^{r_1+r_2} b_i(x_1, x_2)v_i)^2]}{\mathbb{E}_P[(\sum_{i=1}^{r_1+r_2} b_i(x_1, x_2)v_i)^2]} \tag{30}$$

$$= \sup_{v \in \mathbb{R}^{r_1+r_2}} \frac{\sum_{i,j=1}^{r_1+r_2} v_i v_j \mathbb{E}_Q[b_i(x_1, x_2)b_j(x_1, x_2)]}{\sum_{i,j=1}^{r_1+r_2} v_i v_j \mathbb{E}_P[b_i(x_1, x_2)b_j(x_1, x_2)]}. \tag{31}$$

The definition of $b_t(x_1, x_2)$ implies that for any distribution P,

$$\mathbb{E}_P[b_i(x_1, x_2)b_j(x_1, x_2)] = \begin{cases} \mathbb{I}[i = j]\, P(x_1 = i), & \text{when } 1 \le i, j \le r_1, \\ \mathbb{I}[i = j]\, P(x_2 = j), & \text{when } r_1 < i, j \le r_1 + r_2, \\ P(x_1 = i, x_2 = j), & \text{when } 1 \le i \le r_1 < j \le r_1 + r_2, \\ P(x_1 = j, x_2 = i), & \text{when } 1 \le j \le r_1 < i \le r_1 + r_2. \end{cases} \tag{32}$$

Then we have $[K_P]_{i,j} = \mathbb{E}_P[b_i(x_1, x_2)b_j(x_1, x_2)]$. Consequently,

$$\tau = \sup_{v \in \mathbb{R}^{r_1+r_2}} \frac{\sum_{i,j=1}^{r_1+r_2} v_i v_j \mathbb{E}_Q[b_j(x_1, x_2)b_j(x_1, x_2)]}{\sum_{i,j=1}^{r_1+r_2} v_i v_j \mathbb{E}_P[b_j(x_1, x_2)b_j(x_1, x_2)]} = \sup_{v \in \mathbb{R}^{r_1+r_2}} \frac{v^\top K_Q v}{v^\top K_P v}, \tag{33}$$

Let $u \in \mathbb{R}^{r_1+r_2}$ such that

$$[u]_i = \begin{cases} 1, & \text{when } i \le r_1, \\ -1, & \text{when } r_1 < i \le r_2. \end{cases}$$

We claim that u is a eigenvector to both K_Q and K_P with eigenvalue 0. To see this, for any distribution P and $i \in [r_1]$ we have

$$[K_P u]_i = P(x_1 = i)[u]_i + \sum_{j=1}^{r_2} P(x_1 = i, x_2 = j)[u]_{r_1+j} \tag{34}$$

$$= P(x_1 = i) - \sum_{j=1}^{r_2} P(x_1 = i, x_2 = j) = 0. \tag{35}$$

Similarly for $i \in [r_2]$ we have $[K_P u]_{r_1+i} = 0$. Combining these two cases together we prove $K_P u = 0$.

Then, by algebraic manipulation,

$$\sup_{v\in\mathbb{R}^{r_1+r_2}} \frac{v^\top K_Q v}{v^\top K_P v} = \sup_{v\in\mathbb{R}^{r_1+r_2}, v\perp u} \frac{v^\top K_Q v}{v^\top K_P v} = \sup_{v\in\mathbb{R}^{r_1+r_2}, v\perp u} \frac{v^\top D_Q^{1/2} \bar{K}_Q D_Q^{1/2} v}{v^\top D_P^{1/2} \bar{K}_P D_P^{1/2} v} \tag{36}$$

$$\leq \frac{\lambda_{\max}(\bar{K}_Q)}{\lambda_2(\bar{K}_P)} \sup_v \frac{\|D_Q^{1/2} v\|_2^2}{\|D_P^{1/2} v\|_2^2} \leq \frac{\lambda_{\max}(\bar{K}_Q)}{\lambda_2(\bar{K}_P)} \max_{i\in[2], t\in[r_i]} \frac{P(x_i = t)}{Q(x_i = t)}. \tag{37}$$

As a result, we only need to prove $\lambda_{\max}(\bar{K}_Q) \leq 2$. To this end, note that

$$\lambda_{\max}(\bar{K}_Q) = \sup_{u\in\mathbb{R}^{r_1+r_2}} \frac{u^\top \bar{K}_Q u}{u^\top u} = \sup_{v\in\mathbb{R}^{r_1+r_2}} \frac{v^\top K_Q v}{v^\top D_Q v} \tag{38}$$

$$= \sup_{v\in\mathbb{R}^{r_1+r_2}} \frac{\sum_{i\in[r_1], j\in[r_2]} P(x_1 = i, x_2 = j)([v]_i + [v]_{r_1+j})^2}{\sum_{i\in[r_1]} P(x_1 = i)[v]_i^2 + \sum_{j\in[r_2]} P(x_2 = j)[v]_{r_1+j})^2} \tag{39}$$

$$\leq \sup_{v\in\mathbb{R}^{r_1+r_2}} \frac{2\sum_{i\in[r_1], j\in[r_2]} P(x_1 = i, x_2 = j)([v]_i^2 + [v]_{r_1+j}^2)}{\sum_{i\in[r_1]} P(x_1 = i)[v]_i^2 + \sum_{j\in[r_2]} P(x_2 = j)[v]_{r_1+j})^2} \tag{40}$$

$$\leq 2. \tag{41}$$

□

F.4 Proof of Corollary 3

In this section, we prove Corollary 3.

Proof of Corollary 3. Recall that Theorem 2 proves

$$\tau \leq 2\lambda_2(\bar{K}_P)^{-1} \max_{i\in[2], t\in[r_i]} \frac{P(x_i = t)}{Q(x_i = t)}. \tag{42}$$

As a result, we only need to prove conditions (a) and (b) implies the RHS of Eq. (42) is finite.

First of all, by definition condition (a) implies $\max_{i\in[2],t\in[r_i]} \frac{P(x_i=t)}{Q(x_i=t)} < \infty$. Secondly, we prove that $\lambda_2(\bar{K}_P) = 0$ implies the negation of condition (b), which is equivalent to the statement "condition (b) implies $\lambda_2(\bar{K}_P) > 0$". By Cheeger's inequality (Eq. (7)), $\lambda_2(\bar{K}_P) = 0$ implies that $\phi(G_P) = 0$. By the definition of $\phi(G_P)$, there exists subsets $U_1 \subseteq [r_1], U_2 \subseteq [r_2]$ such that

1. $P(x_1 = u_1, x_2 = u_2) = 0$ when $u_1 \in U_1, u_2 \notin U_2$, and
2. $P(x_1 = u_1, x_2 = u_2) = 0$ when $u_1 \notin U_1, u_2 \in U_2$.

Then, we can reorganize $P(x_1, x_2)$ by listing the rows corresponding to U_1 first, and the columns corresponding to U_2 first. As a result, $P(x_1, x_2)$ will be a block diagonal matrix where the first block has dimension $|U_1|\times|U_2|$ and the second block $(r_1 - |U_1|) \times (r_2 - |U_2|)$. Consequently, we prove that $\lambda_2(\bar{K}_P) = 0$ implies the negation of condition (b).

Hence, with condition (b) we get $\lambda_{\bar{K}_P} > 0$. It follows that $\lambda_2(\bar{K}_P)^{-1} < \infty$.

Combining everything together we have $\tau < \infty$. □

F.5 Proof of Theorem 4

In this section, we prove Theorem 4.

Proof of Theorem 4. Now consider any fixed model pairs $f, f' \in \mathcal{F}$. Let $g_i(x_i) = f_i(x_i) - f'_i(x_i)$ and $g = \sum_{i=1}^d g_i(x_i)$. Recall that we assume the marginals satisfies $x_i \sim \mathcal{N}(0,1)$ for every $i \in [d]$. As a result, Lemma 9 implies that there exists coefficients $\{\alpha_i^{(n)}\}_{i\in[d],n\in\mathbb{Z}_+}$ with

$$g_i(x_i)\sqrt{P(x_i)} = \sum_{n\geq 0} \alpha_i^{(n)} \psi_n(x_i), \quad \forall i \in [d], \tag{43}$$

where $\{\psi_n(\cdot)\}_{n\in\mathbb{Z}_+}$ are a set of orthonormal basis of $L^2(\mathbb{R})$ defined by

$$\psi_n(x) \triangleq H_n\left(\frac{1}{\sqrt{2}}x\right)\exp\left(-\frac{1}{4}x^2\right)(2\pi)^{-1/4}(2^n n!)^{-1/2}.$$

As a result,

$$\mathbb{E}_P\left[\left(\sum_{i=1}^d g_i(x_i)\right)^2\right] = \int \sum_{i,j=1}^d g_i(x_i)g_j(x_j)P(x_1,\cdots,x_d)\mathrm{d}x_1\cdots x_d \tag{44}$$

$$= \sum_{i=1}^d \int g_i(x_i)^2 P(x_i)\mathrm{d}x_i + 2\sum_{1\leq i<j\leq d}\int g_i(x_i)g_j(x_j)P(x_i,x_j)\mathrm{d}x_i x_j \tag{45}$$

$$= \sum_{i=1}^d \int \Big(\sum_{n\geq 0}\alpha_i^{(n)}\psi_n(x_i)\Big)^2 \mathrm{d}x_i \tag{46}$$

$$+ 2\sum_{1\leq i<j\leq d}\int (g_i(x_i)\sqrt{P(x_i)})(g_j(x_j)\sqrt{P(x_j)})\frac{P(x_i,x_j)}{\sqrt{P(x_i)P(x_j)}}\mathrm{d}x_i x_j \tag{47}$$

$$= \sum_{i=1}^d\sum_{n\geq 0}(\alpha_i^{(n)})^2 + 2\sum_{1\leq i<j\leq d}\int \Big(\sum_{n\geq 0}\alpha_i^{(n)}\psi_n(x_i)\Big)\Big(\sum_{n\geq 0}\alpha_j^{(n)}\psi_n(x_j)\Big)\frac{P(x_i,x_j)}{\sqrt{P(x_i)P(x_j)}}\mathrm{d}x_i x_j. \tag{48}$$

By Theorem 11, we have

$$\frac{P(x_i,x_j)}{\sqrt{P(x_i)P(x_j)}} = \sum_{n\geq 0}([\Sigma_P]_{i,j})^n \psi_n(x_i)\psi_n(x_j). \tag{49}$$

Continuing Eq. (48) we get

$$\sum_{1\leq i<j\leq d}\int \Big(\sum_{n\geq 0}\alpha_i^{(n)}\psi_n(x_i)\Big)\Big(\sum_{n\geq 0}\alpha_j^{(n)}\psi_n(x_i)\Big)\frac{P(x_i,x_j)}{\sqrt{P(x_i)P(x_j)}}\mathrm{d}x_i x_j \tag{50}$$

$$= \sum_{1\leq i<j\leq d}\int \Big(\sum_{n\geq 0}\alpha_i^{(n)}\psi_n(x_i)\Big)\Big(\sum_{n\geq 0}\alpha_j^{(n)}\psi_n(x_i)\Big)\Big(\sum_{n\geq 0}([\Sigma_P]_{i,j})^n\psi_n(x_i)\psi_n(x_j)\Big)\mathrm{d}x_i x_j \tag{51}$$

$$= \sum_{1\leq i<j\leq d}\sum_{n\geq 0}\alpha_i^{(n)}\alpha_j^{(n)}([\Sigma_P]_{i,j})^n. \tag{52}$$

As a result,

$$\mathbb{E}_P\left[\left(\sum_{i=1}^d g_i(x_i)\right)^2\right] = \sum_{i=1}^d\sum_{n\geq 0}(\alpha_i^{(n)})^2 + 2\sum_{1\leq i<j\leq d}\sum_{n\geq 0}\alpha_i^{(n)}\alpha_j^{(n)}([\Sigma_P]_{i,j})^n. \tag{53}$$

Define $\kappa = \sup_{n\geq 0}\sup_{v\in\mathbb{R}^d}\frac{v^\top(\Sigma_\mathrm{Q})^{\odot n}v}{v^\top(\Sigma_\mathrm{P})^{\odot n}v}$. In the following, we prove

$$\mathbb{E}_Q\left[\left(\sum_{i=1}^d g_i(x_i)\right)^2\right] \leq \kappa\mathbb{E}_P\left[\left(\sum_{i=1}^d g_i(x_i)\right)^2\right]. \tag{54}$$

For any fixed $n \geq 0$, let $\Sigma_P^{(n)} \in \mathbb{R}^{d\times d}$ be the matrix where $[\Sigma_P^{(n)}]_{i,j} = ([\Sigma_P]_{i,j})^n$ (define $\Sigma_Q^{(n)}$ similarly), and $\vec{\alpha}^{(n)} \in \mathbb{R}^d$ the vector where $[\vec{\alpha}^{(n)}]_i = \alpha_i^{(n)}$. Then

$$\mathbb{E}_P\left[\left(\sum_{i=1}^d g_i(x_i)\right)^2\right] = \sum_{n\geq 0}(\vec{\alpha}^{(n)})^\top \Sigma_P^{(n)}\vec{\alpha}^{(n)}, \quad \mathbb{E}_Q\left[\left(\sum_{i=1}^d g_i(x_i)\right)^2\right] = \sum_{n\geq 0}(\vec{\alpha}^{(n)})^\top \Sigma_Q^{(n)}\vec{\alpha}^{(n)}.$$

By the definition of κ we get $\Sigma_Q^{(n)} \preceq \kappa\Sigma_P^{(n)}$ for all $n \geq 0$. Consequently,

$$\sum_{n\geq 0}(\vec{\alpha}^{(n)})^\top \Sigma_Q^{(n)}\vec{\alpha}^{(n)} \leq \kappa\Big(\sum_{n\geq 0}(\vec{\alpha}^{(n)})^\top \Sigma_P^{(n)}\vec{\alpha}^{(n)}\Big), \tag{55}$$

which implies Eq. (54).

Now we prove $\kappa \leq \frac{d}{\lambda_{\min}(\Sigma_P)}$. Since Σ_Q is a covariance matrix with $\mathrm{diag}(\Sigma_Q) = I$, we have $\|\Sigma_Q\|_2 \leq d\|\Sigma_Q\|_\infty \leq d$. In addition, Lemma 13 proves that $\lambda_{\min}(\Sigma_P^{\odot n}) \geq \lambda_{\min}(\Sigma_P)$ for every $n \geq 1$. As a result,

$$\inf_{n\geq 1}\inf_{v\in\mathbb{R}^d}(v^\top(\Sigma_{\mathrm{P}})^{\odot n}v)^{-1} \geq \lambda_{\min}(\Sigma_P^{\odot n}) \geq \lambda_{\min}(\Sigma_P).$$

Since $n = 0 \implies (\Sigma_{\mathrm{P}})^{\odot n} = (\Sigma_{\mathrm{Q}})^{\odot n}$, we get

$$\kappa = \sup_{n\geq 0}\sup_{v\in\mathbb{R}^d}\frac{v^\top(\Sigma_{\mathrm{Q}})^{\odot n}v}{v^\top(\Sigma_{\mathrm{P}})^{\odot n}v} = \max\left\{1, \sup_{n\geq 1}\sup_{v\in\mathbb{R}^d}\frac{v^\top(\Sigma_{\mathrm{Q}})^{\odot n}v}{v^\top(\Sigma_{\mathrm{P}})^{\odot n}v}\right\} \leq \max\left\{1, \frac{d}{\lambda_{\min}(\Sigma_P)}\right\}. \tag{56}$$

By the fact that $\lambda_{\min}(\Sigma_P) \leq d\|\Sigma_P\|_\infty \leq d$, we prove the desired result. □

In the following, we state and proof Lemma 8.

Lemma 8. *For any distributions P, Q with matching marginals, define P' to be the density of the random variable $t = (t_1, \cdots, t_d)$ where $t_i \triangleq \mathrm{var}_P(x_i)^{-1/2}(x_i - \mathbb{E}_P[x_i])$, $x \sim P$ (and define Q' similarly). Then we have*

$$\tau(P, Q, \mathcal{F}) = \tau(P', Q', \mathcal{F}). \tag{57}$$

Proof. Let $\{h_i : \mathbb{R} \to \mathbb{R}\}_{i=1,2,\cdots,d}$ be any set of invertible functions. Since $f = \sum_i f_i(x_i) \in \mathcal{F}$ is equivalent to $f \circ h \triangleq \sum_i f_i(h_i(x_i)) \in \mathcal{F}$, we get

$$\sup_{f\in\mathcal{F}}\frac{\mathbb{E}_Q[\sum_i f_i(x_i)^2]}{\mathbb{E}_P[\sum_i f_i(x_i)^2]} = \sup_{f\in\mathcal{F}}\frac{\mathbb{E}_Q[\sum_i f_i(h_i(x_i))^2]}{\mathbb{E}_P[\sum_i f_i(h_i(x_i))^2]}. \tag{58}$$

Let P' be the density of the random variable $(h_1(x_1), \cdots, h_d(x_d))$ when $x \sim P$ (and define Q' similarly). Then we have

$$\frac{\mathbb{E}_{x\sim Q}[\sum_i f_i(h_i(x_i))^2]}{\mathbb{E}_{x\sim P}[\sum_i f_i(h_i(x_i))^2]} = \frac{\mathbb{E}_{t\sim Q'}[\sum_i f_i(t_i)^2]}{\mathbb{E}_{t\sim P'}[\sum_i f_i(t_i)^2]}. \tag{59}$$

Consequently,

$$\tau(P, Q, \mathcal{F}) = \sup_{f\in\mathcal{F}}\frac{\mathbb{E}_Q[\sum_i f_i(x_i)^2]}{\mathbb{E}_P[\sum_i f_i(x_i)^2]} = \sup_{f\in\mathcal{F}}\frac{\mathbb{E}_Q[\sum_i f_i(h_i(x_i))^2]}{\mathbb{E}_P[\sum_i f_i(h_i(x_i))^2]} \tag{60}$$

$$= \sup_{f\in\mathcal{F}}\frac{\mathbb{E}_{t\sim Q'}[\sum_i f_i(t_i)^2]}{\mathbb{E}_{t\sim P'}[\sum_i f_i(t_i)^2]} = \tau(P', Q', \mathcal{F}). \tag{61}$$

Finally, this lemma is proved by taking $h_i(x) = \mathrm{var}_P(x_i)^{-1/2}(x - \mathbb{E}_P[x_i])$. □

F.6 Proof of Theorem 5

In this section, we prove Theorem 5.

Proof of Theorem 5. Without loss of generality, we assume $d_2 \le d_1$. Since $\mathcal{F}$ is closed under addition, we have

$$\tau = \max_{g\in\mathcal{F}} \frac{\mathbb{E}_Q[(g_1(x_1)+g_2(x_2))^2]}{\mathbb{E}_P[(g_1(x_1)+g_2(x_2))^2]}.$$

In the following, we first prove

$$\tau \le \frac{2}{1-\sigma_{\max}(\Sigma_{12})} \tag{62}$$

by considering two cases separately.

Case 1: when $g=(g_1,g_2)\in\mathcal{F}$ **satisfies** $\mathbb{E}_P[g_1(x_1)]=\mathbb{E}_P[g_2(x_2)]=0$**.** Since P and Q have matching marginals on x_1 and x_2 respectively, we get

$$\mathbb{E}_Q[g_1(x_1)^2+g_2(x_2)^2] = \mathbb{E}_Q[g_1(x_1)^2]+\mathbb{E}_Q[g_2(x_2)^2] \tag{63}$$

$$= \mathbb{E}_P[g_1(x_1)^2]+\mathbb{E}_P[g_2(x_2)^2] = \mathbb{E}_P[(g_1(x_1)+g_2(x_2))^2]. \tag{64}$$

As a result, by the definition of τ we have

$$\begin{aligned}&\frac{\mathbb{E}_Q[(g_1(x_1)+g_2(x_2))^2]}{\mathbb{E}_P[(g_1(x_1)+g_2(x_2))^2]} = \frac{\mathbb{E}_Q[(g_1(x_1)+g_2(x_2))^2]}{\mathbb{E}_Q[g_1(x_1)^2+g_2(x_2)^2]}\frac{\mathbb{E}_P[g_1(x_1)^2+g_2(x_2)^2]}{\mathbb{E}_P[(g_1(x_1)+g_2(x_2))^2]}\\ \le\ & \frac{\mathbb{E}_Q[2(g_1(x_1)^2+g_2(x_2)^2)]}{\mathbb{E}_Q[g_1(x_1)^2+g_2(x_2)^2]}\frac{\mathbb{E}_P[g_1(x_1)^2+g_2(x_2)^2]}{\mathbb{E}_P[(g_1(x_1)+g_2(x_2))^2]} = 2\frac{\mathbb{E}_P[g_1(x_1)^2+g_2(x_2)^2]}{\mathbb{E}_P[(g_1(x_1)+g_2(x_2))^2]}.\end{aligned}$$

Hence, we only need to prove

$$(1-\sigma_{\max}(\Sigma_{12}))\mathbb{E}_P[g_1(x_1)^2+g_2(x_2)^2] \le \mathbb{E}_P[(g_1(x_1)+g_2(x_2))^2]. \tag{65}$$

Let $\Sigma_{12}=U\Lambda V^\top$ be the singular value decomposition of Σ_{12}, where $\Lambda\in\mathbb{R}^{d_1\times d_2}$ is a diagonal matrix with entries $\sigma_1,\cdots,\sigma_{d_2}$.

Following Theorem 12, we define $t_1=U^\top x_1\in\mathbb{R}^{d_1}, t_2=V^\top x_2\in\mathbb{R}^{d_2}$. Consequently, $\mathbb{E}_P[t_1 t_2^\top]=\Lambda$ and

$$(t_1,t_2)\sim\mathcal{N}\left(0,\begin{pmatrix} I_{d_1} & \Lambda\\ \Lambda^\top & I_{d_2}\end{pmatrix}\right). \tag{66}$$

Let $P'(t_1,t_2)$ be the density of the distribution of (t_1,t_2), and $P'(t_1),P'(t_2)$ the density of corresponding marginals. As a result,

$$\mathbb{E}_P[(g_1(x_1)+g_2(x_2))^2] = \mathbb{E}_{P'}[(g_1(t_1)+g_2(t_2))^2] \tag{67}$$

$$= \sum_{i=1}^2\int g_i(t_i)^2P'(t_i)\mathrm{d}t_i + 2\int g_1(t_1)g_2(t_2)P'(t_1,t_2)\mathrm{d}t_1t_2. \tag{68}$$

Since $g_i(t_i)\sqrt{P'(t_i)}$ is square integrable, there exists weights $\{\alpha_i^{(n_1,\cdots,n_{d_i})}\}_{n_1,\cdots,n_{d_i}\ge 0}$ such that

$$\forall i\in[2],\quad g_i(t_i)\sqrt{P'(t_i)} = \sum_{n_1,\cdots,n_{d_i}\ge 0}\alpha_i^{(n_1,\cdots,n_{d_i})}\prod_{j=1}^{d_i}\psi_{n_j}([t_i]_j), \tag{69}$$

where $\{\prod_{j=1}^{d_i}\psi_{n_j}([t_i]_j)\}_{n_1,\cdots,n_{d_i}\ge 0}$ forms an orthonormal basis of $L^2(\mathbb{R}^{d_i})$. Consequently,

$$\mathbb{E}_P[g_1(x_1)^2+g_2(x_2)^2] = \mathbb{E}_{P'}[g_1(t_1)^2+g_2(t_2)^2] \tag{70}$$

$$= \sum_{n_1,\cdots,n_{d_1}}\left(\alpha_1^{(n_1,\cdots,n_{d_1})}\right)^2 + \sum_{n_1,\cdots,n_{d_2}}\left(\alpha_2^{(n_1,\cdots,n_{d_2})}\right)^2. \tag{71}$$

Now we turn to the RHS of Eq. (65). Continuing Eq. (68) and apply Theorem 12 we get,

$$\sum_{i=1}^{2} \int g_i(t_i)^2 P'(t_i) \mathrm{d}t_i = \sum_{i=1}^{2} \sum_{n_1,\cdots,n_{d_i}} \left(\alpha_i^{(n_1,\cdots,n_{d_i})}\right)^2$$

and

$$\begin{aligned}
&\int g_1(t_1) g_2(t_2) P'(t_1,t_2) \mathrm{d}t_1 t_2 \\
= &\int \prod_{i=1}^{2} \left(\sum_{n_1,\cdots,n_{d_1}\geq 0} \alpha_i^{(n_1,\cdots,n_{d_i})} \prod_{j=1}^{d_i} \psi_{n_j}([t_i]_j) \right) \prod_{i=1}^{d_2} \left(\sum_{n=0}^{\infty} \sigma_i^n \psi_n([t_1]_i) \psi_n([t_2]_i) \right) \prod_{i=d_2+1}^{d_1} \psi_0([t_1]_i) \mathrm{d}t_1 t_2 \\
= &\sum_{n_1,\cdots,n_{d_2}\geq 0} \alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)} \alpha_2^{(n_1,\cdots,n_{d_2})} \prod_{i=1}^{d_2} \sigma_i^{n_i}.
\end{aligned}$$

As a result,

$$\begin{aligned}
&\mathbb{E}_{P'}[(g_1(t_1)+g_2(t_2))^2] &(72)\\
=&\sum_{i=1}^{2} \sum_{n_1,\cdots,n_{d_i}} \left(\alpha_i^{(n_1,\cdots,n_{d_i})}\right)^2 + 2 \sum_{n_1,\cdots,n_{d_2}\geq 0} \alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)} \alpha_2^{(n_1,\cdots,n_{d_2})} \prod_{i=1}^{d_2} \sigma_i^{n_i} &(73)\\
=&\sum_{n_1,\cdots,n_{d_2}} \begin{pmatrix} \alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)} \\ \alpha_2^{(n_1,\cdots,n_{d_2})} \end{pmatrix}^\top \begin{pmatrix} 1 & \prod_{i=1}^{d_2} \sigma_i^{n_i} \\ \prod_{i=1}^{d_2} \sigma_i^{n_i} & 1 \end{pmatrix} \begin{pmatrix} \alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)} \\ \alpha_2^{(n_1,\cdots,n_{d_2})} \end{pmatrix} \\
&+ \sum_{n_1,\cdots,n_{d_1}} \mathbb{I}\left[n_{d_2+1}+\cdots+n_{d_1} \neq 0\right] \left(\alpha_i^{(n_1,\cdots,n_{d_i})}\right)^2. &(74)
\end{aligned}$$

Now for any $n_1,\cdots,n_{d_2} \geq 0$, consider the matrix

$$\begin{pmatrix} 1 & \prod_{i=1}^{d_2} \sigma_i^{n_i} \\ \prod_{i=1}^{d_2} \sigma_i^{n_i} & 1 \end{pmatrix}.$$

When $n_1+\cdots+n_{d_2} \neq 0$, we have $\prod_{i=1}^{d_2} \sigma_i^{n_i} \leq \max_{i\in[d_2]} \sigma_i$ because Lemma 14 implies $\sigma_i \leq 1$ for every $i \in [d_2]$. Therefore in this case,

$$\begin{aligned}
&\begin{pmatrix} \alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)} \\ \alpha_2^{(n_1,\cdots,n_{d_2})} \end{pmatrix}^\top \begin{pmatrix} 1 & \prod_{i=1}^{d_2} \sigma_i^{n_i} \\ \prod_{i=1}^{d_2} \sigma_i^{n_i} & 1 \end{pmatrix} \begin{pmatrix} \alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)} \\ \alpha_2^{(n_1,\cdots,n_{d_2})} \end{pmatrix} &(75)\\
\geq &(1-\sigma_{\max}(\Sigma_{12})) \left(\left(\alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)}\right)^2 + \left(\alpha_2^{(n_1,\cdots,n_{d_2})}\right)^2 \right). &(76)
\end{aligned}$$

When $n_1+\cdots+n_{d_2} = 0$, by Eq. (69) we get

$$\begin{aligned}
\alpha_1^{(0,\cdots,0)} &= \int g_1(t_1) \sqrt{P'(t_1)} \prod_{j=1}^{d_1} \psi_0([t_1]_j) \mathrm{d}t_1 &(77)\\
&= \int g_1(t_1) \sqrt{P'(t_1)} \prod_{j=1}^{d_i} \sqrt{P'([t_1]_j)} \mathrm{d}t_1 &(78)\\
&= \int g_1(t_1) P'(t_1) \mathrm{d}t_1 = \mathbb{E}_{P'}[g_1(t_1)] = \mathbb{E}_P[g_1(x_1)] = 0. &(79)
\end{aligned}$$

Similarly, $\alpha_2^{(0,\cdots,0)} = 0$. As a result, when $n_1+\cdots+n_{d_2} = 0$ we have

$$\begin{pmatrix} \alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)} \\ \alpha_2^{(n_1,\cdots,n_{d_2})} \end{pmatrix}^\top \begin{pmatrix} 1 & \prod_{i=1}^{d_2} \sigma_i^{n_i} \\ \prod_{i=1}^{d_2} \sigma_i^{n_i} & 1 \end{pmatrix} \begin{pmatrix} \alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)} \\ \alpha_2^{(n_1,\cdots,n_{d_2})} \end{pmatrix} \quad (80)$$

$$= \left(\left(\alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)}\right)^2 + \left(\alpha_2^{(n_1,\cdots,n_{d_2})}\right)^2\right) = 0. \tag{81}$$

Combining these two cases together, we get

$$\sum_{n_1,\cdots,n_{d_2}} \begin{pmatrix} \alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)} \\ \alpha_2^{(n_1,\cdots,n_{d_2})} \end{pmatrix}^\top \begin{pmatrix} 1 & \prod_{i=1}^{d_2} \sigma_i^{n_i} \\ \prod_{i=1}^{d_2} \sigma_i^{n_i} & 1 \end{pmatrix} \begin{pmatrix} \alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)} \\ \alpha_2^{(n_1,\cdots,n_{d_2})} \end{pmatrix} \tag{82}$$

$$\geq (1 - \sigma_{\max}(\Sigma_{12})) \left(\sum_{n_1,\cdots,n_{d_2}} \left(\alpha_1^{(n_1,\cdots,n_{d_2},0,\cdots,0)}\right)^2 + \left(\alpha_2^{(n_1,\cdots,n_{d_2})}\right)^2\right). \tag{83}$$

Plug in to Eq. (74) we get

$$\mathbb{E}_P[(g_1(x_1) + g_2(x_2))^2] = \mathbb{E}_{P'}[(g_1(t_1) + g_2(t_2))^2] \tag{84}$$

$$\geq (1 - \sigma_{\max}(\Sigma_{12})) \left(\sum_{n_1,\cdots,n_{d_1}} \left(\alpha_1^{(n_1,\cdots,n_{d_1})}\right)^2 + \sum_{n_1,\cdots,n_{d_2}} \left(\alpha_2^{(n_1,\cdots,n_{d_2})}\right)^2\right) \tag{85}$$

$$= (1 - \sigma_{\max}(\Sigma_{12}))\mathbb{E}_P[g_1(x_1)^2 + g_2(x_2)^2], \tag{86}$$

which finished the proof for Eq. (65).

Case 2: when $g = (g_1, g_2) \in \mathcal{F}$ **satisfies** $\mathbb{E}_P[g_1(x_1) + g_2(x_2)] = 0$**.** Let $\mu = \mathbb{E}_P[g_1(x_1)]$. Define $g' = (g'_1, g'_2)$ where $g'_1(x_1) = g_1(x_1) - \mu, g'_2(x_2) = g_2(x_2) + \mu$. By the definition of $\mathcal{F}$ we get $g' \in \mathcal{F}$, and g' satisfies $\mathbb{E}_P[g_1(x_1)] = \mathbb{E}_P[g_2(x_2)] = 0$. Because $g_1(x_1)+g_2(x_2) = g'_1(x_1)+g'_2(x_2)$ for all $x_1 \in \mathbb{R}^{d_1}, x_2 \in \mathbb{R}^{+d_2}$, plugging in the result of case 1 we get

$$\frac{\mathbb{E}_Q[(g_1(x_1) + g_2(x_2))^2]}{\mathbb{E}_P[(g_1(x_1) + g_2(x_2))^2]} = \frac{\mathbb{E}_Q[(g'_1(x_1) + g'_2(x_2))^2]}{\mathbb{E}_P[(g'_1(x_1) + g'_2(x_2))^2]} \leq \frac{1}{1 - \sigma_{\max}(\Sigma_{12})}. \tag{87}$$

Case 3: when $g = (g_1, g_2) \in \mathcal{F}$**.** Now we consider the most general case. Let $\nu = \mathbb{E}_P[g_1(x_1) + g_2(x_2)]$. Since P and Q have matching marginals on x_1 and x_2 respectively, we get

$$\mathbb{E}_Q[g_1(x_1) + g_2(x_2)] = \mathbb{E}_Q[g_1(x_1)] + \mathbb{E}_Q[g_2(x_2)] \tag{88}$$

$$= \mathbb{E}_P[g_1(x_1)] + \mathbb{E}_P[g_2(x_2)] = \mathbb{E}_P[g_1(x_1) + g_2(x_2)] = \nu. \tag{89}$$

Define $g' = (g'_1, g'_2)$ where $g'_1(x_1) = g_1(x_1) - \nu, g'_2(x_2) = g_2(x_2)$. It follows that

$$\frac{\mathbb{E}_Q[(g_1(x_1) + g_2(x_2))^2]}{\mathbb{E}_P[(g_1(x_1) + g_2(x_2))^2]} = \frac{\mathbb{E}_Q[(g'_1(x_1) + g'_2(x_2))^2] + \nu^2}{\mathbb{E}_P[(g'_1(x_1) + g'_2(x_2))^2] + \nu^2} \tag{90}$$

$$\leq \max\left\{\frac{\mathbb{E}_Q[(g'_1(x_1) + g'_2(x_2))^2]}{\mathbb{E}_P[(g'_1(x_1) + g'_2(x_2))^2]}, 1\right\} \leq \frac{1}{1 - \sigma_{\max}(\Sigma_{12})}, \tag{91}$$

where the last inequality comes from applying case 2's result to g'.

Combining these three cases together we prove Eq. (62). Then Eq. (19) follows directly from Eq. (62) and Lemma 15. □

F.7 Proof of Proposition 6

In this section, we prove Proposition 6.

Proof of Proposition 6. First, for any $t \in S^{d-1}$ and $\epsilon > 0$ we construct a two-layer neural network $h_{t,\epsilon}(x)$ such that

1. for all $x \in S^{d-1}$, $h_{t,\epsilon}(x) = 0$ if $\|x - t\|_2 \geq \epsilon$,
2. for all $x \in S^{d-1}$, $h_{t,\epsilon}(x) \geq 1$ if $\|x - t\|_2 \leq \epsilon/2$, and
3. for all $x \in S^{d-1}$, $h_{t,\epsilon}(x) \geq 0$.

Recall that a two layer neural network is parameterized as $\sum_i a_i \mathrm{ReLU}(w_i^\top x + b_i)$. Then $h_{t,\epsilon}(x)$ can be constructed using one neuron by setting $w_1 = t, b_1 = -1 + \epsilon^2/2$ and $w_1 = 8/(3\epsilon^2)$.

We can verify the construction as follows. When $\|x - t\|_2 \geq \epsilon$, we get

$$\|x - t\|_2^2 \geq \epsilon^2 \implies 2 - 2x^\top t \geq \epsilon^2 \implies x^\top t \leq 1 - \frac{\epsilon^2}{2}. \tag{92}$$

As a result,

$$a_1 \mathrm{ReLU}(w_1^\top x + b_1) = \frac{8}{3\epsilon^2} \mathrm{ReLU}(t^\top x - 1 + \epsilon^2/2) = 0.$$

When $\|x - t\|_2 \leq \epsilon/2$, we get

$$\|x - t\|_2^2 \leq \epsilon^2/4 \implies 2 - 2x^\top t \leq \epsilon^2/4 \implies x^\top t \geq 1 - \frac{\epsilon^2}{8}. \tag{93}$$

As a result,

$$a_1 \mathrm{ReLU}(w_1^\top x + b_1) = \frac{8}{3\epsilon^2} \mathrm{ReLU}(t^\top x - 1 + \epsilon^2/2) \geq 1.$$

Now we construct a two layer neural network f such that $\|f\|_P = 0$ but $\|f\|_Q = 0$. Let $X = \{x : x \in S^{d-1}, \mathrm{dist}(x, \mathrm{supp}(P)) \geq \epsilon\}$. By the condition of this proposition we have $Q(X) > 0$.

Let C be the minimum $\epsilon/2$-covering of the set X, and thus $|C| < \infty$. Consider the 2-layer neural network $f(x) = \sum_{c \in C} h_{c,\epsilon}$. It follows from the definition of X that $\|x - c\|_2 \geq \epsilon$ for every $x \in \mathrm{supp}(P)$. Consequently, $f(x) = \sum_{c \in C} h_{c,\epsilon} = 0$ for every $x \in \mathrm{supp}(P)$, which implies that $\|f\|_P = 0$.

Now for every $x \in \mathcal{X}$, there exists $t \in \mathcal{C}$ such that $\|x - t\|_2 \leq \epsilon/2$. As a result, $h_{t,\epsilon}(x) \geq 1$, which implies that

$$f(x) = \sum_{c \in C} h_{c,\epsilon} \geq h_{t,\epsilon}(x) \geq 1. \tag{94}$$

Hence,

$$\mathbb{E}_Q[f(x)^2] \geq \mathbb{E}_Q[\mathbb{I}[x \in X]] = Q(X) > 0. \tag{95}$$

Since $cf(x) \in \mathcal{F}$ for every $c > 0$, we prove the desired result. □

G Hermite polynomial and Gaussian kernel

The Hermite polynomial $H_n(x)$ is an degree n polynomial defined as follows,

$$H_n(x) = (-1)^n \exp(x^2) \frac{\mathrm{d}^n \exp(-t^2)}{\mathrm{d}t^n}, \quad \forall n = 1, 2 \cdots. \tag{96}$$

with the following orthogonality property [Poularikas, 2018]:

$$\int \exp(-x^2/2) H_n(x/\sqrt{2}) H_m(x/\sqrt{2}) \mathrm{d}x = \mathbb{I}[n = m]\, 2^n n!\, \sqrt{2\pi}. \tag{97}$$

As a result, we can construct a set of orthonormal basis of square integrable functions $L^2(\mathbb{R})$ using the Hermite polynomial.

Lemma 9 (see e.g., Celeghini et al. [2021]). *The set of functions $\{\psi_n(x)\}_{n \geq 0}$ form an orthonormal basis for $L^2(\mathbb{R})$, where*

$$\psi_n(x) \triangleq H_n\left(\frac{1}{\sqrt{2}} x\right) \exp\left(-\frac{1}{4} x^2\right) (2\pi)^{-1/4} (2^n n!)^{-1/2}.$$

Theorem 10 (Chapter 6.2 of Fasshauer [2011], also see Zhu et al. [1997]). *For any $\epsilon > 0, \alpha > 0$, the eigenfunction expansion for the Gaussian $x, z \in \mathbb{R}$ is*

$$\exp\left(-\epsilon^2 (x - z)^2\right) = \sum_{n=0}^{\infty} \lambda_n \phi_n(x) \phi_n(z), \tag{98}$$

where

$$\lambda_n = \frac{\alpha\epsilon^{2n}}{\left(\frac{\alpha^2}{2}\left(1+\sqrt{1+(2\epsilon/\alpha)^2}\right)+\epsilon^2\right)^{n+1/2}}, \quad n = 0, 1, 2, \cdots,$$

$$\phi_n(x) = \frac{\left(1+(2\epsilon/\alpha)^2\right)^{1/8}}{\sqrt{2^n n!}} \exp\left(-\left(\sqrt{1+(2\epsilon/\alpha)^2}-1\right)\frac{\alpha^2 x^2}{2}\right) H_n\left(\left(1+(2\epsilon/\alpha)^2\right)^{1/4}\alpha x\right).$$

And the eigenfunctions ϕ_n forms an orthonormal basis under weighted $L^2(\mathbb{R})$ space:

$$\int \phi_m(x)\phi_n(x)\frac{\alpha}{\sqrt{\pi}}\exp(-\alpha^2 x^2)\mathrm{d}x = \mathbb{I}\left[m=n\right]. \tag{99}$$

Theorem 11. *For any $\rho \in [-1,1]$, let $P(x_1, x_2)$ be the density of $(x_1, x_2) \sim \mathcal{N}\left(0, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$ and $P(x_1), P(x_2)$ the density of corresponding marginals. Then the Gaussian kernel*

$$K_\rho(x_1, x_2) = \frac{P(x_1, x_2)}{\sqrt{P(x_1)P(x_2)}} \tag{100}$$

$$= \frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left(-\frac{1}{2(1-\rho^2)}(x_1^2+x_2^2-2\rho x_1 x_2)+\frac{x_1^2}{4}+\frac{x_2^2}{4}\right) \tag{101}$$

has the following eigen-decomposition

$$K_\rho(x_1, x_2) = \sum_{k=0}^{\infty}\rho^k\psi_k(x_1)\psi_k(x_2), \tag{102}$$

where $\psi_k(\cdot)$ is defined in Lemma 9.

Proof. We prove this theorem by considering the following two cases separately.

Case 1: $\rho \geq 0$. By algebraic manipulation we get

$$\sqrt{2\pi}K_\rho(x_1, x_2) = \frac{1}{(1-\rho^2)^{1/2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\rho(x_1-x_2)^2+\frac{(\rho-1)^2}{2}x_1^2+\frac{(\rho-1)^2}{2}x_2^2\right)\right).$$

Let $\epsilon^2 = \frac{\rho}{2(1-\rho^2)}$ and $\alpha^2 = \frac{(\rho-1)^2}{2(1-\rho^2)}$, we can equivalent write

$$\sqrt{2\pi}K_\rho(x_1, x_2) = \frac{1}{(1-\rho^2)^{1/2}}\exp\left(-\epsilon^2(x_1-x_2)^2\right)\exp\left(-\frac{\alpha^2}{2}x_1^2\right)\exp\left(-\frac{\alpha^2}{2}x_2^2\right). \tag{103}$$

By Theorem 10 we get

$$\exp\left(-\epsilon^2(x_1-x_2)^2\right) = \sum_n \lambda_n\phi_n(x_1)\phi_n(x_2) \tag{104}$$

with $\int \phi_m(x)\phi_n(x)\frac{\alpha}{\sqrt{\pi}}\exp(-\alpha^2x^2)\mathrm{d}x = \mathbb{I}\left[m=n\right]$. Then we can define the function

$$\psi_n(x) = \left(\frac{\alpha}{\sqrt{\pi}}\right)^{1/2}\phi_n(x)\exp\left(-\frac{\alpha^2}{2}x^2\right) \tag{105}$$

such that $\int \psi_n(x)\psi_m(x)\mathrm{d}x = \mathbb{I}\left[m=n\right]$. Combining Eqs. (103),(104), and (105) we get

$$K_\rho(x_1, x_2) = \sum_n\left(\lambda_n\frac{1}{\sqrt{2}\alpha(1-\rho^2)^{1/2}}\right)\psi_n(x_1)\psi_n(x_2). \tag{106}$$

Now we only need to prove that $\psi_n(\cdot)$ defined in Eq. (105) has the same form as those in Lemma 9, and $\lambda_n\frac{1}{\sqrt{2}\alpha(1-\rho^2)^{1/2}} = \rho^n$.

Recall that

$$\phi_n(x) = \frac{\left(1 + (2\epsilon/\alpha)^2\right)^{1/8}}{\sqrt{2^n n!}} \exp\left(-\left(\sqrt{1 + (2\epsilon/\alpha)^2} - 1\right)\frac{\alpha^2 x^2}{2}\right) H_n\left(\left(1 + (2\epsilon/\alpha)^2\right)^{1/4}\alpha x\right).$$

Plugin $\epsilon^2 = \frac{\rho}{2(1-\rho)^2}$ and $\alpha^2 = \frac{(\rho-1)^2}{2(1-\rho)^2}$ we get

$$\left(1 + (2\epsilon/\alpha)^2\right)^{1/8}\left(\frac{\alpha}{\sqrt{\pi}}\right)^{1/2} = \frac{1}{\pi^{1/4}}\left(\alpha^4 + 4\epsilon^2\alpha^2\right)^{1/8} \tag{107}$$

$$= \frac{1}{\pi^{1/4}}\left(\frac{(\rho-1)^4 + 4\rho(\rho-1)^2}{4(1-\rho^2)^2}\right)^{1/8} = \frac{1}{(2\pi)^{1/4}} \tag{108}$$

$$\exp\left(-\frac{\alpha^2}{2}x^2\right)\exp\left(-\left(\sqrt{1 + (2\epsilon/\alpha)^2} - 1\right)\frac{\alpha^2 x^2}{2}\right) \tag{109}$$

$$= \exp\left(-\sqrt{\alpha^4 + 2\epsilon^2\alpha^2}\frac{x^2}{2}\right) = \exp\left(-\sqrt{\frac{(\rho-1)^4 + 4\rho(\rho-1)^2}{4(1-\rho^2)^2}}\frac{x^2}{2}\right) \tag{110}$$

$$= \exp\left(-\frac{x^2}{4}\right), \tag{111}$$

$$\left(1 + (2\epsilon/\alpha)^2\right)^{1/4}\alpha = \left(1 + \frac{4\rho}{(\rho-1)^2}\right)^{1/4}\left(\frac{(\rho-1)^2}{2(1-\rho^2)}\right)^{1/2} \tag{112}$$

$$= \left(\frac{(1+\rho)^2}{(\rho-1)^2}\frac{(\rho-1)^4}{(1-\rho^2)^2}\right)^{1/4}\frac{1}{\sqrt{2}} = \frac{1}{\sqrt{2}}. \tag{113}$$

As a result,

$$\psi_n(x) = \left(\frac{\alpha}{\sqrt{\pi}}\right)^{1/2}\phi_n(x)\exp\left(-\frac{\alpha^2}{2}x^2\right) \tag{114}$$

$$= H_n\left(\frac{1}{\sqrt{2}}x\right)\exp\left(-\frac{1}{4}x^2\right)(2\pi)^{-1/4}(2^n n!)^{-1/2}. \tag{115}$$

Now we turn to the eigenvalues. Recall that

$$\lambda_n = \frac{\alpha\epsilon^{2n}}{\left(\frac{\alpha^2}{2}\left(1 + \sqrt{1 + (2\epsilon/\alpha)^2}\right) + \epsilon^2\right)^{n+1/2}}.$$

Plugin $\epsilon^2 = \frac{\rho}{2(1-\rho)^2}$ and $\alpha^2 = \frac{(\rho-1)^2}{2(1-\rho)^2}$ we get

$$\frac{\alpha^2}{2}\left(1 + \sqrt{1 + (2\epsilon/\alpha)^2}\right) + \epsilon^2 = \frac{1}{2}\left(\alpha^2 + \sqrt{\alpha^4 + 4\epsilon^2\alpha^2}\right) + \epsilon^2 \tag{116}$$

$$= \frac{1}{2}\left(\frac{(\rho-1)^2}{2(1-\rho^2)} + \frac{1}{2}\right) + \frac{\rho}{2(1-\rho^2)} = \frac{1}{2(1-\rho^2)}. \tag{117}$$

Consequently,

$$\lambda_n \frac{1}{\sqrt{2}\alpha(1-\rho^2)^{1/2}} = \frac{\left(\frac{\rho}{2(1-\rho^2)}\right)^n}{\left(\frac{1}{2(1-\rho^2)}\right)^{n+1/2}}\frac{1}{\sqrt{2}(1-\rho^2)^{1/2}} = \rho^n. \tag{118}$$

Case 2: $\rho < 0$. Recall that

$$K_\rho(x_1, x_2) = \frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left(-\frac{1}{2(1-\rho^2)}(x_1^2 + x_2^2 - 2\rho x_1 x_2) + \frac{x_1^2}{4} + \frac{x_2^2}{4}\right).$$

In this case we reuse the results from Case 1 to get

$$K_{-\rho}(x_1, x_2) = \sum_{k=0}^{\infty}(-\rho)^k\psi_k(x_1)\psi_k(x_2). \tag{119}$$

By definition, we get $K_\rho(x_1, x_2) = K_{-\rho}(x_1, -x_2)$. As a result,

$$K_\rho(x_1, x_2) = \sum_{k=0}^{\infty} (-\rho)^k \psi_k(x_1)\psi_k(-x_2) \tag{120}$$

$$= \sum_{k=0}^{\infty} \rho^k \psi_k(x_1)\psi_k(x_2), \tag{121}$$

where the last equation follows from the fact that $\psi_k(x_2) = (-1)^k \psi_k(-x_2)$. $\square$

Theorem 12. *For any $\Sigma \in \mathbb{R}^{d_1 \times d_2}$, let $P(x_1, x_2)$ be the density of $(x_1, x_2) \sim \mathcal{N}\left(0, \begin{pmatrix} I_{d_1} & \Sigma \\ \Sigma^\top & I_{d_2} \end{pmatrix}\right)$ and $P(x_1), P(x_2)$ the density of corresponding marginals, where $x_1 \in \mathbb{R}^{d_1}, x_2 \in \mathbb{R}^{d_2}$. Without loss of generality, assume $d_1 \geq d_2$. Let $\Sigma = U\Lambda V^\top$ be the singular value decomposition of Σ, where $U \in \mathbb{R}^{d_1 \times d_2}, \Lambda \in \mathbb{R}^{d_1 \times d_2}, V \in \mathbb{R}^{d_2 \times d_2}$. Let $\sigma_1, \cdots, \sigma_{d_2}$ be the singular values of Σ (i.e., $[\Lambda]_{i,i} = \sigma_i$). Then the Gaussian kernel*

$$K_\Sigma(x_1, x_2) = \frac{P(x_1, x_2)}{\sqrt{P(x_1)P(x_2)}} \tag{122}$$

has the following eigen-decomposition

$$K_\Sigma(x_1, x_2) = \prod_{i=1}^{d_2} \left(\sum_{n=0}^{\infty} \sigma_i^n \psi_n([t_1]_i)\psi_n([t_2]_i) \right) \prod_{i=d_2+1}^{d_1} \psi_0([t_1]_i). \tag{123}$$

where $t_1 = U^\top x_1, t_2 = V^\top x_2$, and $\psi_k(\cdot)$ is defined in Lemma 9.

Proof. We prove this theorem by decomposing the multi-dimensional kernel into products of several one-dimensional kernels described in Theorem 11.

Recall that $t_1 = U^\top x_1 \in \mathbb{R}^{d_1}, t_2 = V^\top x_2 \in \mathbb{R}^{d_2}$, where $\Sigma = U\Lambda V^\top$ is the singular value decomposition of Σ. Let $t \in \mathbb{R}^{d_1+d_2}$ be the concatenation of t_1, t_2. Then we have

$$\mathbb{E}_P[tt^\top] = \begin{pmatrix} I_{d_1} & \Lambda \\ \Lambda^\top & I_{d_2} \end{pmatrix}. \tag{124}$$

Let $P'(t_1, t_2)$ be the density of $(t_1, t_2) \sim \mathcal{N}\left(0, \begin{pmatrix} I_{d_1} & \Lambda \\ \Lambda^\top & I_{d_2} \end{pmatrix}\right)$. By the fact that Λ is a diagonal matrix, we know that variables $\{[t_1]_i, [t_2]_i\}$ is independent from $\{[t_1]_j\}_{j\neq i} \cup \{[t_2]_j\}_{j\neq i}$ for every $1 \leq i \leq d_2$, and variable $[t_1]_i$ is independent from $\{[t_1]_j\}_{j\neq i} \cup \{[t_2]_j\}_{1\leq j\leq d_2}$ for every $d_2 + 1 \leq i \leq d_1$. In addition, $([t_1]_i, [t_2]_i) \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \sigma_i \\ \sigma_i & 1 \end{pmatrix}\right)$. As a result,

$$K_\Sigma(x_1, x_2) = \frac{P(x_1, x_2)}{\sqrt{P(x_1)P(x_2)}} = \frac{P'(t_1, t_2)}{\sqrt{P'(t_1)P'(t_2)}} \tag{125}$$

$$= \prod_{i=1}^{d_2} \left(\frac{P'([t_1]_i, [t_2]_i)}{\sqrt{P'([t_1]_i)P'([t_2]_i)}} \right) \prod_{i=d_2+1}^{d_1} \sqrt{P'([t_1]_i)} \tag{126}$$

$$= \prod_{i=1}^{d_2} \left(\sum_{n=0}^{\infty} \sigma_i^n \psi_n([t_1]_i)\psi_n([t_2]_i) \right) \prod_{i=d_2+1}^{d_1} \sqrt{P'([t_1]_i)} \tag{127}$$

$$= \prod_{i=1}^{d_2} \left(\sum_{n=0}^{\infty} \sigma_i^n \psi_n([t_1]_i)\psi_n([t_2]_i) \right) \prod_{i=d_2+1}^{d_1} \psi_0([t_1]_i), \tag{128}$$

where Eq. (127) follows from Theorem 12, and Eq. (128) follows from the definition of ψ_0. $\square$

H Helper Lemmas

Lemma 13. *Let $\Sigma \in \mathbb{R}^{d \times d}$ be a positive semi-definite matrix with $\operatorname{diag}(\Sigma) = I$. Then we have*

$$\forall k \geq 1, \quad \lambda_{\min}(\Sigma^{\odot k}) \geq \lambda_{\min}(\Sigma), \tag{129}$$

where $\Sigma^{\odot k}$ denotes the element-wise k-th power of the matrix Σ.

Proof. Let $\lambda = \lambda_{\min(\Sigma)}$ and we have $\Sigma - \lambda I \succeq 0$. As a result, $(\Sigma - \lambda I)^{\odot k} = \Sigma^{\odot k} + (1-\lambda)^k I - I \succeq 0$. Note that $0 \leq \lambda \leq 1$ because $\operatorname{Tr}(\Sigma - I) = 0$ implies $\lambda_{\min}(\Sigma - I) \leq 0$. It follows that for any $k \geq 1$,

$$\Sigma^{\odot k} \succeq I - (1-\lambda)^k I \succeq \lambda I, \tag{130}$$

where the last inequality follows from the fact that $1 - \lambda \geq (1-\lambda)^k$ when $k \geq 1$. Consequently,

$$\lambda_{\min}(\Sigma^{\odot k}) \geq \lambda_{\min}(\Sigma). \tag{131}$$

□

Lemma 14. *Let $\Sigma \in \mathbb{R}^{d_1 \times d_2}$ be a matrix such that*

$$\begin{pmatrix} I_{d_1} & \Sigma \\ \Sigma^\top & I_{d_2} \end{pmatrix} \succeq 0. \tag{132}$$

Then we have $\lambda_{\max}(\Sigma^\top \Sigma) \leq 1$.

Proof. We prove by contradiction. Suppose otherwise $\lambda_{\max}(\Sigma^\top \Sigma) > 1$. Let $v \in \mathbb{R}^{d_2}$ be the eigenvector of $\Sigma^\top \Sigma$ corresponds to its maximum eigenvalue. Then

$$\begin{pmatrix} -\Sigma v \\ v \end{pmatrix}^\top \begin{pmatrix} I_{d_1} & \Sigma \\ \Sigma^\top & I_{d_2} \end{pmatrix} \begin{pmatrix} -\Sigma v \\ v \end{pmatrix} = \begin{pmatrix} -\Sigma v \\ v \end{pmatrix}^\top \begin{pmatrix} 0 \\ -\Sigma^\top \Sigma v + v \end{pmatrix}^\top = -v^\top \Sigma^\top \Sigma v + v^\top v. \tag{133}$$

By the assumption that $\lambda_{\max}(\Sigma^\top \Sigma) > 1$, we get

$$-v^\top \Sigma^\top \Sigma v + v^\top v = \|v\|_2^2 (1 - \lambda_{\max}(\Sigma^\top \Sigma)) < 0, \tag{134}$$

which contradicts to Eq. (132). Therefore, we must have $\lambda_{\max}(\Sigma^\top \Sigma) \leq 1$. □

Lemma 15. *Let M be a positive semi-definite matrix with the following form*

$$M = \begin{pmatrix} I_{d_1} & \Sigma \\ \Sigma^\top & I_{d_2} \end{pmatrix} \tag{135}$$

for some $d_1, d_2 > 0$. Then we have

$$\lambda_{\min}(M) = 1 - \sigma_{\max}(\Sigma), \tag{136}$$

where $\sigma_{\max}$ is the largest singular value of Σ.

Proof. Without loss of generality, we assume $d_1 \geq d_2$. Let $\Sigma = U \Lambda V^\top$ be the singular value decomposition of Σ, with $U \in \mathbb{R}^{d_1 \times d_1}, V \in \mathbb{R}^{d_2 \times d_2}, \Lambda \in \mathbb{R}^{d_1 \times d_2}$. Then we have

$$M = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} I_{d_1} & \Lambda \\ \Lambda^\top & I_{d_2} \end{pmatrix} \begin{pmatrix} U^\top & 0 \\ 0 & V^\top \end{pmatrix}. \tag{137}$$

Let $\bar{M} = \begin{pmatrix} I_{d_1} & \Lambda \\ \Lambda^\top & I_{d_2} \end{pmatrix}$. Because U, V are orthonormal matrices, we get $\lambda_{\min}(M) = \lambda_{\min}(\bar{M})$. Note that $[\Lambda]_{i,j} = 0$ whenever $i \neq j$ and $[\Lambda]_{i,i} = \sigma_i(\Sigma)$. Consequently, the eigenvalues of $\bar{M}$ is

$$1 \pm \sigma_1(\Sigma), 1 \pm \sigma_2(\Sigma), \cdots, 1 \pm \sigma_{d_2}(\Sigma) \tag{138}$$

with multiplicity 1, and 1 with multiplicity $d_1 - d_2$. As a result,

$$\lambda_{\min}(M) = \lambda_{\min}(\bar{M}) 1 - \sigma_{\max}(\Sigma).$$

□