Rethinking Memorization–Generalization Trade-Off in Generative Models

Jiseok Chae* JSCH@KAIST.AC.KR
KAIST, Daejeon, Republic of Korea

Kyuwon Kim* KKW4053@KAIST.AC.KR
KAIST, Daejeon, Republic of Korea

Donghwan Kim DONGHWANKIM@KAIST.AC.KR
KAIST, Daejeon, Republic of Korea

Abstract

Existing generative models exhibit a memorization–generalization trade-off, and thus, avoiding memorization is a common strategy to promote generalization. In supervised learning, this long-accepted trade-off is being challenged, as recent studies show modern overparametrized models can achieve benign overfitting; that is, they generalize well even while exactly fitting, or memorizing, the training data. This raises the question of whether overparametrized generative models can similarly bypass this trade-off and achieve superior generalization alongside memorization. We address this with an empirical risk formulation that uses presampled latent variables instead of integrating over the entire latent distribution. We then recast the generative modeling problem as a supervised learning task of learning an optimal transport map, enabling us to leverage the concept of benign overfitting. In the one-dimensional setting, we show for the first time that benign overfitting can occur in generative models. We further expand and empirically validate our approach to higher dimensions, illustrating that benign overfitting extends more broadly across generative models.

1. Introduction

Generative models, such as generative adversarial networks [12] and diffusion models [15, 36], have become increasingly influential across various domains, from image synthesis [17, 31, 38, 41] to novel material generation [3, 25, 44]. Accordingly, interest in understanding their *generalizability*, the ability to generate diverse realistic data unseen during training, has grown significantly in recent years. A conventional wisdom in generative models is that there exists a trade-off between memorization (*i.e.*, overfitting) and generalization [27, 42, 45]. Hence, techniques such as weight decay, early stopping, and underparametrization, are typically incorporated [4, 14] to suppress memorization and thereby improve generalization.

Comparably, in the context of supervised learning, the *memorization–generalization trade-off*, based on the bias–variance trade-off [13], was generally accepted as true, until just a few years ago, and techniques to prevent memorization were commonly employed [19, 40]. However, recent studies have shown that overparametrized supervised learning models trained to achieve zero empirical risk, memorizing the training data, can exhibit superior generalization performance [6, 28]. This empirically observed and theoretically supported phenomenon is now known as *benign overfitting* [5, 24].

* Equal contribution

In contrast, to the best of our knowledge, there is currently neither theoretical nor even experimental evidence showing that generative models can achieve superior generalization beyond the conventional trade-off. This motivates the following central question of our study:

Can generative models exhibit superior generalization via benign overfitting, similar to supervised learning?

To explore this question, we first note that the empirical risk in training generative models is often set as the distance between the generated distribution and the *empirical* target data distribution [20, 46]. However, this forces all generated outputs to match training data, resulting in poor generalization [22, 30]. Therefore, for the model to generate outputs not in the training data, we avoid utilizing the entire latent distribution during training. Instead, we consider minimizing an empirical risk that only uses a fixed subset of latent vectors presampled from the latent distribution.

While this alternative choice opens the possibility to challenge the conventional trade-off, it remains unclear whether it can lead to superior generalization performance. As a first step toward this direction, we recast generative modeling as a regression task enabling us to leverage the theory of benign overfitting in regression. This leads to the following contributions.

1) In Section 3, we recast generative modeling as a regression task of learning the optimal transport map from the latent distribution to the true data distribution, using presampled latent variables and training data.

2) In Section 4, we show that generative models can exhibit benign overfitting, based on the random feature model framework [35]. We chose this framework, among several relevant works [8, 16, 34], as it presents the "more is better" principle—increasing the model size improves performance—in the context of regression problems. This suggests that a similar phenomenon may happen in generative modeling, despite the common belief that increasing model size without proper regularization impairs performance.

2. Preliminaries

2.1. Regression: population and empirical risks

Consider a standard regression setting with a training dataset $\mathcal{D} = \{(\boldsymbol{z}_i, x_i)\}_{i=1}^n$ where $\{\boldsymbol{z}_i\}_{i=1}^n$ are i.i.d. samples from a probability distribution ξ over $\mathbb{R}^d$, and

$$x_i = f^*(\boldsymbol{z}_i) + \epsilon_i \tag{1}$$

for a target function $f^* : \mathbb{R}^d \to \mathbb{R}$ and ϵ_i denoting the additive noise component. We further consider regression with feature maps, where $f(\cdot) = \boldsymbol{\psi}(\cdot)^\top \boldsymbol{\beta}$ with a feature map $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^p$ and model parameters $\boldsymbol{\beta} \in \mathbb{R}^p$. Then the regression problem reduces to estimating $\boldsymbol{\beta}^*$ such that $f^*(\cdot) = \boldsymbol{\psi}(\cdot)^\top \boldsymbol{\beta}^*$. Moreover, we set $\boldsymbol{\psi}$ to be a *random* feature map, which is of the form $\boldsymbol{\psi}(z) = \big(g(\boldsymbol{w}_1, z), \ldots, g(\boldsymbol{w}_p, z)\big)$ for some function g, and $\boldsymbol{w}_i$ are i.i.d. samples from some distribution ρ. For later purposes, let us further assume that g admits a singular value decomposition $g(\boldsymbol{w}, z) = \sum_i \sqrt{\lambda_i}\zeta_i(\boldsymbol{w})\phi_i(z)$, where $\{\zeta_i\}_i \subset L^2(\rho)$ and $\{\phi_i\}_i \subset L^2(\xi)$ are each orthonormal sequences in their respective spaces.

Ideally, we would like to minimize the *population risk*

$$\mathcal{R}(\boldsymbol{\beta}) := \mathbb{E}_{\boldsymbol{z}\sim\xi}[\|\boldsymbol{\psi}(\boldsymbol{z})^\top \boldsymbol{\beta} - \boldsymbol{\psi}(\boldsymbol{z})^\top \boldsymbol{\beta}^*\|^2] = \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}^2, \tag{2}$$

where $\boldsymbol{\Sigma} \coloneqq \mathbb{E}_{\boldsymbol{z}\sim\xi}[\boldsymbol{\psi}(\boldsymbol{z})\boldsymbol{\psi}(\boldsymbol{z})^\top]$, but typically ν, and thus $\boldsymbol{\beta}^*$, are inaccessible. Hence, the *empirical risk*

$$\hat{\mathcal{R}}(\boldsymbol{\beta}) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\psi}(\boldsymbol{z}_i)^\top\boldsymbol{\beta} - x_i\right)^2 \tag{3}$$

is used as a proxy to be minimized. The generalization performance of the learned model $\hat{f} = \boldsymbol{\psi}^\top\hat{\boldsymbol{\beta}}$, obtained by minimizing the empirical risk, is still formulated by the population risk $\mathcal{R}(\hat{\boldsymbol{\beta}})$.

Classical statistical learning theory suggests a memorization–generalization trade-off, based on bias–variance trade-off [13]. That is, models with high bias tend to underfit the data, failing to capture underlying patterns, whereas models with high variance tend to overfit, undesirably capturing noise in the training data. However, recent studies have shown the opposite, that overparametrized models, despite being trained to perfectly memorize the training data, can still exhibit strong generalization performance [5, 6, 28, 35, 39].

2.2. Generative modeling: population and empirical risks

Now consider when the dataset $\{\boldsymbol{x}_i\}_{i=1}^n$ consists of n i.i.d. samples from a distribution ν on $\mathbb{R}^m$. Generative models aim to learn a process which approximates the target distribution ν by transforming the latent distribution ξ. We focus on the models whose generative process is represented by a deterministic map G. In this case, the goal is to approximate ν by $G_\sharp\xi$, the pushforward of ξ by G.

The population risk of a generative model is defined by a divergence between the generated distribution and target data distribution. In this paper, we set the squared 2-Wasserstein distance $\mathcal{R}(G) \coloneqq \mathcal{W}_2^2(G_\sharp\xi, \nu)$ as the population risk.

In practice, ν is not directly accessible, but only by its empirical distribution $\hat{\nu} \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i}$. Thus,the *semi-discrete empirical risk* $\hat{\mathcal{R}}_{\text{semi}}(G) \coloneqq \mathcal{W}_2^2(G_\sharp\xi, \hat{\nu})$ is often minimized instead as a proxy. A few works, *e.g.*, [2, 37], have also explored an alternative of minimizing

$$\hat{\mathcal{R}}_{\text{fully}}(G) \coloneqq \mathcal{W}_2^2(G_\sharp\hat{\xi}, \hat{\nu}) = \inf_{\pi\in S_n}\frac{1}{n}\sum_{i=1}^{n}\|G(\boldsymbol{z}_i) - \boldsymbol{x}_{\pi(i)}\|^2, \tag{4}$$

which we call the *fully-discrete empirical risk*, where $\hat{\xi} = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{z}_i}$ denotes the empirical latent distribution on the i.i.d. samples $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \sim \xi$, and S_n is the set of all permutations of $\{1, \ldots, n\}$.

3. Recasting Generative Modeling as Regression via Optimal Transport

Generative modeling and supervised learning are typically viewed as two distinct paradigms. Here, we bridge the two by recasting generative modeling as a regression problem via optimal transport.

3.1. The first step toward bridging generative modeling and regression

Since generative modeling in general admits infinitely many mappings minimizing $\mathcal{R}(G)$, linking it with regression requires choosing a particular generator G^*. In this paper, we choose G^* to be the optimal transport map. This approach not only offers a well-defined structure but also enables us to exploit the rich, existing results in optimal transport theory. To this end, the rest of this paper considers the equal-dimensional case $d = m$, leaving extensions to unequal dimensions for future work.

Given a cost function c, an *optimal transport (OT) map* G^* from ξ to ν is an optimal solution to

$$\inf_{G:\, G_\sharp \xi = \nu} \int c(\boldsymbol{z}, G(\boldsymbol{z}))\, d\xi(\boldsymbol{z}). \tag{5}$$

If ξ and ν are both discrete measures, then there exists an OT map G^* where for some permutation $\tilde{\pi} \in S_n$ it holds that $G^*(z_i) = x_{\tilde{\pi}(i)}$ for all $i = 1, \ldots, n$, regardless of the cost function [29, Proposition 2.1]. So, for these cases, without loss of generality we always assume that the OT map G^* is associated with a permutation $\tilde{\pi}$ as such. Meanwhile, for the quadratic cost $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2$, if ξ is absolutely continuous with respect to the Lebesgue measure, then Brenier's theorem [7] ensures that there uniquely exists an optimal transport map G^*.

Learning the OT map in the context of generative modeling is not new; see, *e.g.*, [18, 23, 32]. However, there were no discussions on the generalization behavior when learning from finite training data. In contrast, this paper systemically analyzes how an OT map can be learned from finite training data $\{\boldsymbol{x}_i\}_{i=1}^n$ and presampled latent variables $\{\boldsymbol{z}_i\}_{i=1}^n$, by casting the problem as a regression task.

3.2. Optimal transport map based regression model for generative modeling

We now formulate a regression model, analogous to (1), for generative modeling based on the OT map. This requires pairing the finite training data with presampled latent variables. We naturally choose the OT map $\tilde{\pi} \in S_n$ between them, a permutation minimizing $\min_{\pi \in S_n} \sum_{i=1}^n \|\boldsymbol{z}_i - \boldsymbol{x}_{\pi(i)}\|^2$. Having paired the dataset as $\{(\boldsymbol{z}_i, \boldsymbol{x}_{\tilde{\pi}(i)})\}_{i=1}^n$, we obtain the regression model

$$\boldsymbol{x}_{\tilde{\pi}(i)} = G^*(\boldsymbol{z}_i) + \boldsymbol{\epsilon}_i, \tag{6}$$

where $\boldsymbol{\epsilon}_i \in \mathbb{R}^m$ denotes noise resulting from recasting generative modeling as regression. That said, these noises are not i.i.d., and this complicates using the existing benign overfitting theories directly.

3.2.1. OPTIMAL TRANSPORT MAP BASED REGRESSION MODELS: ONE-DIMENSIONAL SETTING

By particularly focusing on the one-dimensional setting, we can get a more refined characterization of the noises. Let us denote by $x_{i:n}$ the ith order statistic, *i.e.*, the ith smallest one among $\{x_i\}_{i=1}^n$. For empirical distributions $\hat{\xi}$ and $\hat{\nu}$ on $\mathbb{R}$, the OT map pairs $z_{i:n}$ to $x_{i:n}$, for each $i = 1, \ldots, n$ [29, Remark 2.28]. Hence, the regression model (6) becomes[1]

$$x_{i:n} = G^*(z_{i:n}) + \epsilon_i. \tag{7}$$

However, for arbitrary target distribution ν, complete characterization of the summary statistics of the order statistics remains unavailable. Fortunately, for our subsequent benign overfitting analyses, we only need the variances of ϵ_is to be bounded. This turns out to be achievable, under a mild assumption on the target distribution ν.

Assumption 1 *The target distribution ν has finite variance; i.e., for $x \sim \nu$, we have $\mathrm{Var}(x) \le \sigma^2$.*

Lemma 1 *Suppose that Assumption 1 holds, and ξ is absolutely continuous with respect to the Lebesgue measure. Then for any $n \ge 1$, the noise in (7) satisfies $\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(\epsilon_i) \le 2\sigma^2$.*

1. For simplicity, we continue to use ϵ_i, where the index i is assumed to have been reordered.

4. Benign Overfitting in Generative Models

In this section, we demonstrate benign overfitting of generative models, built upon the results of [35].

4.1. Theoretical results in the one-dimensional setting

Let us first study the simpler case, of when $d = m = 1$. To facilitate a plausible yet theoretical analysis, we adopt the following ansatz, introduced therein.

Assumption 2 (35, *Gaussian Universality Ansatz*) *The expected population risk remains unchanged even if we replace $\{\tilde{\phi}_i(z)\}$ with i.i.d. samples from $\mathcal{N}(0,1)$ when $z \sim \xi$, and $\{\tilde{\zeta}_i(\boldsymbol{w})\}$ with i.i.d. samples from $\mathcal{N}(0,1)$ when $\boldsymbol{w} \sim \rho$.*

As in kernel regression, we approximate G^* by minimizing the least squares objective,

$$\min_{G(\,\cdot\,)=\boldsymbol{\psi}(\,\cdot\,)^\top\boldsymbol{\beta}} \ \frac{1}{n}\sum_{i=1}^{n} \|G(\boldsymbol{z}_i) - \boldsymbol{x}_{\tilde{\pi}(i)}\|^2. \tag{8}$$

We focus on its ridge regression version, in which we solve (8) with an additional term $\frac{\delta}{n}\|\boldsymbol{\beta}\|^2$ with $\delta > 0$ in the objective. The optimal solution is then $\hat{\boldsymbol{\beta}}(\delta) = \boldsymbol{\Psi}^\top(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \delta \boldsymbol{I})^{-1}\boldsymbol{x}$, where $\boldsymbol{\Psi}$ is a matrix whose ith row is $\boldsymbol{\psi}(\boldsymbol{z}_i)^\top$. Accordingly, the learned model to be studied is $\hat{G}(\,\cdot\,) = \boldsymbol{\psi}(\,\cdot\,)^\top\hat{\boldsymbol{\beta}}(\delta)$.

Let $G^*(z) = \sum_i v_i\phi_i(z)$ be the expansion of the target function with respect to $\{\phi_i\}_i$. In the random feature eigenframework [35], which deals with classical ridge regression problems under the traditional assumption that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Gaussians, it is proposed that the expected population risk can be well approximated by

$$\mathbb{E}_{\mathcal{D}}\big[\mathcal{R}(\hat{\boldsymbol{\beta}}(\delta))\big] \approx \mathcal{E}_{\text{te}} \coloneqq \frac{1}{1 - \frac{q(p-2s)+s^2}{n(p-q)}}\left(\sum_i \left(\frac{\gamma}{\lambda_i+\gamma} - \frac{\kappa\lambda_i}{(\lambda_i+\gamma)^2}\frac{p}{p-q}\right)v_i^2 + \check{\sigma}^2\right), \tag{9}$$

where, for $s \coloneqq \sum_i \frac{\lambda_i}{\lambda_i+\gamma}$ and $q \coloneqq \sum_i\big(\frac{\lambda_i}{\lambda_i+\gamma}\big)^2$, κ and γ are unique nonnegative numbers such that $n = s + \frac{\delta}{\kappa}$ and $p = s + \frac{p\kappa}{\gamma}$. It is stated in Appendix I.3 of [34], the work from which [35] is developed, that $\check{\sigma}^2$ is the term corresponding to what is "effectively noise". In their setting, that quantity is the mean squared error of $x_i - G^*(z_i)$. We argue that this reasoning also applies to our setting, and $\check{\sigma}^2$ should be set as the expected residual variance $\check{\sigma}^2 = \frac{1}{n}\,\mathbb{E}\big[\sum_{i=1}^{n}(x_{i:n} - G^*(z_{i:n}))^2\big]$.

Founded on these specifications, as in [35], we can show that increasing the number of features reduces the risk, so long as we are free to choose the ridge parameter.

Theorem 2 *Let $\mathcal{E}_{\text{te}}(n,p,\delta)$ denote the value of $\mathcal{E}_{\text{te}}$ with dataset size n, number of features p, and ridge parameter δ. Suppose that $p \le p'$. Then, under Assumptions 1 and 2, it holds that*

$$\min_{\delta} \mathcal{E}_{\text{te}}(n,p',\delta) \le \min_{\delta} \mathcal{E}_{\text{te}}(n,p,\delta).$$

Moreover, if we further assume that $p > n$, then denoting $\mathcal{E}_{\text{te},\,0} = \lim_{\delta\to 0+}\mathcal{E}_{\text{te}}$, it holds that

$$\mathcal{E}_{\text{te},\,0}(n,p') \le \mathcal{E}_{\text{te},\,0}(n,p).$$

4.2. Extending the theory to higher dimensions

The theoretical results in Section 4.1 can be extended to higher dimensions, under assumptions appropriately modified. Let $\{\boldsymbol{z}_i\}_{i=1}^n \subset \mathbb{R}^m$ be the latent vectors, and $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^m$ be the given dataset. Let $\tilde{\pi}$ be the optimal pairing between $\{\boldsymbol{z}_i\}_{i=1}^n$ and $\{\boldsymbol{x}_i\}_{i=1}^n$, as described in (6). We aim to estimate $\boldsymbol{\beta}$ under the linear model generalizing (1); for $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^{p\times m}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and $\boldsymbol{\epsilon}_i \in \mathbb{R}^m$,

$$\boldsymbol{x}_{\tilde{\pi}(i)} = \boldsymbol{\psi}(\boldsymbol{z}_i)^\top \boldsymbol{\beta} + \boldsymbol{\epsilon}_i. \tag{10}$$

In particular, the change in $\boldsymbol{\psi}$ amounts to g now being an $\mathbb{R}^m$-valued function. Then, we can show that Theorem 2 holds almost verbatim; the exact statements are in Theorems 14 and 15. For further details, with an overview on kernel regression with vector-valued targets, see Appendices C and D.

4.3. Experiments

We present experimental results in Appendix E to validate our analyses of the random feature model, particularly the approximation of the population risk by $\mathcal{E}_{\text{te}}$, thereby empirically supporting the principle that using more features is better.

5. Conclusion

We showed that overparametrized generative models can generalize despite memorizing training data, contrary to conventional belief that memorization undermines generalization performance. Our approach demonstrated this through learning the optimal transport map. By reformulating generative modeling into a regression task, quantitative generalization analyses became possible, leading to theoretical demonstrations that benign overfitting can occur in generative models. Specifically, we showed that increasing the number of features improves performance. Our work hence frames a new approach for studying the interplay between overparameterization and generalization in generative models.

Our work is yet limited to the latent and target distributions defined on the same space. Future works could explore alternative mappings beyond the optimal transport map, such as those based on the Gromov–Wasserstein theory or optimal transport maps composed with embeddings, which would enable analyses when the distributions lie on different underlying spaces.

References

[1] Barry C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in Order Statistics*. Society for Industrial and Applied Mathematics, 2008.

[2] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, 2017.

[3] Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, and Jian Zhou. Dirichlet diffusion score model for biological sequence generation. In *International Conference on Machine Learning*, pages 1276–1301. PMLR, 2023.

[4] Ricardo Baptista, Agnimitra Dasgupta, Nikola B. Kovachki, Assad Oberai, and Andrew M. Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785*, 2025.

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

[6] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[7] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *Comptes Rendus de l'Académie des Sciences - Série I - Mathématique*, 305:805–808, 1987.

[8] Abdulkadir Canatar, Brendan Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12:2914, 2021.

[9] Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanità. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(1):19–61, 2010.

[10] Daniel K. Crane. *The singular value expansion for compact and non-compact operators*. PhD thesis, Michigan Technological University, 2020.

[11] J. D. Esary, F. Proschan, and D. W. Walkup. Association of random variables, with applications. *The Annals of Mathematical Statistics*, 38(5):1466–1474, 1967.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014.

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[14] Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. In *International Conference on Learning Representations*, 2019.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[16] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clement Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. In *Advances in Neural Information Processing Systems*, 2020.

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[18] Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[20] Dohyun Kwon, Ying Fan, and Kangwook Lee. Score-based generative modeling secretly minimizes the Wasserstein distance. *Advances in Neural Information Processing Systems*, 35: 20205–20217, 2022.

[21] Peter D. Lax. *Functional Analysis*. John Wiley and Sons, Inc., 2002.

[22] Lorenzo Luzi, Yehuda Dar, and Richard Baraniuk. Double descent and other interpolation phenomena in GANs. *arXiv preprint arXiv:2106.04003*, 2021.

[23] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[24] Neil Mallinar, James Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting. *Advances in Neural Information Processing Systems*, 35:1182–1195, 2022.

[25] Yunwei Mao, Qi He, and Xuanhe Zhao. Designing complex architectured materials with generative adversarial networks. *Science Advances*, 6(17), 2020.

[26] Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

[27] Vaishnavh Nagarajan, Colin Raffel, and Ian J. Goodfellow. Theoretical insights into memorization in GANs. In *Integration of Deep Learning Theories Workshop, NeurIPS*, 2018.

[28] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

[29] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Now Publishers, 2019.

[30] Jakiw Pidstrigach. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[32] Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022.

[33] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Birkhäuser, 2015.

[34] James B. Simon, Madeline Dickens, Dhruva Karkada, and Michael R. DeWeese. The eigenlearning framework: A conservation law perspective on kernel ridge regression and wide neural networks. *Transactions on Machine Learning Research*, 2023.

[35] James B. Simon, Dhruva Karkada, Nikhil Ghosh, and Mikhail Belkin. More is better in modern machine learning: when infinite overparameterization is optimal and overfitting is obligatory. In *International Conference on Learning Representations*, 2024.

[36] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015.

[37] Ruoyu Sun, Tiantian Fang, and Alexander Schwing. Towards a better global loss landscape of GANs. *Advances in Neural Information Processing Systems*, 33:10186–10198, 2020.

[38] Yuhta Takida, Masaaki Imaizumi, Takashi Shibuya, Chieh-Hsin Lai, Toshimitsu Uesaka, Naoki Murata, and Yuki Mitsufuji. SAN: Inducing metrizability of GAN with discriminative normalized linear layer. In *International Conference on Learning Representations*, 2024.

[39] Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

[40] Vladimir Vapnik. *Statistical learning theory*. John Wiley & Sons, 1998.

[41] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-GAN: Training GANs with diffusion. In *International Conference on Learning Representations*, 2023.

[42] TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K. Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 workshop on structured probabilistic inference & generative modeling*, 2023.

[43] Kôsaku Yosida. *Functional analysis*. Springer, 6th edition, 1995.

[44] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025.

[45] Huijie Zhang, Jinfan Zhou, Yifu Lu, Minzhe Guo, Peng Wang, Liyue Shen, and Qing Qu. The emergence of reproducibility and consistency in diffusion models. In *International Conference on Machine Learning*, 2024.

[46] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in GANs. In *International Conference on Learning Representations*, 2018.

Contents

Appendix A. Missing Details for Section 3

For a distribution ν on $\mathbb{R}$, let $F_\nu(x) \coloneqq \nu((-\infty, x])$ be its *cumulative distribution function* (CDF), and $F_\nu^{[-1]}(x) \coloneqq \inf\{t \in \mathbb{R} : F_\nu(t) \geq x\}$ be its *quantile function*, the generalized inverse of F_ν.

Let us begin by recalling the following well-known fact.

Theorem 3 (33, Theorem 2.9) *Let ξ and ν be probability distributions on $\mathbb{R}$. If ξ is absolutely continuous with respect to the Lebesgue measure, then a nondecreasing function $G^* = F_\nu^{[-1]} \circ F_\xi$ is the OT map under the quadratic cost.*

A.1. Characterizing the noise model in the one-dimensional setting

We study the distributions of the noise ϵ_i in (7). Our main focus is to show that they are bounded on expectation. In proving this boundedness, the following lemma plays a pivotal role.

Lemma 4 *Suppose that ξ is absolutely continuous with respect to the Lebesgue measure. Then, $x_{i:n}$ and $G^*(z_{i:n})$ are identically distributed.*

Proof To show that $G^*(z_{i:n})$ and $x_{i:n}$ are identically distributed, it suffices to show that their CDFs are identical. Notice that by the definition of the generalized inverse of F_μ it holds that

$$F_\mu^{[-1]}(p) \leq x \quad \iff \quad p \leq F_\mu(x). \tag{11}$$

It is well known that the CDF of the order statistic $z_{i:n}$ is

$$F_{z_{i:n}}(t) = \mathbb{P}(z_{i:n} \leq t) = \sum_{j=i}^{n} \binom{n}{j} \big(F_\xi(t)\big)^j \big(1 - F_\xi(t)\big)^{n-j}.$$

By Theorem 3, we know that $G^*(t) = \big(F_\nu^{[-1]} \circ F_\xi\big)(t)$. Thus, with (11), one can observe that the CDF of $G^*(z_{i:n})$ satisfies

$$\begin{aligned} F_{G^*(z_{i:n})}(t) &= \mathbb{P}(G^*(z_{i:n}) \leq t) \\ &= \mathbb{P}\big(\big(F_\nu^{[-1]} \circ F_\xi\big)(z_{i:n}) \leq t\big) \\ &= \mathbb{P}(F_\xi(z_{i:n}) \leq F_\nu(t)). \end{aligned}$$

As we assume that ξ is absolutely continuous with respect to the Lebesgue measure, F_ξ has a well-defined inverse. Hence, continuing from the above, we get

$$\begin{aligned} F_{G^*(z_{i:n})}(t) &= \mathbb{P}(F_\xi(z_{i:n}) \leq F_\nu(t)) \\ &= \mathbb{P}(z_{i:n} \leq (F_\xi^{-1} \circ F_\nu)(t)) \\ &= \sum_{j=i}^{n} \binom{n}{j} \big((F_\xi \circ F_\xi^{-1} \circ F_\nu)(t)\big)^j \big(1 - (F_\xi \circ F_\xi^{-1} \circ F_\nu)(t)\big)^{n-j} \\ &= \sum_{j=i}^{n} \binom{n}{j} \big(F_\nu(t)\big)^j \big(1 - F_\nu(t)\big)^{n-j} \\ &= \mathbb{P}(x_{i:n} \leq t) \\ &= F_{x_{i:n}}(t). \end{aligned}$$

Therefore, the distributions of $G^*(z_{i:n})$ and $x_{i:n}$ are identical. ■

A.2. Proof of Lemma 1

Let us begin with some simple observations. As usual, we denote $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \nu$, and let $x_{1:n}, \dots, x_{n:n}$ be their order statistics.

Proposition 5 *Let ν be a probability distribution with $\mathbb{E}_{x\sim\nu}[|x|] < \infty$. Then for any $i = 1, \dots, n$, it holds that $\mathbb{E}[x_{i:n}] \le n\,\mathbb{E}[|x_1|]$.*

Proof From the chain of inequalities

$$\mathbb{E}[x_{i:n}] \le \mathbb{E}[|x_{i:n}|] \le \mathbb{E}\left[\sum_{i=1}^{n} |x_{i:n}|\right] = \mathbb{E}\left[\sum_{i=1}^{n} |x_i|\right] = n\,\mathbb{E}\big[|x_1|\big]$$

the bound is immediate. ■

Proposition 6 *Let ν be a probability distribution with a finite second moment. Then for any $1 \le i, j \le n$, it holds that $\mathbb{E}[x_{i:n}x_{j:n}] \le n\,\mathbb{E}[x_1^2]$.*

Proof In a similar manner to the preceding proposition, it holds for any $i = 1, \dots, n$ that

$$\mathbb{E}[x_{i:n}^2] \le \mathbb{E}\left[\sum_{i=1}^{n} x_{i:n}^2\right] = n\,\mathbb{E}\big[x_1^2\big].$$

Therefore, by the Cauchy-Schwarz inequality

$$\mathbb{E}[x_{i:n}x_{j:n}] \le \sqrt{\mathbb{E}[x_{i:n}^2]\,\mathbb{E}[x_{j:n}^2]} \le n\,\mathbb{E}\big[x_1^2\big]$$

and we are done. ■

With these results, we can prove the following lemma, which will play a key role in the proof of Lemma 1.

Lemma 7 *Under Assumption 1, for any $1 \le i, j \le n$, the covariance of $x_{i:n}$ and $x_{j:n}$ is nonnegative. That is,*

$$\mathrm{Cov}(x_{i:n}, x_{j:n}) \ge 0. \tag{12}$$

Proof For notational convenience, let $\boldsymbol{x} = (x_1, \dots, x_n)$. As $x_1, \dots, x_n$ are i.i.d. samples, they are *associated* [11, Theorem 2.1], in the sense that for any two nondecreasing functions f, g with all of $\mathbb{E}[f(\boldsymbol{x})]$, $\mathbb{E}[g(\boldsymbol{x})]$, and $\mathbb{E}[f(\boldsymbol{x})g(\boldsymbol{x})]$ finite, it holds that

$$\mathrm{Cov}(f(\boldsymbol{x}), g(\boldsymbol{x})) \ge 0. \tag{13}$$

Now for each $i = 1, \dots, n$, define

$$h_i(a_1, \dots, a_n) = (\text{the } i\text{th smallest value among } a_1, \dots, a_n)$$

so that $h_i(x_1, \dots, x_n) = x_{i:n}$. If we can take $f = h_i$ and $g = h_j$ in (13), then (12) will follow. To this end, observe that for any i and j, by Proposition 5 and Jensen's inequality,

$$\mathbb{E}[h_i(\boldsymbol{x})] = \mathbb{E}[x_{i:n}] \le n\,\mathbb{E}[|x_1|] \le n\sqrt{\mathbb{E}[x_1^2]} < \infty,$$

and by Proposition 6,

$$\mathbb{E}[h_i(\boldsymbol{x})h_j(\boldsymbol{x})] = \mathbb{E}[x_{i:n}x_{j:n}] \leq n\,\mathbb{E}[x_1^2] < \infty.$$

Hence, it suffices to show that h_i is nondecreasing for each i. But this is immediate from the definition of h_i; indeed, if any x_k increases (while all the other arguments are held fixed), then the ith order statistic cannot decrease. This completes the proof. ■

We are now ready to prove Lemma 1.

Lemma 8 (Lemma 1) *Suppose that Assumption 1 holds, and ξ is absolutely continuous with respect to the Lebesgue measure. Then for any $n \geq 1$, the noise in (7) satisfies*

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\epsilon_i) \leq 2\sigma^2.$$

Proof By Lemma 4, $x_{i:n}$ and $G^*(z_{i:n})$ are identically distributed. In particular, $\mathbb{E}[x_{i:n}] = \mathbb{E}[G^*(z_{i:n})]$. Hence, by defining $\epsilon_{x_{i:n}} = x_{i:n} - \mathbb{E}[x_{i:n}]$ and $\epsilon_{z_{i:n}} = G^*(z_{i:n}) - \mathbb{E}[G^*(z_{i:n})]$, as z_i and x_j are independent for any pair of i and j, we can decompose the variances into

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\epsilon_i) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(x_{i:n} - G^*(z_{i:n})) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(x_{i:n} - \mathbb{E}[x_{i:n}] + \mathbb{E}[G^*(z_{i:n})] - G^*(z_{i:n})) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\epsilon_{x_{i:n}} - \epsilon_{z_{i:n}}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\epsilon_{x_{i:n}}) + \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\epsilon_{z_{i:n}}).
\end{aligned}$$

Thus, it suffices to show that both noise components induced by $x_{i:n}$ and $z_{i:n}$ have bounded variances. Bounding the variance of the noises induced by $x_{i:n}$ can be done as

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\epsilon_{x_{i:n}}) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(x_{i:n} - \mathbb{E}[x_{i:n}]) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(x_{i:n}) \\
&= \frac{1}{n}\mathrm{Var}\left(\sum_{i=1}^{n}x_{i:n}\right) - \frac{2}{n}\sum_{1\leq i<j\leq n}\mathrm{Cov}(x_{i:n}, x_{j:n}) \\
&= \frac{1}{n}\mathrm{Var}\left(\sum_{i=1}^{n}x_{i}\right) - \frac{2}{n}\sum_{1\leq i<j\leq n}\mathrm{Cov}(x_{i:n}, x_{j:n}) \\
&\leq \sigma^2 - \frac{2}{n}\sum_{1\leq i<j\leq n}^{n}\mathrm{Cov}(x_{i:n}, x_{j:n}) \\
&\leq \sigma^2
\end{aligned}$$

where the first inequality follows from ν being a distribution with a bounded variance (Assumption 1), and the second inequality is a direct consequence of Lemma 7.

For the other sum, observe that

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\epsilon_{i,z}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(G^*(z_{i:n}) - \mathbb{E}[G^*(z_{i:n})]) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(G^*(z_{i:n})).$$

By Lemma 4, for each $i = 1, \dots, n$, we know that $G^*(z_{i:n})$ and $x_{i:n}$ are identically distributed. Hence, we have

$$\sum_{i=1}^{n}\mathrm{Var}(G^*(z_{i:n})) = \sum_{i=1}^{n}\mathrm{Var}(x_{i:n}),$$

and therefore

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\epsilon_i) = \frac{2}{n}\sum_{i=1}^{n}\mathrm{Var}(x_{i:n}) \le 2\sigma^2 \tag{14}$$

which is exactly the claimed bound. ■

Appendix B. Proof for Section 4.1

B.1. Proof of Theorem 2

Theorem 9 (Theorem 2) *Let $\mathcal{E}_{\text{te}}(n, p, \delta)$ denote the value of $\mathcal{E}_{\text{te}}$ with dataset size n, number of features p, and ridge parameter $\delta \ge 0$. Suppose that $p \le p'$. Then, under Assumptions 1 and 2, it holds that*

$$\min_{\delta} \mathcal{E}_{\text{te}}(n, p', \delta) \le \min_{\delta} \mathcal{E}_{\text{te}}(n, p, \delta).$$

Moreover, if we further assume that $p > n$, then denoting $\mathcal{E}_{\text{te, 0}} = \lim_{\delta \to 0+} \mathcal{E}_{\text{te}}$, it holds that

$$\mathcal{E}_{\text{te, 0}}(n, p') \le \mathcal{E}_{\text{te, 0}}(n, p).$$

Proof The first part of the statement is an immediate consequence of Theorem 1 in [35], which states that if σ^2 is a fixed constant then $p \mapsto \min_\delta \mathcal{E}_{\text{te}}(n, p, \delta)$ is nonincreasing. In our regression model, the noise may depend on the dataset size n, but it is clear that it does not depend on the number of features p. Hence, if n is kept fixed, $\breve{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\epsilon_i)$ in (9) will also be a fixed constant, and therefore, the nonincreasing property of $\mathcal{E}_{\text{te}}$ implies the claimed inequality.

The second part of the statement follows from Proposition 2 in [35], under a similar logic. Indeed, if n is kept fixed, $\breve{\sigma}^2$ will also be fixed, so in the limit as $\delta \to 0+$, we can apply [35, Proposition 2] to conclude that $\frac{\partial \mathcal{E}_{\text{te, 0}}}{\partial p} \le 0$ when $p > n$. The claimed inequality is then immediate. ■

Appendix C. Vector-Valued Kernel Regression: Prerequisites for Section 4.2

C.1. Vector-valued linear models

As a preliminary step, let us briefly review how the kernel method can be extended to when the targets, or labels, are multidimensional. For further details, see, for example, [26].

Suppose we are given a dataset $\{(\boldsymbol{z}_i, \boldsymbol{x}_i)\}_{i=1,\dots,n} \subset \mathbb{R}^d \times \mathbb{R}^m$. In finding a regressor, we consider linear models, which are functions of the form $\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{\psi}(\boldsymbol{z})^\top \boldsymbol{\beta}$ for some *matrix-valued* feature map $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^{p\times m}$, so that the model is linear on the p-dimensional parameter $\boldsymbol{\beta} \in \mathbb{R}^p$.

Such a feature map induces a *matrix-valued* kernel $\boldsymbol{k}(\boldsymbol{z}, \boldsymbol{z}') = \boldsymbol{\psi}(\boldsymbol{z})^\top \boldsymbol{\psi}(\boldsymbol{z}') \in \mathbb{R}^{m\times m}$. By mimicking the construction of the reproducing Hilbert kernel space (RKHS) associated with scalar-valued kernels, one can show the existence of the RKHS $\mathcal{H}$ associated with $\boldsymbol{k}$, in the sense that

(i) $\mathcal{H}$ is a Hilbert space consisting of functions $\mathbb{R}^d \to \mathbb{R}^p$,

(ii) there exists a function $\kappa : \mathbb{R}^d \to \mathcal{H}$ such that $\boldsymbol{k}(\boldsymbol{z}, \boldsymbol{z}') = \langle \kappa(\boldsymbol{z}), \kappa(\boldsymbol{z}') \rangle_{\mathcal{H}}$ $\forall \boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^d$, where $\langle \cdot\,, \cdot \rangle_{\mathcal{H}}$ is the inner product defined on $\mathcal{H}$, and

(iii) for any $\boldsymbol{h} \in \mathcal{H}$ and $\boldsymbol{z} \in \mathbb{R}^d$, denoting by h_j the jth component function of $\boldsymbol{h}$ and by $\boldsymbol{e}_j$ the jth standard basis vector of $\mathbb{R}^m$, it holds that

$$\langle \boldsymbol{h}, \boldsymbol{k}(\,\cdot\,, \boldsymbol{z})\boldsymbol{e}_j \rangle_{\mathcal{H}} = h_j(\boldsymbol{z}). \tag{15}$$

Notice that the empirical kernel matrix (or the *Gram matrix*) becomes a block matrix of the form

$$\hat{\boldsymbol{K}} = \begin{bmatrix} \boldsymbol{k}(\boldsymbol{z}_1, \boldsymbol{z}_1) & \cdots & \boldsymbol{k}(\boldsymbol{z}_1, \boldsymbol{z}_n) \\ \vdots & \ddots & \vdots \\ \boldsymbol{k}(\boldsymbol{z}_n, \boldsymbol{z}_1) & \cdots & \boldsymbol{k}(\boldsymbol{z}_n, \boldsymbol{z}_n) \end{bmatrix} \in \mathbb{R}^{nm\times nm}.$$

Now we consider the empirical risk minimization problem, applied to this setting. For a loss function $\ell : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$, one can either consider an empirical risk minimization problem of finding an optimal parameter,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n \ell(\boldsymbol{x}_i, \boldsymbol{\psi}(\boldsymbol{z}_i)^\top \boldsymbol{\beta}) + \frac{\delta}{2} \|\boldsymbol{\beta}\|^2, \tag{16}$$

or minimizing the empirical risk directly over the RKHS,

$$\min_{\boldsymbol{f}\in\mathcal{H}} \sum_{i=1}^n \ell(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{z}_i)) + \frac{\delta}{2} \|\boldsymbol{f}\|_{\mathcal{H}}^2. \tag{17}$$

The representer theorem also holds for vector-valued outputs. In other words, there exist vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n \in \mathbb{R}^m$ such that the empirical risk minimization problem (16) has an optimal solution $\hat{\boldsymbol{\beta}}$ which can be written as

$$\hat{\boldsymbol{\beta}} = \sum_{j=1}^n \boldsymbol{\psi}(\boldsymbol{z}_j)\boldsymbol{a}_j,$$

and moreover, for the same vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$, the function

$$\hat{\boldsymbol{f}} = \sum_{j=1}^n \boldsymbol{k}(\,\cdot\,, \boldsymbol{z}_j)\boldsymbol{a}_j$$

becomes an optimal solution $\hat{\boldsymbol{f}}$ to (17). In particular, for our purpose, we set ℓ to be the mean squared error $\ell(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2$. Then, (16) becomes an unconstrained convex quadratic programming, for which we readily have a closed-form solution

$$\mathfrak{a} = (\hat{\boldsymbol{K}} + \delta \boldsymbol{I})^{-1} \mathfrak{x}.$$

Here $\mathfrak{a}$ and $\mathfrak{x}$ are concatenations of $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, respectively, into vectors in $\mathbb{R}^{nm}$.

C.2. Extending the eigenlearning framework

To simplify our narration, we act as if the data points $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are sampled without replacement from a discrete set $\mathcal{X} \subset \mathbb{R}^d$ with $|\mathcal{X}| = M$ very large. This is a strategy also taken in [34], and the arguments supporting the validity of this choice therein still apply.

Under very mild assumptions, the kernel $\boldsymbol{k}$ admits a Mercer-type decomposition

$$\boldsymbol{k}(\boldsymbol{z}, \boldsymbol{z}') = \sum_i \lambda_i \boldsymbol{\phi}_i(\boldsymbol{z}) \boldsymbol{\phi}_i(\boldsymbol{z}')^\top. \tag{18}$$

For example, only by assuming that $\boldsymbol{\psi}$ is continuous and $\boldsymbol{z} \mapsto \|\boldsymbol{k}(\boldsymbol{z}, \boldsymbol{z})\|$ is locally bounded one can get a decomposition with each eigenvector $\boldsymbol{\phi}_i$ being continuous [9].

The decomposition (18) of the kernel $\boldsymbol{k}$ leads to a decomposition $\boldsymbol{K} = \boldsymbol{\Phi}^\top \boldsymbol{\Lambda} \boldsymbol{\Phi}$ of the empirical kernel matrix, where $\boldsymbol{\Phi}$ is the design matrix having a block matrix form

$$\boldsymbol{\Phi} = \begin{bmatrix} — \, \boldsymbol{\phi}_1(\boldsymbol{z}_1)^\top \, — & \cdots & — \, \boldsymbol{\phi}_1(\boldsymbol{z}_n)^\top \, — \\ \vdots & \ddots & \vdots \\ — \, \boldsymbol{\phi}_M(\boldsymbol{z}_1)^\top \, — & \cdots & — \, \boldsymbol{\phi}_M(\boldsymbol{z}_n)^\top \, — \end{bmatrix}$$

and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_M)$.

Let $\boldsymbol{f} = \sum_i v_i \boldsymbol{\phi}_i$ and $\hat{\boldsymbol{f}} = \sum_i \hat{v}_i \boldsymbol{\phi}_i$ be the expansions of the target function $\boldsymbol{f}$ and the estimator $\hat{\boldsymbol{f}}$, respectively, with respect to the eigenbasis. With some algebra (which we omit here, as the derivations are mostly identical to what we detail in Appendix C.3), one can verify that the identity

$$\hat{\boldsymbol{v}} = \underbrace{\boldsymbol{\Lambda} \boldsymbol{\Phi} (\boldsymbol{\Phi}^\top \boldsymbol{\Lambda} \boldsymbol{\Phi} + \delta \boldsymbol{I})^{-1} \boldsymbol{\Phi}^\top}_{=: \boldsymbol{T}} \boldsymbol{v}$$

still holds *mutatis mutandis*, with the *learning transfer matrix* $\boldsymbol{T}$.

C.3. Extending the RF eigenframework

Let us now move on to random feature models, where the feature map $\boldsymbol{\psi}$ is possibly nondeterministic. More precisely, we assume that for some function $\boldsymbol{g} : \mathbb{R}^h \times \mathbb{R}^d \to \mathbb{R}^m$ and a probability distribution ρ on $\mathbb{R}^h$, the feature map $\boldsymbol{\psi}$ is given by

$$\boldsymbol{\psi}(\boldsymbol{z}) = \begin{bmatrix} — \, \boldsymbol{g}(\boldsymbol{w}_1, \boldsymbol{z})^\top \, — \\ \vdots \\ — \, \boldsymbol{g}(\boldsymbol{w}_p, \boldsymbol{z})^\top \, — \end{bmatrix} \tag{19}$$

where $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p$ are i.i.d. samples from ρ.

Due to the targets being multidimensional, we have to deal with vector spaces of $\mathbb{R}^m$-valued functions. In particular, we have to work with $L^2(\xi; \mathbb{R}^m)$, the vector space of $\mathbb{R}^m$-valued functions that are "square-integrable with respect to ξ". Strictly speaking, to define measurability and integrability of Banach-space-valued functions in general, one has to introduce the concept of *Bochner integrals*; see, *e.g.*, [43, Sections V.4–5]. Fortunately, when the codomain is a Euclidean space $\mathbb{R}^m$, the measurability (resp., integrability) of a function reduces to the measurability (resp., integrability)

of each of the component functions. Under this notion of integrability, $L^2(\xi;\mathbb{R}^m)$ becomes a Hilbert space once we give the inner product as

$$\langle \boldsymbol{f}, \boldsymbol{h} \rangle_{L^2(\xi;\mathbb{R}^m)} \coloneqq \int \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{h}(\boldsymbol{z})\, d\xi(\boldsymbol{z}).$$

In the RF eigenlearning framework [35] which considers the case where $m = 1$, the singular value decomposition (SVD) of $\boldsymbol{g}(\boldsymbol{w}, \boldsymbol{z})$, a scalar-valued function in that case, plays an important role. Thus, a key step in extending this framework is to show that SVD is also possible for vector-valued functions.

Theorem 10 (Singular Value Decomposition of Vector-Valued Functions) *Assume that $L^2(\rho)$ and $L^2(\xi;\mathbb{R}^m)$ are separable, and $\boldsymbol{g} \in L^2(\rho \times \xi;\mathbb{R}^m)$. Then, there exist positive numbers $\sigma_1 \geq \sigma_2 \geq \cdots$ and orthonormal sequences $\{\zeta_i\}_i \subset L^2(\rho)$ and $\{\boldsymbol{\phi}_i\}_i \subset L^2(\xi;\mathbb{R}^m)$ such that*

$$\boldsymbol{g}(\boldsymbol{w}, \boldsymbol{z}) = \sum_i \sigma_i \zeta_i(\boldsymbol{w}) \boldsymbol{\phi}_i(\boldsymbol{z}) \tag{20}$$

where the sum is either finite or converges in the L^2 norm.

Proof Consider an integral operator $\mathcal{I} : L^2(\rho) \to L^2(\xi\,;\mathbb{R}^m)$ defined as

$$\mathcal{I}(f) = \int \boldsymbol{g}(\boldsymbol{w}, \cdot) f(\boldsymbol{w})\, d\rho(\boldsymbol{w}),$$

and denote the component operators of $\mathcal{I}$ by $\mathcal{I}(f) = (\mathcal{I}_1(f), \ldots, \mathcal{I}_m(f))$.

We first claim that $\mathcal{I}$ is a compact operator. As $\boldsymbol{g}$ is coordinatewise square-integrable, each of $\mathcal{I}_1, \ldots, \mathcal{I}_m$ is compact [21, Theorem 22.4]. Now suppose that $f_1, f_2, \ldots$ is a bounded sequence in $L^2(\rho)$. By passing into a subsequence m times, with at the jth pass ensuring that the image of the obtained subsequence under $\mathcal{I}_j$ converges, we can find a subsequence, say $\{f_{k_j}\}_j$, of $\{f_k\}_k$ such that all of $\{\mathcal{I}_1(f_{k_j})\}_j, \ldots, \{\mathcal{I}_m(f_{k_j})\}_j$ converge. This is equivalent to the convergence of $\{\mathcal{I}(f_{k_j})\}_j$, from which we can conclude that $\mathcal{I}$ itself is also indeed a compact operator.

The compactness of $\mathcal{I}$ implies that it admits a singular value decomposition [10, Theorem 1.6]; there exist positive numbers $\sigma_1 \geq \sigma_2 \geq \ldots$ and orthonormal sequences $\{\zeta_i\}_i \subset L^2(\rho)$ and $\{\boldsymbol{\phi}_i\}_i \subset L^2(\xi;\mathbb{R}^m)$ such that

$$\mathcal{I}(\,\cdot\,) = \sum_i \sigma_i \,\langle \zeta_i, \cdot \rangle_{L^2(\rho)}\, \boldsymbol{\phi}_i \tag{21}$$

where the sum converges in the operator norm.

Let $\{\eta_i\}_i$ an orthonormal sequence in $L^2(\rho)$ so that $\{\zeta_i\} \cup \{\eta_i\}$ is an orthonormal basis of $L^2(\rho)$. Similarly, let $\{\boldsymbol{\chi}_i\}_i \subset L^2(\xi;\mathbb{R}^m)$ be so that $\{\boldsymbol{\phi}_i\}_i \cup \{\boldsymbol{\chi}_i\}_i$ is an orthonormal basis of $L^2(\xi;\mathbb{R}^m)$. For the moment, let us assume that the right hand side of (21) is an infinite sum. For each $k = 1, 2, \ldots$, let $\boldsymbol{g}_k(\boldsymbol{w}, \boldsymbol{z}) = \sum_{i=1}^k \sigma_i \zeta_i(\boldsymbol{w}) \boldsymbol{\phi}_i(\boldsymbol{z})$, $\boldsymbol{r}_{k+1} = \boldsymbol{g} - \boldsymbol{g}_k$, and

$$\mathcal{J}_k(f) = \int \boldsymbol{r}_{k+1}(\boldsymbol{w}, \cdot) f(\boldsymbol{w})\, d\rho(\boldsymbol{w}).$$

Then by (21), we have

$$\begin{aligned}\mathcal{J}_k(\zeta_i) &= \int \boldsymbol{g}(\boldsymbol{w},\,\cdot\,)\zeta_i(\boldsymbol{w})\,d\rho(\boldsymbol{w}) - \int \boldsymbol{g}_k(\boldsymbol{w},\,\cdot\,)\zeta_i(\boldsymbol{w})\,d\rho(\boldsymbol{w}) \\ &= \mathcal{I}(\zeta_i) - \sum_{j=1}^{k}\int \zeta_i(\boldsymbol{w})\zeta_j(\boldsymbol{w})\boldsymbol{\phi}_j\,d\rho(\boldsymbol{w}) \\ &= \sigma_i\boldsymbol{\phi}_i - \sigma_i\boldsymbol{\phi}_i\mathbb{1}_{\{i\le k\}} \\ &= \sigma_i\boldsymbol{\phi}_i\mathbb{1}_{\{i>k\}}.\end{aligned} \tag{22}$$

For any $i, j = 1, 2, \ldots,$ using $\zeta_i\boldsymbol{\phi}_j$ to denote $(\boldsymbol{w}, \boldsymbol{z}) \mapsto \zeta_i(\boldsymbol{w})\boldsymbol{\phi}_j(\boldsymbol{z})$, it follows that

$$\begin{aligned}\langle \boldsymbol{r}_{k+1}, \zeta_i\boldsymbol{\phi}_j\rangle_{L^2(\rho\times\xi;\mathbb{R}^m)} &= \iint \boldsymbol{r}_{k+1}(\boldsymbol{w},\boldsymbol{z})^\top\left(\zeta_i(\boldsymbol{w})\boldsymbol{\phi}_j(\boldsymbol{z})\right)\,d(\rho\times\xi)(\boldsymbol{w},\boldsymbol{z}) \\ &= \int\left(\int \boldsymbol{r}_{k+1}(\boldsymbol{w},\boldsymbol{z})\zeta_i(\boldsymbol{w})\,d\rho(\boldsymbol{w})\right)^\top\boldsymbol{\phi}_j(\boldsymbol{z})\,d\xi(\boldsymbol{z}) \\ &= \int \mathcal{J}_k(\zeta_i)^\top\boldsymbol{\phi}_j\,d\xi \\ &= \langle \mathcal{J}_k(\zeta_i), \boldsymbol{\phi}_j\rangle_{L^2(\xi;\mathbb{R}^m)} \\ &= \sigma_i\delta_{ij}\mathbb{1}_{\{i>k\}}.\end{aligned} \tag{23}$$

Let us use analogous notations for $\eta_i\boldsymbol{\phi}_j$, $\zeta_i\boldsymbol{\chi}_j$, and $\eta_i\boldsymbol{\chi}_j$. As $\langle\zeta_I, \eta_i\rangle_{L^2(\rho)} = 0$ for any pair (I, i), it is immediate from (21) that $\mathcal{I}(\eta_i) = \mathbf{0}$. Following the same steps as in (22) and (23), we also get $\mathcal{J}_k(\eta_i) = \mathbf{0}$ and hence

$$\langle \boldsymbol{r}_{k+1}, \eta_i\boldsymbol{\phi}_j\rangle_{L^2(\rho\times\xi;\mathbb{R}^m)} = 0. \tag{24}$$

Moreover, as $\langle\boldsymbol{\phi}_I, \boldsymbol{\chi}_i\rangle_{L^2(\xi;\mathbb{R}^m)} = 0$ for any pair (I, i), for any $\theta \in \{\zeta_i\}\cup\{\eta_i\}$, from (21) we get $\langle\mathcal{I}(\theta), \boldsymbol{\chi}_j\rangle_{L^2(\xi;\mathbb{R}^m)} = 0$, and by the construction of $\boldsymbol{g}_k$ we further have $\langle\mathcal{J}_k(\theta), \boldsymbol{\chi}_j\rangle_{L^2(\xi;\mathbb{R}^m)} = 0$. So it follows that

$$\langle \boldsymbol{r}_{k+1}, \theta\boldsymbol{\chi}_j\rangle_{L^2(\rho\times\xi;\mathbb{R}^m)} = \langle\mathcal{J}_k(\theta), \boldsymbol{\chi}_j\rangle_{L^2(\xi;\mathbb{R}^m)} = 0. \tag{25}$$

As $\{\zeta_i\boldsymbol{\phi}_j\}_{i,j}\cup\{\eta_i\boldsymbol{\phi}_j\}_{i,j}\cup\{\zeta_i\boldsymbol{\chi}_j\}_{i,j}\cup\{\eta_i\boldsymbol{\chi}_j\}_{i,j}$ is an orthonormal basis of $L^2(\rho\times\xi;\mathbb{R}^m)$ with (24) and (25), we conclude that

$$\begin{aligned}\|\boldsymbol{r}_{k+1}\|^2_{L^2(\rho\times\xi;\mathbb{R}^m)} &= \sum_{i,j}\left|\langle \boldsymbol{r}_{k+1}, \zeta_i\boldsymbol{\phi}_j\rangle_{L^2(\rho\times\xi;\mathbb{R}^m)}\right|^2 \\ &= \sum_{i,j}\sigma_i^2\delta_{ij}\mathbb{1}_{\{i>k\}} \\ &= \sum_{i>k}\sigma_i^2.\end{aligned}$$

In particular, by considering when $k = 0$, because $\boldsymbol{r}_1 = \boldsymbol{g}$ is square-integrable we get $\sum_i\sigma_i^2 < \infty$, and consequently, $\|\boldsymbol{g} - \boldsymbol{g}_k\|^2_{L^2(\rho\times\xi;\mathbb{R}^m)} = \|\boldsymbol{r}_{k+1}\|^2_{L^2(\rho\times\xi;\mathbb{R}^m)} \to 0$ as $k\to\infty$. That is,

$$\boldsymbol{g}(\boldsymbol{w},\boldsymbol{z}) = \sum_i \sigma_i\zeta_i(\boldsymbol{w})\boldsymbol{\phi}_i(\boldsymbol{z}), \tag{26}$$

where the sum converges in the L^2 norm. Now, if the right hand side of (21) were a finite sum, say i ranging from 1 to I, the same logic applies, but it suffices to consider k ranging only up to I, and we would have $\boldsymbol{g} - \boldsymbol{g}_I = \mathbf{0}$. Still, we do have (26) in this case also, hence the proof is complete. ■

To avoid unnecessary complexities, it is desirable that $\{\zeta_i\}_i$ and $\{\boldsymbol{\phi}_i\}$ are orthonormal bases of their respective ambient spaces. To this end, from now on let us assume that the dimensions of $L^2(\rho)$ and $L^2(\xi; \mathbb{R}^m)$ are equal, so that even if we extend $\{\zeta_i\}_i$ and $\{\boldsymbol{\phi}_i\}_i$ into orthonormal bases, we still have the expansion of $\boldsymbol{g}$ as in (20) by allowing $\sigma_i = 0$ if necessary. For example, under the mild assumption that ρ and ξ each assign nonzero measures to a countably infinite number of disjoint sets in $\mathbb{R}^h$ and $\mathbb{R}^d$ respectively, the dimensions of both spaces will be countably infinite.

Equipped with the SVD of vector-valued functions, we have the empirical kernel

$$\begin{aligned}\hat{\boldsymbol{k}}(\boldsymbol{z}, \boldsymbol{z}') = \frac{1}{p}\boldsymbol{\psi}(\boldsymbol{z})^\top \boldsymbol{\psi}(\boldsymbol{z}') &= \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{z})\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{z}')^\top \\ &= \frac{1}{p}\sum_{j,j'}\sum_{i=1}^{p} \sigma_j \sigma_{j'} \zeta_j(\boldsymbol{w}_i)\zeta_{j'}(\boldsymbol{w}_i)\boldsymbol{\phi}_j(\boldsymbol{z})\boldsymbol{\phi}_{j'}(\boldsymbol{z}')^\top\end{aligned} \tag{27}$$

and the deterministic kernel

$$\boldsymbol{k}(\boldsymbol{z}, \boldsymbol{z}') = \mathbb{E}_{\boldsymbol{w}}[\boldsymbol{g}(\boldsymbol{w}, \boldsymbol{z})\boldsymbol{g}(\boldsymbol{w}, \boldsymbol{z}')^\top] = \sum_j \lambda_j \boldsymbol{\phi}_j(\boldsymbol{z})\boldsymbol{\phi}_j(\boldsymbol{z}')^\top \tag{28}$$

where we set $\lambda_j = \sigma_j^2$. Thanks to the similarity of the equations (27) and (28) to their counterparts in the scalar-valued regression RF eigenframework [35], the remaining steps are mostly straightforward, resembling what we did in Appendix C.2 to extend the scalar-valued framework into the vector-valued setting. Define $\boldsymbol{\Lambda} \coloneqq \operatorname{diag}(\lambda_1, \lambda_2, \dots)$ and

$$\boldsymbol{Z} \coloneqq \begin{bmatrix} \zeta_1(\boldsymbol{w}_1) & \cdots & \zeta_1(\boldsymbol{w}_p) \\ \zeta_2(\boldsymbol{w}_1) & \cdots & \zeta_2(\boldsymbol{w}_p) \\ \vdots & \vdots & \vdots \end{bmatrix}$$

so that $\tilde{\boldsymbol{\Lambda}} \coloneqq \frac{1}{p}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{\Lambda}^{1/2}$ is a matrix whose entries are $\tilde{\Lambda}_{jj'} = \frac{1}{p}\sum_{i=1}^{p}\sigma_j\sigma_{j'}\zeta_j(\boldsymbol{w}_i)\zeta_{j'}(\boldsymbol{w}_i)$. Then for a design matrix

$$\boldsymbol{\Phi} = \begin{bmatrix} — \boldsymbol{\phi}_1(\boldsymbol{z}_1)^\top — & \cdots & — \boldsymbol{\phi}_1(\boldsymbol{z}_n)^\top — \\ — \boldsymbol{\phi}_2(\boldsymbol{z}_1)^\top — & \cdots & — \boldsymbol{\phi}_2(\boldsymbol{z}_n)^\top — \\ \vdots & \vdots & \vdots \end{bmatrix},$$

it is clear from (27) that the empirical kernel matrix $\hat{\boldsymbol{K}} = \left[\hat{\boldsymbol{k}}(\boldsymbol{z}_i, \boldsymbol{z}_j)\right]_{i,j}$ can be written as

$$\hat{\boldsymbol{K}} = \boldsymbol{\Phi}^\top \tilde{\boldsymbol{\Lambda}} \boldsymbol{\Phi}.$$

Again, let $\boldsymbol{f} = \sum_i v_i \boldsymbol{\phi}_i$ and $\hat{\boldsymbol{f}} = \sum_i \hat{v}_i \boldsymbol{\phi}_i$ be the expansions with respect to the eigenbasis of the target function $\boldsymbol{f}$ and the estimator $\hat{\boldsymbol{f}}$, respectively. Then for each $J = 1, 2, \dots$, noting that the

kernel we actually use in the regression is the empirical kernel $\hat{\boldsymbol{k}}$, one can observe that

$$\begin{aligned}\hat{v}_J = \left\langle \boldsymbol{\phi}_J, \hat{\boldsymbol{f}} \right\rangle_{L^2(\xi;\mathbb{R}^d)} &= \left\langle \boldsymbol{\phi}_J, \sum_{s=1}^{n} \hat{\boldsymbol{k}}(\,\cdot\,, \boldsymbol{z}_s)\boldsymbol{a}_s \right\rangle_{L^2(\xi;\mathbb{R}^d)} \\ &= \sum_{s=1}^{n} \frac{1}{p} \sum_{j'} \sum_{i=1}^{p} \sigma_J \sigma_{j'} \zeta_J(\boldsymbol{w}_i)\zeta_{j'}(\boldsymbol{w}_i)\boldsymbol{\phi}_{j'}(\boldsymbol{z}_s)^\top \boldsymbol{a}_s \\ &= \sum_{s=1}^{n} \sum_{j'} \tilde{\boldsymbol{\Lambda}}_{Jj'} \boldsymbol{\phi}_{j'}(\boldsymbol{z}_s)^\top \boldsymbol{a}_s \\ &= \left[\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Phi}\mathbb{a}\right]_J\end{aligned}$$

where $[\,\cdot\,]_J$ denotes the Jth component of a vector. Therefore, when there is no noise in the targets (*i.e.,* labels), the identity

$$\hat{\boldsymbol{v}} = \underbrace{\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Phi} + \delta\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}^\top}_{=:\tilde{\boldsymbol{T}}}\boldsymbol{v}$$

still holds *mutatis mutandis*, with the *learning transfer matrix* $\tilde{\boldsymbol{T}}$.

For further analyses, it would be desirable to have an assumption similar to the Gaussian universality ansatz (Assumption 2). The orthonormality equations

$$\mathbb{E}_{\boldsymbol{w}\sim\rho}[\zeta_i(\boldsymbol{w})\zeta_j(\boldsymbol{w})] = \delta_{ij} \tag{29a}$$

$$\mathbb{E}_{\boldsymbol{z}\sim\xi}[\boldsymbol{\phi}_i(\boldsymbol{z})^\top\boldsymbol{\phi}_j(\boldsymbol{z})] = \delta_{ij} \tag{29b}$$

suggest a natural multivariate rendition of the ansatz as follows.

Assumption M2 (*Multivariate Gaussian Universality Ansatz*) *The expected population risk remains unchanged even if we replace $\{\boldsymbol{\phi}_i\}$ and $\{\zeta_i\}$ each with random Gaussian functions $\{\tilde{\boldsymbol{\phi}}_i\}$ and $\{\tilde{\zeta}_i\}$, in the sense that $\{\tilde{\zeta}_i(\boldsymbol{w})\}$ become i.i.d. samples from $\mathcal{N}(0,1)$ when $\boldsymbol{w}\sim\mu$, and $\{\tilde{\boldsymbol{\phi}}_i(\boldsymbol{z})\}$ become i.i.d. samples from $\mathcal{N}(\boldsymbol{0}, \frac{1}{m}\boldsymbol{I})$ when $\boldsymbol{z}\sim\xi$.*

Remark 11 *The covariance matrix associated to $\boldsymbol{\phi}$ is $\frac{1}{m}\boldsymbol{I}$ instead of $\boldsymbol{I}$, as we wish to have $\mathbb{E}\left[\left\|\tilde{\boldsymbol{\phi}}_i(\boldsymbol{z}_j)\right\|^2\right] = 1$, based on (29b).*

Under the multivariate Gaussian universality ansatz, $\boldsymbol{\Phi}$ is a matrix whose entries are i.i.d. samples from $\mathcal{N}(0, \frac{1}{m})$, or equivalently, $\frac{1}{\sqrt{m}}\mathcal{N}(0,1)$. Hence, rewriting the learning transfer matrix in terms of $\boldsymbol{\Omega} = \sqrt{m}\boldsymbol{\Phi}$ as

$$\tilde{\boldsymbol{T}} = \tilde{\boldsymbol{\Lambda}}\boldsymbol{\Omega}\left(\boldsymbol{\Omega}^\top\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Omega} + m\delta\boldsymbol{I}\right)^{-1}\boldsymbol{\Omega}^\top$$

shows that $\tilde{\boldsymbol{T}}$ is statistically equivalent to the learning transfer matrix appearing in the scalar-valued RF eigenframework, up to the number of samples n and the ridge parameter δ being scaled by a factor of m. That is, we can apply the results from the RF eigenframework [35], with necessary modifications, as follows.

Let $s \coloneqq \sum_i \frac{\lambda_i}{\lambda_i+\gamma}$, $q \coloneqq \sum_i \left(\frac{\lambda_i}{\lambda_i+\gamma}\right)^2$, and $\kappa, \gamma \geq 0$ be the unique nonnegative scalars such that

$$nm = s + \frac{m\delta}{\kappa} \qquad \text{and} \qquad p = s + \frac{p\kappa}{\gamma}.$$

The test error of the vector-valued RF regression is then given approximately by

$$\mathbb{E}_{\mathcal{D}}\big[\mathcal{R}(\hat{\boldsymbol{\beta}})\big] \approx \mathcal{E}_{\text{te}} \coloneqq \frac{1}{1 - \frac{q(p-2s)+s^2}{nm(p-q)}} \left(\sum_i \left(\frac{\gamma}{\lambda_i+\gamma} - \frac{\kappa\lambda_i}{(\lambda_i+\gamma)^2} \frac{p}{p-q} \right) v_i^2 + \breve{\sigma}^2 \right) \tag{30}$$

where $\breve{\sigma}^2$ is the mean squared error of the noise, or in other words, the part of the data that is effectively noise.

C.4. Neural-network-like random feature models

When the labels are scalar-valued, a noteworthy interpretation of linear models that reveals their connection to modern machine learning models is as 2-layer (fully-connected feed-forward) neural networks, with a fixed first layer and a trainable second layer. Thus, it is natural to ask if this is also the case in the vector-valued case. The answer is yes, but with a slight caveat, which we now detail.

A 2-layer network with p hidden neurons, taking d-dimensional inputs and producing m-dimensional outputs, can be modeled as $\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{B}\sigma(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{c})$, for weights $\boldsymbol{W} \in \mathbb{R}^{p\times d}$, $\boldsymbol{c} \in \mathbb{R}^p$, $\boldsymbol{B} \in \mathbb{R}^{m\times p}$, and an activation function σ applied elementwise. As $\boldsymbol{f}$ is linear on $\boldsymbol{B}$, it indeed is a linear model, and this becomes more apparent from the reformulation

$$\boldsymbol{f}(\boldsymbol{z}) = \begin{bmatrix} \sigma(\boldsymbol{W}\boldsymbol{z}+\boldsymbol{b})^\top & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \sigma(\boldsymbol{W}\boldsymbol{z}+\boldsymbol{b})^\top & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \sigma(\boldsymbol{W}\boldsymbol{z}+\boldsymbol{b})^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix}$$

where $\boldsymbol{\beta}_i$ is the ith row vector of $\boldsymbol{B}$ in the column vector form. In other words, $\boldsymbol{f}$ is a linear model with a block diagonal feature map.

That said, this is exactly where the caveat arises. The issue is that the feature map is not of the form of (19), as that formulation cannot produce the precise pattern of zeros we now require. This shows that a separate analysis from Appendix C.3 is required for this case.

To this end, let us consider a setting where the feature map now is a block diagonal matrix

$$\boldsymbol{\psi}(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{\psi}_1(\boldsymbol{z}) & & \\ & \ddots & \\ & & \boldsymbol{\psi}_1(\boldsymbol{z}) \end{bmatrix} \in \mathbb{R}^{mp\times m}. \tag{31}$$

The notation $\boldsymbol{\psi}_1$ is used to emphasize that $\boldsymbol{\psi}_1(\boldsymbol{z})$ takes the same form as feature maps (19) for when $m = 1$. In other words, as in the scalar-valued labels setting, for some scalar valued map $g : \mathbb{R}^h \times \mathbb{R}^d \to \mathbb{R}$ we consider when $\boldsymbol{\psi}_1(\boldsymbol{z}) = \big(g(\boldsymbol{w}_1, \boldsymbol{z}), \dots, g(\boldsymbol{w}_p, \boldsymbol{z})\big)$ with $\boldsymbol{w}_1, \dots, \boldsymbol{w}_p \overset{\text{i.i.d.}}{\sim} \rho$. Let us call the linear models that arise from (random) feature maps having such block diagonal form the *neural-network-like linear models*. For convenience, we also let $\boldsymbol{\Psi}_1 \coloneqq \big[\boldsymbol{\psi}_1(\boldsymbol{z}_1) \ \cdots \ \boldsymbol{\psi}_1(\boldsymbol{z}_n)\big]^\top$.

Similar to before, we define the empirical kernel as $\hat{\boldsymbol{k}}(\boldsymbol{z}, \boldsymbol{z}') = \frac{1}{p}\boldsymbol{\psi}(\boldsymbol{z})^\top \boldsymbol{\psi}(\boldsymbol{z}')$,[2] then we get

$$\begin{aligned}\hat{\boldsymbol{k}}(\boldsymbol{z}, \boldsymbol{z}') &= \frac{1}{p}\boldsymbol{\psi}_1(\boldsymbol{z})^\top \boldsymbol{\psi}_1(\boldsymbol{z}')\boldsymbol{I} \\ &= \frac{1}{p}\sum_{i=1}^{p} g(\boldsymbol{w}_i, \boldsymbol{z})g(\boldsymbol{w}_i, \boldsymbol{z}')\boldsymbol{I}.\end{aligned}$$

Observe that $\frac{1}{p}\sum_{i=1}^{p} g(\boldsymbol{w}_i, \boldsymbol{z})g(\boldsymbol{w}_i, \boldsymbol{z}')$ is in fact what we would get as the empirical kernel when $m = 1$ so that $\boldsymbol{\psi}_1$ is used as the feature map. Thus, for $\hat{\boldsymbol{K}}_1$ denoting the matrix whose entries are $\hat{K}_{j,j'} = \frac{1}{p}\sum_{i=1}^{p} g(\boldsymbol{w}_i, \boldsymbol{z}_j)g(\boldsymbol{w}_i, \boldsymbol{z}_{j'})$, the empirical kernel matrix $\hat{\boldsymbol{K}} = \left[\hat{\boldsymbol{k}}(\boldsymbol{z}_i, \boldsymbol{z}_j)\right]_{i,j}$ is a block diagonal matrix whose block entries are all constant multiples of the identity matrix, that is, $\hat{\boldsymbol{K}} = \hat{\boldsymbol{K}}_1 \otimes \boldsymbol{I}$ where $\otimes$ denotes the Kronecker product of two matrices.

Recall that $\mathbb{x} = (x_{11}, \ldots, x_{1m}, \ldots, x_{n1}, \ldots, x_{nm})$. Define a permutation matrix $\boldsymbol{P} \in \mathbb{R}^{nm\times nm}$ such that $\boldsymbol{P}\mathbb{x} = (x_{11}, \ldots, x_{n1}, \ldots, x_{1m}, \ldots, x_{nm})$. For $j = 1, \ldots, m$, denote $\boldsymbol{y}_j = (x_{1j}, \ldots, x_{nj})$. Then as permutation matrices satisfy $\boldsymbol{P}^\top = \boldsymbol{P}^{-1}$, it holds that

$$\begin{aligned}(\hat{\boldsymbol{K}} + \delta\boldsymbol{I})^{-1}\mathbb{x} &= \boldsymbol{P}^\top \boldsymbol{P}(\hat{\boldsymbol{K}} + \delta\boldsymbol{I})^{-1}\boldsymbol{P}^\top \boldsymbol{P}\mathbb{x} \\ &= \boldsymbol{P}^\top (\boldsymbol{P}\hat{\boldsymbol{K}}\boldsymbol{P}^\top + \delta\boldsymbol{I})^{-1}\boldsymbol{P}\mathbb{x} \\ &= \boldsymbol{P}^\top \begin{bmatrix} \hat{\boldsymbol{K}}_1(\boldsymbol{z}_1, \boldsymbol{z}_1) + \delta\boldsymbol{I} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \hat{\boldsymbol{K}}_1(\boldsymbol{z}_1, \boldsymbol{z}_1) + \delta\boldsymbol{I} \end{bmatrix}^{-1} \boldsymbol{P}\mathbb{x} \\ &= \boldsymbol{P}^\top \begin{bmatrix} \left(\hat{\boldsymbol{K}}_1(\boldsymbol{z}_1, \boldsymbol{z}_1) + \delta\boldsymbol{I}\right)^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \left(\hat{\boldsymbol{K}}_1(\boldsymbol{z}_1, \boldsymbol{z}_1) + \delta\boldsymbol{I}\right)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix},\end{aligned}$$

and thus denoting by $\hat{\mathbb{\beta}}$ the concatenation of $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_n$ into a vector in $\mathbb{R}^{mp}$, we obtain

$$\begin{aligned}\hat{\mathbb{\beta}} &= \begin{bmatrix}\boldsymbol{\psi}(\boldsymbol{z}_1) & \cdots & \boldsymbol{\psi}(\boldsymbol{z}_n)\end{bmatrix}(\hat{\boldsymbol{K}} + \delta\boldsymbol{I})^{-1}\mathbb{x} \\ &= \begin{bmatrix}\boldsymbol{\psi}(\boldsymbol{z}_1) & \cdots & \boldsymbol{\psi}(\boldsymbol{z}_n)\end{bmatrix}\boldsymbol{P}^\top \begin{bmatrix} \left(\hat{\boldsymbol{K}}_1(\boldsymbol{z}_1, \boldsymbol{z}_1) + \delta\boldsymbol{I}\right)^{-1}\boldsymbol{y}_1 \\ \vdots \\ \left(\hat{\boldsymbol{K}}_1(\boldsymbol{z}_1, \boldsymbol{z}_1) + \delta\boldsymbol{I}\right)^{-1}\boldsymbol{y}_m \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Psi}_1^\top & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{\Psi}_1^\top \end{bmatrix} \begin{bmatrix} \left(\hat{\boldsymbol{K}}_1(\boldsymbol{z}_1, \boldsymbol{z}_1) + \delta\boldsymbol{I}\right)^{-1}\boldsymbol{y}_1 \\ \vdots \\ \left(\hat{\boldsymbol{K}}_1(\boldsymbol{z}_1, \boldsymbol{z}_1) + \delta\boldsymbol{I}\right)^{-1}\boldsymbol{y}_m \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Psi}_1^\top \left(\hat{\boldsymbol{K}}_1(\boldsymbol{z}_1, \boldsymbol{z}_1) + \delta\boldsymbol{I}\right)^{-1}\boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{\Psi}_1^\top \left(\hat{\boldsymbol{K}}_1(\boldsymbol{z}_1, \boldsymbol{z}_1) + \delta\boldsymbol{I}\right)^{-1}\boldsymbol{y}_m \end{bmatrix}.\end{aligned}$$

2. As the number of rows in $\boldsymbol{\psi}$ is now mp, one might consider choosing the factor multiplied to the sum to be $1/mp$ instead of $1/p$ to maintain consistency with Appendix C.3. However, as the main functionality of that factor is to convert the sum into a mean, and as each column of $\boldsymbol{\psi}$ now contains only p nonzero elements, we prefer $1/p$ over $1/mp$. It soon turns out that such a choice also allows us to directly apply one-dimensional results to this setting.

That is, for each $i = 1, \dots, m$, each $\hat{\boldsymbol{\beta}}_i$ only depends on $\boldsymbol{y}_i$, the collection of the ith coordinates from $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$. In other words, each $\hat{\boldsymbol{\beta}}_i$ is computed independently from others, as if it is obtained from a kernel regression in a scalar-valued labels setting by only looking at the ith coordinates in the data. On top of this, we also have

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbb{E}_{\boldsymbol{z}\sim\xi}[\boldsymbol{\psi}(\boldsymbol{z})\boldsymbol{\psi}(\boldsymbol{z})^\top] \\ &= \begin{bmatrix} \mathbb{E}_{\boldsymbol{z}\sim\xi}[\boldsymbol{\psi}_1(\boldsymbol{z})\boldsymbol{\psi}_1(\boldsymbol{z})^\top] & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbb{E}_{\boldsymbol{z}\sim\xi}[\boldsymbol{\psi}_1(\boldsymbol{z})\boldsymbol{\psi}_1(\boldsymbol{z})^\top] & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbb{E}_{\boldsymbol{z}\sim\xi}[\boldsymbol{\psi}_1(\boldsymbol{z})\boldsymbol{\psi}_1(\boldsymbol{z})^\top] \end{bmatrix},\end{aligned}$$

allowing us to conclude that the population risk is the sum of the coordinatewise population risk.

Meanwhile, notice that the mean squared error of the noise is the sum of the mean squared errors of the coordinatewise noises, no matter the noise model. Hence, estimating the population risk can be done as computing the sum of coordinatewise estimations by $\mathcal{E}_{\text{te}}$ as in (9). It follows that the proper assumption to make when studying neural-network-like linear models is the singular value decomposition of the scalar-valued function $g(\boldsymbol{w}, \boldsymbol{z}) = \sum_i \sigma_i \zeta_i(\boldsymbol{w})\phi_i(\boldsymbol{z})$ and the Gaussian universality ansatz (Assumption 2), not their multivariate versions.

With these established, let us now consider the expansions $\boldsymbol{f} = \left(\sum_j v_{1j}\phi_j, \dots, \sum_j v_{mj}\phi_j\right)$ and $\hat{\boldsymbol{f}} = \left(\sum_j \hat{v}_{1j}\phi_j, \dots, \sum_j \hat{v}_{mj}\phi_j\right)$. Denote by $\breve{\sigma}_i^2$ the mean squared noise in the ith coordinate. Then, from the coordinatewise approximation using (9), we have

$$\mathbb{E}_{\mathcal{D}}[\mathcal{R}(\hat{\beta})] \approx \sum_{i=1}^{m} \frac{1}{1 - \frac{q(p-2s)+s^2}{n(p-q)}} \left(\sum_j \left(\frac{\gamma}{\lambda_j + \gamma} - \frac{\kappa\lambda_j}{(\lambda_j+\gamma)^2} \frac{p}{p-q} \right) v_{ij}^2 + \breve{\sigma}_i^2 \right) \tag{32}$$

where, for $s \coloneqq \sum_j \frac{\lambda_j}{\lambda_j+\gamma}$ and $q \coloneqq \sum_j \left(\frac{\lambda_j}{\lambda_j+\gamma}\right)^2$, κ and γ are unique nonnegative numbers such that $n = s + \frac{\delta}{\kappa}$ and $p = s + \frac{p\kappa}{\gamma}$. As n, p, q, and s do not depend on the summation index i in the above, denoting by $\breve{\sigma}^2 = \sum_{i=1}^n \breve{\sigma}_i^2$ the overall mean squared noise, we conclude that

$$\mathbb{E}_{\mathcal{D}}[\mathcal{R}(\hat{\beta})] \approx \frac{1}{1 - \frac{q(p-2s)+s^2}{n(p-q)}} \left(\sum_{i=1}^{m} \sum_j \left(\frac{\gamma}{\lambda_j + \gamma} - \frac{\kappa\lambda_j}{(\lambda_j+\gamma)^2} \frac{p}{p-q} \right) v_{ij}^2 + \breve{\sigma}^2 \right). \tag{33}$$

Remark 12 *The same conclusion can be derived by repeating the logic developed in Appendix C.3 but with $\boldsymbol{\psi}$ as described in (31). In particular, the same expansions $\boldsymbol{f} = \left(\sum_j v_{1j}\phi_j, \dots, \sum_j v_{mj}\phi_j\right)$ and $\hat{\boldsymbol{f}} = \left(\sum_j \hat{v}_{1j}\phi_j, \dots, \sum_j \hat{v}_{mj}\phi_j\right)$ will be obtained by considering $\{\phi_j \boldsymbol{e}_i\}_{i,j}$ as the orthonormal basis of $L^2(\xi; \mathbb{R}^m)$, where $\boldsymbol{e}_i$ denotes the ith standard basis vector of $\mathbb{R}^m$. This also shows that the multivariate Gaussian universality ansatz is not quite appropriate for neural-network-like linear models; $\phi_j \boldsymbol{e}_i$ is a sparse vector, whereas $\mathcal{N}(0, \frac{1}{p}\boldsymbol{I})$ is a full-dimensional distribution. We omit the details of the derivations in this direction.*

Appendix D. Missing Details for Section 4.2

D.1. Problem settings in higher dimensions

Let us further elaborate the problem setting in higher dimensions, which was briefly mentioned in Section 4.2. Brenier's theorem holds regardless of the data dimension m, so an optimal transport map G^* from ξ to ν exists. Thus, the regression model described in (6), namely

$$\boldsymbol{x}_{\tilde{\pi}(i)} = G^*(\boldsymbol{z}_i) + \boldsymbol{\epsilon}_i, \tag{34}$$

is still valid. The linear model we consider is as stated in (10); for a feature map $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^{p\times m}$, we wish to find the model parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ such that

$$\boldsymbol{x}_{\tilde{\pi}(i)} \approx \boldsymbol{\psi}(\boldsymbol{z}_i)^\top \boldsymbol{\beta} + \boldsymbol{\epsilon}_i.$$

Moreover, as in (19), we assume that $\boldsymbol{\psi}$ is a random feature map, in the sense that there exists a function $\boldsymbol{g} : \mathbb{R}^h \times \mathbb{R}^d \to \mathbb{R}^m$ and a probability distribution ρ on $\mathbb{R}^h$ so that $\boldsymbol{\psi}$ is given by

$$\boldsymbol{\psi}(\boldsymbol{z}) = \begin{bmatrix} — \, \boldsymbol{g}(\boldsymbol{w}_1, \boldsymbol{z})^\top \, — \\ \vdots \\ — \, \boldsymbol{g}(\boldsymbol{w}_p, \boldsymbol{z})^\top \, — \end{bmatrix}$$

for i.i.d. samples $\boldsymbol{w}_1, \dots, \boldsymbol{w}_p \sim \rho$. Let us further assume that $\boldsymbol{g} \in L^2(\rho \times \xi; \mathbb{R}^m)$ so that $\boldsymbol{g}$, by Theorem 10, admits a singular value decomposition $\boldsymbol{g}(\boldsymbol{w}, \boldsymbol{z}) = \sum_i \sigma_i \zeta_i(\boldsymbol{w}) \boldsymbol{\phi}_i(\boldsymbol{z})$.

D.2. Statements on the random feature models in the high-dimensional setting

In order to analyze random feature models in the high-dimensional setting, we impose an assumption that is analogous to Assumption 1.

Assumption M1 *The target distribution ν and the latent distribution ξ have finite second moments. That is, both $\mathbb{E}_{\boldsymbol{x}\sim\nu}[\boldsymbol{x}\boldsymbol{x}^\top] := \boldsymbol{N}$ and $\mathbb{E}_{\boldsymbol{z}\sim\xi}[\boldsymbol{z}\boldsymbol{z}^\top] := \boldsymbol{\Xi}$ are finite.*

By the linearity of expectations and the invariance of the trace under cyclic permutations, we have

$$\operatorname{tr}\left(\mathbb{E}_{\boldsymbol{x}\sim\nu}[\boldsymbol{x}\boldsymbol{x}^\top]\right) = \mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\operatorname{tr}\left(\boldsymbol{x}\boldsymbol{x}^\top\right)\right] = \mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\boldsymbol{x}^\top\boldsymbol{x}\right] = \mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\|\boldsymbol{x}\|^2\right]. \tag{35}$$

Thus, Assumption M1 implies that $\mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\|\boldsymbol{x}\|^2\right] < \infty$. We also note that, for any $i = 1, \dots, m$, Jensen's inequality gives us

$$\mathbb{E}_{\boldsymbol{x}\sim\nu}[x_i] \le \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\nu}[x_i^2]} \le \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\|\boldsymbol{x}\|^2\right]}.$$

From this, we can conclude that ν also has a finite first moment.

With imposing Assumption M1, we can show a boundedness result analogous to Lemma 1.

Lemma 13 *Under Assumption M1, let $\sigma^2 := \operatorname{tr}(\boldsymbol{N}) + \operatorname{tr}(\boldsymbol{\Xi})$. Then for any $n \ge 1$, the noises in (34) satisfy $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\boldsymbol{\epsilon}_i\|^2] \le 4\sigma^2$.*

Proof By Young's inequality, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\boldsymbol{\epsilon}_i\|^2] = \frac{1}{n}\,\mathbb{E}\left[\sum_{i=1}^{n}\|\boldsymbol{x}_{\tilde{\pi}(i)} - G^*(\boldsymbol{z}_i)\|^2\right]$$
$$\leq \frac{2}{n}\,\mathbb{E}\left[\sum_{i=1}^{n}\|\boldsymbol{x}_{\tilde{\pi}(i)} - \boldsymbol{z}_i\|^2\right] + \frac{4}{n}\,\mathbb{E}\left[\sum_{i=1}^{n}\|\boldsymbol{z}_i\|^2\right] + \frac{4}{n}\,\mathbb{E}\left[\sum_{i=1}^{n}\|G^*(\boldsymbol{z}_i)\|^2\right]. \quad (36)$$

For the first term, recall that $\tilde{\pi} = \arg\min_{\pi\in S_n}\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_{\pi(i)} - \boldsymbol{z}_i\|^2$. Hence, with explicitly writing out the source of randomness over which the expectation is taken, it holds that

$$\mathbb{E}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \overset{\text{i.i.d.}}{\sim}\xi,\ \boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\overset{\text{i.i.d.}}{\sim}\nu}\left[\sum_{i=1}^{n}\|\boldsymbol{x}_{\tilde{\pi}(i)} - \boldsymbol{z}_i\|^2\right]$$
$$\leq \mathbb{E}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \overset{\text{i.i.d.}}{\sim}\xi,\ \boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\overset{\text{i.i.d.}}{\sim}\nu,\ \pi\sim\text{Uniform}(S_n)}\left[\sum_{i=1}^{n}\|\boldsymbol{x}_{\pi(i)} - \boldsymbol{z}_i\|^2\right]$$
$$= \mathbb{E}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \overset{\text{i.i.d.}}{\sim}\xi}\left[\mathbb{E}_{\pi\sim\text{Uniform}(S_n)}\left[\mathbb{E}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\overset{\text{i.i.d.}}{\sim}\nu}\left[\sum_{i=1}^{n}\|\boldsymbol{x}_{\pi(i)} - \boldsymbol{z}_i\|^2\middle|\pi,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\right]\right]\right]$$

where the last line follows from the law of total expectation. Now here, notice that, as $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ are i.i.d. samples, their joint distribution is invariant under a permutation. Hence, continuing from the above, we obtain

$$\mathbb{E}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \overset{\text{i.i.d.}}{\sim}\xi}\left[\mathbb{E}_{\pi\sim\text{Uniform}(S_n)}\left[\mathbb{E}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\overset{\text{i.i.d.}}{\sim}\nu}\left[\sum_{i=1}^{n}\|\boldsymbol{x}_{\pi(i)} - \boldsymbol{z}_i\|^2\middle|\pi,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\right]\right]\right]$$
$$= \mathbb{E}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \overset{\text{i.i.d.}}{\sim}\xi}\left[\mathbb{E}_{\pi\sim\text{Uniform}(S_n)}\left[\mathbb{E}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\overset{\text{i.i.d.}}{\sim}\nu}\left[\sum_{i=1}^{n}\|\boldsymbol{x}_{i} - \boldsymbol{z}_i\|^2\middle|\pi,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\right]\right]\right]$$
$$= \mathbb{E}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \overset{\text{i.i.d.}}{\sim}\xi,\ \boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\overset{\text{i.i.d.}}{\sim}\nu}\left[\sum_{i=1}^{n}\|\boldsymbol{x}_{i} - \boldsymbol{z}_i\|^2\right]$$
$$= n\,\mathbb{E}_{\boldsymbol{z}\sim\xi,\ \boldsymbol{x}\sim\nu}\left[\|\boldsymbol{x} - \boldsymbol{z}\|^2\right]$$
$$\leq 2n\,\mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\|\boldsymbol{x}\|^2\right] + 2n\,\mathbb{E}_{\boldsymbol{z}\sim\xi}\left[\|\boldsymbol{z}\|^2\right]$$

where the inequality step we once again apply Young's inequality. Meanwhile, for the third term of (36), as G^* is an OT map such that $\nu = (G^*)_\sharp\xi$, it holds that

$$\mathbb{E}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n \overset{\text{i.i.d.}}{\sim}\xi}\left[\sum_{i=1}^{n}\|G^*(\boldsymbol{z}_i)\|^2\right] = n\,\mathbb{E}_{\boldsymbol{z}\sim\xi}\left[\|G^*(\boldsymbol{z})\|^2\right] = n\,\mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\|\boldsymbol{x}\|^2\right].$$

Collecting the bounds, from (36) we obtain that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\boldsymbol{\epsilon}_i\|^2] \leq 4\,\mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\|\boldsymbol{x}\|^2\right] + 4\,\mathbb{E}_{\boldsymbol{z}\sim\xi}\left[\|\boldsymbol{z}\|^2\right].$$

The conclusion is now immediate from (35). ■

Then we can prove the following, which states that using more features is better for random feature models, detailed in Appendix C.3, in high dimensions also.

Theorem 14 *Consider random feature models with targets in $\mathbb{R}^m$. Let $\mathcal{E}_{\text{te}}(n,p,\delta)$ denote the value of $\mathcal{E}_{\text{te}}$ in (33) with dataset size n, number of features p, and ridge parameter δ. Under Assumptions M1 and M2, if $p \le p'$ then $\min_\delta \mathcal{E}_{\text{te}}(n,p',\delta) \le \min_\delta \mathcal{E}_{\text{te}}(n,p,\delta)$, and if moreover $p > nm$, then $\mathcal{E}_{\text{te},\,0}(n,p') \le \mathcal{E}_{\text{te},\,0}(n,p)$.*

Proof Let $\mathcal{E}'_{\text{te}}(n,p,\delta)$ denote the value of $\mathcal{E}_{\text{te}}$ in the one-dimensional setting (9), then it holds that $\mathcal{E}_{\text{te}}(n,p,\delta) = \mathcal{E}'_{\text{te}}(nm,p,\delta m)$, given that both sides involve the same $\breve{\sigma}^2$. Hence, the statements in this theorem are direct consequences of Theorem 2. ■

It is clear that a similar result holds for neural-network-like random feature models, in particular because, as we saw in Appendix C.4 the population risk is essentially the aggregation of the coordinatewise risks. For completeness, we formalize this result as follows.

Theorem 15 *Consider neural-network-like random feature models. Let $\mathcal{E}_{\text{te}}(n,p,\delta)$ denote the value of $\mathcal{E}_{\text{te}}$ in (33) with dataset size n, number of features p, and ridge parameter δ. Under Assumptions 2 and M1, if $p \le p'$ then $\min_\delta \mathcal{E}_{\text{te}}(n,p',\delta) \le \min_\delta \mathcal{E}_{\text{te}}(n,p,\delta)$, and if moreover $p > n$, then $\mathcal{E}_{\text{te},\,0}(n,p') \le \mathcal{E}_{\text{te},\,0}(n,p)$.*

Proof Clearly, each $\breve{\sigma}_i^2$ is finite, as we assume that $\breve{\sigma}^2$ is finite. Recalling (32), we know that $\mathcal{E}_{\text{te}}$ in (33) is a sum of m quantities of the form specified as $\mathcal{E}_{\text{te}}$ in the one-dimensional setting (9). Hence, applying Theorem 2, the results on the one-dimensional setting, to each of those m quantities, the conclusions follow. ■

Appendix E. Experimental Results

In this section, we discuss the claims regarding random feature models, in particular focusing on the approximation of the population risk with $\mathcal{E}_{\text{te}}$ as stated in (9) for the one-dimensional case and (30) for higher dimensions, with empirical results.

E.1. One dimensional examples

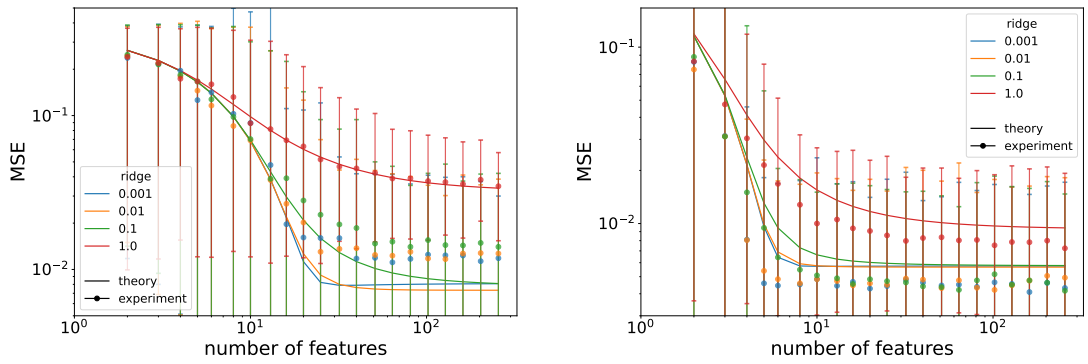
By choosing $\nu = \text{Uniform}(0,1)$, we can exploit the following well-known fact on order statistics. As a clarification, by the *bivariate Beta distribution* $\text{BivariateBeta}(i, j-i, n-j+1)$ with integer parameters $1 \le i < j < n$, we are referring to a bivariate probability distribution whose probability density function is

$$f(x,y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}\, x^{i-1}(y-x)^{j-i-1}(1-y)^{n-j}\mathbb{1}_{0<x<y<1}(x,y).$$

Proposition 16 (1, Examples 2.2.1 and 2.3.1) *In the case of $\nu = \text{Uniform}(0,1)$, the order statistics satisfy $x_{i:n} \sim \text{Beta}(i, n-i+1)$ and $(x_{i:n}, x_{j:n}) \sim \text{BivariateBeta}(i, j-i, n-j+1)$. In particular, we have $\mathbb{E}[x_{i:n}] = \frac{i}{n+1}$, $\text{Var}(x_{i:n}) = \frac{i(n-i+1)}{(n+1)^2(n+2)}$, and $\text{Cov}(x_{i:n}, x_{j:n}) = \frac{i(n-j+1)}{(n+1)^2(n+2)}$.*

Proposition 16 allows us to explicitly characterize the noise model. In specific, with recalling (14), we have

$$\frac{1}{n}\sum_{i=1}^{n} \text{Var}(\epsilon_i) = \frac{2}{n}\sum_{i=1}^{n} \text{Var}(x_{i:n}) = \frac{2}{n}\sum_{i=1}^{n} \frac{i(n-i+1)}{(n+1)^2(n+2)} = \frac{1}{3(n+1)}.$$



(*a*) $\mathcal{N}(0,1)$ to $\mathrm{Uniform}(0,1)$ (*b*) $\mathrm{Uniform}(0,1)$ to $\mathrm{Uniform}(0,1)$

Figure 1: **Learning the optimal transport maps with random feature models in 1D.** We plot the computed theoretical and experimental population risks. For the meaning of the curves and scatter plots, see the discussions in Appendix E.

For the latent distribution ξ, we consider two options, $\mathcal{N}(0,1)$ and $\mathrm{Uniform}(0,1)$. As F_ν is the identity function, when ξ is a standard Gaussian, by Theorem 3 the optimal transport map is $G^*(x) = F_\xi(x) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}(\frac{x}{\sqrt{2}})$. Meanwhile, when ξ is also $\mathrm{Uniform}(0,1)$, we have $\xi = \nu$, so the identity function itself is clearly a solution of (5), hence it is the optimal transport map.

The OT maps are learned by random feature models with $g(\boldsymbol{w}, z) = \sin(wz + b)$, where $\boldsymbol{w}$ denotes the pair (w, b) of random samples $w \sim \mathcal{N}(0, 4^2)$ and $b \sim \mathcal{N}(0,1)$.

Figure 1 shows the plots of the results. The curves drawn with solid lines, labeled theory, are the plots of the estimated population risk $\mathcal{E}_{\text{te}}$, computed by (9) and (30). The scatter plots with vertical error bars, labeled experiment, represent the "true" population risks computed by 200 iterations of Monte Carlo integration. The circle markers are the means, and the error bars range from the 2.5th percentile to the 97.5th percentile of the results from those 200 iterations.

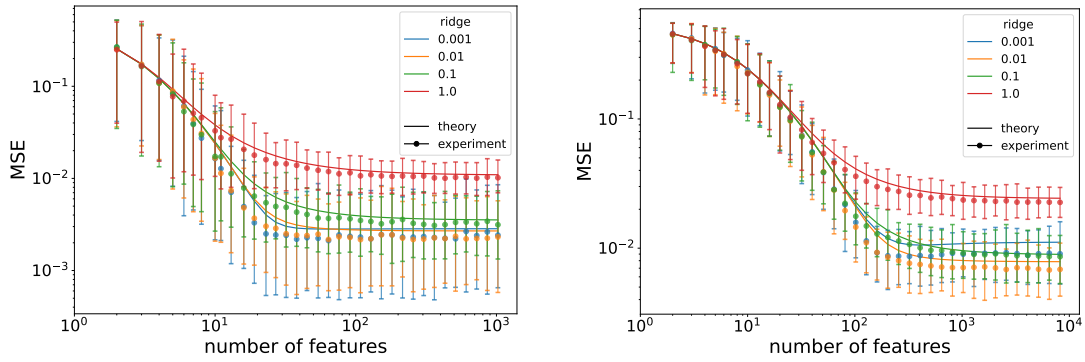
As the captions indicate, Figure 1(*a*) shows the results when $\xi = \mathcal{N}(0,1)$, and Figure 1(*b*) shows the results when $\xi = \mathrm{Uniform}(0,1)$. In both experiments, we set $n = 64$.

From the plots, we can observe that $\mathcal{E}_{\text{te}}$ reasonably approximates the population risk $\mathbb{E}_{\mathcal{D}}\big[\mathcal{R}(\hat{\boldsymbol{\beta}})\big]$, concurrently capturing the overall tendency of decreasing with respect to the number of features. These results demonstrate that $\mathcal{E}_{\text{te}}$ serves as a reliable proxy for the expected population risk, thereby empirically supporting the results regarding "using more features is better" derived via $\mathcal{E}_{\text{te}}$.

E.2. Random feature models in higher dimensions

As an empirical validation of our extension of the theories to the setting of multidimensional targets, we consider the setting where both the latent and the target distributions are multivariate Gaussians, $\xi = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\nu = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$.

One advantage of working with Gaussians is that we have an explicit formula for the OT map, allowing us to seamlessly apply our theories.



(*a*) Between 2D Gaussians (*b*) Between 3D Gaussians

Figure 2: **Learning the optimal transport maps between multivariate Gaussians with random feature models.** We plot the computed theoretical and experimental population risks. The meanings of the curves and scatter plots are the same as Figure 1.

Theorem 17 (29, Remark 2.31) *For $\xi = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\nu = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ with $\boldsymbol{\Sigma}_0$ invertible, the optimal transport map G^* from ξ to ν is an affine function of the form*

$$G^*(\boldsymbol{z}) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_0^{-1/2}\left(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2}\right)^{1/2}\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu}_0).$$

As an instantiation of the manifold hypothesis, we set $\boldsymbol{\Sigma}_1$ to be degenerate, so that ν is supported on an affine subspace whose dimensionality is lower than its ambient space.

The noise part, that is, the term $\check{\sigma}^2$ in the formula (30) of $\mathcal{E}_{\text{te}}$, is estimated in a way described in Section 3.2. More precisely, given samples $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n \overset{\text{i.i.d.}}{\sim} \xi$ and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \nu$, let $\tilde{\pi} \in S_n$ be the permutation of $\{1, \dots, n\}$ such that the optimal transport map from $\hat{\xi}$ to $\hat{\nu}$ is $\boldsymbol{z}_i \mapsto \boldsymbol{x}_{\tilde{\pi}(i)}$, and for such $\tilde{\pi}$ we set

$$\check{\sigma}^2 = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|G^*(\boldsymbol{z}_i) - \boldsymbol{x}_{\tilde{\pi}(i)}\right\|^2\right].$$

In the experiments, the values of $\check{\sigma}^2$ are approximated by Monte Carlo integration using fresh samples that are not used in the regression phase. While we do not have a theoretical proof that Assumption M1 holds in this setting, the results of the Monte Carlo integration serve as empirical evidence supporting its validity.

For the function $\boldsymbol{g}$ determining the (random) feature map $\boldsymbol{\psi}$ in (19), we chose

$$\boldsymbol{g}(\boldsymbol{w}, \boldsymbol{x}) = \begin{bmatrix} \sin(\boldsymbol{\omega}_1^\top \boldsymbol{x} + b_1) \\ \vdots \\ \sin(\boldsymbol{\omega}_p^\top \boldsymbol{x} + b_p) \end{bmatrix} \tag{37}$$

for $\boldsymbol{w}$ here denoting the collection of all the weights $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_p \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, 4\boldsymbol{I})$ and all the biases $b_1, \dots, b_p \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. We remark that this is essentially ρ being a Gaussian distribution on $\mathbb{R}^{dp+p}$.

Figure 2(*a*) is a result when the ambient space is $\mathbb{R}^2$, with

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad \boldsymbol{\Sigma}_0^{1/2} = \begin{bmatrix} 0.16 & 0 \\ 0 & 0.08 \end{bmatrix},$$
$$\boldsymbol{\mu}_1 = \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1^{1/2} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 0.16 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}^\top.$$

In this case, we set $n = 64$. In each of the Monte Carlo integrations to compute the "true" population risks, 256 samples were used.

Figure 2(*b*) is a result when the ambient space is $\mathbb{R}^3$, with

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}, \quad \boldsymbol{\Sigma}_0^{1/2} = \frac{1}{4} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0 & 0.5 \end{bmatrix},$$
$$\boldsymbol{\mu}_1 = \begin{bmatrix} -0.3 \\ -0.4 \\ -0.5 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1^{1/2} = \frac{1}{4} \boldsymbol{R} \begin{bmatrix} 0.75 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{R}^\top$$

where $\boldsymbol{R}$ is a rotation matrix

$$\boldsymbol{R} = \begin{bmatrix} 1/2 & 0 & -\sqrt{3}/2 \\ 0 & 1 & 0 \\ \sqrt{3}/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

In this case, we set $n = 128$. In each of the Monte Carlo integrations to compute the "true" population risks, 1024 samples were used.

Both results in Figure 2 show that $\mathcal{E}_{\text{te}}$ is a fairly accurate approximation of the population risk $\mathbb{E}_{\mathcal{D}}\big[\mathcal{R}(\hat{\boldsymbol{\beta}})\big]$. This empirically validates the discussions we made in Appendix C.

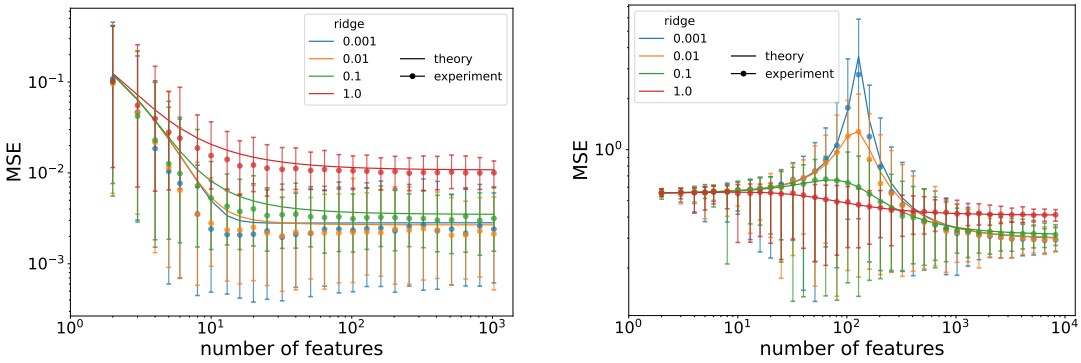
E.3. Neural-network-like model experiments

We also experimented with neural-network-like models, testing the potency of the formula (33) for $\mathcal{E}_{\text{te}}$ as the estimator of the population risk.

To exhibit a comparison to the previous section, we performed an experiment in two dimensions, using the same target and latent distributions as in the two-dimensional experiment in Appendix E.2. Meanwhile, to provide insight into more practical settings involving modern generative models where the neural networks are trained with the standard Gaussian latent distribution, we conducted the three-dimensional experiment with the same target distribution as in Appendix E.2 but changed the latent distribution to $\xi = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

The same $\boldsymbol{g}$ is used as in (37), but the random feature map follows the configuration of (31). All other details remain identical to Appendix E.2.

Figure 3 shows the plots of the results. The curves and the scatter plots are drawn in the exact same way as in the previous experiments. Again, we can observe an agreement between actual experimental results and the theoretical estimation by $\mathcal{E}_{\text{te}}$, this time computed using (33).



(*a*) Between 2D Gaussians (*b*) Between 3D Gaussians

Figure 3: **Learning the optimal transport maps with neural-network-like models.** We plot the computed theoretical and experimental population risks. The meanings of the curves and scatter plots are the same as Figures 1 and 2.