AUGMENTATION ALONE LEADS TO GENERALIZATION

Runtian Zhai, Bingbin Liu, Andrej Risteski, Zico Kolter, Pradeep Ravikumar
School of Computer Science, Carnegie Mellon University
{rzhai,bingbinl,aristesk,zkolter,pradeepr}@cs.cmu.edu

ABSTRACT

We study self-supervised representation learning with data augmentation, such as contrastive learning and masked image/language modeling. Our main result is that a sufficiently good data augmentation technique alone can lead to good generalization, for which we prove generalization bounds for an arbitrary encoder with a model-free analysis. Our results model the upstream stage as RKHS approximation and the downstream stage as RKHS regression, where the RKHS is fully determined by the augmentation. We identify *augmentation complexity* as a key ingredient that replaces the model complexity and additionally use it to quantitatively analyze augmentations on real datasets. For the full paper, see Zhai et al. (2024).

1 A MODEL-FREE APPROACH TO WHY FOUNDATION MODELS GENERALIZE

One of the most important and classic open problems in machine learning is why big models generalize. However, long before the advent of foundation models, classical generalization bounds have been well-known to be vacuous in deep learning. Yet, generalization guarantees remain relevant, perhaps even more so as we look for reliable and responsible deployment on test data. Over the years, there have been a number of hypotheses about *what factor helps big models generalize*, such as spectral normalization (Bartlett et al., 2017; Neyshabur et al., 2018), model overparameterization (Du & Lee, 2018; Arora et al., 2019b), linearization or kernalization (Jacot et al., 2018; Lee et al., 2019), interpolation induced benign overfitting (Belkin et al., 2018; Bartlett et al., 2020), and the implicit bias of optimization, especially GD or SGD (Arora et al., 2019a; Damian et al., 2022).

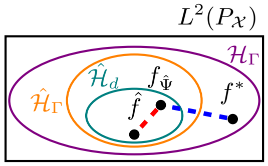
However, there are two major caveats. First, some assumptions in these papers have been reported to be at odds with empirical observations (Nagarajan & Kolter, 2019; Chizat et al., 2019), thereby questioning the relevance of their results in practice. Second, most of these works either require or suggest constraining the model, such as simplified architectures, lazy training, bounded norms, freezing a layer at training, etc. This seems to contradict with the modern practitioners' guideline that bigger and more complex models come with better generalization.

In contrast, we use a model-free approach to show that *a good data augmentation alone can lead to good generalization*. We present two sets of results, the first permitting an arbitrary encoder, and the second focusing on a near-optimal encoder. By decoupling the effect of the model and the augmentation, our approach allows us to *better understand the role of data augmentation in self-supervised learning* without worrying about the complexity of foundation models. A limitation, however, is that we cannot leverage the model inductive bias, which is also important for generalization. Consequently, our generalization bounds are still far from being realistic, but we believe that this work can shed light on how data augmentation contributes to pretraining a good foundation model.

2 THE AUGMENTATION INDUCED RKHS AND THE ISOMETRY PROPERTY

Let $\mathcal{X} \subset \mathbb{R}^{d_\mathcal{X}}$ denote the data and $P_\mathcal{X}$ the distribution. Let $f^* \in L^2(P_\mathcal{X})$ be the target function. Denote $\langle f_1, f_2 \rangle_{P_\mathcal{X}} = \int f_1 f_2 dP_\mathcal{X}$, and $\|f\|_{P_\mathcal{X}}^2 = \langle f, f \rangle_{P_\mathcal{X}}$. Our task is the regression problem:

Problem. *Given unlabeled samples $x_1, \cdots, x_N$ and labeled samples $\tilde{x}_1, \cdots, \tilde{x}_n$ i.i.d. sampled from $P_\mathcal{X}$, and labels $\tilde{y}_k = f^*(\tilde{x}_k) + \nu_k$ for $k \in [n]$ and random noise ν_k, find a predictor $\hat{f} \in L^2(P_\mathcal{X})$ with a low prediction error* $\text{err}(\hat{f}, f^*) := \|\hat{f} - f^*\|_{P_\mathcal{X}}^2 = \mathbb{E}_{P_\mathcal{X}}[(\hat{f}(X) - f^*(X))^2]$.



$L^2(P_{\mathcal{X}})$: Large space of L^2 bounded functions, containing the target function f^*
Induced RKHS $\mathcal{H}_\Gamma$: Functions with soft invariance to augmentation, $f^* \in \mathcal{H}_\Gamma$
Empirical RKHS $\hat{\mathcal{H}}_\Gamma$: Obtained with N samples to approximate inaccessible $\mathcal{H}_\Gamma$
Empirical top-d RKHS $\hat{\mathcal{H}}_d$: Induced by the pretrained d-dimensional encoder
Approximation error $\|f^* - f_{\hat{\Psi}}\|$: Distance from f^* to $\hat{\mathcal{H}}_d$
Estimation error $\|f_{\hat{\Psi}} - \hat{f}\|$: Entailed by downstream regression

Figure 1: Overall RKHS approximation/regression framework illustration and commentary.

We study how data augmentation helps with self-supervised pretraining of a good encoder. Let $\mathcal{A}$ be the space of augmented samples, and $P_{\mathcal{AX}}$ be a joint distribution with marginals $P_{\mathcal{A}}$ and $P_{\mathcal{X}}$. Define the *augmentation operator* $\Gamma = \Gamma_{x\to a} : L^2(P_{\mathcal{X}}) \to L^2(P_{\mathcal{A}})$ as $(\Gamma_{x\to a} f)(a) = \mathbb{E}[f(X)|a]$. Denote its adjoint by $\Gamma^* = \Gamma_{a\to x} : L^2(P_{\mathcal{A}}) \to L^2(P_{\mathcal{X}})$ with $(\Gamma_{a\to x} g)(x) = \mathbb{E}[g(A)|x]$, such that $\langle \Gamma_{x\to a} f, g\rangle_{P_{\mathcal{A}}} = \iint f(x)g(a)p(a,x)dadx = \langle f, \Gamma_{a\to x} g\rangle_{P_{\mathcal{X}}}$ for all f, g.

Example: Consider BERT with 15% random masking. Then, $\mathcal{X}$ is the space of original sentences, and $\mathcal{A}$ is the space of 15% masked sentences; $P_{\mathcal{X}}$ is the distribution over original sentences, $A \sim p(\cdot|x)$ is the 15% randomly masked version of an original sentence x, and $P_{\mathcal{AX}}(a, x) = P_{\mathcal{X}}(x)p(a|x)$. Thus, $(\Gamma_{a\to x} g)(x)$ is essentially the mean of g over all 15% randomly masked sentences of x.

For $\epsilon > 0$, we say f^* is ϵ-*coherent* with augmentation Γ, if $\exists g^* \in L^2(P_{\mathcal{A}})$, such that $f^* = \Gamma_{a\to x} g^* = E[g^*(A)|\cdot]$ and $\frac{1}{2}\mathbb{E}_{X\sim P_{\mathcal{X}}}\mathbb{E}_{A,A'\sim p(\cdot|X)}\left[(g^*(A) - g^*(A'))^2\right] \le \epsilon \|g^*\|^2_{P_{\mathcal{A}}}$. It has an additional $\|g^*\|^2_{P_{\mathcal{A}}}$ term on the right compared to Assumption 1.1 in Johnson et al. (2023), so it is homogeneous. We assume $f^* \in \mathcal{F}_B(\Gamma;\epsilon)$, where $\mathcal{F}_B(\Gamma;\epsilon)$ contains all f that are ϵ-coherent and satisfy $\|f\|_{P_{\mathcal{X}}} \le B$. This condition can be shown to be equivalent to the *isometry property*:

$$(1-\epsilon)\|f^*\|^2_{\mathcal{H}_\Gamma} \le \|f^*\|^2_{P_{\mathcal{X}}} \le \|f^*\|^2_{\mathcal{H}_\Gamma}, \tag{1}$$

where $\mathcal{H}_\Gamma$ is the *(augmentation) induced RKHS*, which depends on the augmentation only and nothing else. To define $\mathcal{H}_\Gamma$, let the positive-pair kernel K_A on $\mathcal{A}\times\mathcal{A}$ (Johnson et al., 2023) be

$$K_A(a_1, a_2) := \frac{dP_A^+}{d(P_{\mathcal{A}} \otimes P_{\mathcal{A}})} = \frac{P_A^+(a_1, a_2)}{P_{\mathcal{A}}(a_1)P_{\mathcal{A}}(a_2)}, \; P_A^+(a_1, a_2) := \int p(a_1|x)p(a_2|x)dP_{\mathcal{X}}(x),$$

which uses the augmentation graph (HaoChen et al., 2021). Then, define a *dual kernel* on $\mathcal{X}\times\mathcal{X}$ as

$$K_X(x_1, x_2) := \frac{dP_X^+}{d(P_{\mathcal{X}} \otimes P_{\mathcal{X}})} = \frac{P_X^+(x_1, x_2)}{P_{\mathcal{X}}(x_1)P_{\mathcal{X}}(x_2)} = \int \frac{p(a|x_1)p(a|x_2)}{P_{\mathcal{A}}(a)} da.$$

In fact, $(\Gamma\Gamma^* g)(a) = \int K_A(a, a')g(a')P_{\mathcal{A}}(a')da'$, *i.e.* $\Gamma\Gamma^*$ is the integral operator of K_A. Likewise, $\Gamma^*\Gamma$ is the integral operator of K_X; and $\boldsymbol{\mathcal{H}_\Gamma}$ **is defined as the RKHS associated with** $\boldsymbol{K_X}$. Let $\psi_1, \psi_2, \cdots$ be eigenfunctions of $\Gamma^*\Gamma$ with decreasing eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$, such that $\Gamma^*\Gamma\psi_i = \lambda_i\psi_i$. Suppose $\int K_X(x, x')^2 dP_{\mathcal{X}}(x)dP_{\mathcal{X}}(x') < \infty$. By Hilbert-Schmidt theorem, we can choose $\psi_1, \psi_2, \cdots$ that form an orthonormal basis of $L^2(P_{\mathcal{X}})$, such that $\langle \psi_i, \psi_j\rangle_{P_{\mathcal{X}}} = \delta_{i,j}$, and any $f \in L^2(P_{\mathcal{X}})$ can be written as $f = \sum_i u_i\psi_i$ for some u_i. Then, we can show the following properties:

(i) Operators $\Gamma\Gamma^*$ and $\Gamma^*\Gamma$ share the same non-zero eigenvalues, and there exist eigenfunctions $\{\phi_i\}$ of $\Gamma\Gamma^*$ that form an orthonormal basis of $L^2(P_{\mathcal{A}})$, such that for any $\lambda_i > 0$,

$$\psi_i = \lambda_i^{-1/2}\Gamma^*\phi_i = \lambda_i^{-1/2}\Gamma_{a\to x}\phi_i \quad \textit{and} \quad \phi_i = \lambda_i^{-1/2}\Gamma\psi_i = \lambda_i^{-1/2}\Gamma_{x\to a}\psi_i.$$

(ii) Range $R(\Gamma^*) = \left\{f = \Gamma_{a\to x} g \;\middle|\; g \in L^2(P_{\mathcal{A}})\right\}$ is the induced RKHS $\mathcal{H}_\Gamma$ associated with K_X.

With these, we can show that ϵ-coherence is the same as the isometry property Eqn. (1), which essentially says that $\Gamma^*\Gamma$ preserves most variance of target function f^*. Thus, the optimal d-dimensional encoder should keep the most variance, which we will show consists of the top-d eigenfunctions. This is analogous to PCA for a finite-dimensional vector space, where the top-d eigenvectors of a linear transformation keeps the most variance.

3 Generalization Bounds

Our general proof framework is illustrated in Figure 1. The upstream stage pretrains a d-dimensional encoder $\hat{\Psi}$, which we model as learning a d-dimensional subspace $\hat{\mathcal{H}}_d$ that approximates the induced

RKHS $\mathcal{H}_\Gamma$ and incurs an approximation error. The downstream stage fits a linear layer (linear probe) on top of the encoder, which we model as RKHS regression on $\hat{\mathcal{H}}_d$ and entails an estimation error. By Eqn. (1) and $\|f^*\|_{P_\mathcal{X}} \le B$, we have $\|f^*\|_{\mathcal{H}_\Gamma} \le \frac{B}{\sqrt{1-\epsilon}}$. Given $\hat{\Psi}$, we use the following predictor:

$$\hat{f} := \underset{f: f = w^\top \hat{\Psi} \in \hat{\mathcal{H}}_d, \|f\|_{\mathcal{H}_\Gamma} \le \frac{B}{\sqrt{1-\epsilon}}}{\arg\min} \left\{ \frac{1}{n} \sum_{k=1}^n (\tilde{y}_k - f(\tilde{x}_k))^2 \right\}, \tag{2}$$

where $\hat{\mathcal{H}}_d$ is the linear span of $\hat{\Psi} = [\hat{\psi}_1, \cdots, \hat{\psi}_d]$. In practice though, $\hat{\Psi}$ is likely not directly obtained from pretraining, as people often first pretrain an encoder $\hat{\Phi} = [\hat{\phi}_1, \cdots, \hat{\phi}_d]$ on $\mathcal{A}$, and then convert it into $\hat{\Psi}$. For example, BERT is pretrained on masked sentences but used on unmasked ones. While in practice the pretrained encoder is usually directly applied to downstream, theoretical analyses require explicitly writing out the relationship between $\hat{\Phi}$ and $\hat{\Psi}$. We use the *average encoder*

$$\hat{\Psi}(x) = \mathbb{E}[\hat{\Phi}(A)|x] = \int \hat{\Phi}(a) p(a|x) da, \tag{3}$$

which is equivalent to $\hat{\Psi} = \Gamma^* \hat{\Phi}$, and thus $\hat{\psi}_i \in R(\Gamma^*) = \mathcal{H}_\Gamma$ for all $i \in [d]$. The average encoder has been widely studied in prior art, such as Saunshi et al. (2022, Eqn. (4)). We now derive generalization bounds for two cases: (i) $\hat{\Phi}$ is an arbitrary function; (ii) $\hat{\Phi}$ is a near optimal d-dimensional encoder.

3.1 CASE I: ARBITRARY ENCODER

There are two critical ingredients: (i) Augmentation complexity $\kappa := \|K_X\|_\infty^{1/2}$, which replaces the model complexity in our bounds and make them model-free; (ii) Trace gap τ^2, which is smaller for "better" $\hat{\Phi}$; see definitions in Appendix A. Denote $S_\lambda(d) := \lambda_1 + \cdots + \lambda_d$. Then, we have:

Theorem 1. *Let $\nu_1, \cdots, \nu_n$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ variates, and $\hat{f}$ be given by Eqn. (2). If $\hat{\Phi}$ has d dimensions (d can be ∞) and $\tau < 1$, then there are universal constants c_0, c_1, c_2 such that with probability at least $1 - c_1 \exp\left(-\frac{c_2\sqrt{2nS_\lambda(d+1)}}{\kappa}\right) - \exp\left(-\sqrt{\frac{2n\kappa^2B^2}{1-\epsilon}}\right)$, there is*

$$\|\hat{f} - f^*\|_{P_\mathcal{X}}^2 \le \frac{9\tau^2(\tau+\epsilon)B^2}{(1-\tau^2)(1-\epsilon)} + \frac{c_0\kappa(B^2+\sigma B)}{1-\epsilon}\sqrt{\frac{S_\lambda(d+1)}{n}} \quad \textit{for all } f^* \in \mathcal{F}_B(\Gamma;\epsilon). \tag{4}$$

Note that this bound does not constrain the form and dimension that $\hat{\Phi}$ takes. The first term in the bound controls the approximation error, and the second controls the estimation error. While the second term vanishes as the number of unlabeled and labeled samples $N, n \to \infty$, the first term may not: With d output dimensions, if $\lambda_{d+1} > 0$, then the first term won't vanish since $\tau^2 \ge \lambda_{d+1}$. This could happen, for example, when d is so small that $\hat{\Phi}$ doesn't have enough capacity to represent f^*.

3.2 CASE II: NEAR OPTIMAL D-DIMENSIONAL ENCODER

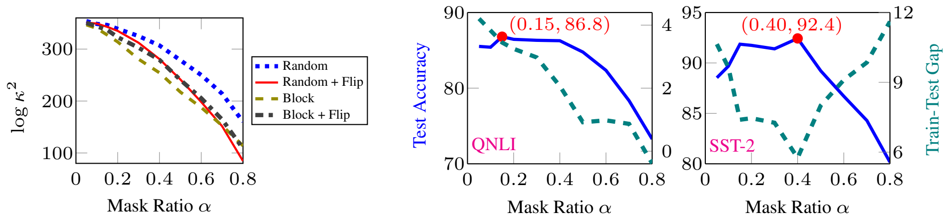
We define the optimal encoder in a minimax sense. It minimizes the *worst-case approximation error* over $\mathcal{F}_B(\Gamma;\epsilon)$, defined as $\mathrm{err}(\hat{\Psi}; \mathcal{F}_B(\Gamma;\epsilon)) := \sup_{f \in \mathcal{F}_B(\Gamma;\epsilon)} \min_{w \in \mathbb{R}^d} \|w^\top \hat{\Psi} - f\|_{P_\mathcal{X}}^2$. We now show that $\hat{\Psi}$ is optimal if it spans the top-d eigenspace, *i.e.* the linear span of $\psi_1, \cdots, \psi_d$:

Proposition 1 (Approximation error, lower bound)**.** *For any $\hat{\Psi} = [\hat{\psi}_1, \cdots, \hat{\psi}_d]$ where $\hat{\psi}_i \in L^2(P_\mathcal{X})$,*

$$\mathrm{err}(\hat{\Psi}; \mathcal{F}_B(\Gamma;\epsilon)) \ge \frac{\lambda_{d+1}}{1-\lambda_{d+1}} \frac{\epsilon}{1-\epsilon} B^2 \quad \textit{given that} \quad \frac{\lambda_{d+1}}{1-\lambda_{d+1}} \frac{\epsilon}{1-\epsilon} \le \frac{1}{2}. \tag{5}$$

To attain equality, it is sufficient for $\hat{\Psi}$ to span the top-d eigenspace, and also necessary if $\lambda_{d+1} < \lambda_d$.

The optimal d-dimensional $\hat{\Psi}$ achieves the smallest trace gap $\tau^2 = \lambda_{d+1}$. We consider its Monte-Carlo approximation as we only have access to finite samples. Given unlabeled samples $x_1, \cdots, x_N$, we define the empirical augmentation operator as $(\bar{\Gamma} f)(a) = \frac{1}{N}\sum_{k=1}^N \frac{f(x_k)p(a|x_k)}{\hat{P}_\mathcal{A}(a)}$, where $\hat{P}_\mathcal{A}(a) =$



(a) Estimated $\log \kappa^2$.

(b) Results on QNLI and SST-2.

Figure 2: Plots for Section 4. In (a), $\log \kappa^2$ is estimated on `wikipedia-simple`.

$\frac{1}{N}\sum_{k=1}^{N} p(a|x_k)$. The adjoint of $\bar{\Gamma}$ is still Γ^*. Let $\{(\bar{\lambda}_i, \bar{\psi}_i)\}$ be the eigenvalues and eigenfunctions of $\Gamma^*\bar{\Gamma}$, and $\bar{\phi}_i$ the eigenfunctions of $\bar{\Gamma}\Gamma^*$. We consider the empirical top-d eigenfunctions $[\bar{\phi}_1, \cdots, \bar{\phi}_d]$, which is a Monte-Carlo approximation of the real top-d eigenfunctions. We have:

Theorem 2. *Let $\hat{\phi}_i = \bar{\phi}_i$ for $i \in [d]$. Define covariance matrix $\boldsymbol{G}$ as $\boldsymbol{G}(i,j) = \langle \hat{\phi}_i, \hat{\phi}_j \rangle_{P_{\mathcal{A}}}$ for $i, j \in [d]$. Let $\gamma_{\boldsymbol{G}} := \lambda_{\max}(\boldsymbol{G})/\lambda_{\min}(\boldsymbol{G})$ be the condition number of $\boldsymbol{G}$. Then, for any $\delta > 0$, it holds with probability at least $1 - \delta$ that*

$$\tau^2 \le \lambda_{d+1} + \left(2 + \sqrt{2\log\frac{2}{\delta}}\right) \frac{(\lambda_d^{-1} + \bar{\lambda}_d^{-1}\gamma_{\boldsymbol{G}}^{1/2} + 2)\kappa^2}{\sqrt{N}} d.$$

Combining Theorem 1 and 2 leads to the bound for this near optimal encoder. We can see that this bound is near tight by comparing the upper bound in Theorem 2 to the lower bound in Proposition 1; the only difference is $\frac{\tau+\epsilon}{1-\epsilon}$ instead of $\frac{\epsilon}{1-\epsilon}$ in Eqn. (4). Note also that τ^2 can be arbitrarily close to λ_{d+1}, and Theorem 2 does not require a gap between λ_d and λ_{d+1} unlike prior work.

4 Estimating and Exploiting the Augmentation Complexity

In our model-free bounds, the augmentation complexity κ completely replaces the model complexity. In fact, κ can be a practical tool for analyzing augmentations. As a demonstration, in Figure 2a, we plot the size of κ for four types of random masking augmentations *w.r.t.* different mask ratios on `wikipedia-simple`. Our bounds suggest that a smaller κ leads to good generalization, and one natural way to reduce κ is via a stronger augmentation, which has indeed been helpful in practice (Chen et al., 2020; Wettig et al., 2023).

We also study how the mask ratio α affects the downstream performance using QNLI (Wang et al., 2018) and SST-2 (Socher et al., 2013). For pretraining, we train `roberta-large` models with random masking, using different mask ratios following the fast pretraining recipe in Wettig et al. (2023). For downstream, we fine-tune the encoder together with the linear head following common practice. We use the average encoder (Eqn. (3)) estimated by sampling 16 augmentations a per x.

We evaluate the train/test accuracies of the models, and plot the test accuracy (blue solid) and the train-test accuracy gap (green dashed) in Figure 2b. The highest test accuracy is achieved at $\alpha = 0.15$ on QNLI and at $\alpha = 0.40$ on SST-2 (marked in red). The test accuracy is low when α is too small due to the large generalization gap, and also low when α is too large due to low training accuracy. Regarding the train-test gap, QNLI shows a monotonic decrease in the gap as the mask ratio grows, but the gap on SST-2 is U-shaped, with the lowest point at $\alpha = 0.40$. This is likely because with $\alpha > 0.40$ is too strong an augmentation for SST-2 that breaks the isometry property, in which case our theoretical results will not hold. Thus, these results align with our theory that while augmentations should be sufficiently robust, they must not be so strong that breaks isometry property. This suggests the presence of a "sweet spot", which is also supported by evidence in prior work (Tian et al., 2020).

Discussions: In this work, we showed that a sufficiently good augmentation alone leads to good generalization. However, "sufficiently good" is a strong constraint hardly realizable in practice, hence our bounds are yet to be made more practical. Indeed, in Figure 2a, $\log \kappa^2$ can be as large as 300. We suspect this to be a manifestation of the typical curse of dimensionality in high-dimensional statistics in the absence of strong inductive bias (Bengio et al., 2013). Moreover, we postulate that even though our worst-case bounds come with an exponential dependency on data dimension, the empirical success of existing augmentation-based self-supervised learning suggest that they implicitly adapt to the inherent low-dimensional manifold structure in real-world data. We conjecture that the curse can be evaded if the augmentation captures such a low-dimensional structure.

References

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019a.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019b.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

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.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pp. 5413–5452. PMLR, 2022.

Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2018.

Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Daniel D. Johnson, Ayoub El Hanchi, and Chris J. Maddison. Contrastive learning can find an optimal basis for approximately view-invariant functions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=AjC0KBjiMu.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pp. 19250–19286. PMLR, 2022.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *BLACKBOXNLP@EMNLP*, 2018. doi: 10.18653/v1/W18-5446.

Huan Wang, Shuicheng Yan, Dong Xu, Xiaoou Tang, and Thomas Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2007.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2023.

Runtian Zhai, Bingbin Liu, Andrej Risteski, Zico Kolter, and Pradeep Ravikumar. Understanding augmentation-based self-supervised representation learning via rkhs approximation and regression. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ax2yRhCQr1.

A AUGMENTATION COMPLEXITY AND TRACE GAP

Definition 1. *Define the* ***augmentation complexity*** *as* $\kappa := \|K_X\|_\infty^{1/2}$*, i.e. for* $P_\mathcal{X}$*-almost all* x*,*

$$K_X(x, x) = \sum_i \lambda_i \psi_i(x)^2 = \int \frac{p(a|x)^2}{P_\mathcal{A}(a)} da = D_{\chi^2}(P_\mathcal{A}(\cdot|x) \parallel P_\mathcal{A}) + 1 \le \kappa^2.$$

Here, $D_{\chi^2}(P \parallel Q) := \int (\frac{dP}{dQ} - 1)^2 dQ$ is the χ^2-divergence. It is non-negative, so $\kappa \ge 1$. Next, to define the trace gap, we first define the *ratio trace* for a given encoder $\hat{\Phi}$.

Definition 2. *Define covariance matrices* $\boldsymbol{F}, \boldsymbol{G}$ *as* $\boldsymbol{F}(i,j) = \langle \hat{\psi}_i, \hat{\psi}_j \rangle_{P_\mathcal{X}} = \langle \Gamma^* \hat{\phi}_i, \Gamma^* \hat{\phi}_j \rangle_{P_\mathcal{X}}$ *and* $\boldsymbol{G}(i,j) = \langle \hat{\phi}_i, \hat{\phi}_j \rangle_{P_\mathcal{A}}$*. Then, the* ***ratio trace*** *is defined as* $\mathrm{Tr}(\boldsymbol{G}^{-1}\boldsymbol{F})$*, if* $\boldsymbol{G}^{-1}$ *is well-defined.*

Ratio trace is a classical quantity in linear discriminant analysis (LDA) (Wang et al., 2007) and, as we will show, controls the approximation error. The largest ratio trace of any d-dimensional $\hat{\Phi}$ is $\lambda_1 + \cdots + \lambda_d$, and can be achieved by the top-d eigenspace of $\mathcal{H}_\Gamma$. Then, define the *learned kernel* as

$$\hat{K}_{\hat{\Psi}}(x, x') = \langle \Gamma^*(\boldsymbol{G}^{-1/2}\hat{\Phi})(x), \Gamma^*(\boldsymbol{G}^{-1/2}\hat{\Phi})(x') \rangle,$$

which is the reproducing kernel of $\mathcal{H}_{\hat{\Psi}} = \text{span}(\hat{\psi}_1, \hat{\psi}_2, \cdots)$, a subspace of $\mathcal{H}_\Gamma$. Here $\boldsymbol{G}^{-1/2}$ is used for normalization. The ratio trace can be viewed as the trace of $\mathcal{H}_{\hat{\Psi}}$. Then, define the ***trace gap*** as:

$$\tau^2 := \inf_{d' \le d} \inf_{h_1, \cdots, h_{d'}} S_\lambda(d'+1) - \text{Tr}(\boldsymbol{G}_h^{-1}\boldsymbol{F}_h),$$

where $\tau \ge 0$, $h_i = w_i^\top \hat{\Phi}$, $\boldsymbol{G}_h = (\langle h_i, h_j \rangle_{P_\mathcal{A}})_{i,j \in [d']}$, and $\boldsymbol{F}_h = (\langle \Gamma^* h_i, \Gamma^* h_j \rangle_{P_\mathcal{X}})_{i,j \in [d']}$. Note that for any $d' \le d$ there is $\text{Tr}(\boldsymbol{G}_h^{-1}\boldsymbol{F}_h) \le S_\lambda(d')$, so τ^2 is always lower bounded by λ_{d+1}. And by choosing $h_i = \hat{\phi}_i$ for $i \in [d]$, we can see that $\tau^2 \le S_\lambda(d+1) - \text{Tr}(\boldsymbol{G}^{-1}\boldsymbol{F})$.

B Proof of the Isometry Property

$\Gamma^*\Gamma$ and $\Gamma\Gamma^*$ are integral operators.

$$\begin{cases} (\Gamma_{a\to x}\Gamma_{x\to a} f)(x) = (\Gamma^*\Gamma f)(x) = \int K_X(x,x') f(x') p(x') dx'; \\ (\Gamma_{x\to a}\Gamma_{a\to x} g)(a) = (\Gamma\Gamma^* g)(a) = \int K_A(a,a') g(a') p(a') da'. \end{cases} \tag{6}$$

Proof. We only show the first equation, and the second one can be proved in the same way.

$$\begin{aligned} (\Gamma^*\Gamma f)(x) &= \Gamma^* \left(\int f(x') p(x'|a) dx' \right) = \int \left(\int f(x') p(x'|a) dx' \right) p(a|x) da \\ &= \iint f(x') p(a|x) p(x'|a) da dx' = \iint f(x') \frac{p(a|x)p(a|x')}{p(a)} p(x') da dx' \\ &= \int K_X(x,x') f(x') p(x') dx'. \end{aligned}$$

$\square$

Duality. $\Gamma\Gamma^*$ *shares the same non-zero eigenvalues as* $\Gamma^*\Gamma$*, and there exist eigenfunctions* $\{\phi_i\}$ *of* $\Gamma\Gamma^*$ *that form an orthonormal basis of* $L^2(P_\mathcal{A})$*, such that for any* $\lambda_i > 0$*,*

$$\psi_i = \lambda_i^{-1/2} \Gamma^* \phi_i \quad \textit{and} \quad \phi_i = \lambda_i^{-1/2} \Gamma \psi_i, \tag{7}$$

and we also have the following spectral decomposition of the Radon-Nikodym derivative:

$$\frac{dP_{\mathcal{A}\mathcal{X}}}{d(P_\mathcal{A} \otimes P_\mathcal{X})} = \frac{p(a,x)}{p(a)p(x)} = \sum_i \lambda_i^{1/2} \phi_i(a) \psi_i(x). \tag{8}$$

Proof. Suppose $\lambda_i, \psi_i(x)$ is a pair of eigenvalue and eigenfunction of $\Gamma^*\Gamma$, and $\lambda_i > 0$. Then, we have $\Gamma\Gamma^*\Gamma\psi_i = \lambda_i \Gamma\psi_i$, which means that $\Gamma\psi_i$ is an eigenfunction of $\Gamma\Gamma^*$ with eigenvalue λ_i. The $\lambda_i^{-1/2}$ is used for normalization. To see this, let $\phi_i = \lambda_i^{-1/2}\Gamma\psi_i$. Then, we have

$$\begin{aligned} \langle \phi_i, \phi_j \rangle_{P_\mathcal{A}} &= \lambda_i^{-1/2} \lambda_j^{-1/2} \langle \Gamma\psi_i, \Gamma\psi_j \rangle_{P_\mathcal{A}} \\ &= \lambda_i^{-1/2} \lambda_j^{-1/2} \langle \Gamma^*\Gamma\psi_i, \psi_j \rangle_{P_\mathcal{X}} \\ &= \lambda_i^{-1/2} \lambda_j^{-1/2} \langle \lambda_i \psi_i, \psi_j \rangle_{P_\mathcal{X}} = \delta_{i,j}. \end{aligned}$$

We can prove the reverse direction similarly. And for any fixed x, there is

$$\left\langle \frac{p(a,x)}{p(a)p(x)}, \phi_i \right\rangle_{P_\mathcal{A}} = \int \frac{p(a,x)}{p(a)p(x)} \phi_i(a) p(a) da = \int p(a|x) \phi_i(a) da = \sqrt{\lambda_i} \psi_i(x). \tag{9}$$

which implies Eqn. (8). $\square$

Basic properties of $\mathcal{H}_\Gamma$.

(i) *K_X is the reproducing kernel of $\mathcal{H}_\Gamma$, such that for all $f \in \mathcal{H}_\Gamma$, $f(x) = \langle f, K_X(x, \cdot) \rangle_{\mathcal{H}_\Gamma}$.*

(ii) *$\mathcal{H}_\Gamma = R(\Gamma^*)$.*

(iii) *$\mathcal{H}_\Gamma$ is isometric to $\text{span}(\{\phi_i\}_{\lambda_i > 0})$, a subspace of $L^2(P_\mathcal{A})$, and $\|f\|_{\mathcal{H}_\Gamma} = \inf_{g: f = \Gamma^* g} \|g\|_{P_\mathcal{A}}$.*

(iv) *For any $f^* \in \mathcal{F}_B(\Gamma; \epsilon) \subset R(\Gamma^*)$, let $f^* = \sum_i u_i \psi_i$. Define $g_0 := \sum_i \lambda_i^{-1/2} u_i \phi_i$. Then, we can choose $g^* = g_0$, in which case ϵ-coherence is equivalent to:*

$$\langle g^*, (I - \Gamma\Gamma^*) g^* \rangle_{P_\mathcal{A}} \le \epsilon \|g^*\|_{P_\mathcal{A}}^2 \;\Leftrightarrow\; \sum_i \frac{1 - \lambda_i}{\lambda_i} u_i^2 \le \epsilon \sum_i \frac{1}{\lambda_i} u_i^2, \tag{10}$$

and this is equivalent to Eqn. (1).

Proof. (i) First, note that $\mathcal{H}_\Gamma = \left\{ \sum_{i:\lambda_i > 0} a_i \boldsymbol{e}_i \;\middle|\; \sum_i a_i^2 < \infty \right\}$ where $\boldsymbol{e}_i = \lambda_i^{-1/2} \psi_i$, so it is isomorphic to $\ell^2((a_i)_{i:\lambda_i > 0})$ and is thus a Hilbert space. Then, $K_X(x, x') = \sum_i \lambda_i \psi_i(x) \psi_i(x')$. For any $f \in \mathcal{H}_\Gamma$, let $f = \sum_i u_i \psi_i$, then

$$\langle f(x'), K_X(x, x') \rangle_{\mathcal{H}_\Gamma} = \sum_i \frac{1}{\lambda_i} u_i (\lambda_i \psi_i(x)) = \sum_i u_i \psi_i(x) = f(x).$$

(ii) For any $f = \sum_i u_i \psi_i \in \mathcal{H}_\Gamma$, there is $\sum_i \lambda_i^{-1} u_i^2 < \infty$ by definition. So for any $\lambda_i = 0$, there must be $u_i = 0$. Let $g = \sum_i \lambda_i^{-1/2} u_i \psi_i$. Then, $\|g\|_{P_\mathcal{A}}^2 = \sum_i \lambda_i^{-1} u_i^2 < \infty$, meaning that $g \in L^2(P_\mathcal{A})$. And there is $f = \Gamma^* g$, so $f \in R(\Gamma^*)$, which implies that $\mathcal{H}_\Gamma \subseteq R(\Gamma^*)$. Meanwhile, for any $f = \Gamma^* g \in R(\Gamma^*)$, let $g = \sum_i v_i \phi_i$, then $\sum_i v_i^2 < \infty$. Then, $f = \sum_i \lambda_i^{1/2} v_i \psi_i$ by duality, so $\sum_i \lambda_i^{-1} (\lambda_i^{1/2} v_i)^2 < \infty$, meaning that $f \in \mathcal{H}_\Gamma$, so $R(\Gamma^*) \subseteq \mathcal{H}_\Gamma$.

(iii) For any $f = \sum_i u_i \psi_i \in \mathcal{H}_\Gamma$, let $g = \sum_i \lambda_i^{-1/2} u_i \psi_i$. By the proof of (ii) we know that $f \mapsto g$ is bijective, and $\|f\|_{\mathcal{H}_\Gamma} = \|g\|_{P_\mathcal{A}}$. Moreover, $g \in \text{span}(\{\phi_i\}_{\lambda_i > 0})$.

(iv) Let $g^* = \sum_i v_i \phi_i$. Then, since we have $f^* = \Gamma^* g^* = \sum_i \lambda_i^{1/2} v_i \psi_i$, for any $\lambda_i > 0$, there is $v_i = \lambda^{-1/2} u_i$; and for any $\lambda_i = 0$, there is $u_i = 0$. Let $g^* = g_0 + g_1$, where $g_0 = \sum_i \lambda_i^{-1/2} u_i \phi_i$, and $g_1 \perp g_0$ and $\Gamma^* g_1 = 0$. By duality, $\Gamma\Gamma^* g_0$ belongs to the linear span of $\{\phi_i\}_{\lambda_i > 0}$, so $g_1 \perp \Gamma\Gamma^* g_0$. Later we will show the equivalence between ϵ-coherence and Eqn. (10), which is the random walk normalized Laplacian over the augmentation graph (Chung, 1997, Section 1.2). This is equivalent to $\langle g^*, (I - \Gamma\Gamma^*) g^* \rangle_{P_\mathcal{A}} \le \epsilon \|g^*\|_{P_\mathcal{A}}^2$, which is further equivalent to $\langle g_0, (I - \Gamma\Gamma^*) g_0 \rangle_{P_\mathcal{A}} + \|g_1\|_{P_\mathcal{A}}^2 \le \epsilon (\|g_0\|_{P_\mathcal{A}}^2 + \|g_1\|_{P_\mathcal{A}}^2)$ (note that $\Gamma\Gamma^* g_1 = 0$). This implies that $\langle g_0, (I - \Gamma\Gamma^*) g_0 \rangle_{P_\mathcal{A}} \le \epsilon \|g_0\|_{P_\mathcal{A}}^2$, *i.e.* g_0 satisfies ϵ-coherence. Since we also have $f^* = \Gamma^* g_0$, we can choose $g^* = g_0$.

Next, to show the equivalence to Eqn. (10), We just need to show that $\langle g^*, (I - \Gamma\Gamma^*) g^* \rangle_{P_\mathcal{A}} = \frac{1}{2} \mathbb{E}_{X \sim P_\mathcal{X}} \mathbb{E}_{A, A' \sim p(\cdot|X)} \left[(g^*(A) - g^*(A'))^2 \right]$. And indeed, we have:

$$\begin{aligned}
\langle g^*, (I - \Gamma\Gamma^*) g^* \rangle_{P_\mathcal{A}} &= \left\langle g^*, g^* - \int g^*(a') K_A(\cdot, a') p(a') da' \right\rangle_{P_\mathcal{A}} \\
&= \|g^*\|_{P_\mathcal{A}}^2 - \iint g(a) g(a') \frac{\int p(a|x) p(a'|x) p(x) dx}{p(a) p(a')} p(a') p(a) da da' \\
&= \frac{1}{2} \mathbb{E}[g^*(A)^2] + \frac{1}{2} \mathbb{E}[g^*(A')^2] - \frac{1}{2} \mathbb{E}_{X \sim P_\mathcal{X}} \mathbb{E}_{A, A' \sim p(\cdot|X)} [2 g^*(A) g^*(A')] \\
&= \frac{1}{2} \mathbb{E}_{X \sim P_\mathcal{X}} \mathbb{E}_{A, A' \sim p(\cdot|X)} \left[(g^*(A) - g^*(A'))^2 \right],
\end{aligned}$$

as desired. $\square$

C PROOF OF THEOREM 1

C.1 LOCAL GAUSSIAN COMPLEXITY AND LOCALIZED RADEMACHER COMPLEXITY

We first provide the definition of the two complexities we will use in our analysis. For a function f, let $\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^n f(\tilde{x}_i)^2$ be its mean on the downstream samples.

Definition 3. *(Wainwright, 2019, Eqns. (13.16) & (14.3)) For any $B, \epsilon > 0$, define*

$$\mathcal{F}_0 := \left\{ f_1 - f_2 \;\middle|\; f_i \in \mathcal{H}_{\hat{\Psi}}, \|f_i\|_{\mathcal{H}_\Gamma} \le \frac{B}{\sqrt{1-\epsilon}} \right\} = \left\{ f \in \mathcal{H}_{\hat{\Psi}} \;\middle|\; \|f\|_{\mathcal{H}_\Gamma} \le \frac{2B}{\sqrt{1-\epsilon}} \right\}. \tag{11}$$

Then, the local Gaussian complexity around $f_{\hat{\Psi}}$ at scale $\delta > 0$ is given by

$$\mathcal{G}_n(\delta; \mathcal{F}_0) := \mathop{\mathbb{E}}_{\omega_1,\cdots,\omega_n}\left[\sup_{f\in\mathcal{F}_0, \|f\|_n \le \delta} \left| \frac{1}{n}\sum_{i=1}^n \omega_i f(\tilde{x}_i) \right| \right], \tag{12}$$

where $\omega_1, \cdots, \omega_n$ are i.i.d. $\mathcal{N}(0,1)$ variates. And define

$$\mathcal{F}_* := \left\{ f = f_1 + \alpha f^* \;\middle|\; \alpha \in [-1,1], f_1 \in \mathcal{H}_{\hat{\Psi}}, \|f_1\|_{\mathcal{H}_\Gamma} \le \frac{B}{\sqrt{1-\epsilon}} \right\}. \tag{13}$$

Then, the localized population Rademacher complexity of radius $\delta > 0$ is given by

$$\bar{\mathfrak{R}}_n(\delta; \mathcal{F}_*) := \mathop{\mathbb{E}}_{\sigma_1,\cdots,\sigma_n,x_1,\cdots,x_n}\left[\sup_{f\in\mathcal{F}_*, \|f\|_{P_\mathcal{X}} \le \delta} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \right| \right], \tag{14}$$

where $\sigma_1, \cdots, \sigma_n$ are i.i.d. Rademacher variables taking values in $\{-1,+1\}$ equiprobably.

Our master plan is to apply Theorems 13.13 and 14.1 of Wainwright (2019) to $f_{\hat{\Psi}} = \Gamma^*(\Pi_{\hat{\Phi}} g^*)$, where $\Pi_{\hat{\Phi}}$ is the projection operator onto $\hat{\Phi}$ in $L^2(P_\mathcal{X})$, and $f_{\hat{\Psi}}$ is the projection of f^* onto $\mathcal{H}_{\hat{\Psi}}$ *w.r.t.* $\langle \cdot, \cdot \rangle_{\mathcal{H}_\Gamma}$. Therefore, we need to bound $\mathcal{G}_n(\delta; \mathcal{F}_0)$ and $\bar{\mathfrak{R}}_n(\delta; \mathcal{F}_*)$. We start with the following uniform bound:

Proposition 2. *If $f = \Gamma^* g$, and $\|g\|_{P_\mathcal{A}} \le T$, then $|f(x)| \le \kappa T$ for all x.*

Proof. By Eqn. (8), we have $p(a|x) = \sum_i \sqrt{\lambda_i}\phi_i(a)\psi_i(x)p(a)$. For any $g = \sum_i u_i \phi_i \in L^2(P_\mathcal{A})$ such that $\|g\|_{P_\mathcal{A}} \le T$, $(\Gamma^* g)(x) = \int g(a)p(a|x)da = \sum_i \sqrt{\lambda_i} u_i \psi_i(x)$. Then, by Cauchy-Schwarz inequality, we have for all x, $f(x)^2 = (\Gamma^* g)(x)^2 \le (\sum_i \lambda_i \psi_i(x)^2)(\sum_i u_i^2) \le \kappa^2 T^2$. □

This proposition immediately implies that f^* and $f_{\hat{\Psi}}$ are uniformly bounded:

Corollary 3. *For any $f^* \in \mathcal{F}_B(\Gamma; \epsilon)$, Eqn. (1) ensures that $\|g^*\|_{P_\mathcal{A}}^2 \le \frac{B^2}{1-\epsilon}$, so $|f^*(x)| \le \frac{\kappa B}{\sqrt{1-\epsilon}}$ for all x. Moreover, $\|\Pi_{\hat{\Phi}} g^*\|_{P_\mathcal{A}} \le \|g^*\|_{P_\mathcal{A}}$ implies that $\|f_{\hat{\Psi}}\|_{\mathcal{H}_\Gamma} \le \frac{B}{\sqrt{1-\epsilon}}$, and $|f_{\hat{\Psi}}(x)| \le \frac{\kappa B}{\sqrt{1-\epsilon}}$ for all x.*

We will also use the following simple result in linear algebra:

Lemma 4. *Let $\boldsymbol{D}_\lambda = \operatorname{diag}(\lambda_1, \lambda_2, \cdots)$ where $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ and $\lambda_i \to 0$. Let $\boldsymbol{Q}$ be a matrix with d rows that are unit vectors. Then, $\operatorname{Tr}(\boldsymbol{Q}\boldsymbol{D}_\lambda\boldsymbol{Q}^\top) \le \lambda_1 + \cdots + \lambda_d$.*

Proof. Let $\boldsymbol{q}_i$ be the i-th column of $\boldsymbol{Q}$. Then for all $j \in [d]$, there is $\sum_{i=1}^j \boldsymbol{q}_i^\top \boldsymbol{q}_i \le j$. And for $j > d$, $\sum_{i=1}^j \boldsymbol{q}_i^\top \boldsymbol{q}_i \le d$. Thus, using Abel transformation, we have

$$\operatorname{Tr}(\boldsymbol{Q}\boldsymbol{D}_\lambda\boldsymbol{Q}^\top) = \operatorname{Tr}(\boldsymbol{D}_\lambda\boldsymbol{Q}^\top\boldsymbol{Q}) = \sum_{i=1}^\infty \lambda_i \boldsymbol{q}_i^\top \boldsymbol{q}_i = \sum_{j=1}^\infty \left(\sum_{i=1}^j \boldsymbol{q}_i^\top \boldsymbol{q}_i \right)(\lambda_j - \lambda_{j+1}) \le \sum_{i=1}^d \lambda_i,$$

which proves the assertion. □

This result has many implications. For instance, for any rank-d subspace of $\mathcal{H}_\Gamma$, its trace (the sum of its eigenvalues) is at most $S_\lambda(d)$.

Now, let us bound $\mathcal{G}_n(\delta; \mathcal{F}_0)$ with the following result:

Lemma 5. *(Application of Wainwright (2019, Lemma 13.22)) Let $\mathcal{H}$ be an RKHS with reproducing kernel K. Given samples $\tilde{x}_1, \cdots, \tilde{x}_n$, let $\boldsymbol{K}$ be the normalized kernel matrix with entries $\boldsymbol{K}(i,j) = K(\tilde{x}_i, \tilde{x}_j)/n$. Let $\mu_1 \geq \cdots \geq \mu_n \geq 0$ be the eigenvalues of $\boldsymbol{K}$. Then for all $\delta > 0$, we have*

$$\mathbb{E}\left[\sup_{\|f\|_{\mathcal{H}} \leq T, \|f\|_n \leq \delta} \left| \frac{1}{n} \sum_{i=1}^n \omega_i f(\tilde{x}_i) \right| \right] \leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \mu_j T^2\}}, \tag{15}$$

where $\omega_1, \cdots, \omega_n$ are i.i.d. $\mathcal{N}(0,1)$ variates. We apply this result to $K = K_X$. By Definition 1, all elements on the diagonal of $\boldsymbol{K}$ are at most κ^2/n, so $\sum_j \mu_j = \operatorname{Tr}(\boldsymbol{K}) \leq \kappa^2$. Thus, we have

$$\mathcal{G}_n(\delta; \mathcal{F}_0) \leq \sqrt{\frac{8\kappa^2 B^2}{n(1-\epsilon)}} \qquad \text{for any } \mathcal{H}_{\hat{\Psi}}. \tag{16}$$

Regarding $\bar{\mathfrak{R}}_n(\delta; \mathcal{F}_*)$, $\mathcal{F}_*$ is also a subset of RKHS $\hat{\mathcal{H}}_*$, which is the linear span of $\hat{\Psi}$ and f^*, and is a subspace of $\mathcal{H}_\Gamma$ whose rank is at most $(d+1)$. By Lemma 4, the sum of eigenvalues of $\hat{\mathcal{H}}_*$ is at most $S_\lambda(d+1)$. Since $\|f^*\|_{\mathcal{H}_\Gamma} \leq \frac{B}{\sqrt{1-\epsilon}}$, all $f \in \mathcal{F}_*$ satisfy $\|f\|_{\mathcal{H}_\Gamma} \leq \frac{2B}{\sqrt{1-\epsilon}}$. So we have the following bound for $\bar{\mathfrak{R}}_n(\delta; \mathcal{F}_*)$:

Lemma 6. *(Application of Wainwright (2019, Corollary 14.5)) Let $\mu_1, \mu_2, \cdots$ be the eigenvalues of the RKHS $\hat{\mathcal{H}}_*$. Since $\operatorname{rank}(\hat{\mathcal{H}}_*) \leq \operatorname{rank}(\mathcal{H}_{\hat{\Psi}}) + 1$, we have*

$$\bar{\mathfrak{R}}_n(\delta; \mathcal{F}_*) \leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^\infty \min\left\{\delta^2, \frac{4\mu_j B^2}{1-\epsilon}\right\}} \leq \sqrt{\frac{8B^2}{n(1-\epsilon)} S_\lambda(d+1)} \quad \text{if } \operatorname{rank}(\mathcal{H}_{\hat{\Psi}}) \leq d, \tag{17}$$

and for an arbitrary $\mathcal{H}_{\hat{\Psi}}$, we can simply replace $S_\lambda(d+1)$ with S_λ.

C.2 Proofs

Lemma 7. *Suppose $\nu_1, \cdots, \nu_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ variates. If $\hat{\Phi}$ has d dimensions (d can be ∞), then we have the following uniform bound over all $f^* = \Gamma^* g^* \in \mathcal{F}_B(\Gamma; \epsilon)$:*

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\tilde{x}_i, \nu_i}\left[\forall f^* \in \mathcal{F}_B(\Gamma; \epsilon), \|\hat{f} - f^*\|_{P_\mathcal{X}}^2 \leq 9\|f_{\hat{\Psi}} - f^*\|_{P_\mathcal{X}}^2 + \frac{c_0 \kappa(B^2 + \sigma B)}{1-\epsilon}\sqrt{\frac{S_\lambda(d+1)}{n}}\right] \\
&\geq 1 - c_1 \exp\left(-\frac{c_2\sqrt{2nS_\lambda(d+1)}}{\kappa}\right) - \exp\left(-\sqrt{\frac{2n\kappa^2 B^2}{1-\epsilon}}\right),
\end{aligned}$$

where $f_{\hat{\Psi}} = \Gamma^*(\Pi_{\hat{\Phi}} g^*)$ is the projection of f^* onto $\mathcal{H}_{\hat{\Psi}}$ w.r.t. $\langle\cdot,\cdot\rangle_{\mathcal{H}_\Gamma}$, and c_0, c_1, c_2 are universal constants. Moreover, $S_\lambda(d+1) \leq \min\left\{d+1, \kappa^2\right\}$.

Proof. By Proposition 2, all functions in $\mathcal{F}_*$ are b-uniformly bounded, with $b = \frac{2\kappa B}{\sqrt{1-\epsilon}}$. And obviously $\mathcal{F}_*$ is star-shaped, meaning that for all $f \in \mathcal{F}_*$ and all $\beta \in [0,1]$, $\beta f \in \mathcal{F}_*$. Let $t^2 = b \cdot \sqrt{\frac{8B^2}{n(1-\epsilon)} S_\lambda(d+1)} \geq b\bar{\mathfrak{R}}_n(\delta; \mathcal{F}_*)$. Then, by Wainwright (2019, Theorem 14.1), we have

$$\mathbb{P}\left[\left|\|f\|_n^2 - \|f\|_{P_\mathcal{X}}^2\right| \geq \frac{1}{2}\|f\|_{P_\mathcal{X}}^2 + \frac{t^2}{2}\right] \leq c_1 \exp\left(-c_2 \frac{nt^2}{b^2}\right) \qquad \text{for all } f \in \mathcal{F}_* \tag{18}$$

for universal constant c_1, c_2. We know that $\hat{f} - f^* \in \mathcal{F}_*$ and $f_{\hat{\Psi}} - f^* \in \mathcal{F}_*$, which means that

$$\begin{aligned}
&\mathbb{P}\left[\left(\|\hat{f} - f^*\|_{P_\mathcal{X}}^2 \geq 2\|\hat{f} - f^*\|_n^2 + t^2\right) \vee \left(\|f_{\hat{\Psi}} - f^*\|_n^2 \geq \frac{3}{2}\|f_{\hat{\Psi}} - f^*\|_{P_\mathcal{X}}^2 + \frac{t^2}{2}\right)\right] \\
&\leq c_1 \exp\left(-c_2 \frac{nt^2}{b^2}\right).
\end{aligned} \tag{19}$$

Let $\delta_n^2 = 2\sigma\sqrt{\frac{8\kappa^2 B^2}{n(1-\epsilon)}}$. By Lemma 5, we have $\delta_n^2 \geq 2\sigma\mathcal{G}_n(\delta_n; \mathcal{F}_0)$. And $\mathcal{F}_0$ is also star-shaped. Thus, by setting $\gamma = 1/2$ in Wainwright (2019, Theorem 13.13), we have*

$$\mathbb{P}\left[\|\hat{f} - f^*\|_n^2 \geq 3\|f_{\hat{\Psi}} - f^*\|_n^2 + 32\delta_n^2\right] \leq \exp\left(-\frac{n\delta_n^2}{2\sigma^2}\right). \tag{20}$$

Combining the two inequalities above with the union bound, we obtain the result. □

Now we prove Lemma 9. Without loss of generality, suppose $h_1, \cdots, h_{d'}$ are linearly independent. Let $\hat{\mathcal{H}}_{d'} := \text{span}\{h_1, \cdots, h_{d'}\}$. Let $g^* = g_0 + \beta g_1$, where $g_0 = \Pi_{\hat{\mathcal{H}}_{d'}} g^*$, $g_1 \perp g_0$, and $\|g_1\|_{P_\mathcal{A}} = 1$. So by Lemma 4, we have:

Proposition 8. $\|\Gamma^*(\boldsymbol{G}_h^{-1/2}h_1)\|_{P_\mathcal{X}}^2 + \cdots + \|\Gamma^*(\boldsymbol{G}_h^{-1/2}h_{d'})\|_{P_\mathcal{X}}^2 + \|\Gamma^* g_1\|_{P_\mathcal{X}}^2 \leq \lambda_1 + \cdots + \lambda_{d'+1}$.

Proof. Let $\left[\boldsymbol{G}_h^{-1/2}h_1, \cdots, \boldsymbol{G}_h^{-1/2}h_{d'}, g_1\right] = \boldsymbol{Q}\boldsymbol{\Phi}^*$, where $\boldsymbol{Q}$ is a matrix with $(d'+1)$ orthonormal rows. Then, $\left[\Gamma^*(\boldsymbol{G}_h^{-1/2}h_1), \cdots, \Gamma^*(\boldsymbol{G}_h^{-1/2}h_{d'}), \Gamma^* g_1\right] = \boldsymbol{Q}\boldsymbol{D}_\lambda{}^{1/2}\boldsymbol{\Phi}^*$. Thus, we have

$$\|\Gamma^*(\boldsymbol{G}_h^{-1/2}h_1)\|_{P_\mathcal{X}}^2 + \cdots + \|\Gamma^*(\boldsymbol{G}_h^{-1/2}h_{d'})\|_{P_\mathcal{X}}^2 + \|\Gamma^* g_1\|_{P_\mathcal{X}}^2 = \text{Tr}(\boldsymbol{Q}\boldsymbol{D}_\lambda\boldsymbol{Q}^\top).$$

Then, applying Lemma 4 completes the proof. □

Remark. This proposition is the functional version of Fan (1949, Theorem 1).

Notice that $\|\Gamma^*(\boldsymbol{G}_h^{-1/2}h_1)\|_{P_\mathcal{X}}^2 + \cdots + \|\Gamma^*(\boldsymbol{G}_h^{-1/2}h_{d'})\|_{P_\mathcal{X}}^2 = \text{Tr}(\boldsymbol{G}_h^{-1/2}\boldsymbol{F}_h\boldsymbol{G}_h^{-1/2}) = \text{Tr}(\boldsymbol{G}_h^{-1}\boldsymbol{F}_h)$. With this, we can prove the following:

Lemma 9. *For any $f^* \in \mathcal{F}_B(\Gamma; \epsilon)$, there is*

$$\|f_{\hat{\Psi}} - f^*\|_{P_\mathcal{X}}^2 \leq \frac{\tau^2}{1-\tau^2}\frac{\tau+\epsilon}{1-\epsilon}B^2.$$

Proof. Let $\alpha^2 = \|g_0\|_{P_\mathcal{A}}^2$, and $\beta^2 = \|g_0 - g^*\|_{P_\mathcal{A}}^2$. By Corollary 3, $\alpha^2 + \beta^2 \leq \frac{B^2}{1-\epsilon}$. Eqn. (1) implies that

$$(1-\epsilon)(\alpha^2+\beta^2) \leq \|\Gamma^*(g_0 + \beta g_1)\|_{P_\mathcal{X}}^2 \leq \alpha^2 + \beta^2\tau^2 + 2\alpha\beta\tau,$$

since $\|\Gamma^* g_0\|_{P_\mathcal{X}}^2 \leq \|g_0\|_{P_\mathcal{A}}^2 = \alpha^2$, and $\|\Gamma^* g_1\|_{P_\mathcal{X}}^2 \leq \tau^2$ by Proposition 8. Thus,

$$(1-\tau^2)\beta^2 \leq \epsilon(\alpha^2+\beta^2) + 2\alpha\beta\tau \leq (\epsilon+\tau)(\alpha^2+\beta^2) \leq (\epsilon+\tau)\frac{B^2}{1-\epsilon}.$$

Thus, we have $\|f_{\hat{\Psi}} - f^*\|_{P_\mathcal{X}}^2 = \|\Gamma^*(g_0 - g^*)\|_{P_\mathcal{X}}^2 = \beta^2\|\Gamma^* g_1\|_{P_\mathcal{X}}^2 \leq \beta^2\tau^2$, which leads to the inequality we need to prove. Finally, by setting $h_i = \hat{\phi}_i$, we can see that $\tau^2 \leq S_\lambda(d+1) - \text{Tr}(\boldsymbol{G}^{-1}\boldsymbol{F})$. And for all $d' \leq d$, $\text{Tr}(\boldsymbol{G}_h^{-1}\boldsymbol{F}_h) \leq S_\lambda(d')$, so $\tau^2 \geq \lambda_{d+1}$. □

D Proof of Theorem 2

Proposition 10. *For any $\hat{\Psi} = [\hat{\psi}_1, \cdots, \hat{\psi}_d]$ where $\hat{\psi}_i \in L^2(P_\mathcal{X})$, it holds that*

$$\text{err}(\hat{\Psi}; \mathcal{F}_B(\Gamma;\epsilon)) \geq \frac{\lambda_{d+1}}{1-\lambda_{d+1}}\frac{\epsilon}{1-\epsilon}B^2 \quad \textit{given that} \quad \frac{\lambda_{d+1}}{1-\lambda_{d+1}}\frac{\epsilon}{1-\epsilon} \leq \frac{1}{2}. \tag{21}$$

To attain equality, it is sufficient for $\hat{\Psi}$ to span the top-d eigenspace, and also necessary if $\lambda_{d+1} < \lambda_d$.

*Please refer to the proof of Wainwright (2019, Theorem 13.13) for removing the universal constants in this theorem.

Proof. *Necessity:* Since $\hat{\Psi}$ is at most rank-d, there must be a function in $\text{span}\{\psi_1, \cdots, \psi_{d+1}\}$ that is orthogonal to $\hat{\Psi}$. Thus, we can find two functions $f_1, f_2 \in \text{span}\{\psi_1, \cdots, \psi_{d+1}\}$ such that: $\|f_1\|_{P_{\mathcal{X}}} = \|f_2\|_{P_{\mathcal{X}}} = 1$, f_1 is orthogonal to $\hat{\Psi}$, $f_2 = \boldsymbol{u}^\top \hat{\Psi}$ (which means that $f_2 \perp f_1$), and $\psi_1 \in \text{span}\{f_1, f_2\}$. Recall that $\lambda_1 = 1$, and $\psi_1 \equiv 1$. Let $\psi_1 = \alpha_1 f_1 + \alpha_2 f_2$, then $\alpha_1^2 + \alpha_2^2 = 1$. Without loss of generality, suppose $\alpha_1, \alpha_2 \in [0, 1]$. Let $f_0 = \alpha_2 f_1 - \alpha_1 f_2$. Then, $\|f_0\|_{P_{\mathcal{X}}} = 1$, $f_0 \perp \psi_1$. Note that we also have $\langle \psi_1, f_0 \rangle_{\mathcal{H}_\Gamma} = 0$ by duality. Let $\beta_1, \beta_2 \in [0, 1]$ be any value such that $f = \beta_1 \psi_1 + \beta_2 f_0$ satisfies $\|f\|_{P_{\mathcal{X}}}^2 = \beta_1^2 + \beta_2^2 = 1$, and $\|f\|_{\mathcal{H}_\Gamma}^2 \le \frac{1}{1-\epsilon}$. This is satisfied as long as $\beta_2^2 \le \frac{\epsilon}{1-\epsilon} \frac{\lambda_{d+1}}{1-\lambda_{d+1}}$, because $\|f\|_{\mathcal{H}_\Gamma}^2 \le \beta_1^2 + \frac{\beta_2^2}{\lambda_{d+1}} = 1 + \frac{1-\lambda_{d+1}}{\lambda_{d+1}} \beta_2^2 \le \frac{1}{1-\epsilon}$. Moreover, we have $Bf \in \mathcal{F}_B(\Gamma; \epsilon)$.

It is easy to show that $F(\alpha_1) = \alpha_1 \beta_1 + \alpha_2 \beta_2 = \alpha_1 \beta_1 + \sqrt{1 - \alpha_1^2} \beta_2$ ($\alpha_1 \in [0, 1]$) first increases then decreases, so $F(\alpha_1)^2 \ge \min\left\{F(0)^2, F(1)^2\right\} = \min\left\{\beta_1^2, \beta_2^2\right\}$, which can be $\frac{\epsilon}{1-\epsilon} \frac{\lambda_{d+1}}{1-\lambda_{d+1}}$ in the worst case given that it is at most $\frac{1}{2}$, in which case the prediction error of Bf is $\|B(\alpha_1 \beta_1 + \alpha_2 \beta_2) f_1\|_{P_{\mathcal{X}}}^2 = F(\alpha_1)^2 B^2 = \frac{\epsilon}{1-\epsilon} \frac{\lambda_{d+1}}{1-\lambda_{d+1}} B^2$. Thus, for any $\hat{\Psi}$, we can find a function $Bf \in \mathcal{F}_B(\Gamma; \epsilon)$ such that $\min_w \text{err}(w^\top \hat{\Psi}, Bf) \ge \frac{\epsilon}{1-\epsilon} \frac{\lambda_{d+1}}{1-\lambda_{d+1}} B^2$.

When $\lambda_d > \lambda_{d+1}$, to attain equality, we need $\alpha_1 = 0$, and $\|f\|_{\mathcal{H}_\Gamma}^2 = \beta_1^2 + \frac{\beta_2^2}{\lambda_{d+1}}$, which means that $f_0 = \psi_{d+1}$. Thus, only $f_1 = f_0 = \psi_{d+1}$ is orthogonal to $\hat{\Psi}$, so $\hat{\Psi}$ must span the top-d eigenspace.

Sufficiency: Suppose $\hat{\Psi}$ spans the top-d eigenspace. For any $f \in \mathcal{F}_B(\Gamma; \epsilon)$ such that $f = \sum_i u_i \psi_i$, we have $\sum_i u_i^2 \le B^2$, and $\sum_i \frac{1-\epsilon-\lambda_i}{\lambda_i} u_i^2 \le 0$. Let $a = \sum_{i \ge d+1} u_i^2$ and $b = \sum_{i=1}^d u_i^2$. Then, $a = \min_w \text{err}(w^\top \hat{\Psi}, f)$, and $a + b \le B^2$. So we have

$$
\begin{aligned}
0 &\ge \sum_i \frac{1-\epsilon-\lambda_i}{\lambda_i} u_i^2 \ge -\epsilon b + \frac{1-\epsilon-\lambda_{d+1}}{\lambda_{d+1}} a \quad \left(\text{since } \frac{1-\epsilon-\lambda}{\lambda} \text{ decreases with } \lambda\right) \\
&\ge -\epsilon(B^2 - a) + \frac{1-\epsilon-\lambda_{d+1}}{\lambda_{d+1}} a \\
&= -\epsilon B^2 + (1-\epsilon) \frac{1-\lambda_{d+1}}{\lambda_{d+1}} a,
\end{aligned}
$$

which combined with the necessity part implies that $\text{err}(\hat{\Psi}; \mathcal{F}_B(\Gamma; \epsilon)) = \frac{\epsilon}{1-\epsilon} \frac{\lambda_{d+1}}{1-\lambda_{d+1}} B^2$. □

Lemma 11. *Suppose there exists a constant $C > 0$ such that $\mathbb{E}_{P_{\mathcal{A}}}[g^4] \le C^2 \|g\|_{P_{\mathcal{A}}}^2$, for all $g = w^\top \hat{\Phi}$ where $\|g\|_{P_{\mathcal{A}}} \le 1$. Then, for any $\delta > 0$, it holds with probability at least $1 - \delta$ that*

$$
|\operatorname{Tr}(\hat{\boldsymbol{G}}^{-1} \hat{\boldsymbol{F}}) - \operatorname{Tr}(\boldsymbol{G}^{-1} \boldsymbol{F})| \le \left(2 + \sqrt{2 \log \frac{2}{\delta}}\right) \frac{C\kappa + \kappa^2}{\sqrt{N}} d. \tag{22}
$$

Proof. Since multiplying an invertible $d \times d$ matrix to $\hat{\Phi}$ does not change either $\operatorname{Tr}(\hat{\boldsymbol{G}}^{-1} \hat{\boldsymbol{F}})$ or $\operatorname{Tr}(\boldsymbol{G}^{-1} \boldsymbol{F})$, for simplicity let us multiply $\boldsymbol{G}^{-1/2}$ to $\hat{\Phi}$, so that $\langle \hat{\phi}_i, \hat{\phi}_j \rangle_{P_{\mathcal{A}}} = \delta_{i,j}$ for all $i, j \in [d]$ (*i.e.* $\boldsymbol{G} = \boldsymbol{I}$). Define $\mathcal{F}_1 = \{f \in \mathcal{H}_\Gamma \mid \|f\|_{\mathcal{H}_\Gamma} \le 1\}$. Its Rademacher complexity is given by

$$
\mathfrak{R}_N(\mathcal{F}_1) = \mathop{\mathbb{E}}_{x_1, \cdots, x_N} \mathop{\mathbb{E}}_{\sigma_1, \cdots, \sigma_N} \left[\sup_{f \in \mathcal{F}_1} \frac{1}{N} \sum_{k=1}^N \sigma_k f(x_k)\right]. \tag{23}
$$

By Mohri et al. (2018, Theorem 6.12), we have $\mathfrak{R}_N(\mathcal{F}_1) \le \kappa N^{-1/2}$. Moreover, by Proposition 2, all $f \in \mathcal{F}_1$ satisfy $|f(x)| \le \kappa$ for all x. Thus, by Wainwright (2019, Theorem 4.10), for any $\delta > 0$, with probability at least $1 - \delta/2$, it holds for all $f \in \mathcal{F}_1$ that

$$
\left|\frac{1}{N} \sum_{k=1}^N f(x_k) - \mathbb{E}[f(X)]\right| \le 2\mathfrak{R}_N(\mathcal{F}_1) + \kappa \sqrt{\frac{2}{N} \log \frac{2}{\delta}} \le \left(2 + \sqrt{2 \log \frac{2}{\delta}}\right) \frac{\kappa}{\sqrt{N}}. \tag{24}
$$

Define matrix $\boldsymbol{M} = \hat{\boldsymbol{G}}^{-1/2}\hat{\boldsymbol{F}}\hat{\boldsymbol{G}}^{-1/2} = (m_{i,j})_{i,j\in[d]}$. $\|\boldsymbol{M}\|_2 \le 1$, so $\sum_{i=1}^d m_{i,j}^2 \le 1$ for all $j \in [d]$. Consider $\mathrm{Tr}((\boldsymbol{I} - \hat{\boldsymbol{G}})\boldsymbol{M})$. For any $j \in [d]$, we have

$$((\boldsymbol{I} - \hat{\boldsymbol{G}})\boldsymbol{M})(j,j) = \left\langle \hat{\phi}_j, \sum_{i=1}^d m_{i,j}\hat{\phi}_i \right\rangle_{\hat{P}_{\mathcal{A}}} - \left\langle \hat{\phi}_j, \sum_{i=1}^d m_{i,j}\hat{\phi}_i \right\rangle_{P_{\mathcal{A}}}.$$

Note that $\left\|\sum_{i=1}^d m_{i,j}\hat{\phi}_i\right\|_{P_{\mathcal{A}}} \le 1$, so $\left\|\hat{\phi}_j\left(\sum_{i=1}^d m_{i,j}\hat{\phi}_i\right)\right\|_{P_{\mathcal{A}}}^2 \le \sqrt{\mathbb{E}[\hat{\phi}_j^4]\mathbb{E}\left[\left(\sum_{i=1}^d m_{i,j}\hat{\phi}_i\right)^4\right]} \le C^2$, which means that $C^{-1}\Gamma^*\left(\hat{\phi}_j\left(\sum_{i=1}^d m_{i,j}\hat{\phi}_i\right)\right) \in \mathcal{F}_1$. So if Eqn. (24) holds, then for all $j \in [d]$, we have

$$\begin{aligned}((\boldsymbol{I} - \hat{\boldsymbol{G}})\boldsymbol{M})(j,j) &= \left|\frac{1}{N}\sum_{k=1}^N \Gamma^*\left(\hat{\phi}_j\left(\sum_{i=1}^d m_{i,j}\hat{\phi}_i\right)\right)(x_k) - \mathbb{E}\left[\Gamma^*\left(\hat{\phi}_j\left(\sum_{i=1}^d m_{i,j}\hat{\phi}_i\right)\right)(X)\right]\right| \\ &\le \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{C\kappa}{\sqrt{N}},\end{aligned}$$

which implies that

$$\mathrm{Tr}\left(\hat{\boldsymbol{G}}^{-1}\hat{\boldsymbol{F}} - \hat{\boldsymbol{F}}\right) = \mathrm{Tr}\left(\hat{\boldsymbol{G}}^{-1/2}(\boldsymbol{I} - \hat{\boldsymbol{G}})\hat{\boldsymbol{G}}^{-1/2}\hat{\boldsymbol{F}}\right) = \mathrm{Tr}\left((\boldsymbol{I} - \hat{\boldsymbol{G}})\boldsymbol{M}\right) \le \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{C\kappa d}{\sqrt{N}}.$$

Next, define $\mathcal{F}_2 = \{f_1 f_2 \mid f_1, f_2 \in \mathcal{H}_\Gamma, \|f_1\|_{\mathcal{H}_\Gamma} \le 1, \|f_2\|_{\mathcal{H}_\Gamma} \le 1\}$. By Proposition 12 (proved after this lemma), we have $\mathfrak{R}_N(\mathcal{F}_2) \le \kappa^2 N^{-1/2}$. And all $f \in \mathcal{F}_2$ satisfy $|f(x)| \le \kappa^2$ for all x by Proposition 2. So with probability at least $1 - \delta/2$, we have for all $f \in \mathcal{F}_2$,

$$\left|\frac{1}{N}\sum_{k=1}^N f(x_k) - \mathbb{E}[f(X)]\right| \le 2\mathfrak{R}_N(\mathcal{F}_2) + \kappa^2\sqrt{\frac{2}{N}\log\frac{2}{\delta}} \le \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2}{\sqrt{N}}. \tag{25}$$

Note that $\|\hat{\psi}_i\|_{\mathcal{H}_\Gamma} \le 1$. So under Eqn. (25), we have for all $i, j \in [d]$,

$$\left|\langle\hat{\psi}_i, \hat{\psi}_j\rangle_{\hat{P}_{\mathcal{X}}} - \langle\hat{\psi}_i, \hat{\psi}_j\rangle_{P_{\mathcal{X}}}\right| = \left|\frac{1}{N}\sum_{k=1}^N \hat{\psi}_i(x_k)\hat{\psi}_j(x_k) - \mathbb{E}[\hat{\psi}_i\hat{\psi}_j]\right| \le \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2}{\sqrt{N}},$$

which implies that $\mathrm{Tr}\left(\hat{\boldsymbol{F}} - \boldsymbol{G}^{-1}\boldsymbol{F}\right) = \mathrm{Tr}\left(\hat{\boldsymbol{F}} - \boldsymbol{F}\right) \le \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2 d}{\sqrt{N}}$.

Finally, applying the union bound completes the proof. □

Proposition 12. *Let $\mathcal{F}_2 = \{f_1 f_2 \mid f_1, f_2 \in \mathcal{H}_\Gamma, \|f_1\|_{\mathcal{H}_\Gamma} \le 1, \|f_2\|_{\mathcal{H}_\Gamma} \le 1\}$. Then, $\mathfrak{R}_N(\mathcal{F}_2) \le \frac{\kappa^2}{\sqrt{N}}$.*

Proof. For any $h(x) = f_1(x)f_2(x) \in \mathcal{F}_2$, let $f_1 = \Gamma^* g_1$ and $f_2 = \Gamma^* g_2$, where $\|g_1\|_{P_{\mathcal{A}}} \le 1$ and $\|g_2\|_{P_{\mathcal{A}}} \le 1$. Let $g_1 = \sum_i u_i \phi_i$ and $g_2 = \sum_i v_i \phi_i$. Let $\boldsymbol{u} = [u_1, u_2, \cdots]$ and $\boldsymbol{v} = [v_1, v_2, \cdots]$. Then, $\|\boldsymbol{u}\|_2 \le 1$ and $\|\boldsymbol{v}\|_2 \le 1$. And we have $f_1 = \sum_i \lambda_i^{1/2} u_i \psi_i$, and $f_2 = \sum_i \lambda_i^{1/2} v_i \psi_i$.

For any $x \in \mathcal{X}$, let $\Psi(x) = [\lambda_1^{1/2}\psi_1(x), \lambda_2^{1/2}\psi_2(x), \cdots]$. Then, $f_1(x) = \boldsymbol{u}^\top \Psi(x)$ and $f_2(x) = \boldsymbol{v}^\top \Psi(x)$. Denote $\Psi_k = \Psi(x_k)$. Then, $\Psi_k^\top \Psi_k \le \kappa^2$ for all $k \in [N]$. So for any $S = \{x_1, \cdots, x_N\}$, the empirical Rademacher complexity satisfies

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{F}_2) &\le \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\boldsymbol{u}\|_2\le 1, \|\boldsymbol{v}\|_2\le 1}\left|\frac{1}{N}\sum_{k=1}^N \sigma_k \boldsymbol{u}^\top \Psi_k \Psi_k^\top \boldsymbol{v}\right|\right] \\
&\le \frac{1}{N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{k=1}^N \sigma_k \Psi_k \Psi_k^\top\right\|_2\right] \\
&\le \frac{1}{N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{k=1}^N \sigma_k \Psi_k \Psi_k^\top\right\|_F\right] \\
&= \frac{1}{N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\mathrm{Tr}\left(\left(\sum_{k=1}^N \sigma_k \Psi_k \Psi_k^\top\right)^\top \left(\sum_{l=1}^N \sigma_l \Psi_l \Psi_l^\top\right)\right)^{1/2}\right] \\
&\le \frac{1}{N}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\mathrm{Tr}\left(\sum_{k,l=1}^N \sigma_k\sigma_l \Psi_k \Psi_k^\top \Psi_l \Psi_l^\top\right)\right]} \qquad \text{(Jensen)} \\
&= \frac{1}{N}\sqrt{\mathrm{Tr}\left(\sum_{k,l=1}^N \mathbb{E}[\sigma_k\sigma_l] \Psi_k \Psi_k^\top \Psi_l \Psi_l^\top\right)} \\
&= \frac{1}{N}\sqrt{\mathrm{Tr}\left(\sum_{k=1}^N \Psi_k \Psi_k^\top \Psi_k \Psi_k^\top\right)} \\
&\le \frac{1}{N}\sqrt{N\kappa^4} = \frac{\kappa^2}{\sqrt{N}}.
\end{aligned}$$

Then, since $\mathfrak{R}_N(\mathcal{F}_2) = \mathbb{E}_S[\hat{\mathfrak{R}}_S(\mathcal{F}_2)]$, we obtain the result. $\square$

Lemma 13. *Suppose $\hat{\phi}_i = \bar{\phi}_i$ for $i \in [d]$. Let $\gamma_{\boldsymbol{G}} := \lambda_{\max}(\boldsymbol{G})/\lambda_{\min}(\boldsymbol{G})$, which is the condition number of $\boldsymbol{G}$. Then, for any $\delta > 0$, both*

$$\sum_{j=1}^d \bar{\lambda}_j \ge \sum_{i=1}^d \lambda_i - \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{(\lambda_d^{-1}+1)\kappa^2}{\sqrt{N}} d$$

and Eqn. (22) with $C = \kappa \bar{\lambda}_d^{-1}\gamma_{\boldsymbol{G}}^{1/2}$ hold simultaneously for $\mathcal{H}_{\hat{\Psi}} = \hat{\mathcal{H}}_d$ with probability at least $1-\delta$.

Proof. Denote $\boldsymbol{\Phi}_d^* = [\phi_1, \cdots, \phi_d]$ and $\bar{\boldsymbol{\Phi}}_d^* = [\bar{\phi}_1, \cdots, \bar{\phi}_d]$. Let $\bar{\boldsymbol{\Phi}}_d^* = \boldsymbol{P}\boldsymbol{\Phi}^*$, where $\boldsymbol{P}$ is a matrix with d rows. Observe that for any $g = \sum_i u_i \bar{\phi}_i$ such that $\|g\|_{P_{\mathcal{A}}} \le 1$, we have $g = \bar{\Gamma}\Gamma^*\left(\sum_i \bar{\lambda}_i^{-1} u_i \bar{\phi}_i\right)$. Let $\boldsymbol{u} = (u_1, \cdots, u_d)$, then there is $g = \boldsymbol{u}^\top \bar{\boldsymbol{\Phi}}_d^* = \boldsymbol{u}^\top \boldsymbol{P}\boldsymbol{\Phi}^*$, so $\|\boldsymbol{P}^\top \boldsymbol{u}\|_2 \le 1$. Thus, we have $\|\sum_i \bar{\lambda}_i^{-1} u_i \bar{\phi}_i\|_{P_{\mathcal{A}}} = \|\boldsymbol{P}^\top \boldsymbol{D}_{\bar{\lambda}^d}^{-1}\boldsymbol{u}\|_2 = \|\boldsymbol{P}^\top \boldsymbol{D}_{\bar{\lambda}^d}^{-1}(\boldsymbol{P}\boldsymbol{P}^\top)^{-1}\boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{u}\|_2$.

So we just need to show that $\|\boldsymbol{P}^\top \boldsymbol{D}_{\bar{\lambda}^d}^{-1}(\boldsymbol{P}\boldsymbol{P}^\top)^{-1}\boldsymbol{P}\|_2 \le \bar{\lambda}_d^{-1}\gamma_{\boldsymbol{G}}^{1/2}$. $\|\boldsymbol{P}^\top \boldsymbol{D}_{\bar{\lambda}^d}^{-1}(\boldsymbol{P}\boldsymbol{P}^\top)^{-1}\boldsymbol{P}\|_2$ is equal to the square root of the largest eigenvalue of $\boldsymbol{P}^\top \boldsymbol{D}_{\bar{\lambda}^d}^{-1}(\boldsymbol{P}\boldsymbol{P}^\top)^{-1}\boldsymbol{D}_{\bar{\lambda}^d}^{-1}\boldsymbol{P}$, and by using two simple linear algebra exercises: (i) $\lambda_{\max}(\boldsymbol{A}\boldsymbol{B}) \le \lambda_{\max}(\boldsymbol{A})\lambda_{\max}(\boldsymbol{B})$ for positive definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and (ii) $\boldsymbol{A}\boldsymbol{B}$ and $\boldsymbol{B}\boldsymbol{A}$ share the same non-zero eigenvalues (Sylvester's Theorem), and the fact that $\boldsymbol{G} = \boldsymbol{P}\boldsymbol{P}^\top$, we can show that the largest eigenvalue of this matrix is at most $\bar{\lambda}_d^{-2}\gamma_{\boldsymbol{G}}$.

Therefore, we have $\|\boldsymbol{P}^\top \boldsymbol{D}_{\bar{\lambda}^d}^{-1}(\boldsymbol{P}\boldsymbol{P}^\top)^{-1}\boldsymbol{P}\|_2 \le \bar{\lambda}_d^{-1}\gamma_{\boldsymbol{G}}^{1/2}$, which combined with $\|\boldsymbol{P}^\top\boldsymbol{u}\|_2 \le 1$ implies that $\|\sum_i \bar{\lambda}_i^{-1}u_i\bar{\phi}_i\|_{P_{\mathcal{A}}} \le \bar{\lambda}_d^{-1}\gamma_{\boldsymbol{G}}^{1/2}$. By Proposition 2, $|\Gamma^*\left(\sum_i \bar{\lambda}_i^{-1}u_i\bar{\phi}_i\right)(x)| \le \kappa\bar{\lambda}_d^{-1}\gamma_{\boldsymbol{G}}^{1/2}$ for all x, so we have $|\bar{\Gamma}\Gamma^*\left(\sum_i \bar{\lambda}_i^{-1}u_i\bar{\phi}_i\right)(a)| = |\int \Gamma^*\left(\sum_i \bar{\lambda}_i^{-1}u_i\bar{\phi}_i\right)(x)p(x|a)dx| \le \kappa\bar{\lambda}_d^{-1}\gamma_{\boldsymbol{G}}^{1/2}$ for all a. This means that with $C = \kappa\bar{\lambda}_d^{-1}\gamma_{\boldsymbol{G}}^{1/2}$, g satisfies the condition of Lemma 11. Therefore, with probability at least $1-\delta$, both Eqn. (24) and Eqn. (25) hold and they lead to Eqn. (22).

Now let $\boldsymbol{\Phi}_d^* = \boldsymbol{Q}\bar{\boldsymbol{\Phi}}^*$, where $\boldsymbol{Q}$ is a matrix with d rows. Consider two matrices $\boldsymbol{Q}\boldsymbol{Q}^\top, \boldsymbol{Q}\boldsymbol{D}_{\bar{\lambda}}\boldsymbol{Q}^\top \in \mathbb{R}^{d\times d}$ where $\boldsymbol{D}_{\bar{\lambda}} = \text{diag}(\bar{\lambda}_1, \bar{\lambda}_2, \cdots)$, for which we have

$$(\boldsymbol{Q}\boldsymbol{Q}^\top)(i,j) = \langle \phi_i, \phi_j \rangle_{\hat{P}_{\mathcal{A}}} \qquad \text{and} \qquad (\boldsymbol{Q}\boldsymbol{D}_{\bar{\lambda}}\boldsymbol{Q}^\top)(i,j) = \langle \Gamma^*\phi_i, \Gamma^*\phi_j \rangle_{\hat{P}_{\mathcal{X}}}.$$

We have $(\langle \phi_i, \phi_j \rangle_{P_{\mathcal{A}}})_{i,j\in[d]} = \boldsymbol{I}$ and $(\langle \Gamma^*\phi_i, \Gamma^*\phi_j \rangle_{P_{\mathcal{X}}})_{i,j\in[d]} = \boldsymbol{D}_{\lambda^d} := \text{diag}(\lambda_1, \cdots, \lambda_d)$. Moreover, for any $g = \boldsymbol{u}^\top \boldsymbol{\Phi}_d^*$ such that $\|g\|_{P_{\mathcal{A}}} \le 1$, there is $g = \Gamma\Gamma^* \left(\sum_i \lambda_i^{-1} u_i \phi_i\right)$, and obviously $\|\sum_i \lambda_i^{-1} u_i \phi_i\|_{P_{\mathcal{A}}} \le \lambda_d^{-1}$. Thus, we can show that for all a, $|g(a)| \le \kappa\lambda_d^{-1}$, which means that $\boldsymbol{\Phi}_d^*$ satisfies the fourth-moment control assumption in Lemma 11 with $C' = \kappa\lambda_d^{-1}$. So similar to the proof of Lemma 11, for all $\boldsymbol{u} \in \mathbb{R}^d$ such that $\|\boldsymbol{u}\|_2 \le 1$, we can show that

$$\left|\boldsymbol{u}^\top(\boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{I})\boldsymbol{u}\right| = \left|\left\langle \boldsymbol{u}^\top\boldsymbol{\Phi}_d^*, \boldsymbol{u}^\top\boldsymbol{\Phi}_d^* \right\rangle_{\hat{P}_{\mathcal{A}}} - \left\langle \boldsymbol{u}^\top\boldsymbol{\Phi}_d^*, \boldsymbol{u}^\top\boldsymbol{\Phi}_d^* \right\rangle_{P_{\mathcal{A}}}\right| \le \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2\lambda_d^{-1}}{\sqrt{N}},$$

which implies that $\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_2 \le 1 + \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2\lambda_d^{-1}}{\sqrt{N}}$. It is easy to show that all non-zero eigenvalues of $\boldsymbol{Q}^\top\boldsymbol{Q}$ are also eigenvalues of $\boldsymbol{Q}\boldsymbol{Q}^\top$, so $\|\boldsymbol{Q}^\top\boldsymbol{Q}\|_2 \le 1 + \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2\lambda_d^{-1}}{\sqrt{N}}$. Moreover, similar to the proof of Lemma 11, we can show that for all $i, j \in [d]$,

$$\begin{cases} \left|\left(\boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{I}\right)(i,j)\right| \le \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2\lambda_d^{-1}}{\sqrt{N}}; & (26) \\ \left|\left(\boldsymbol{Q}\boldsymbol{D}_{\bar{\lambda}}\boldsymbol{Q}^\top - \boldsymbol{D}_{\lambda^d}\right)(i,j)\right| \le \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2}{\sqrt{N}}. & (27) \end{cases}$$

Let $\boldsymbol{q}_i$ be the i-th column of $\boldsymbol{Q}$. Then for all $i \in [d]$, $\boldsymbol{q}_i^\top\boldsymbol{q}_i \le 1 + \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2\lambda_d^{-1}}{\sqrt{N}}$. And we also have $\sum_{i=1}^\infty \boldsymbol{q}_i^\top\boldsymbol{q}_i = \text{Tr}(\boldsymbol{Q}^\top\boldsymbol{Q}) = \text{Tr}(\boldsymbol{Q}\boldsymbol{Q}^\top) \le d + \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2\lambda_d^{-1}d}{\sqrt{N}}$. Thus, we have

$$\begin{aligned} &\sum_{i=1}^d \lambda_i - \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2}{\sqrt{N}}d \le \text{Tr}(\boldsymbol{Q}\boldsymbol{D}_{\bar{\lambda}}\boldsymbol{Q}^\top) = \text{Tr}(\boldsymbol{D}_{\bar{\lambda}}\boldsymbol{Q}^\top\boldsymbol{Q}) \\ =& \sum_{i=1}^\infty \bar{\lambda}_i \boldsymbol{q}_i^\top\boldsymbol{q}_i = \sum_{j=1}^\infty \left(\sum_{i=1}^j \boldsymbol{q}_i^\top\boldsymbol{q}_i\right)(\bar{\lambda}_j - \bar{\lambda}_{j+1}) \\ \le& \sum_{i=1}^d \bar{\lambda}_i \left[1 + \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2\lambda_d^{-1}}{\sqrt{N}}\right] \le \sum_{i=1}^d \bar{\lambda}_i + \left(2 + \sqrt{2\log\frac{2}{\delta}}\right)\frac{\kappa^2\lambda_d^{-1}d}{\sqrt{N}}, \end{aligned}$$

which proves the assertion. □