THE THREE STAGES OF LEARNING DYNAMICS IN HIGH-DIMENSIONAL KERNEL METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

To understand how deep learning works, it is crucial to understand the training dynamics of neural networks. Several interesting hypotheses about these dynamics have been made based on empirically observed phenomena, but there exists a limited theoretical understanding of when and why such phenomena occur.

In this paper, we consider the training dynamics of gradient flow on kernel least-squares objectives, which is a limiting dynamics of SGD trained neural networks. Using precise high-dimensional asymptotics, we characterize the dynamics of the fitted model in two "worlds": in the *Oracle World* the model is trained on the population distribution and in the *Empirical World* the model is trained on an i.i.d finite dataset. We show that under mild conditions on the kernel and L^2 target regression function the training dynamics have three stages that are based on the behaviors of the models in the two worlds. Our theoretical results also mathematically formalize some interesting deep learning phenomena. Specifically, in our setting we show that SGD progressively learns more complex functions and that there is a "deep bootstrap" phenomenon: during the second stage, the test error of both worlds remain close despite the empirical training error being much smaller. Finally, we give a concrete example comparing the dynamics of two different kernels which shows that faster training is not necessary for better generalization.

1 INTRODUCTION

In order to fundamentally understand how and why deep learning works, there has been much effort to understand the dynamics of neural networks trained by gradient descent based algorithms. This effort has led to the discovery of many intriguing empirical phenomena (e.g. Frankle et al. (2020); Fort et al. (2020); Nakkiran et al. (2019a;b; 2020)) that help shape our conceptual framework for understanding the learning process in neural networks. Nakkiran et al. (2019b) provides evidence that SGD starts by first learning a linear classifier and over time learns increasingly complex functions. Nakkiran et al. (2020) introduces the "deep bootstrap" phenomenon: for some deep learning tasks the empirical world test error remains close to the oracle world error[1] for many SGD iterations, even if the empirical training and test errors display a large gap. To better understand such phenomena, it is useful to study training dynamics in related but mathematically tractable settings.

One approach for theoretically investigation is to study kernel methods, which were recently shown to have a tight connection with over-parameterized neural networks (Jacot et al., 2018; Du et al., 2018). Indeed, given a sequence of neural networks $(f_N)_{N\in\mathbb{N}}$ with the widths of the layers increasing with N, we can consider the SGD dynamics on the training dataset

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \eta\nabla[(\mathcal{Y}_k - f_N(\mathcal{X}_k;\boldsymbol{\theta}^k))^2],$$

where $(\mathcal{Y}_k - f_N(\mathcal{X}_k;\boldsymbol{\theta}^k))^2$ is the square loss function evaluated on a fresh batch of data $(\mathcal{X}_k, \mathcal{Y}_k)$. Assuming proper parametrization and initialization, such dynamics have been shown (Jacot et al., 2018; Du et al., 2018) to be well approximated by a continuous time linear dynamics in the large N and small η limit,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\beta}^t = -\nabla\hat{\mathbb{E}}[(\boldsymbol{y} - \langle\boldsymbol{\beta}^t, \nabla_{\boldsymbol{\theta}} f_N(\boldsymbol{x};\boldsymbol{\theta}^0)\rangle)^2].$$

[1] Their paper uses "Ideal World" for "Oracle World" and "Real World" for "Empirical World".

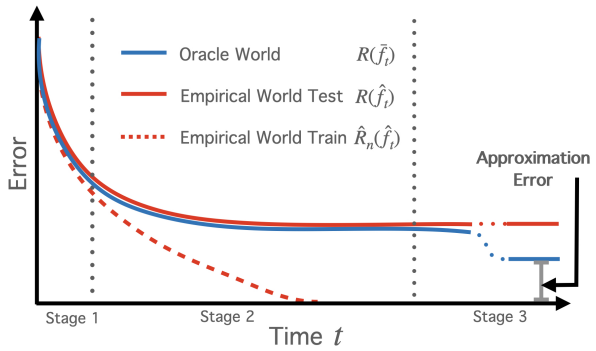


Figure 1: A conceptual drawing of empirical and oracle world learning curves. **Stage 1:** all curves are together. **Stage 2:** training error goes to zero while test and oracle error stay together. **Stage 3:** test error remains constant while oracle error decays to the RKHS approximation error. See Section 1.1 for a more detailed discussion. (Dotted lines in stage 3 indicate compressed time interval.)

The large t limit of such a solution is the kernel least-squares solution with the neural tangent kernel $K_N(\boldsymbol{x}, \boldsymbol{x}') = \langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}; \boldsymbol{\theta}^0), \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}'; \boldsymbol{\theta}^0) \rangle / N$. As a consequence, deep neural networks in the over-parameterized regime can provably find zero training error solutions.

Over the past few years, researchers have used kernel machines as a tractable model to investigate many neural network phenomena including benign overfitting i.e. generalization despite the interpolation of noisy data (Bartlett et al., 2020; Liang & Rakhlin, 2020) and double-descent i.e. non-classically U-shaped risk curves (Belkin et al., 2020; Liu et al., 2021). Kernels have also been studied to better understand certain aspects of neural network architectures such as invariance and stability (Bietti & Mairal, 2017; Mei et al., 2021b). Although kernel methods cannot be used to explain some phenomena such as feature learning, they can still be conceptually useful for understanding other neural networks properties.

1.1 Three stages of kernel dynamics

Despite much classical work in the study of gradient descent training of kernel machines (e.g. Yao et al. (2007); Raskutti et al. (2014)) there has been limited work understanding the high-dimensional setting, which is the setting of interest in this paper. Although solving the linear dynamics of gradient flow is simple, the statistical analysis of the fitted model requires involved random matrix theory arguments. In our analysis we study the dynamics of the *Oracle World*, where training is done on the (usually inaccessible) population risk, and the *Empirical World*, where training is done on the empirical risk (as is done in practice). Associated with the oracle world model f_t^{or} and the empirical world model $\hat{f}_t$ are the following quantities of interest: the **empirical training error** $\widehat{R}_n(\hat{f}_t)$, the **empirical test error** $R(\hat{f}_t)$, and the **oracle error** $R(f_t^{\mathrm{or}})$ defined in Eq. (1), (2), (3) for which we derive expressions that are accurate in high dimensions.

Informally, our main results show that under reasonable conditions on the regression function and the kernel the training dynamics undergo the following three stages:

- **Stage one**: the empirical training error, the empirical test error, and the oracle error are all close.
- **Stage two**: the empirical training error decays to zero, but the empirical test error and the oracle error stay close and keep approximately constant.
- **Stage three**: the empirical training error is still zero, the empirical test error stays approximately constant, but the oracle test error decays to the approximation error.

We conceptually illustrate the error curves of the oracle and empirical world in Figure 1 and provide intuition for the evolution of the learned models in Figure 2. The existence of the first and third stages are not unexpected: at the beginning of training the model has not fit the dataset enough to distinguish the oracle and empirical world and at the end of training an expressive enough model with infinite samples will outperform one with finitely many. The most interesting stage is the

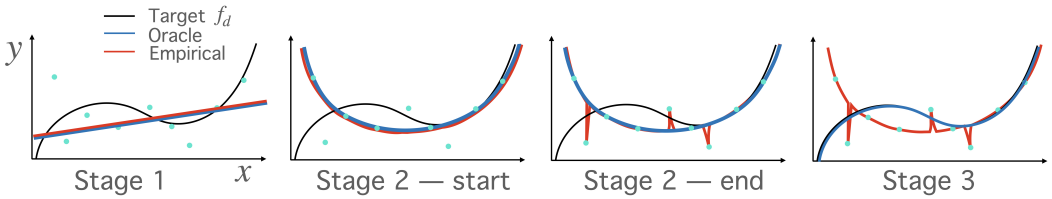


Figure 2: A conceptual drawing of the evolution of the empirical and oracle models $\hat{f}_t$ and f_t^{or}. In stage 1, $\hat{f}_t$ and f_t^{or} learn the best linear approximation of f_d. At the start of stage 2, $\hat{f}_t$ and f_t^{or} learn the best quadratic approximation. At the end of stage 2, $\hat{f}_t$ interpolates the training set but is close to f_t^{or} in the L^2 sense. Lastly in stage 3, f_t^{or} learns f_d while $\hat{f}_t$ stays the same as the end of stage 2.

second one where the empirical model begins to "overfit" the training set while still remaining close to the non-interpolating oracle model in the L^2 sense (see Fig. 2).

In Section 2 we discuss some related work. In Section 3 we elaborate our description of the three stages and give a detailed mathematical characterization for two specific settings in Theorem 1 and 2. Although the three stages arise fairly generally, we remark that certain stages will vanish if the problem parameters are chosen in a special way (c.f. Remark 1). We connect our theoretical results to related empirical deep learning phenomena in Remark 3 and discuss the relation to deep learning in practice in Remark 4. In Section 4 we provide numerical simulations to illustrate the theory more concretely and in Section 5 we end with a summary and discussion of the results.

2 Related Literature

The generalization error of the kernel ridge regression (KRR) solution has been well-studied in both the fixed dimension regime (Wainwright, 2019, Chap. 13), (Caponnetto & De Vito, 2007) and the high-dimensional regime (El Karoui, 2010; Liang & Rakhlin, 2020; Liu et al., 2021; Ghorbani et al., 2020; 2021; Mei et al., 2021a;b). Most closely related to our results is the setting of (Ghorbani et al., 2021; Mei et al., 2021a;b). Analysis of the entire KRR training trajectory has also been done (Yao et al., 2007; Raskutti et al., 2014; Cao et al., 2019) but only for the fixed dimensional setting.

Another line of work directly studies the dynamics of learning in linear neural networks (Saxe et al., 2013; Li et al., 2018; Arora et al., 2019; Vaskevicius et al., 2019). Similar to us, these works show that some notion of complexity (typically effective rank or sparsity) increases in the linear network over the course of optimization.

The relationship between the speed of iterative optimization and gap between population and empirical quantities has been studied before in the context of algorithmic stability (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Chen et al., 2018). These analyses certify good empirical generalization by using stability in the first few iterations to upper bound the gap between train and test error. In contrast, our analysis directly computes the errors at an arbitrary time t (c.f. Remark 2). The relationship between oracle and empirical training dynamics has been considered before in Bottou & LeCun (2004) and Pillaud-Vivien et al. (2018).

3 Results

In this section we introduce the problem and present a specialization of our results to two concrete settings: dot product and group invariant kernels on the sphere (Theorems 1 and 2 respectively). The more general version of our results is described in Appendix A.3.

3.1 Problem setup

We consider the supervised learning problem where we are given i.i.d data $(\boldsymbol{x}_i, y_i)_{i\le n}$. The covariate vectors $(\boldsymbol{x}_i)_{i\le n} \sim_{iid} \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and the real-valued noisy responses $y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$ for some unknown target function $f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ and $(\varepsilon_i)_{i\le n} \sim_{iid} \mathcal{N}(0, \sigma_\varepsilon^2)$. Given a function

$f \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$, we define its test error $R(f)$ and its training error $\widehat{R}_n(f)$ as

$$R(f) \equiv \mathbb{E}_{(\boldsymbol{x}_{\text{new}}, y_{\text{new}})}\{(y_{\text{new}} - f(\boldsymbol{x}_{\text{new}}))^2\}, \qquad \widehat{R}_n(f) \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 \tag{1}$$

where $(x_{\text{new}}, y_{\text{new}})$ is iid with $(\boldsymbol{x}_i, y_i)_{i \le n}$. The test error $R(f)$ measures the goodness-of-fit of f on the population distribution and the training error $\widehat{R}_n(f)$ measures the fit of f to the training set.

For a kernel function $H_d : \mathbb{S}^{d-1}(\sqrt{d}) \times \mathbb{S}^{d-1}(\sqrt{d}) \to \mathbb{R}$, we analyse the dynamics of the following two fitted models indexed by time t: the *oracle model* f_t^{or} and the *empirical model* $\hat{f}_t$, which are given by the gradient flow on R and $\widehat{R}_n$ over the associated RKHS $\mathcal{H}_d$ respectively

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t^{\text{or}}(\boldsymbol{x}) = -\boldsymbol{\nabla} R(f_t^{\text{or}}(\boldsymbol{x})) = \mathbb{E}[H_d(\boldsymbol{x}, \boldsymbol{z})(f_d(\boldsymbol{z}) - f_t^{\text{or}}(\boldsymbol{z}))], \tag{2}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \hat{f}_t(\boldsymbol{x}) = -\boldsymbol{\nabla} \widehat{R}_n(\hat{f}_t(\boldsymbol{x})) = \frac{1}{n}\sum_{i=1}^{n} H_d(\boldsymbol{x}, \boldsymbol{x}_i)(y_i - \hat{f}_t(\boldsymbol{x}_i)), \tag{3}$$

with zero initialization $f_0^{\text{or}} \equiv \hat{f}_0 \equiv 0$. These dynamics are motivated from the neural tangent kernel perspective of over-parameterized neural networks (Jacot et al., 2018; Du et al., 2018). A precise mathematical definition and derivation of these two dynamics are provided in Appendix E.1.

For our results we make some assumptions on the spectral properties of the kernels H_d similar to those in Mei et al. (2021a) that are discussed in detail in Section A.2. At a high-level we require that the diagonal elements of the kernel concentrate, that the kernel eigenvalues obey certain spectral gap conditions, and that the top eigenfunctions obey a hyperconctractivity condition which says they are "delocalized". For the specific settings of Theorems 1 and 2 we give more specific conditions on the kernels that are more easily verified and imply the required spectral properties.

3.2 Dot Product Kernels

In our first example, we consider dot product kernels H_d of the form

$$H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = h_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d), \qquad \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{S}^{d-1}(\sqrt{d}), \tag{4}$$

for some function $h_d : [-1, 1] \to \mathbb{R}$. Our results apply to general dot product kernels under weak conditions on h_d given in Appendix B.2. In particular they apply to the random feature and neural tangent kernels associated to certain fully connected neural networks (Jacot et al., 2018).

Before presenting our results for this setting we introduce some notation. Denote by $\overline{\mathsf{P}}_{\le \ell}$ the orthogonal projection onto the subspace of $L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ spanned by polynomials of degree less than or equal to ℓ. The projectors $\overline{\mathsf{P}}_{\ell}$ and $\overline{\mathsf{P}}_{>\ell}$ are defined analogously (see Appendix G for details). We use $o_d(\cdot)$ for standard little-o relations, where the subscript d emphasizes the asymptotic variable. We use $o_{d,\mathbb{P}}(\cdot)$ in probability relations. Namely for two sequences of random variables $Z_1(d)$ and $Z_2(d)$, $Z_1(d) = o_{d,\mathbb{P}}(Z_2(d))$ if for any $\varepsilon, C_\varepsilon > 0$ there exists $d_\varepsilon \in \mathbb{Z}_{>0}$, such that $\mathbb{P}(|Z_1(d)/Z_2(d)| < C_\varepsilon) \le \varepsilon$ for all $d \ge d_\varepsilon$. The notations $O_d, O_{d,\mathbb{P}}$ etc. are defined analogously.

Theorem 1 (Dot Product Kernels)**.** *Let $\{f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))\}_{d \ge 1}$ be a sequence of functions such that for some $\eta > 0$, $\|f_d\|_{L^{2+\eta}} = O_d(1)$ and let $\{H_d\}_{d \ge 1}$ be a sequence of dot product kernels satisfying Assumption 9. Assume that for some fixed integers $\mathsf{j}, \mathsf{s} \ge 0$ and some $\delta > 0$ that*

$$d^{\mathsf{j}+\delta} \le t \le d^{\mathsf{j}+1-\delta}, \qquad d^{\mathsf{s}+\delta} \le n \le d^{\mathsf{s}+1-\delta}.$$

Then we have the following characterizations,

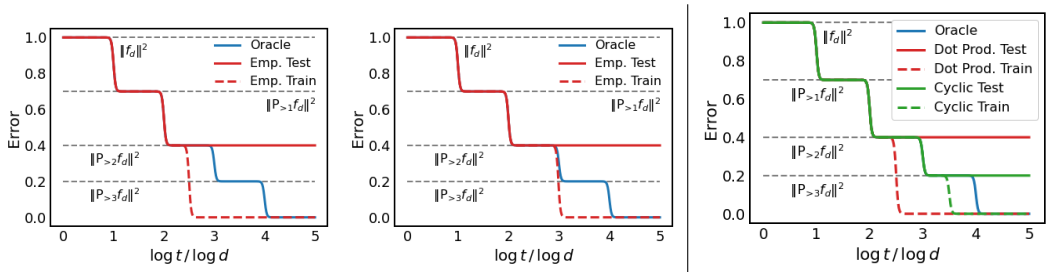
(a) (Oracle World) The oracle model learns every degree component of f_d as time progresses

$$R(f_t^{\text{or}}) = \left\|\overline{\mathsf{P}}_{>\mathsf{j}} f_d\right\|_{L^2}^2 + \sigma_\varepsilon^2 + o_d(1), \qquad \left\|f_t^{\text{or}} - \overline{\mathsf{P}}_{\le \mathsf{j}} f_d\right\|_{L^2}^2 = o_d(1).$$

(b) (Empirical World – Train) Empirical training error follows oracle error then goes to zero

$$\widehat{R}_n(\hat{f}_t) = \left\|\overline{\mathsf{P}}_{>\mathsf{j}} f_d\right\|_{L^2}^2 + \sigma_\varepsilon^2 + o_{d,\mathbb{P}}(1) \qquad \text{if } t/n = o_d(1),$$

$$\widehat{R}_n(\hat{f}_t) = o_{d,\mathbb{P}}(1) \qquad \text{if } t/n = \omega_d(1).$$



(a) Dot prod. kernel, $n \asymp d^{2.5}$ (b) Dot prod. kernel, $n \asymp d^{2.99}$ (c) Dot prod. vs cyclic, $n \asymp d^{2.5}$

Figure 3: Schematic drawings of the conclusions of Theorems 1 and 2 for three different noiseless settings. Panels (3a) and (3b) illustrate the performance of a dot product kernel H_d for two different scalings $n(d)$. The full three stages appear in (3a), but the second stages disappears in (3b) since $\log n / \log d$ is nearly an integer (see Remark 1). Panel (3c) compares the performance of a dot product kernel H_d (red) and corresponding cyclic kernel $H_{d,\mathrm{inv}}$ (green) for a cyclic target function f_d. The cyclic kernel in (3c) generalizes better but optimizes more slowly (see Remark 2).

(c) (Empirical World – Test) Empirical test error follows oracle error until the empirical model learns the degree-s component of f_d

$$R(\hat{f}_t) = \left\|\overline{\mathsf{P}}_{>\min\{\mathsf{j},\mathsf{s}\}} f_d\right\|_{L^2}^2 + \sigma_\varepsilon^2 + o_{d,\mathbb{P}}(1), \qquad \left\|\hat{f}_t - \overline{\mathsf{P}}_{\leq \min\{\mathsf{j},\mathsf{s}\}} f_d\right\|_{L^2}^2 = o_{d,\mathbb{P}}(1).$$

The results are conceptually illustrated in an example in Figure 3a which shows the stair-case phenomenon in high-dimensions and the three learning stages. We see that in both the oracle world and the empirical world, the prediction model increases in complexity over time. More precisely, the model learns the best polynomial fit to the target function (in an L^2 sense) of increasingly higher degree. In the empirical world the maximum complexity is determined by the sample size n, which is in contrast to the oracle world where there are effectively infinite samples.

The results imply that generally (but not always c.f. Remark 1) there will be three stages of learning. In the first stage the oracle and empirical world models are close in L^2 and fit a polynomial with degree determined by t. The first stage lasts from $t = 0$ to $t = nd^{-\varepsilon} \ll n$ for some small $\varepsilon > 0$. As t approaches n, there is a phase transition and the empirical world training error goes to zero at $t = nd^{\varepsilon} \gg n$. From time $nd^{-\varepsilon}$ till at least $d^{\mathsf{s}+1}$ is the second stage where the empirical and oracle models remain close in L^2 but the gap between test and train error can be large. If the sample size n is not large enough for $\hat{f}_t$ to learn the target function, then at some large enough t we will enter a third stage where f_t^{or} improves in performance, outperforming $\hat{f}_t$ which remains the same. On synthetic data in finite dimensions we can see a resemblance of the staircase shape which becomes sharper with increasing d (c.f. Appendix F).

Remark 1 (Degenerate stages of Learning)**.** *For special problem parameters we will not observe the second and/or third stages. The second stage will disappear if $n \asymp d^q$ for some $q \in \mathbb{N}$ (see Fig. 3b), or if f_d is a degree-s polynomial. The third stage will not occur if $\overline{\mathsf{P}}_{>\mathsf{s}} f_d$ lies in the orthogonal complement of the RKHS $\mathcal{H}_d$.*

3.3 Group Invariant Kernels

We now consider our second setting which concerns the invariant function estimation problem introduced in Mei et al. (2021b). As before, we are given i.i.d data $(y_i, \boldsymbol{x}_i)_{i \leq n}$ where the feature vectors $(\boldsymbol{x}_i)_{i \leq n} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and noisy responses $y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$. We now assume that the target function f_d satisfies an invariant property. We consider a general type of invariance, defined by a group $\mathcal{G}_d$ that is represented as a subgroup of the orthogonal group in d dimensions. The group element $g \in \mathcal{G}_d$ acts on a vector $\boldsymbol{x} \in \mathbb{R}^d$ via $\boldsymbol{x} \mapsto g \cdot \boldsymbol{x}$. We say that $f_\star$ is $\mathcal{G}_d$-invariant if $f_\star(\boldsymbol{x}) = f_\star(g \cdot \boldsymbol{x})$ for all $g \in \mathcal{G}_d$. We denote the space of square integrable $\mathcal{G}_d$-invariant functions by $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mathcal{G}_d)$. We focus on groups $\mathcal{G}_d$ that are *groups of degeneracy* α as defined below. As an example, we consider the cyclic group

$\mathrm{Cyc}_d = \{g_0, g_1, \ldots, g_{d-1}\}$ where for any $\boldsymbol{x} = (x_1, \ldots, x_d)^\mathsf{T} \in \mathbb{S}^{d-1}(\sqrt{d})$, the group action is defined by $g_i \cdot \boldsymbol{x} = (x_{i+1}, x_{i+2}, \ldots, x_d, x_1, x_2, \ldots, x_i)^\mathsf{T}$. The cyclic group has degeneracy 1.

Definition 1 (Groups of degeneracy α)**.** *Let $V_{d,k}$ be the subspace of degree-k polynomials that are orthogonal to polynomials of degree at most $(k-1)$ in $L^2(\mathbb{S}^{d-1}(\sqrt{d}))$, and denote by $V_{d,k}(\mathcal{G}_d)$ the subspace of $V_{d,k}$ formed by polynomials that are $\mathcal{G}_d$-invariant. We say that $\mathcal{G}_d$ has degeneracy α if for any integer $k \geq \alpha$ we have $\dim(V_{d,k}/V_{d,k}(\mathcal{G}_d)) \asymp d^\alpha$.*

To encode invariance in our kernel we consider $\mathcal{G}_d$-invariant kernels H_d of the form

$$H_{d,\mathrm{inv}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_{\mathcal{G}_d} h(\langle \boldsymbol{x}_1, g \cdot \boldsymbol{x}_2 \rangle / d) \pi_d(\mathrm{d}g) \tag{5}$$

where π_d is the Haar measuare on $\mathcal{G}_d$. Such kernels satisfy the following invariance property: for all $g, g' \in \mathcal{G}_d$ and for $H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = H_d(g \cdot \boldsymbol{x}_1, g' \cdot \boldsymbol{x}_2)$ for every $\boldsymbol{x}_1, \boldsymbol{x}_2$. For the cyclic group, π_d is the uniform measure.

We now present our results for the group invariant setting.

Theorem 2 (Group Invariant Kernels)**.** *Let $\mathcal{G}_d$ be a group of degeneracy $\alpha \leq 1$. Let $\{f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mathcal{G}_d)\}_{d \geq 1}$ a sequence of $\mathcal{G}_d$-invariant functions such that for some $\eta > 0$, $\|f_d\|_{L^{2+\eta}} = O_d(1)$ and let $\{H_d\}_{d \geq 1}$ be a sequence of $\mathcal{G}_d$-invariant kernels with activation σ satisfying Assumption 10. Assume that for some fixed integers $\mathsf{j} \geq 0, \mathsf{s} \geq 1$ and some $\delta > 0$ that*

$$d^{\mathsf{j}+\delta} \leq t \leq d^{\mathsf{j}+1-\delta}, \quad d^{\mathsf{s}-\alpha+\delta} \leq n \leq d^{\mathsf{s}+1-\alpha-\delta}.$$

Then we have the following characterizations,

(a) (Oracle World) The oracle model learns every degree component of f_d as time progresses

$$R(f_t^{\mathrm{or}}) = \left\|\overline{\mathsf{P}}_{>\mathsf{j}} f_d\right\|_{L^2}^2 + \sigma_\varepsilon^2 + o_d(1), \qquad \left\|f_t^{\mathrm{or}} - \overline{\mathsf{P}}_{\leq \mathsf{j}} f_d\right\|_{L^2}^2 = o_d(1).$$

(b) (Empirical World – Train) Empirical training error follows oracle error then goes to zero

$$\begin{aligned} \widehat{R}_n(\hat{f}_t) &= \left\|\overline{\mathsf{P}}_{>\mathsf{j}} f_d\right\|_{L^2}^2 + \sigma_\varepsilon^2 + o_{d,\mathbb{P}}(1) && \text{if } t/n = o_d(d^\alpha), \\ \widehat{R}_n(\hat{f}_t) &= o_{d,\mathbb{P}}(1) && \text{if } t/n = \omega_d(d^\alpha). \end{aligned}$$

(c) (Empirical World – Test) Empirical test error follows oracle error until the empirical model learns the degree-s component of f_d

$$R(\hat{f}_t) = \left\|\overline{\mathsf{P}}_{>\min\{\mathsf{j},\mathsf{s}\}} f_d\right\|_{L^2}^2 + \sigma_\varepsilon^2 + o_{d,\mathbb{P}}(1), \qquad \left\|\hat{f}_t - \overline{\mathsf{P}}_{\leq \min\{\mathsf{j},\mathsf{s}\}} f_d\right\|_{L^2}^2 = o_{d,\mathbb{P}}(1).$$

With respect to the dot product kernel setting (c.f. Theorem 1), in this setting the behavior of the oracle world is unchanged, but the empirical world behaves as if it has d^α times as many samples. This is illustrated graphically in Figure 3c. It can be shown that using an invariant kernel is equivalent to using a dot product kernel and augmenting the dataset to $\{(g \cdot \boldsymbol{x}_i, y_i) : g \in \mathcal{G}_d, i \in [n]\}$ (c.f. Appendix E.2). Hence for the cyclic group which has size $d^\alpha = d$, we arrive at the following intriguing conclusion: if the target function is invariant, then using a dot product kernel and augmenting n i.i.d samples to nd many samples is asymptotically equivalent to using nd i.i.d samples.

Remark 2 (Optimization Speed versus Generalization)**.** *Interestingly, training with an invariant kernel is slower than with a dot product kernel and takes longer to interpolate the dataset despite eventually generalizing better on invariant function estimation tasks (c.f. Figure 3c). This conclusion is not an artifact of the continuous time analysis (c.f. Appendix E.3) and is observed empirically in Section 4.2 for discrete-time SGD. This example highlights the limitation of stability based analyses (e.g. Hardt et al. (2016)) which argue that faster SGD training leads to better generalization. While a faster rate leads to better generalization in the first stage when stability can control the gap between train and test error, our analysis shows that the duration length of the first stage also impacts the final generalization error.*

Remark 3 (Connection with Deep Phenomena)**.** *The dynamics of high-dimensional kernel regression display behaviors that parallel some empirically observed phenomena in deep learning. For*

kernel regression, we have shown that the complexity of the empirical model, measured as the number of learned eigenfunctions, depends on the time optimized when $t \ll n$ and the sample size when $t \gg n$. At a high-level, we also expect a similar story for neural networks but for some other notion of complexity. It is believed that neural networks first learn simple functions and then progressively more complex ones, until the complexity saturates after interpolating at some time proportional to n (Nakkiran et al., 2019b). We have also shown that in kernel regression there is a non-trivial "deep boostrap" phenomenon (Nakkiran et al., 2020) during the second learning stage: the gap between the oracle world and empirical world test errors is negligible whereas the train and test errors exhibit a substantial gap. The gradient flow results for kernel regression can also provide insight into the deep bootstrap for random feature networks SGD as these results can approximately predict their behavior (see Section 4.2).

Remark 4 (Connection with Deep Learning Practice)**.** *Although we believe our results conceptually shed light on some of the interesting behaviors observed in the training dynamics of deep learning, due to our stylized setting we may not exactly see the predicted phenomena in practice. Accurately observing the three stages of kernel regression requires sufficiently high-dimensional data in order for the kernel eigenvalues to obey a staircase-like decay and for training to be sufficiently long as the time axis should be in log-scale. Our results hold for regression whereas for classification the empirical model may continue improving after classifying the train set correctly. Despite these caveats, certain conclusions can be observed in some realistic settings (c.f. Appendix E.4).*

4 Numerical Simulations

As mentioned previously, Figure 3 is a "cartoon" of the conclusions stated in Theorem 1 and 2. In this section, we verify the qualitative predictions of our theorems using synthetic data. Concretely, throughout this section we take $d = 400$ and $n = d^{1.5} = 8000$, and following our theoretical setup (c.f. Section 3.1) generate covariates $(\boldsymbol{x}_i)_{i \le n} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and responses $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$ with $(\varepsilon_i)_{i \le n} \sim_{iid} \mathcal{N}(0, \sigma_\varepsilon^2)$, for different choices of target function $f_\star$. All simulations in this section are for the noiseless case $\sigma_\varepsilon^2 = 0$ but a noisy example is given in Appendix F.

In Section 4.1, we simulate the gradient flows of kernel least-squares with dot product kernels and cyclic kernels (Figure 4) to reproduce the three stages as shown in Figure 3. In Section 4.2 we show that SGD of (dot product and cyclic) random-feature models (Figure 5) exhibit similar three stages phenomena, in which the second stage behaviors are consistent with the deep bootstrap phenomena observed in deep learning experiments (Nakkiran et al., 2020). Empirical quantities are averaged over 10 trials and the shaded regions indicate one standard deviation from the mean.

4.1 Gradient Flow of Kernel Least-Squares

Under the synthetic data set-up mentioned earlier, we first simulate the oracle world and empirical world errors curves of gradient flow dynamics using dot product kernels of the form

$$H(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}_{\boldsymbol{w} \sim \mathbb{S}^{d-1}}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_2 \rangle)] \tag{6}$$

for some activation function σ. We will examine a few different choices of $f_\star$ and kernel H, which are specified in the descriptions of each figure.

The oracle world error is computed analytically but the empirical world errors require sampling train and test datasets. We compute empirical world errors by averaging over 10 trials. The results are visualized both in log-scale and linear scale on the time axis. The log-scale plots allow for direct comparison with the cartoons in Figure 3. The linear scale plots are zoomed into the region $0 < t \le nd^{0.4}$ since 1) plotting the full interval squeeze all curves to the left boundary which is uninformative; 2) in practice one would not optimize for very long after interpolation.

In Figures 4a and 4b we take the target function to be a polynomial of the form

$$f_\star(\boldsymbol{x}) = a_0 \,\mathrm{He}_0(x_1) + \ldots + a_k \,\mathrm{He}_k(x_1), \quad \left\|\overline{\mathsf{P}}_j f_\star\right\|_{L^2}^2 \approx a_j^2 j!, \tag{7}$$

where $\mathrm{He}_i(t)$ is the ith Hermite polynomial (c.f. Appendix G.4) and the approximate equality in (7) holds in high-dimensions. In panel (4a) we consider use the ReLU activation function $\sigma(t) = \max(t, 0)$, and take $f_\star$ to be a quadratic polynomial with $(a_0, a_1, a_2) = (1/2, 1/\sqrt{2}, 1/\sqrt{8})$. With

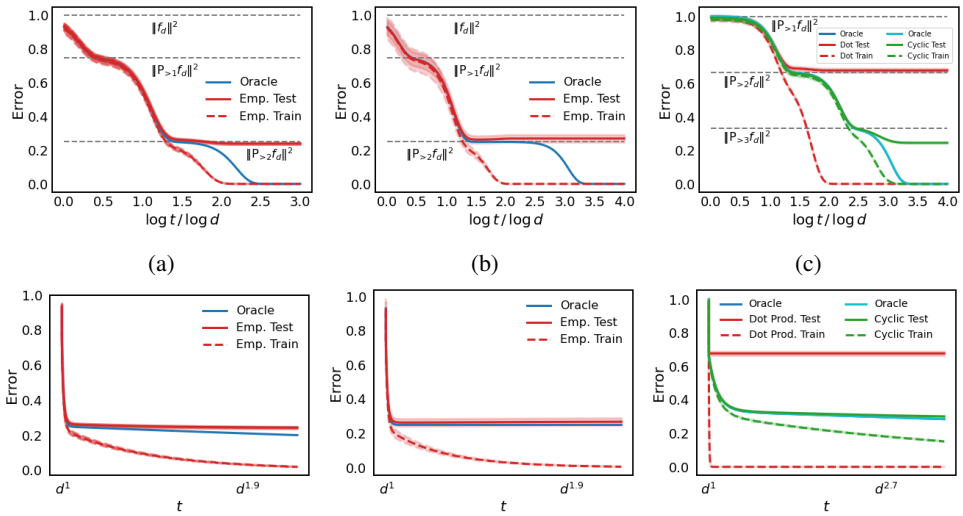


Figure 4: **Top row:** Log-scale plot of errors versus training time for dot product kernel (4a, 4b) and {dot product, cyclic} kernels in (4c). In (4a) $\sigma = \mathrm{ReLU}$ and $(a_0, a_1, a_2) = (1/2, 1/\sqrt{2}, 1/\sqrt{8})$. In (4b), $\sigma = \mathrm{ReLU} + 0.1\,\mathrm{He}_3$ and $(a_0, a_1, a_2, a_3) = (1/2, 1/\sqrt{2}, 0, 1/\sqrt{24})$. In (4c), $f_\star$ is Eq. (8) and $\sigma = \mathrm{ReLU} + 0.1\,\mathrm{He}_3$. **Bottom row:** Same as the top row but with linear scale time and zoomed in.

such a choice of parameters, we can see the three stages phenomenon. In panel (4b) we choose $f_\star$ to be a cubic polynomial with $(a_0, a_1, a_2, a_3) = (1/2, 1/\sqrt{2}, 0, 1/\sqrt{24})$ and $\sigma(t) = \max(t, 0) + 0.1\,\mathrm{He}_3(t)$ (we need the third Hermite coefficient of σ to be non-zero for stage 3 to occur). This choice of coefficients for $f_\star$ is such that $\|\overline{\mathsf{P}}_{>1} f_\star\|_{L^2}^2 \approx \|\overline{\mathsf{P}}_{>2} f_\star\|_{L^2}^2$, so that the second stage in (4b) is longer compared to (4a).

In Figure 4c, we take the target function to be a cubic cyclic polynomial

$$f_\star(\boldsymbol{x}) = \frac{1}{\sqrt{3d}} \left(\sum_{i=1}^{d} x_i + \sum_{i=1}^{d} x_i x_{i+1} + \sum_{i=1}^{d} x_i x_{i+1} x_{i+2} \right) \tag{8}$$

where the subindex addition in x_{i+k} is understood to be taken modulo d. We compare the performance of the dot product kernel H and its invariant version H_{inv} (c.f. Eq. (5)) with activation function $\sigma(t) = \max(t, 0) + 0.1\,\mathrm{He}_3(t)$. The kernel H_{inv} with n samples performs equivalently to H with nd samples (c.f. Remark 2), but is more computationally efficient since the size of the kernel matrix is still $n \times n$. Using H_{inv} elongates the first stage by a factor d, delaying the later stages and ensuring that the empirical world model improves longer.

Although in the simulations, the dimension d is not yet high enough to see a totally sharp staircase phenomenon as in the illustration Figure 3, even for at this d we are still able to clearly see the three predicted learning stages and deep bootstrap phenomenon across a range of settings. To better understand the effect of dimension we show similar plots with varying d in Appendix F.

4.2 SGD for Two-Layer Random-Feature Models

To make a better comparison with deep learning practice and connect to the deep bootstrap phenomena (Nakkiran et al., 2020), we simulate the error curves of SGD training on random-feature (RF) models (i.e. two-layer networks with random first-layer weights and trainable second-layer weights), in the same synthetic data setup as before. In particular, we look at dot product RF models

$$\hat{f}_{\text{dot}}(\boldsymbol{x}; \boldsymbol{a}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle), \qquad \boldsymbol{w}_i \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}),$$

and cyclic invariant RF models

$$\hat{f}_{\text{cyc}}(\boldsymbol{x}; \boldsymbol{a}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \int_{\mathcal{G}_d} \sigma(\langle \boldsymbol{w}_i, g \cdot \boldsymbol{x} \rangle) \pi_d(\mathrm{d}g), \qquad \boldsymbol{w}_i \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}).$$

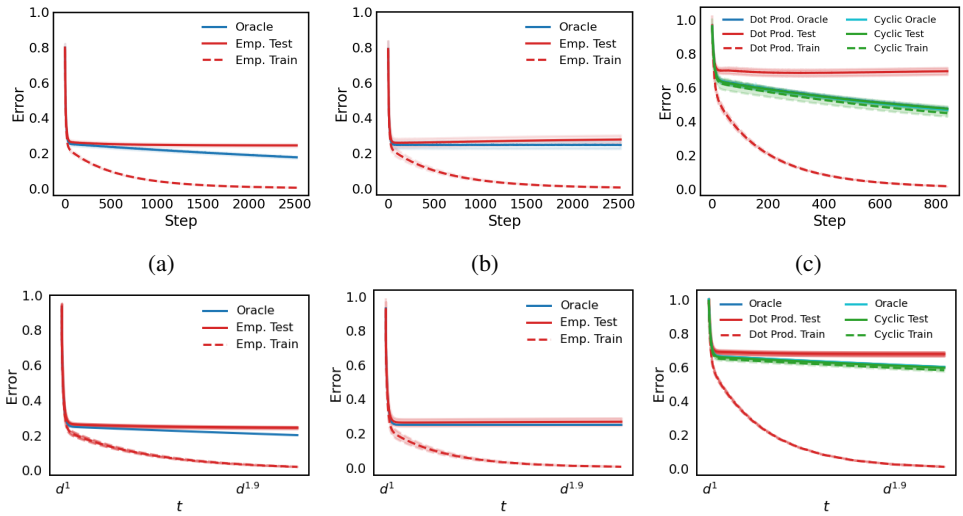


Figure 5: **Top Row:** SGD dynamics of random-feature models as described in Section 4.2. The target function and data distribution of (5a), (5b), (5c) are that of (4a), (4b), (4c) respectively. **Bottom Row:** The corresponding linear scale errors of kernel least-squares gradient flow for comparison.

For all following experiments we take the activation σ to be ReLU and $N = 4 \times 10^5 \approx n^{1.4}$.

For a given data distribution and RF model we train two fitted functions, one on a finite dataset (empirical world) and the other on the data distribution (oracle world). More specifically, the training of the empirical model is done using multi-pass SGD on a finite training set of size n with learning rate $\eta = 0.1$ and batch size $b = 50$. The training of the oracle model is done using online SGD with the same learning rate η and batch size b, but at each iteration a fresh batch is sampled from the population distribution. Both models $\hat{f}_t, f_t^{\text{or}}$ are initialized with $a_i = 0$ for $i \in [N]$. To speed up and stabilize optimization we use momentum $\beta = 0.9$. Note that if we took $N \to \infty$, $\eta \to 0$, and $\beta = 0$ we would be exactly in the dot product kernel gradient flow setting.

In Figure 5, the data generating distributions of panels (5a), (5b), (5c) are respectively the same as that of panels (4a), (4b), and (4c) (c.f. Section 4.1). The top row of Figure 5 shows SGD for {dot product, cyclic} RF models, and the bottom row shows the corresponding gradient flow for {dot product, cyclic} kernel least-squares. We see that the corresponding curves in these two rows exhibit qualitatively the same behaviors. Additionally, the results in panel (5c) show that as predicted, even for discrete SGD dynamics the dot product RF optimizes faster but fails to generalize, whereas the invariant RF optimizes slower but generalizes better.

5 Summary and Discussion

In this paper, we used precise asymptotics to study the oracle world and empirical world dynamics of gradient flow on kernel least-squares objectives for high-dimensional regression problems. Under reasonable conditions on the target function and kernel, we showed that in this setting there are three learning stages based on the behaviors of the empirical and oracle models and also connected our results to some empirical deep learning phenomena.

Although our setting already captures some interesting aspects of deep learning training dynamics, there are some limitations which would be interesting to resolve in future work. We require very high-dimensional data in order for the asymptotics to be accurate, but real data distributions have low-dimensional structure. We work in a limiting regime of neural network training where the dynamics are linear and the step-size is infinitesimal. It is an important direction to extend this analysis to the non-linear feature learning regime and to consider discrete step-size minibatch SGD, as these are considered important aspects of network training. Our results hold for the square-loss in regression problems, but many deep learning problems involve classification using cross-entropy loss. Lastly, our analysis holds specifically for gradient flow, so it would be also interesting to consider other iterative learning algorithms such as boosting.

REFERENCES

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32:7413–7424, 2019.

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 Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Alberto Bietti and Julien Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *arXiv preprint arXiv:1706.03078*, 2017.

Léon Bottou and Yann LeCun. Large scale online learning. *Advances in neural information processing systems*, 16:217–224, 2004.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1912.01198*, 2019.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1): 1–50, 2010.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *arXiv preprint arXiv:2010.15110*, 2020.

Jonathan Frankle, David J Schwab, and Ari S Morcos. The early phase of neural network training. *arXiv preprint arXiv:2002.10365*, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *arXiv preprint arXiv:2006.13409*, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234. PMLR, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pp. 2–47. PMLR, 2018.

Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 649–657. PMLR, 2021.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration. *arXiv preprint arXiv:2101.10588*, 2021a.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. *arXiv preprint arXiv:2102.13219*, 2021b.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019a.

Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in Neural Information Processing Systems*, 32:3496–3506, 2019b.

Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The deep bootstrap framework: Good online learners are good offline generalizers. *arXiv preprint arXiv:2010.08127*, 2020.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *arXiv preprint arXiv:1805.10074*, 2018.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1): 335–366, 2014.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32:2972–2983, 2019.

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

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

A GENERAL SETTING

In this section we present our theory for training dynamics of kernel regression in an abstract setting similar to that of Mei et al. (2021a). We first introduce the setting of interest, then state the relevant assumptions, and finally we provide our theoretical results. We provide proofs of these results in Appendix D.

A.1 PROBLEM SETUP

Consider a sequence of Polish probability spaces $(\mathcal{X}_d, \nu_d)$ indexed by an integer d. We denote by $L^2(\mathcal{X}_d) = L^2(\mathcal{X}_d, \nu_d)$ the space of square integrable functions on $(\mathcal{X}_d, \nu_d)$. For $p \geq 1$, we denote $\|f\|_{L^p(\mathcal{X})} = \mathbb{E}_{\boldsymbol{x}\sim\nu}[|f(\boldsymbol{x})^p|]^{1/p}$ the L^p norm of f. Let $\mathcal{D}_d \subseteq L^2(\mathcal{X}_d)$ be a closed linear subspace. In some simple applications we will consider $\mathcal{D}_d = L^2(\mathcal{X}_d)$, but the extra generality will be useful in certain applications.

We are concerned with a supervised learning problem where we are given i.i.d data $(y_i, \boldsymbol{x}_i)_{i\leq n}$. The feature vectors $\boldsymbol{x}_i \sim_{iid} \nu_d$ are in $\mathcal{X}_d$ and the empirical-valued noisy responses y_i are given by

$$y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$$

for some unknown target function $f_d \in \mathcal{D}_d$ and $\varepsilon_i \sim_{iid} \mathcal{N}(0, \sigma_\varepsilon^2)$.

We consider a general RKHS defined on $(\mathcal{X}_d, \nu_d)$ via the compact self-adjoint positive definite operator $\mathbb{H}_d : \mathcal{D}_d \to \mathcal{D}_d$ which admits the representation

$$\mathbb{H}_d g(\boldsymbol{x}) = \int_{\mathcal{X}_d} H_d(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') \nu_d(\mathrm{d}\boldsymbol{x}'),$$

where $H_d \in L^2(\mathcal{X}_d \times \mathcal{X}_d)$ with the property that $\int_{\mathcal{X}_d} H_d(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') \nu_d(\mathrm{d}\boldsymbol{x}') = 0$ for $g \in \mathcal{D}_d^\perp$.

By the spectral theorem of compact operators, there exists an orthonormal basis $(\psi_j)_{j\geq 1}$, $\mathrm{span}(\psi_j, j \geq 1) = \mathcal{D}_d \subseteq L^2(\mathcal{X}_d)$ and empirical eigenvalues $(\lambda_{d,j})_{j\geq 1}$ with nonincreasing absolute values $|\lambda_{d,1}| \geq |\lambda_{d,2}| \geq \cdots$ and $\sum_{j\geq 1} \lambda_{d,j}^2 < \infty$ such that

$$H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{\infty} \lambda_{d,j}^2 \psi_j(\boldsymbol{x}_1) \psi_j(\boldsymbol{x}_2)$$

where convergence holds in $L^2(\mathcal{X}_d)$.

For $S \subseteq \{1, 2, \ldots\}$ we denote P_S to be the projection operator from $L^2(\mathcal{X}_d)$ onto $\mathcal{D}_{d,S} := \mathrm{span}(\psi_j, j \in S)$. We denote $\mathbb{H}_{d,S}$ to be the operator

$$\mathbb{H}_{d,S} = \sum_{j=1}^{\infty} \lambda_{d,j}^2 \psi_j \psi_j^\star$$

and $H_{d,S}$ the corresponding kernel.

If $S = \{j \in \mathbb{N} : j \leq \ell\}$ we will write as short-hand $\mathbb{H}_{d,\leq \ell}$ and analogously for $S = \{j \in \mathbb{N} : j > \ell\}$. The trace of this operator is given by

$$\mathrm{Tr}(\mathbb{H}_{d,S}) \equiv \sum_{j\in S} \lambda_{d,j}^2 = \mathbb{E}_{\boldsymbol{x}\sim\nu_d}[H_{d,S}(\boldsymbol{x}, \boldsymbol{x})] < \infty.$$

A.2 GENERAL ASSUMPTIONS

We now state our assumptions on the kernel and the sequence of probability spaces $(\boldsymbol{X}_d, \nu_d)$.

Assumption 3 ($\{n(d), \mathsf{m}(d)\}_{d\geq 1}$-Kernel Concentration Property)**.** *We say that the sequence of operators $\{\mathbb{H}_d\}_{d\geq 1}$ satisfies the Kernel Concentration Property (KCP) with respect to the sequence $\{n(d), \mathsf{m}(d)\}_{d\geq 1}$ if there exists a sequence of integers $\{r(d)\}_{d\geq 1}$ with $r(d) \geq \mathsf{m}(d)$, such that the following conditions hold.*

(a) *(Hypercontractivity of finite eigenspaces.) For any fixed $q \geq 1$, there exists a constant C such that for any $h \in \mathcal{D}_{d,\leq r(d)} = \mathrm{span}(\psi_s, 1 \leq s \leq r(d))$, we have*

$$\|h\|_{L^{2q}} \leq C\|h\|_{L^2}.$$

(b) *(Properly decaying eigenvalues) There exists fixed δ_0, such that, for all d large enough,*

$$n(d)^{2+\delta_0} \leq \frac{(\sum_{j=r(d)+1}^{\infty} \lambda_{d,j}^4)^2}{\sum_{j=r(d)+1}^{\infty} \lambda_{d,j}^8},$$
$$n(d)^{2+\delta_0} \leq \frac{(\sum_{j=r(d)+1}^{\infty} \lambda_{d,j}^2)^2}{\sum_{j=r(d)+1}^{\infty} \lambda_{d,j}^4}.$$

(c) *(Concentration of diagonal elements of kernel) For $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim_{iid} \nu_d$, we have:*

$$\max_{i\in[n(d)]} \left|\mathbb{E}_{\boldsymbol{x}\sim\nu_d}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x}_i,\boldsymbol{x})^2] - \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\nu_d}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x},\boldsymbol{x}')^2]\right| = o_{d,\mathbb{P}}(1)\cdot \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\nu_d}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x},\boldsymbol{x}')^2],$$
$$\max_{i\in[n(d)]} \left|H_{d,>\mathsf{m}(d)}(\boldsymbol{x}_i,\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x},\boldsymbol{x})]\right| = o_{d,\mathbb{P}}(1)\cdot \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x},\boldsymbol{x})].$$

Assumption (a) can be interpreted as requiring that the top eigenfunctions of $\mathbb{H}_d$ are delocalized. Assumption (b) concerns the tail of eigenvalues of $\mathbb{H}_d$ and is very mild in high-dimensions. Lastly, (c) essentially requires that "most points" in $\mathcal{X}_d$ behave similarly in the sense of having similar values of the kernel diagonal $H_d(\boldsymbol{x},\boldsymbol{x})$.

Assumption 4 (Eigenvalue condition at level $\{(n(d),\mathsf{m}(d))\}_{d\geq 1}$)**.** *We say that the sequence of kernel operators $\{\mathbb{H}_d\}_{d\geq 1}$ satisfies the Eigenvalue Condition at level $\{(n(d),\mathsf{m}(d))\}_{d\geq 1}$ if the following conditions hold for all d large enough*

(a) *There exists a fixed $\delta_0 > 0$, such that*

$$n(d)^{1+\delta_0} \leq \frac{1}{\lambda_{d,\mathsf{m}(d)+1}^4} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda_{d,k}^4,$$
$$n(d)^{1+\delta_0} \leq \frac{1}{\lambda_{d,\mathsf{m}(d)+1}^2} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda_{d,k}^2.$$

(b) *There exists a fixed $\delta_0 > 0$, such that*

$$n(d)^{1-\delta_0} \geq \frac{1}{\lambda_{d,\mathsf{m}(d)}^2} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda_{d,k}^2.$$

(c) *There exists a constant C such that*

$$\sum_{k=1}^{\mathsf{m}(d)} \lambda_{d,k}^2 \leq C \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda_{d,k}^2.$$

(d) *There exists a fixed $\delta_0 > 0$, such that*

$$\mathsf{m}(d) \leq n(d)^{1-\delta_0}.$$

Assumptions (a) and (b) can be seen as a spectral gap assumption. This ensures a clear separation between the subspace $\mathcal{D}_{d,\leq \mathsf{m}(d)}$ which is estimated accurately and the subspace $\mathcal{D}_{d,>\mathsf{m}(d)}$ that is estimated trivially by 0. Assumption (c) requires that the tail does not decay too abruptly. As we will see later, these assumptions hold for several high-dimensional examples. The technical requirement (d) is mild and we do not know of any interesting counterexamples.

In the asymptotic setting, we will be interested in understanding the model learned at a time $t = t(d)$ which scales with the dimension. The next set of assumptions give requirements for a valid scaling.

Assumption 5 (Admissible Time at $\{t(d), \mathsf{u}(d), n(d), \mathsf{m}(d)\}_{d\geq 1}$)**.** *We say that the sequence $\{t(d), \mathsf{u}(d), n(d), \mathsf{m}(d)\}_{d\geq 1}$ is an Admissible Time for the sequence of kernel operators $\{\mathbb{H}_d\}_{d\geq 1}$ if the following conditions hold*

(a) There exists a fixed $\delta_0 > 0$, such that for d large enough

$$\frac{1}{\lambda_{d,\mathsf{u}(d)}^2} \leq t(d)^{1-\delta_0} \leq t(d)^{1+\delta_0} \leq \frac{1}{\lambda_{d,\mathsf{u}(d)+1}^2}.$$

(b) If $\mathsf{u}(d) < \mathsf{m}(d)$ for infinitely many d, then

$$\frac{t(d)}{n(d)} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda_{d,k}^2 = o_d(1).$$

Assumption (a) is similar to the spectral gap condition Assumption 4 (a), (b) for $(t(d), \mathsf{u}(d))_{d\geq 1}$. Assumption (b) relates the ordering of the indices $\mathsf{u}(d), \mathsf{m}(d)$ to the relative growth of $t(d), n(d)$.

A.3 Main Results

In this section we give the main theoretical results. The first result characterizes the oracle model.

Theorem 6 (Oracle World)**.** *Let $\{f_d \in \mathcal{D}_d\}_{d\geq 1}$ be a sequence of functions and $\{\mathbb{H}_d\}_{d\geq 1}$ be a sequence of kernel operators such that $\{(\mathbb{H}_d, t(d), \mathsf{u}(d))\}_{d\geq 1}$ satisfies Assumption 5(a), then*

$$R(f_t^{\mathrm{or}}) = \left\|\mathsf{P}_{>\mathsf{u}(d)} f_d\right\|_{L^2}^2 + \sigma_\varepsilon^2 + o_d(1) \cdot \|f_d\|_{L^2}^2 ,$$
$$\left\|f_t^{\mathrm{or}} - \mathsf{P}_{\leq \mathsf{u}(d)} f_d\right\|_{L^2}^2 = o_d(1) \cdot \|f_d\|_{L^2}^2 .$$

The error of the oracle model is determined solely by optimization time t through $\mathsf{u}(d)$. Due to the spectral gap assumption 5(a) learning only occurs along the top $\mathsf{u}(d)$ eigenfunctions.

The next results describe the empirical model. First we characterize the training error.

Theorem 7 (Empirical World - Train)**.** *Let $\{f_d \in \mathcal{D}_d\}_{d\geq 1}$ be a sequence of functions, $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim \nu_d$ independently, and $\{\mathbb{H}_d\}_{d\geq 1}$ be a sequence of kernel operators such that $\{(\mathbb{H}_d, n(d), \mathsf{m}(d), t(d), \mathsf{u}(d))\}_{d\geq 1}$ satisfies $\{(n(d), \mathsf{m}(d))\}_{d\geq 1}$-KPCP (Assumption 3), eigenvalue condition at level $\{(n(d), \mathsf{m}(d))\}_{d\geq 1}$ (Assumption 4), and $\{(n(d), \mathsf{m}(d), t(d), \mathsf{u}(d))\}_{d\geq 1}$ is a valid time (Assumption 5). Define $\kappa_H = \mathrm{Tr}\big(\mathbb{H}_{d,>\mathsf{m}(d)}\big)$ and $\ell(d) = \min\{\mathsf{u}(d), \mathsf{m}(d)\}$. Then for any $\eta > 0$ we have,*

$$\widehat{R}_n(\hat{f}_t) = \left\|\mathsf{P}_{>\ell(d)} f_d\right\|_{L^2}^2 + \sigma_\varepsilon^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2), \qquad \text{if } t = o_d(n/\kappa_H)$$
$$\widehat{R}_n(\hat{f}_t) = o_{d,\mathbb{P}}(1) \cdot (\|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2) \qquad \text{if } t = \omega_d(n/\kappa_H).$$

In the early-time regime $t \ll n/\kappa_H$ the training error may be non-zero and matches the oracle world error if also $\mathsf{u}(d) \leq \mathsf{m}(d)$. In the late-time regime $t \gg n/\kappa_H$ the training error is negligible and the model interpolates the training set. The quantity κ_H arises since the empirical kernel matrix can be decomposed as $\boldsymbol{H} = \boldsymbol{H}_{\leq \mathsf{m}} + \boldsymbol{H}_{>\mathsf{m}}$ and the second component is approximately a multiple of the identity: $\boldsymbol{H}_{>\mathsf{m}} \approx \mathrm{Tr}(\boldsymbol{H}_{>\mathsf{m}}) \cdot \mathbf{I}_n = \kappa_H \cdot \mathbf{I}_n$. This term acts as a self-induced ridge-regularizer.

Our final result characterizes the test error of the empirical model.

Theorem 8 (Empirical World - Test)**.** *Let $\{f_d \in \mathcal{D}_d\}_{d\geq 1}$ be a sequence of functions, $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim \nu_d$ independently, and $\{\mathbb{H}_d\}_{d\geq 1}$ be a sequence of kernel operators such that $\{(\mathbb{H}_d, n(d), \mathsf{m}(d), t(d), \mathsf{u}(d))\}_{d\geq 1}$ satisfies $\{(n(d), \mathsf{m}(d))\}_{d\geq 1}$-KPCP (Assumption 3), eigenvalue condition at level $\{(n(d), \mathsf{m}(d))\}_{d\geq 1}$ (Assumption 4), and $\{(n(d), \mathsf{m}(d), t(d), \mathsf{u}(d))\}_{d\geq 1}$ is a valid time (Assumption 5). Define $\ell(d) = \min\{\mathsf{u}(d), \mathsf{m}(d)\}$. Then for any $\eta > 0$ we have,*

$$R(\hat{f}_t) = \left\|\mathsf{P}_{>\ell(d)} f_d\right\|_{L^2}^2 + \sigma_\varepsilon^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2),$$
$$\left\|\hat{f}_t - \mathsf{P}_{\leq \ell(d)} f_d\right\|_{L^2}^2 = o_{d,\mathbb{P}}(1) \cdot (\|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2).$$

The result shows that the empirical model is essentially the projection of the regression function onto the first $\ell(d)$ eigenfunctions. The quantity $\ell(d)$ controls the complexity of the model which increases with time t up until $\mathsf{u}(d) \geq \mathsf{m}(d)$ after which the complexity is limited by n.

B DOT PRODUCT KERNELS ON $\mathbb{S}^{d-1}(\sqrt{d})$

B.1 SETTING

We now apply our general theorems to the setting of dot product kernels on the sphere. Concretely we take $\mathcal{X}_d = \mathbb{S}^{d-1}(\sqrt{d})$ and $\nu_d = \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and consider dot product kernels H_d which take the form of Eq. (4). Note that any dot product kernel h_d can be decomposed as

$$h_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d) = \mathbb{E}_{\boldsymbol{x}\sim \mathrm{Unif}(\mathbb{S}^{d-1})}[\sigma_d(\langle \boldsymbol{w}, \boldsymbol{x}_1\rangle)\sigma_d(\langle \boldsymbol{w}, \boldsymbol{x}_2\rangle)] \tag{9}$$

for some activation function σ_d, which amounts to taking the square root of the positive semidefinite kernel operator associated to h_d. for some activation function σ_d. We state mild assumptions on σ_d and show that under these conditions we can apply the results in Section A.3.

B.2 ASSUMPTIONS

We state our assumptions on σ_d after giving some definitions. See App. G for additional background. Denote by $\overline{\mathsf{P}}_{\leq \ell}$ the orthogonal projection onto the subspace of $L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ spanned by polynomials of degree less than or equal to ℓ. The projectors $\overline{\mathsf{P}}_{\ell}$ and $\overline{\mathsf{P}}_{>\ell}$ are defined analogously. Let us emphasize that the projectors $\overline{\mathsf{P}}_{\leq \ell}$ are related but distinct from the $\mathsf{P}_{\leq \mathsf{m}}$: while $\overline{\mathsf{P}}_{\leq \ell}$ projects onto the eigenspace of polynomials of degree at most ℓ, $\mathsf{P}_{\leq \mathsf{m}}$ projects onto the top m-eigenfunctions.

Assumption 9 (Assumptions for Dot Product Kernels at level $\mathsf{s} \in \mathbb{N}$)**.** *Let $\{H_d\}_{d\geq 1}$ be a sequence of dot product kernels with associated activation functions $\{\sigma_d\}_{d\geq 1}$ as in equation 9. We assume the following hold*

(a) There exists an integer k and constants $c_1 < 1$ and $c_0 > 0$, such that $|\sigma_d(x)| \leq c_0 \exp\big(c_1 x^2/(4k)\big)$.

(b) We have

$$\min_{k\leq \mathsf{s}} d^{\mathsf{s}-k} \left\|\overline{\mathsf{P}}_k \sigma_d(\langle \boldsymbol{e}, \cdot\rangle)\right\|_{L^2}^2 = \Omega_d(1),$$
$$\left\|\overline{\mathsf{P}}_{>2\mathsf{s}+1} \sigma_d(\langle \boldsymbol{e}, \cdot\rangle)\right\|_{L^2}^2 = \Omega_d(1).$$

where $\boldsymbol{e} \in \mathbb{S}^{d-1}$ is a fixed vector (it is easy to see that these quantities do not depend on $\boldsymbol{e}$).

We now verify that the above assumptions fall under the conditions given in Appendix A.2 for an appropriate choice of $(t(d), \mathsf{u}(d), n(d), \mathsf{m}(d))$ according to Theorem 1. Consider $d^{\mathsf{j}+\delta_0} \leq t \leq d^{\mathsf{j}+1-\delta_0}$ and $d^{\mathsf{s}+\delta_0} \leq n \leq d^{\mathsf{s}+1-\delta_0}$. We set u and m to be the number of eigenvalues associated to spherical harmonics of degree less than or equal to j and s respectively

$$\mathsf{u} = \sum_{k=0}^{\mathsf{j}} B(\mathbb{S}^{d-1}; k) = \Theta_d(d^{\mathsf{j}}), \quad \mathsf{m} = \sum_{k=0}^{\mathsf{s}} B(\mathbb{S}^{d-1}; k) = \Theta_d(d^{\mathsf{s}}).$$

The verification of the Kernel Concentration Property 3 at level $\{(n(d), \mathsf{m}(d)\}$ is exactly the same as the treatment in Theorem 2 of Mei et al. (2021a). The verification of the Eigenvalue Condition is also virtually the same, except for the extra condition (c) which is easily seen to hold since the trace of the kernel operator $\mathrm{Tr}(\mathbb{H}_d) = \Theta_d(1)$. Lastly, we must verify Assumption 5. To see part (a), note that $1/\lambda_{d,\mathsf{u}(d)}^2 = \Theta_d(d^{\mathsf{j}})$ and $1/\lambda_{d,\mathsf{u}(d)}^2 = \Theta_d(d^{\mathsf{j}+1})$. For part (b), the condition holds because $\mathsf{u}(d) < \mathsf{m}(d)$ for large d if and only if $\mathsf{j} < \mathsf{s}$.

C GROUP INVARIANT KERNELS ON $\mathbb{S}^{d-1}(\sqrt{d})$

C.1 SETTING

We now apply our general theorems to the setting of group invariant kernels on the sphere. Concretely we take $\mathcal{X}_d = \mathbb{S}^{d-1}(\sqrt{d})$ and $\nu_d = \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and consider kernels H_d which take the form of Eq. (5) for some activation function σ_d. We state mild assumptions on σ_d and show that under these conditions we can apply the results in Appendix A.3. For additional technical background refer to Appendix G.

C.2 ASSUMPTIONS

We will make the following assumption on σ.

Assumption 10 (Assumption on Group Invariant Kernel at level s)**.** *For $\sigma : \mathbb{R} \to \mathbb{R}$, we assume the following conditions hold*

(a) For $\mathcal{G}_d = \mathrm{Cyc}_d$, we assume σ to be $(\mathsf{s}+1) \vee 3$ differentiable and there exists constants $c_0 > 0$ and $c_1 < 1$ such that $|\sigma^{(k)}| \le c_0 e^{c_1 u^2/2}$ for any $2 \le k \le (\mathsf{s}+1) \vee 3$.

For general $\mathcal{G}_d$, we assume that σ_d is a (finite degree) polynomial function.

(b) The Hermite coefficients $\mu_k(\sigma)$ (c.f. Appendix) verify $\mu_k \neq 0$ for any $0 \le k \le \mathsf{s}$.

(c) We assume that σ is not a polynomial with degrees less than or equal to s.

We now verify that the above assumptions fall under the conditions given in Appendix A.2 for an appropriate choice of $(t(d), \mathsf{u}(d), n(d), \mathsf{m}(d))$ according to Theorem 2. Consider $d^{\mathsf{j}+\delta_0} \le t \le d^{\mathsf{j}+1-\delta_0}$ and $d^{\mathsf{s}-\alpha+\delta_0} \le n \le d^{\mathsf{s}-\alpha+1-\delta_0}$. We set u and m to be the number of eigenvalues invariant polynomials of degree less than or equal to j and s respectively

$$\mathsf{u} = \sum_{k=0}^{\mathsf{j}} D(\mathbb{S}^{d-1}; k) = \Theta_d(d^{\mathsf{j}-\alpha}), \quad \mathsf{m} = \sum_{k=0}^{\mathsf{s}} D(\mathbb{S}^{d-1}; k) = \Theta_d(d^{\mathsf{s}-\alpha}).$$

The verification of the Kernel Concentration Property 3 at level $\{(n(d), \mathsf{m}(d)\}$ is exactly the same as the treatment in Theorem 1 in Mei et al. (2021b). The verification of the Eigenvalue Condition is also virtually the same, except for the extra condition (c) which can be seen from the fact that $\mathrm{Tr}\big(\mathbb{H}_{d,>\mathsf{m}(d)}\big) = \Theta(d^{\mathsf{s}-\alpha})$ and

$$\sum_{j=0}^{\mathsf{m}} \lambda_{d,j}^2 = \sum_{k=0}^{\mathsf{s}} \xi_{d,k}^2 D(\mathbb{S}^{d-1}; k) = \Theta(\mathsf{s} d^{-\alpha}) = \Theta(d^{-\alpha})$$

from which it follows that

$$\sum_{j=0}^{\mathsf{m}} \lambda_{d,j}^2 \le C \sum_{j>\mathsf{m}}^{\infty} \lambda_{d,j}^2$$

for a constant C.

Lastly, we must verify Assumption 5. To see part (a), note that $\frac{1}{\lambda_{d,\mathsf{u}(d)}^2} = \Theta_d(d^{\mathsf{j}})$ and $\frac{1}{\lambda_{d,\mathsf{u}(d)}^2} = \Theta_d(d^{\mathsf{j}+1})$. For part (b), the condition holds because $\mathsf{u}(d) < \mathsf{m}(d)$ for large d if and only if $\mathsf{j} < \mathsf{s}$ in which case

$$\frac{t(d)}{n(d)} \mathrm{Tr}\big(\mathbb{H}_{d,\mathsf{m}(d)}\big) \le \frac{d^{\mathsf{j}+1-\delta_0}}{d^{\mathsf{s}-\alpha+\delta_0}} \Theta(d^{-\alpha}) = O(d^{-2\delta_0} d^{\mathsf{j}-\mathsf{s}+1}) = o_d(1).$$

D PROOF OF GENERAL SETTING

D.1 ORACLE WORLD - PROOF OF THEOREM 6

The oracle model ODE Eq. (2) with initialization $f_0^{\mathrm{or}} \equiv 0$ can be solved (c.f. Appendix E.1) to yield the solution

$$f_t^{\mathrm{or}} = f_d - e^{-t\mathbb{H}_d} f_d.$$

Therefore the excess risk is given by

$$R(f_t^{\mathrm{or}}) - \sigma_\varepsilon^2 = \mathbb{E}_{\boldsymbol{x}\sim\nu_d}(f_d(\boldsymbol{x}) - f_t^{\mathrm{or}}(\boldsymbol{x}))^2 = \int_{\mathcal{X}_d} f_d(\boldsymbol{x}) e^{-2t\mathbb{H}_d} f_d(\boldsymbol{x}) \nu_d(\mathrm{d}\boldsymbol{x}) = \sum_{k=0}^{\infty} \exp\big(-2t\lambda_{d,k}^2\big) \hat{f}_k^2.$$

where $\lambda_{d,k}^2$ are the kernel eigenvalues and $\hat{f}_k := \langle f_d, \psi_k \rangle$ are the Fourier coefficients of f_d in the kernel eigenbasis (c.f. Appendix A.1). We can control this quantity as follows,

$$\begin{aligned}
|R(f_t^{\mathrm{or}}) - \sigma_\varepsilon^2 - \|\mathsf{P}_{>\mathsf{u}} f_d\|_{L^2}^2 |/ \|f_d\|_{L^2}^2 &\le \max\left\{ \max_{k\le \mathsf{u}} \exp\big(-2t\lambda_{d,k}^2\big), \max_{k\ge \mathsf{u}+1} 1 - \exp\big(-2t\lambda_{d,k}^2\big) \right\} \\
&\le \max\left\{ \max_{k\le \mathsf{u}} \exp\big(-\Omega(t\lambda_{d,k}^2)\big), \max_{k\ge \mathsf{u}+1} O(t\lambda_{d,k}^2) \right\} \\
&\le \max\left\{ \exp\big(-\Omega(t^{\delta_0})\big), O(t^{-\delta_0}) \right\} = o_d(1)
\end{aligned}$$

where the last inequality follows from Assumption 5(a). This shows the first theorem statement,

$$\|f_t^{\mathrm{or}} - \mathsf{P}_{\le \mathsf{u}} f_d\|_{L^2}^2 = o_d(1) \cdot \|f_d\|_{L^2}^2 .$$

Now observe that

$$\mathsf{P}_{\le u} f_d - f_t^{\mathrm{or}} = \sum_{k=0}^{\mathsf{u}} e^{-t\lambda_{d,k}^2} \hat{f}_k \psi_k + \sum_{k\ge \mathsf{u}+1} (1 - e^{-t\lambda_{d,k}^2}) \hat{f}_k \psi_k$$

hence

$$\begin{aligned}
\|f_t^{\mathrm{or}} - \mathsf{P}_{\le u} f_d\|_{L^2}^2 / \|f_d\|_{L^2}^2 &\le \max\left\{ \max_{k\le \mathsf{u}} \exp\big(-2t\lambda_{d,k}^2\big), \max_{k\ge \mathsf{u}+1} (1 - e^{-t\lambda_{d,k}^2})^2 \right\} \\
&\le \max\left\{ \max_{k\le \mathsf{u}} \exp\big(-2t\lambda_{d,k}^2\big), \max_{k\ge \mathsf{u}+1} 1 - \exp\big(-2t\lambda_{d,k}^2\big) \right\} = o_d(1)
\end{aligned}$$

where the second inequality follows from the fact that $(1 - e^{-x})^2 \le 1 - e^{-2x}$ and the final equality is from the proof of the first part of the theorem. Thus,

$$\|f_t^{\mathrm{or}} - \mathsf{P}_{\le u} f_d\|_{L^2}^2 = o_d(1) \cdot \|f_d\|_{L^2}^2$$

completing the proof.

D.2 EMPIRICAL WORLD – PRELIMINARIES

We will introduce some useful notations and the objects of analysis for studying the empirical world.

D.2.1 TRAINING DYNAMICS

For the training of the empirical world, if $\boldsymbol{u}(t) = (\hat{f}_t(\boldsymbol{x}_1), \dots, \hat{f}_t(\boldsymbol{x}_n)) \in \mathbb{R}^n$ then from Eq. 3 we see that

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}(t) = -\frac{1}{n}\sum_{i=1}^{n} H_d(\boldsymbol{x}, \boldsymbol{x}_i)(f_t(\boldsymbol{x}_i) - y_i)$$

which has the following solution for $\boldsymbol{u}(0) = \mathbf{0}$

$$\boldsymbol{u}(t) = \boldsymbol{y} - e^{-t\boldsymbol{H}/n}\boldsymbol{y} = (\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{y}, \tag{10}$$

where $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ with the (i,j)'th element given by $H_{ij} = H_d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon} \in \mathbb{R}^{n\times n}$ with $\boldsymbol{f} = (f_d(\boldsymbol{x}_1), \dots, f_d(\boldsymbol{x}_n))^\mathsf{T} \in \mathbb{R}^n$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\mathsf{T} \in \mathbb{R}^n$.

The training and test errors can be written as

$$\widehat{R}_n(\hat{f}_t) = \frac{1}{n}\|\boldsymbol{u}(t) - \boldsymbol{y}\|_2^2 = \frac{1}{n}\boldsymbol{y}^\mathsf{T} e^{-2t\boldsymbol{H}/n}\boldsymbol{y},$$
$$R(\hat{f}_t) = \mathbb{E}_{\boldsymbol{x}}\big[(f(\boldsymbol{x}) - \boldsymbol{u}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{h}(\boldsymbol{x}))^2\big].$$

Expanding $R(\hat{f}_t)$ yields

$$R(\hat{f}_t) = \mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})^2] - 2\boldsymbol{u}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{E} + \boldsymbol{u}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}\boldsymbol{u}(t)$$

where $\boldsymbol{E} = (E_1, \dots, E_n)^\mathsf{T} \in \mathbb{R}^n$, $\boldsymbol{M} = (M_{ij})_{ij\in[n]} \in \mathbb{R}^{n\times n}$, and $\boldsymbol{H} = (H_{ij})_{ij\in[n]} \in \mathbb{R}^{n\times n}$ with

$$E_i = \mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})H_d(\boldsymbol{x}, \boldsymbol{x}_i)],$$
$$M_{ij} = \mathbb{E}_{\boldsymbol{x}}[H_d(\boldsymbol{x}, \boldsymbol{x}_i)H_d(\boldsymbol{x}, \boldsymbol{x}_j)],$$
$$H_{ij} = H_d(\boldsymbol{x}_i, \boldsymbol{x}_i).$$

D.2.2 Decompositions

In this section we recall some useful decompositions of empirical quantities from Mei et al. (2021a). As mentioned earlier the eigendecomposition of H_d is given by

$$H_d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{\infty} \lambda_{d,k}^2 \psi_k(\boldsymbol{x})\psi_k(\boldsymbol{y}).$$

We write the orthogonal decomposition of f_d in the basis $\{\psi_k\}_{k\geq 1}$ as

$$f_d(\boldsymbol{x}) = \sum_{k=1}^{\infty} \hat{f}_{d,k}\psi_k(\boldsymbol{x}).$$

Define

$$\boldsymbol{\psi}_k = (\psi_k(\boldsymbol{x}_1), \dots, \psi_k(\boldsymbol{x}_n))^\mathsf{T} \in \mathbb{R}^n,$$
$$\boldsymbol{D}_{\leq \mathsf{m}} = \mathrm{diag}(\lambda_{d,1}, \lambda_{d,2}, \dots, \lambda_{d,\mathsf{m}}) \in \mathbb{R}^{\mathsf{m}\times\mathsf{m}},$$
$$\boldsymbol{\Psi}_{\leq \mathsf{m}} = (\psi_k(\boldsymbol{x}_i))_{i\in[n], k\in[\mathsf{m}]} \in \mathbb{R}^{n\times\mathsf{m}},$$
$$\widehat{\boldsymbol{f}}_{\leq \mathsf{m}} = (\hat{f}_{d,1}, \hat{f}_{d,2}, \dots, \hat{f}_{d,\mathsf{m}})^\mathsf{T} \in \mathbb{R}^{\mathsf{m}}.$$

We have the following orthogonal basis decompositions of $\boldsymbol{f}$, $\boldsymbol{H}$, $\boldsymbol{E}$ and $\boldsymbol{M}$

$$\boldsymbol{f} = \widehat{\boldsymbol{f}}_{\leq \mathsf{m}} + \widehat{\boldsymbol{f}}_{>\mathsf{m}}, \qquad \widehat{\boldsymbol{f}}_{\leq \mathsf{m}} = \boldsymbol{\Psi}_{\leq \mathsf{m}}\widehat{\boldsymbol{f}}_{\leq \mathsf{m}}, \qquad \widehat{\boldsymbol{f}}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_{d,k}\boldsymbol{\psi}_k,$$

$$\boldsymbol{H} = \boldsymbol{H}_{\leq \mathsf{m}} + \boldsymbol{H}_{>\mathsf{m}}, \qquad \boldsymbol{H}_{\leq \mathsf{m}} = \boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^2\boldsymbol{\Psi}_{\leq \mathsf{m}}^\mathsf{T}, \qquad \boldsymbol{H}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \lambda_k^2\boldsymbol{\psi}_k\boldsymbol{\psi}_k^\mathsf{T},$$

$$\boldsymbol{E} = \boldsymbol{E}_{\leq \mathsf{m}} + \boldsymbol{E}_{>\mathsf{m}}, \qquad \boldsymbol{E}_{\leq \mathsf{m}} = \boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^2\widehat{\boldsymbol{f}}_{\leq \mathsf{m}}, \qquad \boldsymbol{E}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \lambda_k^2\hat{f}_{d,k}\boldsymbol{\psi}_k,$$

$$\boldsymbol{M} = \boldsymbol{M}_{\leq \mathsf{m}} + \boldsymbol{M}_{>\mathsf{m}}, \qquad \boldsymbol{M}_{\leq \mathsf{m}} = \boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^4\boldsymbol{\Psi}_{\leq \mathsf{m}}^\mathsf{T}, \qquad \boldsymbol{M}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \lambda_k^4\boldsymbol{\psi}_k\boldsymbol{\psi}_k^\mathsf{T}.$$

By Lemma 17, under Assumptions 3 and 4(a) the matrices $\boldsymbol{H}$ and $\boldsymbol{M}$ can be written as

$$\boldsymbol{H} = \boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^2\boldsymbol{\Psi}_{\leq \mathsf{m}}^\mathsf{T} + \kappa_H(\mathbf{I}_n + \boldsymbol{\Delta}_H), \tag{11}$$
$$\boldsymbol{M} = \boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^4\boldsymbol{\Psi}_{\leq \mathsf{m}}^\mathsf{T} + \kappa_M(\mathbf{I}_n + \boldsymbol{\Delta}_M), \tag{12}$$

where

$$\kappa_H = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \sum_{k\geq \mathsf{m}+1}^{\infty} \lambda_{d,k}^2,$$
$$\kappa_M = \mathrm{Tr}\big(\mathbb{H}_{d,>\mathsf{m}}^2\big) = \sum_{k\geq \mathsf{m}+1}^{\infty} \lambda_{d,k}^4,$$

and

$$\max\{\|\boldsymbol{\Delta}_H\|_{\mathrm{op}}, \|\boldsymbol{\Delta}_M\|_{\mathrm{op}}\} = o_{d,\mathbb{P}}(1).$$

We can also write

$$\boldsymbol{f}_{\leq \mathsf{m}} = \boldsymbol{\Psi}_{\leq \mathsf{m}} \widehat{\boldsymbol{f}}_{\leq \mathsf{m}},$$
$$\boldsymbol{f}_{>\mathsf{m}} = \sum_{k=\mathsf{m}+1}^{\infty} \hat{f}_{d,k} \boldsymbol{\psi}_k.$$

For a vector or matrix $\boldsymbol{X}$ we will use the notation $\boldsymbol{X}_{\ell\mathsf{m}} := \boldsymbol{X}_{\leq \mathsf{m}} - \boldsymbol{X}_{>\ell}$ for $\ell \leq \mathsf{m}$.

We also introduce the shrinkage matrix defined as

$$\boldsymbol{S}_{\leq \mathsf{m}} = \Big(\mathbf{I}_{\mathsf{m}} + \frac{\kappa_H}{n} \boldsymbol{D}_{\leq \mathsf{m}}^{-2}\Big)^{-1} = \mathrm{diag}((s_j)_{j\in[\mathsf{m}]}) \in \mathbb{R}^{\mathsf{m}\times\mathsf{m}}, \quad \text{where } s_j = \frac{\lambda_{d,j}^2}{\lambda_{d,j}^2 + \frac{\kappa_H}{n}}. \tag{13}$$

D.2.3 Auxiliary Lemmas

Here we collect some lemmas which will be of use to us.

Lemma 11 (Matrix Exponential Perturbation Inequality)**.** *For matrix operator norm $\|\cdot\|$, if $\boldsymbol{A}, \boldsymbol{B}$ are symmetric then*

$$\big\|e^{\boldsymbol{A}} - e^{\boldsymbol{B}}\big\| \leq \|\boldsymbol{A} - \boldsymbol{B}\| \max\{\big\|e^{\boldsymbol{A}}\big\|, \big\|e^{\boldsymbol{B}}\big\|\}.$$

For general $\boldsymbol{A}$ and $\boldsymbol{B}$ we have

$$\big\|e^{\boldsymbol{A}} - e^{\boldsymbol{B}}\big\| \leq \|\boldsymbol{A} - \boldsymbol{B}\| e^{\|\boldsymbol{A}\|} e^{\|\boldsymbol{B}\|}.$$

Lemma 12 (Matrix Exponential Identity)**.** *Let $\boldsymbol{Z} \in \mathbb{R}^{m\times n}, \boldsymbol{\Psi} \in \mathbb{R}^{n\times p}, \boldsymbol{D} \in \mathbb{R}^{p\times p}$ and $t \in \mathbb{R}$. Denote $\boldsymbol{A} = \boldsymbol{Z}^\mathsf{T}\boldsymbol{\Psi}$ and $\boldsymbol{B} = \boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}$. Then,*

$$\boldsymbol{Z}^\mathsf{T} e^{t\boldsymbol{\Psi}\boldsymbol{D}\boldsymbol{\Psi}^\mathsf{T}} \boldsymbol{\Psi} = \boldsymbol{A} e^{t\boldsymbol{D}\boldsymbol{B}}.$$

The following Lemmas 13-18 all use the definitions given in Section D.2.2.

Lemma 13 (Lemma 12 from Mei et al. (2021a))**.**

$$\big\|n\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1} - \boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{S}_{\leq \mathsf{m}}^2\boldsymbol{\Psi}_{\leq \mathsf{m}}^\mathsf{T}/n\big\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Lemma 14 (Theorem 6.(b) from Mei et al. (2021a))**.**

$$\big\|\boldsymbol{\Psi}_{\leq \mathsf{m}}^\mathsf{T}\boldsymbol{\Psi}_{\leq \mathsf{m}}/n - \mathbf{I}_n\big\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Lemma 15 (Lemma 13 from Mei et al. (2021a))**.**

$$\big\|\boldsymbol{\Psi}_{\leq \mathsf{m}}^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^2 - \boldsymbol{S}_{\leq \mathsf{m}}\big\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Lemma 16. *Let S, T be disjoint subsets of $\mathbb{N}$. Let $\boldsymbol{D} \in \mathbb{R}^{|S|\times|S|}$ be a diagonal matrix. Then we have that for any $\eta > 0$, there exists $C(\eta)$ (that is independent of d,*

$$\mathbb{E}[\widehat{\boldsymbol{f}}_T^\mathsf{T}\boldsymbol{\Psi}_T^\mathsf{T}\boldsymbol{\Psi}_S\boldsymbol{D}\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_T\widehat{\boldsymbol{f}}_T]/n^2 \leq C(\eta)\, \|\mathsf{P}_T f_d\|_{L^{2+\eta}}^2\, \mathrm{Tr}(\boldsymbol{D})/n.$$

Proof. Let $i : S \to [|S|]$ be the bijection such that gives the index of the column corresponding to ψ_s, then we have

$$\begin{aligned}
\mathbb{E}[\widehat{\boldsymbol{f}}_T^{\mathsf{T}} \boldsymbol{\Psi}_T^{\mathsf{T}} \boldsymbol{\Psi}_S \boldsymbol{D} \boldsymbol{\Psi}_S^{\mathsf{T}} \boldsymbol{\Psi}_T \widehat{\boldsymbol{f}}_T]/n^2 &= \sum_{u,v \in T} \sum_{s \in S} \sum_{i,j \in [n]} \boldsymbol{D}_{i(s)i(s)} \{\mathbb{E}[\psi_u(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_j)\psi_v(\boldsymbol{x}_j)\} \hat{f}_u \hat{f}_v / n^2 \\
&= \sum_{u,v \in T} \sum_{s \in S} \sum_{i \in [n]} \boldsymbol{D}_{i(s)i(s)} \{\mathbb{E}[\psi_u(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_i)\psi_s(\boldsymbol{x}_i)\psi_v(\boldsymbol{x}_i)\} \hat{f}_u \hat{f}_v / n^2 \\
&= \frac{1}{n} \sum_{s \in S} \boldsymbol{D}_{i(s)i(s)} \mathbb{E}_{\boldsymbol{x}}[(\mathsf{P}_T f_d(\boldsymbol{x}))^2 \psi_s(\boldsymbol{x})^2] \\
&\le \frac{1}{n} \sum_{s \in S} \boldsymbol{D}_{i(s)i(s)} \left\| \mathsf{P}_T f_d \right\|_{L^{2+\eta}}^2 \left\| \psi_s \right\|_{L^{(4+2\eta)/\eta}}^2 \\
&\le C(\eta) \left\| \mathsf{P}_T f_d \right\|_{L^{2+\eta}}^2 \sum_{i=1}^{n} \boldsymbol{D}_{ii}/n.
\end{aligned}$$

where the second to last inequality is by Holder's inequality and the last inequality used the hyper-contractivity assumption as in Assumption 3(a). □

Lemma 17 (Theorem 6 from Mei et al. (2021a))**.** *Let Assumptions 3 and 4(a) hold. Then we can decompose the kernel matrices as follows*

$$\begin{aligned}
\boldsymbol{H} &= \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^2 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} + \kappa_H (\mathbf{I}_n + \boldsymbol{\Delta}_H), \\
\boldsymbol{M} &= \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^4 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} + \kappa_M (\mathbf{I}_n + \boldsymbol{\Delta}_M),
\end{aligned}$$

where

$$\begin{aligned}
\kappa_H &= \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \sum_{k \ge \mathsf{m}+1}^{\infty} \lambda_{d,k}^2, \\
\kappa_M &= \mathrm{Tr}\big(\mathbb{H}_{d,>\mathsf{m}}^2\big) = \sum_{k \ge \mathsf{m}+1}^{\infty} \lambda_{d,k}^4,
\end{aligned}$$

and

$$\max\{\|\boldsymbol{\Delta}_H\|_{\mathrm{op}}, \|\boldsymbol{\Delta}_M\|_{\mathrm{op}}\} = o_{d,\mathbb{P}}(1).$$

Lemma 18 (Exponential Kernel Decomposition)**.** *Assume there exists $\delta_0 > 0$, such that $(t(d), \mathsf{u}(d))$ satisfies the condition*

$$t(d)^{1+\delta_0} \le \frac{1}{\lambda_{d,\mathsf{u}(d)+1}^2}.$$

Let $j(d)$ satisfy $\min\{\mathsf{u}(d), \mathsf{m}(d)\} \le j(d) \le \mathsf{m}(d)$ then

$$\left\| e^{-t\boldsymbol{H}_d/n} - e^{-(t/n)(\boldsymbol{\Psi}_{\le j} \boldsymbol{D}_{\le j}^2 \boldsymbol{\Psi}_{\le j}^{\mathsf{T}} + \kappa_H \mathbf{I}_n)} \right\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Proof. First consider the regime $t = O_d(n/\kappa_H)$. Recall the decomposition,

$$\boldsymbol{H} = \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{D}_{\le \mathsf{m}}^2 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} + \kappa_H (\mathbf{I}_n + \boldsymbol{\Delta}_H)$$

where $\|\Delta_H\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. By lemma 11 and lemma 14,

$$\begin{aligned}
\left\| e^{-t\boldsymbol{H}/n} - e^{-(t/n)(\boldsymbol{\Psi}_{\le j} \boldsymbol{D}_{\le j}^2 \boldsymbol{\Psi}_{\le j}^{\mathsf{T}} + \kappa_H \mathbf{I})} \right\|_{\mathrm{op}} &\le \left\| \sum_{k \ge j+1}^{\mathsf{m}} t \lambda_{d,k}^2 \boldsymbol{\psi}_k \boldsymbol{\psi}_k^{\mathsf{T}}/n \right\|_{\mathrm{op}} + (t/n)\kappa_H \|\boldsymbol{\Delta}_h\|_{\mathrm{op}} \\
&\le (t \cdot \max_{k \ge j+1} \cdot \lambda_{d,k}^2) \left\| \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\le \mathsf{m}}/n \right\|_{\mathrm{op}} + o_{d,\mathbb{P}}(1) \\
&\le O_{d,\mathbb{P}}(t \cdot \max_{k \ge j+1} \lambda_{d,k}^2) + o_{d,\mathbb{P}}(1) \\
&= O_{d,\mathbb{P}}(t^{-\delta_0}) + o_{d,\mathbb{P}}(1) = o_{d,\mathbb{P}}(1).
\end{aligned}$$

Now if $t = \omega_d(n/\kappa_H)$, then it is easy to see that

$$\max\left\{\left\|e^{-t\boldsymbol{H}/n}\right\|_{\text{op}}, \left\|e^{-(t/n)(\boldsymbol{\Psi}_{\le j}\boldsymbol{D}^2_{\le j}\boldsymbol{\Psi}^{\mathsf{T}}_{\le j}+\kappa_H\mathbf{I})}\right\|_{\text{op}}\right\} \le e^{-\omega_{d,\mathbb{P}}(1)} = o_{d,\mathbb{P}}(1),$$

and therefore

$$\begin{aligned}&\left\|e^{-t\boldsymbol{H}/n} - e^{-(t/n)(\boldsymbol{\Psi}_{\le j}\boldsymbol{D}^2_{\le j}\boldsymbol{\Psi}^{\mathsf{T}}_{\le j}+\kappa_H\mathbf{I})}\right\|_{\text{op}}\\ &\le 2\max\left\{\left\|e^{-t\boldsymbol{H}/n}\right\|_{\text{op}}, \left\|e^{-(t/n)(\boldsymbol{\Psi}_{\le j}\boldsymbol{D}^2_{\le j}\boldsymbol{\Psi}^{\mathsf{T}}_{\le j}+\kappa_H\mathbf{I})}\right\|_{\text{op}}\right\} = o_{d,\mathbb{P}}(1).\end{aligned}$$

□

D.3 Empirical World - Train

If $t = \omega_d(n/\kappa_H)$, then by (Eq. 11) it is easy to see that $\left\|e^{-2(t/n)\boldsymbol{H}}\right\|_{\text{op}} = o_{d,\mathbb{P}}(1)$, hence

$$\widehat{R}_n(\hat{f}_t) = o_{d,\mathbb{P}}(1)\cdot\|f_d\|^2_{L^2}.$$

From now on we focus on the case that $t = o_d(n/\kappa_H)$. Let us decompose the training error as follows

$$\widehat{R}_n(\hat{f}_t) = R_1 + R_2 + R_3$$

where

$$\begin{aligned}R_1 &= \frac{1}{n}\boldsymbol{f}^{\mathsf{T}}e^{-2(t/n)\boldsymbol{H}}\boldsymbol{f},\\ R_2 &= \frac{2}{n}\boldsymbol{\varepsilon}^{\mathsf{T}}e^{-2(t/n)\boldsymbol{H}}\boldsymbol{f},\\ R_3 &= \frac{1}{n}\boldsymbol{\varepsilon}^{\mathsf{T}}e^{-2(t/n)\boldsymbol{H}}\boldsymbol{\varepsilon}.\end{aligned}$$

D.3.1 Term R_1

Let us start by analysing R_1. We can write

$$\begin{aligned}R_1 &= \frac{1}{n}(\boldsymbol{f}_{\le\ell}+\boldsymbol{f}_{>\ell})^{\mathsf{T}}e^{-2(t/n)\boldsymbol{H}}(\boldsymbol{f}_{\le\ell}+\boldsymbol{f}_{>\ell})\\ &= T_1 + 2T_2 + T_3\end{aligned}$$

where

$$\begin{aligned}T_1 &= \frac{1}{n}\boldsymbol{f}^{\mathsf{T}}_{\le\ell}e^{-2(t/n)\boldsymbol{H}}\boldsymbol{f}_{\le\ell},\\ T_2 &= \frac{1}{n}\boldsymbol{f}^{\mathsf{T}}_{>\ell}e^{-2(t/n)\boldsymbol{H}}\boldsymbol{f}_{\le\ell},\\ T_3 &= \frac{1}{n}\boldsymbol{f}^{\mathsf{T}}_{>\ell}e^{-2(t/n)\boldsymbol{H}}\boldsymbol{f}_{>\ell}.\end{aligned}$$

Let us analyse the term T_1,

$$T_1 \le \frac{1}{n}\boldsymbol{f}^{\mathsf{T}}_{\le\ell}e^{-2(t/n)\boldsymbol{\Psi}_{\le\ell}\boldsymbol{D}^2_{\le\ell}\boldsymbol{\Psi}^{\mathsf{T}}_{\le\ell}}\boldsymbol{f}_{\le\ell}.$$

By Lemma 12, denoting $\boldsymbol{B} = \boldsymbol{\Psi}^{\mathsf{T}}_{\le\ell}\boldsymbol{\Psi}_{\le\ell}/n$ and $\boldsymbol{D}^2 = \boldsymbol{D}^2_{\le\ell}$

$$\frac{1}{n}\boldsymbol{\Psi}^{\mathsf{T}}_{\le\ell}e^{-2(t/n)\boldsymbol{\Psi}_{\le\ell}\boldsymbol{D}^2_{\le\ell}\boldsymbol{\Psi}^{\mathsf{T}}_{\le\ell}}\boldsymbol{\Psi}_{\le\ell} = \boldsymbol{B}e^{-2t\boldsymbol{D}^2\boldsymbol{B}}.$$

By Lemma 14, $\|\boldsymbol{B}\|_{\text{op}} = \Theta_d(1)$ and by Assumption 5(a) $\left\|t\boldsymbol{D}^2\right\|_{\text{op}} = \Omega(t^{\delta_0})$, hence

$$\left\|\boldsymbol{B}e^{-2t\boldsymbol{D}^2\boldsymbol{B}}\right\|_{\text{op}} = o_{d,\mathbb{P}}(1).$$

Therefore $T_1 = o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\leq \ell} f_d\|_{L^2}^2$.

Now we analyse the term T_3. Observe that by the inequality $1 - x \leq e^{-x} \leq 1$,

$$\frac{1}{n}\|\boldsymbol{f}_{>\ell}\|_2^2 - (2t/n)\frac{1}{n}\boldsymbol{f}_{>\ell}^{\mathsf{T}}(\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2\boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} + \kappa_H(\mathbf{I}_n + \Delta))\boldsymbol{f}_{>\ell} \leq T_3 \leq \frac{1}{n}\|\boldsymbol{f}_{>\ell}\|_2^2.$$

We have by Lemma 16,

$$t\mathbb{E}[\widehat{\boldsymbol{f}}_{>\ell}^{\mathsf{T}}\boldsymbol{\Psi}_{>\ell}^{\mathsf{T}}\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2\boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}\boldsymbol{\Psi}_{>\ell}\widehat{\boldsymbol{f}}_{>\ell}]/n^2 \leq C(\eta)\,\|\mathsf{P}_{>\ell} f_d\|_{L^{2+\eta}}^2 \left(\frac{t}{n}\sum_{s=1}^{\ell}\lambda_{d,s}^2\right)$$

where the last quantity is $o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{>\ell} f_d\|_{L^{2+\eta}}^2$ by Assumption 5(c). Thus we see that

$$T_3 = \|\mathsf{P}_{>\ell} f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{>\ell} f_d\|_{L^{2+\eta}}^2 .$$

Observe that by the Cauchy Schwartz inequality,

$$T_2 \leq (T_1 T_3)^{1/2} \leq o_{d,\mathbb{P}}(1)\,\|\mathsf{P}_{\leq \ell} f_d\|_{L^2}\,\|\mathsf{P}_{>\ell} f_d\|_{L^{2+\eta}} ,$$

Putting everything together we see that,

$$R_1 = \|\mathsf{P}_{>\ell} f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot (\|f_d\|_{L^2}^2 + \|\mathsf{P}_{>\ell} f_d\|_{L^{2+\eta}}^2).$$

D.3.2 Term R_2

Turning to R_2, we take the second-moment with respect to ε

$$\begin{aligned}
\frac{1}{\sigma_\varepsilon^2}\mathbb{E}_\varepsilon[R_2^2] &= \frac{4}{n^2}\mathbb{E}_\varepsilon[\varepsilon^{\mathsf{T}} e^{-2(t/n)\boldsymbol{H}}\boldsymbol{f}\boldsymbol{f}^{\mathsf{T}} e^{-2(t/n)\boldsymbol{H}}\varepsilon] \\
&= \frac{4}{n^2}\boldsymbol{f}^{\mathsf{T}} e^{-4(t/n)\boldsymbol{H}}\boldsymbol{f} \leq \frac{4}{n}(\|\boldsymbol{f}\|_2^2/n) \\
&= o_{d,\mathbb{P}}(1) \cdot \|f_d\|_{L^2}^2 .
\end{aligned}$$

By Markov's inequality $R_2 = o_{d,\mathbb{P}}(1) \cdot \|f_d\|_{L^2}^2\,\sigma_\varepsilon^2$.

D.3.3 Term R_3

Now let us analyse R_3. We can compute the expectation

$$\begin{aligned}
\frac{1}{\sigma_\varepsilon^2}\mathbb{E}_\varepsilon[R_3] &= \frac{1}{n}\operatorname{Tr}\Big(e^{-2(t/n)\boldsymbol{H}}\Big) \\
&\overset{(a)}{=} \frac{1}{n}e^{-\kappa_H t/n}\operatorname{Tr}\Big(e^{-2(t/n)(\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2\boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}})}\Big) + o_{d,\mathbb{P}}(1) \\
&\overset{(b)}{=} \frac{1}{n}e^{-\kappa_H t/n}\Big(\operatorname{Tr}\Big(e^{-2t(\boldsymbol{B}\boldsymbol{D}^2)}\Big) + n - \ell\Big) + o_{d,\mathbb{P}}(1) \\
&\overset{(c)}{=} \frac{1}{n}e^{-\kappa_H t/n}\left(\sum_{j=1}^{\ell} e^{-2t(1+o_{d,\mathbb{P}}(1))\lambda_{d,j}^2} + n - \ell\right) + o_{d,\mathbb{P}}(1) \\
&\overset{(d)}{=} \frac{1}{n}e^{-\kappa_H t/n}(\ell - 1 + n - \ell) + o_{d,\mathbb{P}}(1) \\
&= 1 + o_{d,\mathbb{P}}(1)
\end{aligned}$$

Equality (a) follows from Lemma 18. For (b) we used $\ell \leq n$ by Assumption 4(d) and that for $\boldsymbol{X} \in \mathbb{R}^{n\times\ell}, \boldsymbol{Y} \in \mathbb{R}^{\ell\times n}$, the spectrum of $\boldsymbol{X}\boldsymbol{Y}$ is the same as that of $\boldsymbol{Y}\boldsymbol{X}$ except for an additional $n-\ell$ zero eigenvalues. Equality (c) follows from Lemma 14, for (d) we used Assumption 5(a), and in the last equality we used that $t = o_d(n/\kappa_H)$ and $\ell/n = o_d(1)$ by Assumption 4(d). Turning to

the variance

$$\begin{aligned}\mathrm{Var}_\varepsilon[R_3] &= \mathbb{E}_\varepsilon[R_3^2] - \mathbb{E}_\varepsilon[R_3]^2 \\ &= \frac{1}{n^2}\mathbb{E}_\varepsilon[(\varepsilon^\mathsf{T} e^{-2(t/n)\boldsymbol{H}}\varepsilon)^2] - \mathbb{E}_\varepsilon[R_3]^2 \\ &\le \frac{1}{n^2}\mathbb{E}_\varepsilon[(\varepsilon^\mathsf{T}\varepsilon)^2] - \sigma_\varepsilon^4(1+o_{d,\mathbb{P}}(1)) \\ &= O(1/n)\cdot\sigma_\varepsilon^4 + \sigma_\varepsilon^4 - \sigma_\varepsilon^4(1+o_{d,\mathbb{P}}(1)) \\ &= o_{d,\mathbb{P}}(1)\cdot\sigma_\varepsilon^4.\end{aligned}$$

where the first inequality uses that $\left\|e^{-2(t/n)\boldsymbol{H}}\right\|_{\mathrm{op}} \le 1$. Therefore by Chebyshev's inequality, $R_3 = \sigma_\varepsilon^2(1+o_{d,\mathbb{P}}(1))$.

Putting everything together yields

$$\widehat{R}_n(\hat{f}_t) = R_1 + R_2 + R_3 = \|\mathsf{P}_{>\ell}\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\cdot(\|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2).$$

D.4 Empirical World - Test

Recalling $\boldsymbol{u}(t)$ from (Eq. 23), let

$$\boldsymbol{u}(t) = \boldsymbol{v}(t) + \boldsymbol{\varepsilon}(t)$$

where

$$\begin{aligned}\boldsymbol{v}(t) &= (\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{f}, \\ \boldsymbol{\varepsilon}(t) &= (\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{\varepsilon}.\end{aligned}$$

Consider the expansion of the test error,

$$\begin{aligned}R(\hat{f}_t) &= \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})^2] - 2\boldsymbol{u}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{E} + \boldsymbol{u}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}\boldsymbol{u}(t) \\ &= \|f_d\|_{L^2}^2 - 2T_1 + T_2 + T_3 - 2T_4 + 2T_5,\end{aligned}$$

where

$$\begin{aligned}T_1 &= \boldsymbol{v}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{E}, \\ T_2 &= \boldsymbol{v}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}\boldsymbol{v}(t), \\ T_3 &= \boldsymbol{\varepsilon}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}\boldsymbol{\varepsilon}(t), \\ T_4 &= \boldsymbol{\varepsilon}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{E}, \\ T_5 &= \boldsymbol{\varepsilon}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}\boldsymbol{v}(t).\end{aligned}$$

D.4.1 Term T_2

As before, we will analyze each term separately. We begin with term T_2.

Proposition 19 (Term T_2)**.**

$$T_2 = \boldsymbol{v}(t)^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}\boldsymbol{v}(t) = \left\|\boldsymbol{S}_{\le\ell}\widehat{\boldsymbol{f}}_{\le\ell}\right\|_2^2 + o_{d,\mathbb{P}}(1)\cdot\|f_d\|_{L^{2+\eta}}^2.$$

where we recall the shrinkage matrix $\boldsymbol{S}_{\le\mathsf{m}}$ defined in Eq. (13).

To analyze T_2 we further decompose it into the following terms

$$\begin{aligned}T_2 &= (\boldsymbol{f}_{\le\ell} + \boldsymbol{f}_{>\ell})^\mathsf{T}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})(\boldsymbol{f}_{\le\ell} + \boldsymbol{f}_{>\ell}) \\ &= T_{21} + T_{22} + T_{23},\end{aligned}$$

where

$$T_{21} = \boldsymbol{f}_{\le\ell}^\mathsf{T}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{f}_{\le\ell}, \tag{14}$$

$$T_{22} = 2\boldsymbol{f}_{\le\ell}^\mathsf{T}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{f}_{>\ell}, \tag{15}$$

$$T_{23} = \boldsymbol{f}_{>\ell}^\mathsf{T}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{f}_{>\ell}. \tag{16}$$

We first start with T_{21}.

Lemma 20 (Term T_{21} Eq. (14))**.**

$$T_{21} = \left\| \boldsymbol{S}_{\leq \ell} \widehat{\boldsymbol{f}}_{\leq \ell} \right\|_2^2 + o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\leq \ell} f_d\|_{L^2}^2$$

Proof of Lemma 20. Let $\alpha = e^{-(t/n)\kappa_H} \leq 1$ and define

$$\boldsymbol{K}_{\leq j} := \boldsymbol{I}_n - \alpha e^{-(t/n)\boldsymbol{\Psi}_{\leq j}\boldsymbol{D}_{\leq j}^2 \boldsymbol{\Psi}_{\leq j}^{\mathsf{T}}}.$$

By lemma 18, for some matrix $\boldsymbol{\Delta}$ with $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ we can split T_{21} into

$$T_{21} = T_{211} + T_{212} + T_{213} + T_{214}$$

where

$$\begin{aligned}
T_{211} &= \boldsymbol{f}_{\leq \ell}^{\mathsf{T}} \boldsymbol{K}_{\leq \ell} \boldsymbol{\Psi}_{\leq \ell} \boldsymbol{S}_{\leq \ell}^2 \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} \boldsymbol{K}_{\leq \ell} \boldsymbol{f}_{\leq \ell} \\
T_{212} &= \boldsymbol{f}_{\leq \ell}^{\mathsf{T}} \boldsymbol{K}_{\leq \ell} \boldsymbol{\Psi}_{\leq \ell \mathsf{m}} \boldsymbol{S}_{\leq \ell \mathsf{m}}^2 \boldsymbol{\Psi}_{\leq \ell \mathsf{m}}^{\mathsf{T}} \boldsymbol{K}_{\leq \ell} \boldsymbol{f}_{\leq \ell} \\
T_{213} &= \boldsymbol{f}_{\leq \ell}^{\mathsf{T}} \boldsymbol{K}_{\leq \ell} (\boldsymbol{H}^{-1} \boldsymbol{M} \boldsymbol{H}^{-1} - \boldsymbol{\Psi}_{\leq \mathsf{m}} \boldsymbol{S}_{\leq \mathsf{m}}^2 \boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}) \boldsymbol{K}_{\leq \ell} \boldsymbol{f}_{\leq \ell} \\
T_{214} &= \boldsymbol{f}_{\leq \ell}^{\mathsf{T}} \boldsymbol{\Delta} \boldsymbol{H}^{-1} \boldsymbol{M} \boldsymbol{H}^{-1} \boldsymbol{K}_{\leq \ell} \boldsymbol{f}_{\leq \ell} + \boldsymbol{f}_{\leq \ell}^{\mathsf{T}} \boldsymbol{K}_{\leq \ell} \boldsymbol{H}^{-1} \boldsymbol{M} \boldsymbol{H}^{-1} \boldsymbol{\Delta} \boldsymbol{f}_{\leq \ell} + \boldsymbol{f}_{\leq \ell}^{\mathsf{T}} \boldsymbol{\Delta} \boldsymbol{H}^{-1} \boldsymbol{M} \boldsymbol{H}^{-1} \boldsymbol{\Delta} \boldsymbol{f}_{\leq \ell}.
\end{aligned}$$

We will show that the dominant term is T_{211} and the others are of lower order. By Lemma 13,

$$\left\| n \boldsymbol{H}^{-1} \boldsymbol{M} \boldsymbol{H}^{-1} - \boldsymbol{\Psi}_{\leq \mathsf{m}} \boldsymbol{S}_{\leq \mathsf{m}}^2 \boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}} / n \right\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1),$$

hence $T_{213} = o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\leq \ell} f_d\|_{L^2}^2$. Furthermore, using that $\boldsymbol{S}_{\leq \mathsf{m}} \preceq \mathbf{I}_{\mathsf{m}}$ and Lemma 14 it is easy to see that $T_{214} = o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\leq \ell} f_d\|_{L^2}^2$. Turning to T_{212} we have

$$\begin{aligned}
T_{212} &\leq \frac{1}{n^2} \boldsymbol{f}_{\leq \ell}^{\mathsf{T}} (\mathbf{I}_n - \alpha e^{-(t/n)\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2 \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}}) \boldsymbol{\Psi}_{\ell \mathsf{m}} \boldsymbol{\Psi}_{\ell \mathsf{m}}^{\mathsf{T}} (\mathbf{I}_n - \alpha e^{-(t/n)\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2 \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}}) \boldsymbol{f}_{\leq \ell} \\
&\overset{(a)}{=} \frac{1}{n^2} \widehat{\boldsymbol{f}}_{\leq \ell}^{\mathsf{T}} (\mathbf{I} - \alpha e^{-(t/n)\boldsymbol{B}\boldsymbol{D}_{\leq \ell}^2}) \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} \boldsymbol{\Psi}_{\ell \mathsf{m}} \boldsymbol{\Psi}_{\ell \mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\leq \ell} (\mathbf{I} - \alpha e^{-(t/n)\boldsymbol{B}\boldsymbol{D}_{\leq \ell}^2}) \widehat{\boldsymbol{f}}_{\leq \ell}^{\mathsf{T}} \\
&\leq \left\| \widehat{\boldsymbol{f}}_{\leq \ell} \right\|_2^2 \left\| \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} \boldsymbol{\Psi}_{\ell \mathsf{m}} / n \right\|_{\mathrm{op}}^2 = o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\leq \ell} f_d\|_{L^2}^2 .
\end{aligned}$$

where we used Lemma 12 in (a) and Lemma 14 in the last equality. Finally we look at the main term T_{211}. Defining the terms

$$\begin{aligned}
\boldsymbol{A} &= \frac{1}{n} \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} e^{-(t/n)\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2 \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}} \boldsymbol{\Psi}_{\leq \ell} \\
\boldsymbol{B} &= \frac{1}{n} \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} \boldsymbol{\Psi}_{\leq \ell}
\end{aligned}$$

we can write

$$\begin{aligned}
T_{211} &= \frac{1}{n^2} \widehat{\boldsymbol{f}}_{\leq \ell}^{\mathsf{T}} \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} (\mathbf{I}_n - \alpha e^{-(t/n)\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2 \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}}) \boldsymbol{\Psi}_{\leq \ell} \boldsymbol{S}_{\leq \ell}^2 \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} (\mathbf{I}_n - \alpha e^{-(t/n)\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2 \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}}) \boldsymbol{\Psi}_{\leq \ell} \widehat{\boldsymbol{f}}_{\leq \ell} \\
&= \widehat{\boldsymbol{f}}_{\leq \ell}^{\mathsf{T}} (\boldsymbol{B} \boldsymbol{S}_{\leq \ell}^2 \boldsymbol{B} - 2\alpha \boldsymbol{A} \boldsymbol{S}_{\leq \ell}^2 \boldsymbol{B} + \alpha^2 \boldsymbol{A} \boldsymbol{S}_{\leq \ell}^2 \boldsymbol{A}) \widehat{\boldsymbol{f}}_{\leq \ell}.
\end{aligned}$$

Now observe that by Lemma 12,

$$\boldsymbol{A} = \frac{1}{n} \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} e^{-(t/n)\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2 \boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}} \boldsymbol{\Psi}_{\leq \ell} = \boldsymbol{B} e^{-t\boldsymbol{D}_{\leq \ell}^2 \boldsymbol{B}}$$

hence $\|\boldsymbol{A}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. Since $\boldsymbol{B} = \mathbf{I}_n + \boldsymbol{\Delta}$ by Lemma 14, it follows that

$$T_{211} = \left\| \boldsymbol{S}_{\leq \ell} \widehat{\boldsymbol{f}}_{\leq \ell} \right\|_2^2 + o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\leq \ell} f_d\|_{L^2}^2 .$$

hence

$$T_{21} = T_{211} + T_{212} + T_{213} + T_{214} = \left\| \boldsymbol{S}_{\leq \ell} \widehat{\boldsymbol{f}}_{\leq \ell} \right\|_2^2 + o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{\leq \ell} f_d\|_{L^2}^2 .$$

□

Now let us look at term T_{23}

Lemma 21 (Term T_{23} Eq. (16))**.**

$$T_{23} = \boldsymbol{f}_{>\ell}^{\mathsf{T}}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{f}_{>\ell} = o_{d,\mathbb{P}}(1) \cdot \|f_d\|_{L^{2+\eta}}^2 .$$

Proof of Lemma 21. Let us define the matrix

$$\boldsymbol{G} = (\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n}).$$

Using similar reasoning from Lemma 20 we can write

$$\begin{aligned}\boldsymbol{G} &= \frac{1}{n}\boldsymbol{K}_{\leq \mathsf{m}}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{S}_{\leq \mathsf{m}}^2\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}\boldsymbol{K}_{\leq \mathsf{m}} + \boldsymbol{\Delta} \\ &= \frac{1}{n}\boldsymbol{K}_{\leq \ell}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{S}_{\leq \mathsf{m}}^2\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}\boldsymbol{K}_{\leq \ell} + \boldsymbol{\Delta}'.\end{aligned}$$

for matrices $\boldsymbol{\Delta}, \boldsymbol{\Delta}'$ satisfying $\max\{\|\boldsymbol{\Delta}\|_{\mathrm{op}}, \|\boldsymbol{\Delta}'\|_{\mathrm{op}}\} = o_{d,\mathbb{P}}(1)$. We can split T_{23} into

$$\begin{aligned}T_{23} &= \frac{1}{n}\boldsymbol{f}_{\leq \ell}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{f}_{\leq \ell} \\ &= \frac{1}{n}(\boldsymbol{f}_{>\mathsf{m}} + \boldsymbol{f}_{\ell\mathsf{m}})^{\mathsf{T}}\boldsymbol{G}(\boldsymbol{f}_{>\mathsf{m}} + \boldsymbol{f}_{\ell\mathsf{m}}) \\ &= \frac{1}{n}\boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{f}_{>\mathsf{m}} + \frac{2}{n}\boldsymbol{f}_{\ell\mathsf{m}}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{f}_{>\mathsf{m}} + \frac{1}{n}\boldsymbol{f}_{\ell\mathsf{m}}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{f}_{\ell\mathsf{m}} \\ &= T_{231} + T_{232} + T_{233} + T_{234}\end{aligned}$$

where

$$\begin{aligned}T_{231} &= \frac{1}{n^2}\boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}}\boldsymbol{K}_{\leq \mathsf{m}}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{S}_{\leq \mathsf{m}}^2\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}\boldsymbol{K}_{\leq \mathsf{m}}\boldsymbol{f}_{>\mathsf{m}}, \\ T_{232} &= \frac{2}{n^2}\boldsymbol{f}_{\ell\mathsf{m}}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{f}_{>\mathsf{m}}, \\ T_{233} &= \frac{1}{n^2}\boldsymbol{f}_{\ell\mathsf{m}}^{\mathsf{T}}\boldsymbol{K}_{\leq \ell}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{S}_{\leq \mathsf{m}}^2\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}\boldsymbol{K}_{\leq \ell}\boldsymbol{f}_{\ell\mathsf{m}} \\ T_{234} &= o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{>\ell} f_d\|_{L^2}^2 .\end{aligned}$$

Let us first start with analysing T_{231}, we have

$$\begin{aligned}T_{231} &= \frac{1}{n^2}\boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}}\boldsymbol{\Psi}_{\leq \mathsf{m}}(\mathbf{I}_{\mathsf{m}} - \alpha e^{-(t/n)\boldsymbol{D}_{\leq \mathsf{m}}^2\boldsymbol{B}})\boldsymbol{S}_{\leq \mathsf{m}}^2(\mathbf{I}_{\mathsf{m}} - \alpha e^{-(t/n)\boldsymbol{B}\boldsymbol{D}_{\leq \mathsf{m}}^2})\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{m}} \\ &\leq \boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}\boldsymbol{f}_{>\mathsf{m}} \\ &= o_{d,\mathbb{P}}(1) \cdot \|\mathsf{P}_{>\mathsf{m}} f_d\|_{L^{2+\eta}}^2 .\end{aligned}$$

where the first equality is Lemma 12 and the last line follows from Lemma 16. Now let us turn to T_{233}, which we can further split into

$$T_{233} = T_{2331} + T_{2332} \tag{17}$$

where

$$\begin{aligned}T_{2331} &= \frac{1}{n^2}\boldsymbol{f}_{\ell\mathsf{m}}^{\mathsf{T}}\boldsymbol{K}_{\leq \ell}\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{S}_{\leq \ell}^2\boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}\boldsymbol{K}_{\leq \ell}\boldsymbol{f}_{\ell\mathsf{m}} \\ T_{2332} &= \frac{1}{n^2}\boldsymbol{f}_{\ell\mathsf{m}}^{\mathsf{T}}\boldsymbol{K}_{\leq \ell}\boldsymbol{\Psi}_{\ell\mathsf{m}}\boldsymbol{S}_{\ell\mathsf{m}}^2\boldsymbol{\Psi}_{\ell\mathsf{m}}^{\mathsf{T}}\boldsymbol{K}_{\leq \ell}\boldsymbol{f}_{\ell\mathsf{m}}\end{aligned}$$

and using the usual argument the first term can be seen as

$$T_{2331} \leq \left\|\widehat{\boldsymbol{f}}_{\ell\mathsf{m}}\right\|_2^2 \left\|\boldsymbol{\Psi}_{\ell\mathsf{m}}^{\mathsf{T}}\boldsymbol{\Psi}_{\leq \ell}/n\right\|_{\mathrm{op}}^2 = o_{d,\mathbb{P}}(1) \cdot \left\|\widehat{\boldsymbol{f}}_{\ell\mathsf{m}}\right\|_2^2 \tag{18}$$

turning to the second term let

$$\overline{\boldsymbol{A}} = \frac{1}{n}\boldsymbol{\Psi}_{\ell\mathsf{m}}^{\mathsf{T}}\left(\mathbf{I}_\ell - e^{-(t/n)(\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2\boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} + \kappa_H \mathbf{I}_n)}\right)\boldsymbol{\Psi}_{\ell\mathsf{m}}$$

and observe that for the T_{2332} in Eq. (17)

$$\begin{aligned} T_{2332} &= \widehat{\boldsymbol{f}}_{\ell\mathsf{m}}^{\mathsf{T}} \overline{\boldsymbol{A}} \boldsymbol{S}_{\ell\mathsf{m}}^2 \overline{\boldsymbol{A}} \widehat{\boldsymbol{f}}_{\ell\mathsf{m}} \\ &\le \widehat{\boldsymbol{f}}_{\ell\mathsf{m}}^{\mathsf{T}} \overline{\boldsymbol{A}}^2 \widehat{\boldsymbol{f}}_{\ell\mathsf{m}} = \widehat{\boldsymbol{f}}_{\ell\mathsf{m}}^{\mathsf{T}} \overline{\boldsymbol{A}} \widehat{\boldsymbol{f}}_{\ell\mathsf{m}} + o_{d,\mathbb{P}}(1) \cdot \left\| \widehat{\boldsymbol{f}}_{\ell\mathsf{m}} \right\|_2^2 \end{aligned}$$

In the PSD order we see that

$$\overline{\boldsymbol{A}} \preceq \frac{t}{n^2} \boldsymbol{\Psi}_{\ell\mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\le \ell} \boldsymbol{D}_{\le \ell}^2 \boldsymbol{\Psi}_{\le \ell}^{\mathsf{T}} \boldsymbol{\Psi}_{\ell\mathsf{m}} + \frac{\kappa_H t}{n} \left(\frac{1}{n} \boldsymbol{\Psi}_{\ell\mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\ell\mathsf{m}} \right).$$

For the second term, by Assumption 5(b),

$$\left\| \frac{\kappa_H t}{n} \left(\frac{1}{n} \boldsymbol{\Psi}_{\ell\mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\ell\mathsf{m}} \right) \right\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1).$$

Therefore from Eq. (17) and Eq. (18)

$$T_{233} \le \frac{t}{n^2} \widehat{\boldsymbol{f}}_{\ell\mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\ell\mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\le \ell} \boldsymbol{D}_{\le \ell}^2 \boldsymbol{\Psi}_{\le \ell}^{\mathsf{T}} \boldsymbol{\Psi}_{\ell\mathsf{m}} \widehat{\boldsymbol{f}}_{\ell\mathsf{m}} + o_{d,\mathbb{P}}(1) \cdot \left\| \widehat{\boldsymbol{f}}_{\ell\mathsf{m}} \right\|_2^2$$

and it is not hard to see that for the first term on the right by Lemma 16 and Assumption 5(c),

$$\frac{t}{n^2} \mathbb{E}[\widehat{\boldsymbol{f}}_{\ell\mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\ell\mathsf{m}}^{\mathsf{T}} \boldsymbol{\Psi}_{\le \ell} \boldsymbol{D}_{\le \ell}^2 \boldsymbol{\Psi}_{\le \ell}^{\mathsf{T}} \boldsymbol{\Psi}_{\ell\mathsf{m}} \widehat{\boldsymbol{f}}_{\ell\mathsf{m}}] = o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{>\ell} f_d \|_{L^{2+\eta}}^2 .$$

Thus by Markov's inequality we see that

$$T_{233} = o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{>\ell} f_d \|_{L^{2+\eta}}^2 .$$

By the Cauchy-Schwarz inequality

$$\begin{aligned} T_{232} &= \frac{2}{n^2} \boldsymbol{f}_{\ell\mathsf{m}}^{\mathsf{T}} \boldsymbol{G} \boldsymbol{f}_{>\mathsf{m}} \\ &\le 2 \left(\frac{1}{n^2} \boldsymbol{f}_{\ell\mathsf{m}}^{\mathsf{T}} \boldsymbol{G} \boldsymbol{f}_{\ell\mathsf{m}} \frac{1}{n^2} \boldsymbol{f}_{>\mathsf{m}}^{\mathsf{T}} \boldsymbol{G} \boldsymbol{f}_{>\mathsf{m}} \right)^{1/2} \\ &= 2 \left(T_{231} + o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{>\ell} f_d \|_{L^2}^2 \right)^{1/2} \left(T_{233} + o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{>\ell} f_d \|_{L^2}^2 \right)^{1/2} \\ &\le o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{>\ell} f_d \|_{L^{2+\eta}} \| \mathsf{P}_{>\mathsf{m}} f_d \|_{L^{2+\eta}} . \end{aligned}$$

Putting everything together we see that

$$T_{23} = T_{231} + T_{232} + T_{233} + T_{234} = o_{d,\mathbb{P}}(1) \cdot \| f_d \|_{L^{2+\eta}}^2 .$$

□

Now we can characterize T_2,

Proof of Proposition 19. Using Lemmas 20 and 21, by the Cauchy-Schwarz inequality,

$$T_{22} \le (2 T_{21} T_{23})^{1/2} = o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{\le \mathsf{m}} f_d \|_{L^2} \| f_d \|_{L^{2+\eta}}$$

and hence

$$T_2 = T_{21} + T_{22} + T_{23} = \left\| \boldsymbol{S}_{\le \ell} \widehat{\boldsymbol{f}}_{\le \ell} \right\|_2^2 + o_{d,\mathbb{P}}(1) \cdot \| f_d \|_{L^{2+\eta}}^2 .$$

□

D.4.2 TERM T_1

Now we will analyse term T_1.

Proposition 22 (Term T_1)**.**

$$T_1 = \boldsymbol{f}^{\mathsf{T}} (\mathbf{I}_n - e^{-t\boldsymbol{H}/n}) \boldsymbol{H}^{-1} \boldsymbol{E} = \left\| \boldsymbol{S}_{\le \ell}^{1/2} \widehat{\boldsymbol{f}}_{\le \ell} \right\|_2^2 + o_{d,\mathbb{P}}(1) \cdot \| f_d \|_{L^{2+\eta}}^2 .$$

We break T_1 into the following terms

$$T_1 = T_{11} + T_{12} + T_{13}$$

where

$$\begin{aligned} T_{11} &= \boldsymbol{f}_{\leq \ell}^{\mathsf{T}}(\mathbf{I} - e^{-t\boldsymbol{H}/n})\boldsymbol{H}^{-1}\boldsymbol{E}_{\leq \mathsf{m}}, \\ T_{12} &= \boldsymbol{f}_{>\ell}^{\mathsf{T}}(\mathbf{I} - e^{-t\boldsymbol{H}/n})\boldsymbol{H}^{-1}\boldsymbol{E}_{\leq \mathsf{m}}, \\ T_{13} &= \boldsymbol{f}^{\mathsf{T}}(\mathbf{I} - e^{-t\boldsymbol{H}/n})\boldsymbol{H}^{-1}\boldsymbol{E}_{>\mathsf{m}}. \end{aligned}$$

First we state a useful lemma for handling these terms

Lemma 23. *For a vector $\boldsymbol{v} \in \mathbb{R}^n$,*

$$\boldsymbol{v}^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{E}_{\leq m} \leq (\boldsymbol{v}^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}\boldsymbol{v})\left\|\widehat{\boldsymbol{f}}_{\leq \mathsf{m}}\right\|_2$$

Proof. Recall that

$$\boldsymbol{E}_{\leq \mathsf{m}} = \boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^2\widehat{\boldsymbol{f}}_{\leq \mathsf{m}}.$$

By the Cauchy Schwarz inequality and Lemma 13

$$\begin{aligned} \boldsymbol{v}^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{E}_{\leq \mathsf{m}} &= \boldsymbol{v}^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^2\widehat{\boldsymbol{f}}_{\leq \mathsf{m}}, \\ &\leq (\boldsymbol{v}^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^4\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{v})^{1/2}\left\|\widehat{\boldsymbol{f}}_{\leq \mathsf{m}}\right\|_2, \\ &\leq (\boldsymbol{v}^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}\boldsymbol{v})^{1/2}\left\|\widehat{\boldsymbol{f}}_{\leq \mathsf{m}}\right\|_2, \\ &= \left(\frac{1}{n^2}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{S}_{\leq \mathsf{m}}^2\boldsymbol{\Psi}_{\leq \mathsf{m}}^{\mathsf{T}}\boldsymbol{v} + o_{d,\mathbb{P}}(1)\cdot\frac{1}{n}\|\boldsymbol{v}\|_2\right)^{1/2}\left\|\widehat{\boldsymbol{f}}_{\leq \mathsf{m}}\right\|_2. \end{aligned}$$

□

Let us start with T_{11}. First note that by Lemma 18 for some $\boldsymbol{\Delta}$ such that $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$,

$$T_{11} = T_{111} + T_{112}$$

where

$$\begin{aligned} T_{111} &= \boldsymbol{f}_{\leq \ell}^{\mathsf{T}}(\mathbf{I} - e^{-(t/n)(\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2\boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}} + \kappa_H \mathbf{I}_n)})\boldsymbol{H}^{-1}\boldsymbol{E}_{\leq \mathsf{m}} \\ T_{112} &= \boldsymbol{f}_{\leq \ell}^{\mathsf{T}}\boldsymbol{\Delta}\boldsymbol{H}^{-1}\boldsymbol{E}_{\leq \mathsf{m}}. \end{aligned}$$

By Lemma 23,

$$T_{112} \leq \left(\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2\left(\|\boldsymbol{f}_{\leq \ell}\|_2^2/n\right)\left\|n\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}\right\|_{\mathrm{op}}\right)^{1/2}\left\|\widehat{\boldsymbol{f}}_{\leq \mathsf{m}}\right\|_2 = o_{d,\mathbb{P}}(1)\cdot\|\mathsf{P}_{\leq \mathsf{m}}f_d\|_{L^2}^2.$$

We now consider

$$T_{111} = T_{1111} - T_{1112}$$

where

$$\begin{aligned} T_{1111} &= \boldsymbol{f}_{\leq \ell}^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{E}_{\leq \mathsf{m}} \\ T_{1112} &= \alpha\boldsymbol{f}_{\leq \ell}^{\mathsf{T}}e^{-(t/n)\boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2\boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}}\boldsymbol{H}^{-1}\boldsymbol{E}_{\leq \mathsf{m}}. \end{aligned}$$

Note that by Lemma 15,

$$\left\|\boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^2 - [\boldsymbol{S}_{\leq \ell}; \mathbf{0}]\right\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1),$$

hence the first term

$$T_{1111} = \widehat{\boldsymbol{f}}_{\leq \ell}^{\mathsf{T}}\boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{\Psi}_{\leq \mathsf{m}}\boldsymbol{D}_{\leq \mathsf{m}}^2\widehat{\boldsymbol{f}}_{\leq \mathsf{m}} = \left\|\boldsymbol{S}_{\leq \ell}^{1/2}\widehat{\boldsymbol{f}}_{\leq \ell}\right\|_2^2 + o_{d,\mathbb{P}}(1)\cdot\|\mathsf{P}_{\leq \mathsf{m}}f_d\|_{L^2}^2.$$

Define $\boldsymbol{H}_{\leq \ell} := \boldsymbol{\Psi}_{\leq \ell}\boldsymbol{D}_{\leq \ell}^2\boldsymbol{\Psi}_{\leq \ell}^{\mathsf{T}}$. For the second term, by Lemma 23

$$T_{1112} \leq \left(S + o_{d,\mathbb{P}}(1)\cdot\|\mathsf{P}_{\leq \ell}f_d\|_{L^2}^2\right)^{1/2}\left\|\widehat{\boldsymbol{f}}_{\leq \mathsf{m}}\right\|_2. \tag{19}$$

where

$$S = \frac{1}{n^2} \boldsymbol{f}_{\le \ell}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{S}_{\le \mathsf{m}}^2 \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{f}_{\le \ell}.$$

Define $\boldsymbol{A} := \frac{1}{n} \boldsymbol{\Psi}_{\le \ell}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{\Psi}_{\le \ell}$. We have

$$\begin{aligned}
S &\le \frac{1}{n^2} \boldsymbol{f}_{\le \ell}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{\Psi}_{\le \mathsf{m}} \boldsymbol{\Psi}_{\le \mathsf{m}}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{f}_{\le \ell} \\
&= \frac{1}{n^2} \boldsymbol{f}_{\le \ell}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{\Psi}_{\le \ell} \boldsymbol{\Psi}_{\le \ell}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{f}_{\le \ell} + \frac{1}{n^2} \boldsymbol{f}_{\le \ell}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{\Psi}_{\ell \mathsf{m}} \boldsymbol{\Psi}_{\ell \mathsf{m}}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{f}_{\le \ell} \\
&\le \frac{1}{n^2} \boldsymbol{f}_{\le \ell}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{\Psi}_{\le \ell} \boldsymbol{\Psi}_{\le \ell}^{\mathsf{T}} e^{-(t/n)\boldsymbol{H}_{\le \ell}} \boldsymbol{f}_{\le \ell} + \left\| \widehat{\boldsymbol{f}}_{\le \ell} \right\|_2^2 \cdot \left\| \boldsymbol{\Psi}_{\le \ell}^{\mathsf{T}} \boldsymbol{\Psi}_{\ell \mathsf{m}} / n \right\|_{\mathrm{op}}^2 \\
&= \widehat{\boldsymbol{f}}_{\le \ell}^{\mathsf{T}} \boldsymbol{A}^2 \widehat{\boldsymbol{f}}_{\le \ell} + o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{\le \ell} f_d \|_{L^2}^2 \\
&= o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{\le \ell} f_d \|_{L^2}^2
\end{aligned}$$

since as noted before in Lemma 20, $\|\boldsymbol{A}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$. Therefore

$$T_{11} = T_{1111} + T_{1112} + T_{112} = \left\| \boldsymbol{S}_{\le \ell}^{1/2} \widehat{\boldsymbol{f}}_{\le \ell} \right\|_2^2 + o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{\le \mathsf{m}} f_d \|_{L^2}^2 .$$

Again using Lemma 23 and recalling Lemma 21,

$$T_{12} \le (T_{23})^{1/2} \left\| \widehat{\boldsymbol{f}}_{\le \mathsf{m}} \right\|_2 = o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{\le \mathsf{m}} f_d \|_{L^2} \| f_d \|_{L^{2+\eta}} .$$

Finally turning to T_{13}, we have

$$\begin{aligned}
|T_{13}| &= |\boldsymbol{f}^{\mathsf{T}} (\mathbf{I} - e^{-t\boldsymbol{H}/n}) \boldsymbol{H}^{-1} \boldsymbol{E}_{>\mathsf{m}}| \\
&\le \| \boldsymbol{f} \|_2 \left\| \boldsymbol{H}^{-1} \right\|_{\mathrm{op}} \| \boldsymbol{E}_{>\mathsf{m}} \|_2 \\
&\le o_{d,\mathbb{P}}(1) \cdot \| \mathsf{P}_{\le \mathsf{m}} f_d \|_{L^2} \| f_d \|_{L^2} .
\end{aligned}$$

Combining everything yields Proposition 22

$$T_1 = \left\| \boldsymbol{S}_{\le \ell}^{1/2} \widehat{\boldsymbol{f}}_{\le \ell} \right\|_2^2 + o_{d,\mathbb{P}}(1) \cdot \| f_d \|_{L^{2+\eta}}^2 .$$

D.4.3 TERMS T_3, T_4, T_5

Now adapting the corresponding steps for the proof of Theorem 4 in Mei et al. (2021a) and using the fact that $\mathrm{Tr}(\boldsymbol{AB}) \le \|\boldsymbol{A}\|_{\mathrm{op}} \mathrm{Tr}(\boldsymbol{B})$ for Hermitian $\boldsymbol{A}$ and PSD $\boldsymbol{B}$ we have

$$T_3 = \mathrm{Tr}\Big((\mathbf{I}_n - e^{-t\boldsymbol{H}/n})^2 \boldsymbol{H}^{-1} \boldsymbol{M} \boldsymbol{H}^{-1} \Big) \le \mathrm{Tr}\big(\boldsymbol{H}^{-1} \boldsymbol{M} \boldsymbol{H}^{-1} \big) = o_{d,\mathbb{P}}(1) \cdot \sigma_\varepsilon^2$$

and

$$\frac{1}{\sigma_\varepsilon^2} \mathbb{E}[T_4^2] \le \boldsymbol{E}^{\mathsf{T}} \boldsymbol{H}^{-2} \boldsymbol{E} \le o_{d,\mathbb{P}}(1) \cdot \| f_d \|_{L^2}^2$$

which by Markov's inequality gives

$$T_4 = o_{d,\mathbb{P}}(1) \cdot (\sigma_\varepsilon^2 + \| f_d \|_{L^2}^2).$$

Let us now analyze term T_5 which we can write as

$$T_5 = T_{51} + T_{52}$$

where

$$\begin{aligned}
T_{51} &= \boldsymbol{\varepsilon}(t)^{\mathsf{T}} \boldsymbol{H}^{-1} \boldsymbol{M} \boldsymbol{H}^{-1} (\mathbf{I}_n - e^{-t\boldsymbol{H}/n}) \boldsymbol{f}_{\le \ell}, \\
T_{52} &= \boldsymbol{\varepsilon}(t)^{\mathsf{T}} \boldsymbol{H}^{-1} \boldsymbol{M} \boldsymbol{H}^{-1} (\mathbf{I}_n - e^{-t\boldsymbol{H}/n}) \boldsymbol{f}_{>\ell}.
\end{aligned}$$

Note that by Lemma 13

$$\left\| \boldsymbol{M}^{1/2} \boldsymbol{H}^{-2} \boldsymbol{M}^{1/2} \right\|_{\mathrm{op}} = \left\| \boldsymbol{H}^{-1} \boldsymbol{M} \boldsymbol{H}^{-1} \right\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$$

and

$$\begin{aligned}
\frac{1}{\sigma_\varepsilon^2}\mathbb{E}_{\boldsymbol{\varepsilon}}[T_{51}^2] &\le \frac{1}{\sigma_\varepsilon^2}\mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{f}_{\le\ell}\boldsymbol{f}_{\le\ell}^\mathsf{T}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}\boldsymbol{\varepsilon}] \\
&= \boldsymbol{f}_{\le\ell}^\mathsf{T}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})[\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}]^2(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{f}_{\le\ell} \\
&\le \left\|\boldsymbol{M}^{1/2}\boldsymbol{H}^{-2}\boldsymbol{M}^{1/2}\right\|_{\mathrm{op}}\left\|\boldsymbol{M}^{1/2}\boldsymbol{H}^{-1}(\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{f}_{\le\ell}\right\|_2^2 \\
&= o_{d,\mathbb{P}}(1)\cdot T_{21} \\
&= o_{d,\mathbb{P}}(1)\cdot \|\mathsf{P}_{\le\ell}f_d\|_{L^2}^2 .
\end{aligned}$$

Similarly we get that

$$\frac{1}{\sigma_\varepsilon^2}\mathbb{E}_{\boldsymbol{\varepsilon}}[T_{52}^2] = o_{d,\mathbb{P}}(1)\cdot T_{23} = o_{d,\mathbb{P}}(1)\cdot \|f_d\|_{L^{2+\eta}}^2 .$$

By Markov's inequality we deduce that

$$T_5 = o_{d,\mathbb{P}}(1)\cdot\sigma_\varepsilon\cdot(\|\mathsf{P}_{\le\ell}f_d\|_{L^2} + \|f_d\|_{L^{2+\eta}}) = o_{d,\mathbb{P}}(1)\cdot(\sigma_\varepsilon^2 + \|f_d\|_{L^{2+\eta}}^2).$$

Finally putting everything together finishes the proof of Theorem 8

Proof of Theorem 8.

$$\begin{aligned}
R(\hat{f}_t) &= \|f_d\|_{L^2}^2 - 2T_1 + T_2 + T_3 - 2T_4 + 2T_5 \\
&= \|\mathsf{P}_{>\ell}f_d\|_{L^2}^2 + \left\|\widehat{\boldsymbol{f}}_\ell\right\|_2^2 - 2\left\|\boldsymbol{S}_{\le\ell}\widehat{\boldsymbol{f}}_{\le\ell}\right\|^2 + \left\|\boldsymbol{S}_{\le\ell}\widehat{\boldsymbol{f}}_{\le\ell}\right\|^2 \\
&+ o_{d,\mathbb{P}}(1)\cdot(\|f_d\|_{L^2}^2 + \|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2) \\
&= \left\|(\mathbf{I} - \boldsymbol{S}_{\le\ell})\widehat{\boldsymbol{f}}_{\le\ell}\right\|_2^2 + \|\mathsf{P}_{>\ell}f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\cdot(\|f_d\|_{L^2}^2 + \|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2).
\end{aligned}$$

By Assumption 4, $\kappa_H/n = o_d(1)\cdot\max_{j\le\ell}\lambda_{d,j}^2$ hence

$$\left\|(\mathbf{I} - \boldsymbol{S}_{\le\ell})\widehat{\boldsymbol{f}}_{\le\ell}\right\|_2^2 = o_{d,\mathbb{P}}(1)\cdot\|f_d\|_{L^2}^2$$

and

$$R(\hat{f}_t) = \|\mathsf{P}_{>\ell}f_d\|_{L^2}^2 + o_{d,\mathbb{P}}(1)\cdot(\|f_d\|_{L^2}^2 + \|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2).$$

By carefully considering the proof one can show analogously that,

$$\left\|\hat{f}_t - \mathsf{P}_{\le\ell}f_d\right\|_{L^2}^2 = o_{d,\mathbb{P}}(1)\cdot(\|f_d\|_{L^2}^2 + \|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2).$$

□

E AUXILIARY RESULTS

E.1 SOLUTION TO KERNEL DYNAMICS

Recall that we are interested in the following dynamics given in (Eq. 2 and Eq. 3),

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t^{\mathrm{or}}(\boldsymbol{x}) = \mathbb{E}[H_d(\boldsymbol{x}, \boldsymbol{z})(f_d(\boldsymbol{z}) - f_t^{\mathrm{or}}(\boldsymbol{z}))],$$
$$\frac{\mathrm{d}}{\mathrm{d}t} \hat{f}_t(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} H_d(\boldsymbol{x}, \boldsymbol{x}_i)(y_i - \hat{f}_t(\boldsymbol{x}_i))$$

with zero initialization $f_0^{\mathrm{or}} \equiv \hat{f}_0 \equiv 0$. In this section we clarify the derivation, validity, and solution of these dynamics. Let us consider the maps $R : \mathcal{H}_d \to \mathbb{R}$ and $\widehat{R}_n : \mathcal{H}_d \to \mathbb{R}$ defined in (Eq. 1).

$$R(f) = \int_{\mathcal{X}_d} (f(\boldsymbol{x}) - f_d(\boldsymbol{x}))^2 \, \mathrm{d}\nu_d(\boldsymbol{x}) + \sigma_\varepsilon^2,$$
$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - y_i)^2.$$

Then gradients of these maps are given as

$$\boldsymbol{\nabla} R(f) = \mathbb{H}_d(f - f_d),$$
$$\boldsymbol{\nabla} \widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - y_i) H_{\boldsymbol{x}_i},$$

where $H_{\boldsymbol{x}_i}(\boldsymbol{x}) := H_d(\boldsymbol{x}_i, \boldsymbol{x})$ and $\mathbb{H}_d$ is the kernel operator as in Section A.1. A proof of this fact is given in Proposition 2.1 of Yao et al. (2007). Taking $f_t^{\mathrm{or}}(x)$ as shorthand for $f^{\mathrm{or}}(t, x)$ where $f^{\mathrm{or}}(t, \cdot) \in \mathcal{H}_d$ is the oracle model at time t and similarly for $\hat{f}_t(x)$, the following gradient flows with zero initialization are well-defined for $t \geq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t^{\mathrm{or}} = -\boldsymbol{\nabla} R(f_t^{\mathrm{or}}) = -\mathbb{H}_d(f_t^{\mathrm{or}} - f_d) = \mathbb{E}_{\boldsymbol{z}}[H_d(\cdot, \boldsymbol{z})(f_d(\boldsymbol{z}) - f_t^{\mathrm{or}}(\boldsymbol{z})] \tag{20}$$
$$\frac{\mathrm{d}}{\mathrm{d}t} \hat{f}_t = -\boldsymbol{\nabla} \widehat{R}_n(\hat{f}_t) = -\frac{1}{n} \sum_{i=1}^{n} (\hat{f}_t(\boldsymbol{x}_i) - y_i) H_{\boldsymbol{x}_i} = \frac{1}{n} \sum_{i=1}^{n} H_d(\cdot, \boldsymbol{x}_i)(y_i - \hat{f}_t(\boldsymbol{x}_i)). \tag{21}$$

The oracle model ODE (Eq. 20) is simply a linear differential equation which is known to have the following solution involving the operator exponential $\exp(\boldsymbol{A}) := \sum_{k=0}^{\infty} \boldsymbol{A}^k / k!$

$$f_t^{\mathrm{or}} = f_d + \exp(-t\mathbb{H}_d)(f_0^{\mathrm{or}} - f_d) = f_d - \exp(-t\mathbb{H}_d) f_d. \tag{22}$$

For the empirical model ODE (Eq. 21) we first consider the system of scalar differential equations induced at the points $\{(\boldsymbol{x}_i, y_i)\}_{i \in [n]}$. Letting $\boldsymbol{u}(t) = (\hat{f}_t(\boldsymbol{x}_1), \ldots, \hat{f}_t(\boldsymbol{x}_n))$, $\boldsymbol{y} = (y_1, \ldots, y_n)$, and $\boldsymbol{H} = (H_d(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j \in [n]}$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{u}(t) = -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{H}(\boldsymbol{u}(t) - \boldsymbol{y}),$$

with initial condition $\boldsymbol{u}(0) = \boldsymbol{0}$. As this a linear ODE, the solution is given by

$$\boldsymbol{u}(t) = \boldsymbol{y} - e^{-t\boldsymbol{H}/n}(\boldsymbol{u}(0) - \boldsymbol{y}) = (\mathbf{I}_n - e^{-t\boldsymbol{H}/n})\boldsymbol{y}. \tag{23}$$

For arbitrary $\boldsymbol{x}$, define $\boldsymbol{h}(\boldsymbol{x}) = (H_d(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, H_d(\boldsymbol{x}, \boldsymbol{x}_n))^\mathsf{T} \in \mathbb{R}^n$. Let $\boldsymbol{a}(t) = \boldsymbol{H}^{-1}\boldsymbol{u}(t) \in \mathbb{R}^n$. We will show that the function $\hat{f}_t(\cdot) := \langle \boldsymbol{h}(\cdot), \boldsymbol{a}(t) \rangle \in \mathcal{H}_d$, which satisfies $(\hat{f}_t(\boldsymbol{x}_1), \ldots, \hat{f}_t(\boldsymbol{x}_n))^\mathsf{T} = \boldsymbol{u}(t)$, is a solution to the equation below

$$\frac{\mathrm{d}}{\mathrm{d}t} \hat{f}_t(\boldsymbol{x}) = \frac{1}{n} \langle \boldsymbol{h}(x), \boldsymbol{y} - \boldsymbol{u}(t) \rangle = \frac{1}{n} \langle \boldsymbol{h}(x), e^{-t\boldsymbol{H}/n} \boldsymbol{y} \rangle.$$

which is Eq. 21 at point $\boldsymbol{x}$. By the chain rule

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\hat{f}_t(\boldsymbol{x}) &= \frac{\mathrm{d}}{\mathrm{d}t}\langle \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{a}(t)\rangle \\ &= \langle \boldsymbol{h}(\boldsymbol{x}), \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a}(t)\rangle \\ &= \langle \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{H}^{-1}\frac{1}{n}\boldsymbol{H}e^{-t\boldsymbol{H}/n}\boldsymbol{y}\rangle \\ &= \frac{1}{n}\langle \boldsymbol{h}(x), e^{-t\boldsymbol{H}/n}\boldsymbol{y}\rangle\end{aligned}$$

which is what we wanted to show.

E.2 Equivalence between Invariant Kernels and Data Augmentation

In this section we will show an equivalence between the gradient flows for invariant kernels and augmented dataset. Specifically consider a group $\mathcal{G}$ and a H that is $\mathcal{G}$-equivariant, that is

$$H(g\cdot \boldsymbol{x}_1, g'\cdot \boldsymbol{x}_2) = H(\boldsymbol{x}_1, \boldsymbol{x}_2) \quad \forall g, g' \in \mathcal{G}, \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}.$$

The invariant kernel H_{inv} is defined as a group averaged kernel

$$H_{\text{inv}} = \int_{\mathcal{G}} H(\boldsymbol{x}_1, g\cdot \boldsymbol{x}_2)\pi(\mathrm{d}g)$$

for the Haar measure π on $\mathcal{G}$. Given a dataset $(\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_i, y_i : i \in [n]\}$ consider the augmented dataset $(\boldsymbol{X}_{\mathcal{G}}, \boldsymbol{y}_{\mathcal{G}}) = \{(g\cdot \boldsymbol{x}_i, y_i) : g \in \mathcal{G}, i \in [n]\}$. We consider the empirical dynamics (Eq. 3) of the gradient flow on $(\boldsymbol{X}, \boldsymbol{y})$ using H_{inv} which we denote $\hat{f}_{t,\text{inv}}$ and the empirical dynamics of the gradient flow on $(\boldsymbol{X}_{\mathcal{G}}, \boldsymbol{y}_{\mathcal{G}})$ using H which we denote $\hat{f}_{t,\text{aug}}$.

Proposition 24. *Let $\mathcal{G}$ be a finite group with m elements. Given a $\mathcal{G}$-equivariant kernel H, if π is the uniform measure on $\mathcal{G}$ then*

$$\hat{f}_{mt,\text{inv}} \equiv \hat{f}_{t,\text{aug}}.$$

Proof. Let $\mathcal{G} = \{g_1, \dots, g_m\}$ where g_1 is the identity. Define the output vectors

$$\begin{aligned}\boldsymbol{u}_{\text{inv}}(t) &:= (\hat{f}_{t,\text{inv}}(\boldsymbol{x}_1), \dots, \hat{f}_{t,\text{inv}}(\boldsymbol{x}_n)) \in \mathbb{R}^n, \\ \boldsymbol{u}_g(t) &:= (\hat{f}_{t,\text{aug}}(g\cdot \boldsymbol{x}_1), \dots, \hat{f}_{t,\text{aug}}(g\cdot \boldsymbol{x}_n)) \in \mathbb{R}^n, \\ \boldsymbol{u}_{\text{aug}}(t) &:= (\boldsymbol{u}_{g_1}(t), \dots, \boldsymbol{u}_{g_m}(t)) \in \mathbb{R}^{mn}.\end{aligned}$$

Furthermore, define the kernel matrices

$$\begin{aligned}H_{g,g'} &:= [H(g\cdot \boldsymbol{x}_i, g'\cdot \boldsymbol{x}_j)]_{i,j\in[n]} \in \mathbb{R}^{n\times n} \text{ for } g, g' \in \mathcal{G}, \\ H_{\text{aug}} &:= [H_{g,g'}]_{g,g'\in\mathcal{G}} \in \mathbb{R}^{mn\times mn}.\end{aligned}$$

Note that by definition $H_{\text{inv}} = \frac{1}{m}\sum_{j=1}^{m} H_{g_1,g_j}$. We will show that

$$\boldsymbol{u}_{\text{aug}}(t) = (u_{\text{inv}}(mt), u_{\text{inv}}(mt), \dots, u_{\text{inv}}(mt)) \text{ for all } t \geq 0 \tag{24}$$

From this the result follows by Theorem 4.1 in Li et al. (2019) since $\hat{f}_{mt,\text{inv}}, \hat{f}_{t,\text{aug}}$ are given by kernel regressions with targets $\boldsymbol{u}_{\text{inv}}(mt), \boldsymbol{u}_{\text{aug}}(t)$ and kernels H_{inv}, H respectively.

By (Eq. 21), we can write

$$\begin{aligned}\boldsymbol{u}_{\text{inv}}(t) &= (\mathbf{I} - \exp(-tH_{\text{inv}}/n))\boldsymbol{y}, \\ \boldsymbol{u}_{\text{aug}}(t) &= (\mathbf{I} - \exp(-tH_{\text{aug}}/n))\boldsymbol{y}_{\mathcal{G}},\end{aligned}$$

where $\boldsymbol{y}_{\mathcal{G}} = (\boldsymbol{y}, \dots, \boldsymbol{y}) \in \mathbb{R}^{mn}$. By expanding the matrix exponential series and linearity to show (Eq. 24) it suffices to show that

$$H_{\text{aug}}^k \boldsymbol{y}_{\mathcal{G}} = (m^k H_{\text{inv}}^k \boldsymbol{y}, m^k H_{\text{inv}}^k \boldsymbol{y}, \dots, m^k H_{\text{inv}}^k \boldsymbol{y}) \text{ for all } k \in \mathbb{N}.$$

We prove the above by induction on k. For $k = 1$, observe that

$$\begin{aligned} H_{\text{aug}}\boldsymbol{y}_{\mathcal{G}} &= \left(\sum_{j=1}^{m} H_{g_i,g_j}\boldsymbol{y}\right)_{i=1}^{m} \\ &= \left(\sum_{j=1}^{m} H_{g_1,g_j}\boldsymbol{y}\right)_{i=1}^{m} \\ &= (mH_{\text{inv}}\boldsymbol{y})_{i=1}^{m} \end{aligned}$$

where the second equality follows from $\mathcal{G}$-equivariance of H. Assume the inductive hypothesis holds for k. Then

$$\begin{aligned} H_{\text{aug}}^{k+1}\boldsymbol{y}_{\mathcal{G}} &= H_{\text{aug}}H_{\text{aug}}^{k}\boldsymbol{y}_{\mathcal{G}} \\ &= \left(m^k \sum_{j=1}^{m} H_{g_i,g_j}H_{\text{inv}}^{k}\boldsymbol{y}\right)_{i=1}^{m} \\ &= \left(\sum_{j=1}^{m} H_{g_1,g_j}\left(\sum_{j'=1}^{m} H_{g_1,g_{j'}}\right)^{k}\boldsymbol{y}\right)_{i=1}^{m} \\ &= \left(\left(\sum_{j=1}^{m} H_{g_1,g_j}\right)^{k+1}\boldsymbol{y}\right)_{i=1}^{m} \\ &= (m^{k+1}H_{\text{inv}}^{k+1}\boldsymbol{y})_{i=1}^{m} \end{aligned}$$

where the second equality applies the induction hypothesis and the third equality uses equivariance. Thus the inductive claim is proved and the proof is complete.

□

Corollary 25. *If $\mathcal{G}_d = \mathrm{Cyc}_d$ and $\pi_d = \mathrm{Unif}(\mathrm{Cyc}_d)$, then for a dot product kernel H and corresponding invariant kernel H_{inv} we have*

$$\hat{f}_{t,\text{inv}} \equiv \hat{f}_{dt,\text{aug}}.$$

Proof. This simply follows from the previous proposition since $|\,\mathrm{Cyc}_d\,| = d$ and a dot product kernel is Cyc_d-equivariant. □

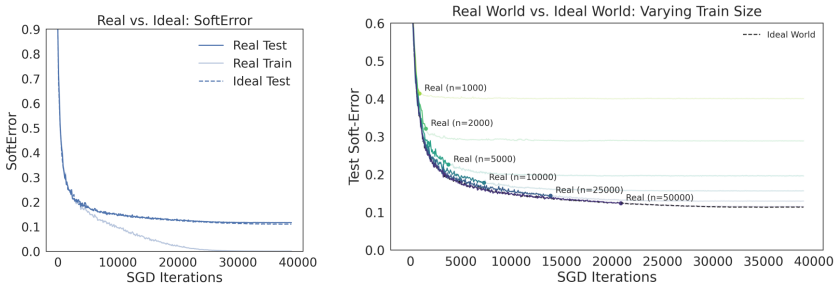
Remark 5. *The two gradient flows $\hat{f}_{t,\text{inv}}$, $\hat{f}_{t,\text{aug}}$ are equivalent up to a rescaling of time. It appears from 24, that $\hat{f}_{t,\text{aug}}$ is m-times faster, however as elaborated in E.3, to compare optimization speed we need to discretize time. As argued in that section, it is natural to take a step-size inversely proportional to the maximum kernel eigenvalue. In the case of high-dimensional dot product kernels, note that*

$$\lambda_{\max}(H_{\text{aug}}) = m\lambda_{\max}(H) \sim m\lambda_{\max}(H_{\text{inv}})$$

hence when discretizing the trajectories, the step-size for the data augmented should be $1/m$ times smaller, thus both trajectories will have roughly the same speed.

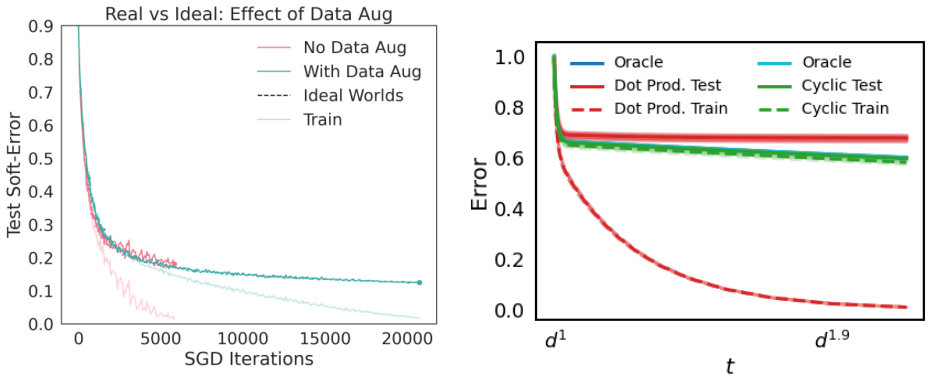
E.3 Discretizing Time

Comparing different "speeds" of optimization algorithms only makes sense for discrete-time algorithms. For kernel least-squares problems, we argue that a natural step-size to consider is $\eta \sim \frac{1}{\lambda_{\max}(\boldsymbol{H})}$ where $\boldsymbol{H}$ is the kernel matrix. This is due to a classical result in convex optimization that states that convergence of gradient descent is guaranteed for step-sizes $\eta < \frac{1}{2L}$, where the



(a) (b)

Figure 6: Soft-error curves of both worlds for Resnet-18 trained on CIFAR-5m. Figures from Nakkiran et al. (2020)). Panel (6a): $n = 5 \times 10^4$ (Figure 6). Panel (6b) Varying n (Figure 4a).



(a) (b)

Figure 7: Panel (7a) Data-augmentation for Resnet-18 on CIFAR-5m. (Fig. 5a from Nakkiran et al. (2020)). Panel (7b) Cyclic versus dot product kernel (Fig. 5c)

smoothness L is the Lipschitz constant of the gradient of the loss. For kernel least-squares problems $L = \lambda_{\max}$. For a dot product kernel H and its corresponding invariant kernel H_{inv} the kernel matrices have operator norms of the same order

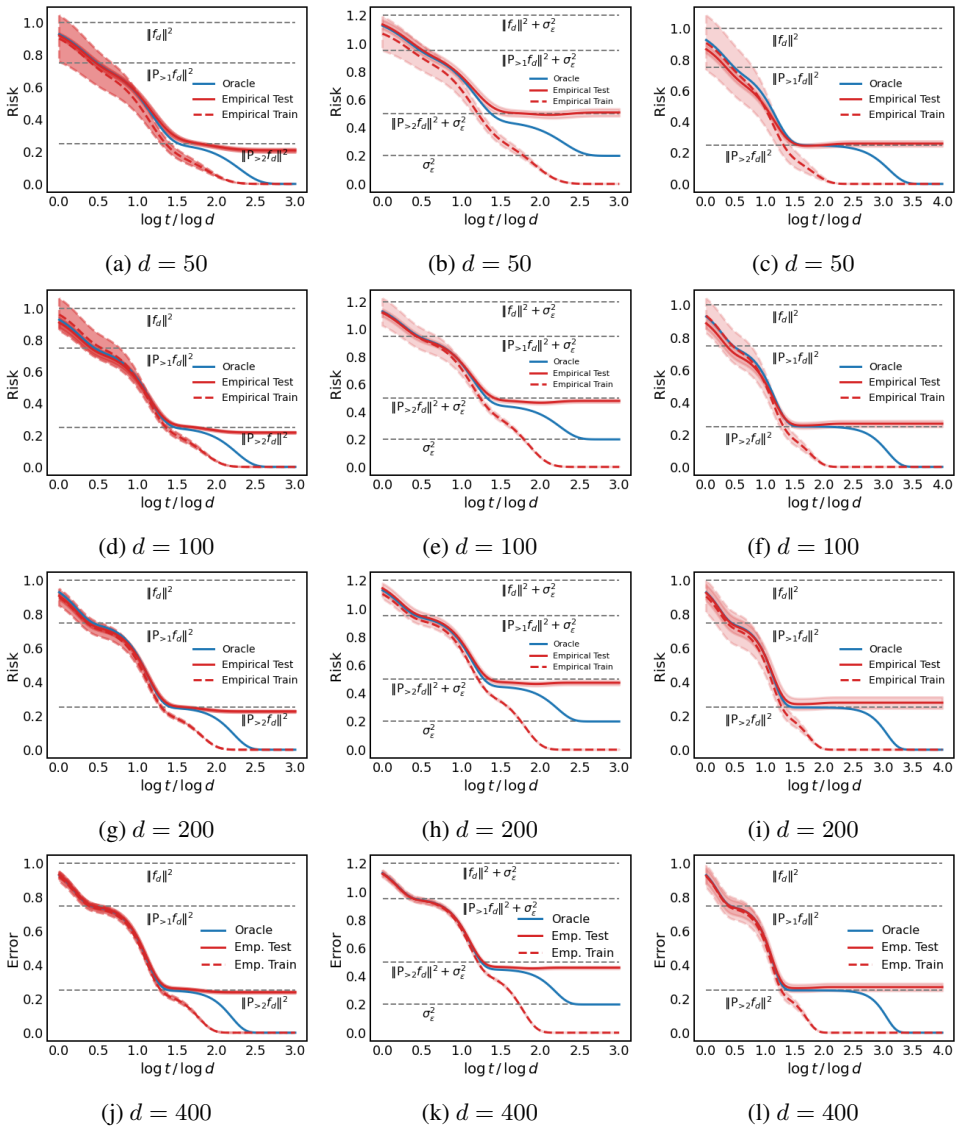
$$\lambda_{\max}(\boldsymbol{H}) \sim \lambda_{\max}(\boldsymbol{H}_{\text{inv}})$$

hence no continuous-time rescaling is need to compare the corresponding optimization speeds.

E.4 Similarities with Empirical Phenomena

In this section we elaborate upon Remark 4 and mention some connections with empirical observations in Nakkiran et al. (2020). Although the metric in our setting is the squared loss, we can still observe three stages in classification problems when measuring the soft error. In Fig. 6a we can observe stage 1 and stage 2. Either training has not continued long enough to observe stage three or n is large enough so that the models have converged to the approximation error of the neural network class (c.f. Remark 1). In Fig. 6b, although the train errors are not plotted, with reference to Fig. 6a, presumably for each n stage 1 and stage 2 occur and for $n < 50000$ stage 3 occurs as well.

In Fig. 7a, we see a parallel between the use of cyclic versus dot product kernels (which is equivalent to data-augmentation c.f. Section E.2) and data augmentation for a Resnet-18 trained on CIFAR-5m. In both cases, the ideal worlds of augmented and non-augmented training are the same, but for augmented training the real world training is slowed down, leading to better generalization.



(a) $d = 50$ (b) $d = 50$ (c) $d = 50$

(d) $d = 100$ (e) $d = 100$ (f) $d = 100$

(g) $d = 200$ (h) $d = 200$ (i) $d = 200$

(j) $d = 400$ (k) $d = 400$ (l) $d = 400$

Figure 8: Each column replicates a kernel gradient flow experiment from Fig. 4 over $d \in \{50, 100, 200, 400\}$. **Column 1:** Replicates Fig. 4a **Column 2:** Replicates Fig. 4a but with $\sigma_\varepsilon^2 = 0.2$. **Column 3:** Replicates Fig. 4b. Averaged over 10 trials.

F Additional Figures

To see the effects of varying the dimension d we replicate the log-scale plots of kernel gradient flow with dot product kernels in Fig. 4. We take $n = d^{1.5}$ and vary $d \in \{50, 100, 200, 400\}$. Each plot is averaged over 10 runs with the shaded region representing one standard deviation around the mean. We can see that as d increasing the standard deviation decreases and the curves approach the theoretical high-dimensional prediction.

G TECHNICAL BACKGROUND

G.1 NOTATIONS

For a positive integer, we denote by $[n]$ the set $\{1, 2, \dots, n\}$. For vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we denote $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = u_1 v_1 + \dots + u_d v_d$ their scalar product, and $\|\boldsymbol{u}\|_2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle^{1/2}$ the ℓ_2 norm. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, we denote $\|\boldsymbol{A}\|_{\mathrm{op}} = \max_{\|\boldsymbol{u}\|_2 = 1} \|\boldsymbol{A}\boldsymbol{u}\|_2$ its operator norm and by $\|\boldsymbol{A}\|_F = \big(\sum_{i,j} A_{ij}^2\big)^{1/2}$ its Frobenius norm. If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a square matrix, the trace of $\boldsymbol{A}$ is denoted by $\mathrm{Tr}(\boldsymbol{A}) = \sum_{i \in [n]} A_{ii}$.

We use $O_d(\,\cdot\,)$ (resp. $o_d(\,\cdot\,)$) for the standard big-O (resp. little-o) relations, where the subscript d emphasizes the asymptotic variable. Furthermore, we write $f = \Omega_d(g)$ if $g(d) = O_d(f(d))$, and $f = \omega_d(g)$ if $g(d) = o_d(f(d))$. Finally, $f = \Theta_d(g)$ if we have both $f = O_d(g)$ and $f = \Omega_d(g)$.

We use $O_{d,\mathbb{P}}(\,\cdot\,)$ (resp. $o_{d,\mathbb{P}}(\,\cdot\,)$) the big-O (resp. little-o) in probability relations. Namely, for $h_1(d)$ and $h_2(d)$ two sequences of random variables, $h_1(d) = O_{d,\mathbb{P}}(h_2(d))$ if for any $\varepsilon > 0$, there exists $C_\varepsilon > 0$ and $d_\varepsilon \in \mathbb{Z}_{>0}$, such that

$$\mathbb{P}(|h_1(d)/h_2(d)| > C_\varepsilon) \le \varepsilon, \qquad \forall d \ge d_\varepsilon,$$

and respectively: $h_1(d) = o_{d,\mathbb{P}}(h_2(d))$, if $h_1(d)/h_2(d)$ converges to 0 in probability. Similarly, we will denote $h_1(d) = \Omega_{d,\mathbb{P}}(h_2(d))$ if $h_2(d) = O_{d,\mathbb{P}}(h_1(d))$, and $h_1(d) = \omega_{d,\mathbb{P}}(h_2(d))$ if $h_2(d) = o_{d,\mathbb{P}}(h_1(d))$. Finally, $h_1(d) = \Theta_{d,\mathbb{P}}(h_2(d))$ if we have both $h_1(d) = O_{d,\mathbb{P}}(h_2(d))$ and $h_1(d) = \Omega_{d,\mathbb{P}}(h_2(d))$.

G.2 FUNCTIONAL SPACES OVER THE SPHERE

For $d \ge 3$, we let $\mathbb{S}^{d-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = r\}$ denote the sphere with radius r in $\mathbb{R}^d$. We will mostly work with the sphere of radius $\sqrt{d}$, $\mathbb{S}^{d-1}(\sqrt{d})$ and will denote by τ_d the uniform probability measure on $\mathbb{S}^{d-1}(\sqrt{d})$. All functions in this section are assumed to be elements of $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \tau_d)$, with scalar product and norm denoted as $\langle\,\cdot\,,\,\cdot\,\rangle_{L^2}$ and $\|\,\cdot\,\|_{L^2}$:

$$\langle f, g \rangle_{L^2} \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} f(\boldsymbol{x})\, g(\boldsymbol{x})\, \tau_d(\mathrm{d}\boldsymbol{x})\,. \tag{25}$$

For $\ell \in \mathbb{Z}_{\ge 0}$, let $\tilde{V}_{d,\ell}$ be the space of homogeneous harmonic polynomials of degree ℓ on $\mathbb{R}^d$ (i.e. homogeneous polynomials $q(\boldsymbol{x})$ satisfying $\Delta q(\boldsymbol{x}) = 0$), and denote by $V_{d,\ell}$ the linear space of functions obtained by restricting the polynomials in $\tilde{V}_{d,\ell}$ to $\mathbb{S}^{d-1}(\sqrt{d})$. With these definitions, we have the following orthogonal decomposition

$$L^2(\mathbb{S}^{d-1}(\sqrt{d}), \tau_d) = \bigoplus_{\ell=0}^{\infty} V_{d,\ell}\,. \tag{26}$$

The dimension of each subspace is given by

$$\dim(V_{d,\ell}) = B(\mathbb{S}^{d-1}; \ell) = \frac{2\ell + d - 2}{d-2} \binom{\ell + d - 3}{\ell}\,. \tag{27}$$

For each $\ell \in \mathbb{Z}_{\ge 0}$, the spherical harmonics $\{Y_{\ell,j}^{(d)}\}_{1 \le j \le B(\mathbb{S}^{d-1};\ell)}$ form an orthonormal basis of $V_{d,\ell}$:

$$\langle Y_{ki}^{(d)}, Y_{sj}^{(d)} \rangle_{L^2} = \delta_{ij}\delta_{ks}.$$

Note that our convention is different from the more standard one, that defines the spherical harmonics as functions on $\mathbb{S}^{d-1}(1)$. It is immediate to pass from one convention to the other by a simple scaling. We will drop the superscript d and write $Y_{\ell,j} = Y_{\ell,j}^{(d)}$ whenever clear from the context.

We denote by $\overline{\mathsf{P}}_k$ the orthogonal projections to $V_{d,k}$ in $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \tau_d)$. This can be written in terms of spherical harmonics as

$$\overline{\mathsf{P}}_k f(\boldsymbol{x}) \equiv \sum_{l=1}^{B(\mathbb{S}^{d-1};k)} \langle f, Y_{kl} \rangle_{L^2} Y_{kl}(\boldsymbol{x}). \tag{28}$$

We also define $\overline{\mathsf{P}}_{\le \ell} \equiv \sum_{k=0}^{\ell} \overline{\mathsf{P}}_k$, $\overline{\mathsf{P}}_{>\ell} \equiv \mathbf{I} - \overline{\mathsf{P}}_{\le \ell} = \sum_{k=\ell+1}^{\infty} \overline{\mathsf{P}}_k$, and $\overline{\mathsf{P}}_{<\ell} \equiv \overline{\mathsf{P}}_{\le \ell-1}$, $\overline{\mathsf{P}}_{\ge \ell} \equiv \overline{\mathsf{P}}_{>\ell-1}$.

G.3 GEGENBAUER POLYNOMIALS

The ℓ-th Gegenbauer polynomial $Q_\ell^{(d)}$ is a polynomial of degree ℓ. Consistently with our convention for spherical harmonics, we view $Q_\ell^{(d)}$ as a function $Q_\ell^{(d)} : [-d, d] \to \mathbb{R}$. The set $\{Q_\ell^{(d)}\}_{\ell \geq 0}$ forms an orthogonal basis on $L^2([-d, d], \tilde{\tau}_d^1)$, where $\tilde{\tau}_d^1$ is the distribution of $\sqrt{d}\langle \boldsymbol{x}, \boldsymbol{e}_1\rangle$ when $\boldsymbol{x} \sim \tau_d$, satisfying the normalization condition:

$$\langle Q_k^{(d)}(\sqrt{d}\langle \boldsymbol{e}_1, \cdot\rangle), Q_j^{(d)}(\sqrt{d}\langle \boldsymbol{e}_1, \cdot\rangle)\rangle_{L^2(\mathbb{S}^{d-1}(\sqrt{d}))} = \frac{1}{B(d;k)}\,\delta_{jk}\,. \tag{29}$$

In particular, these polynomials are normalized so that $Q_\ell^{(d)}(d) = 1$. As above, we will omit the superscript (d) in $Q_\ell^{(d)}$ when clear from the context.

Gegenbauer polynomials are directly related to spherical harmonics as follows. Fix $\boldsymbol{v} \in \mathbb{S}^{d-1}(\sqrt{d})$ and consider the subspace of V_ℓ formed by all functions that are invariant under rotations in $\mathbb{R}^d$ that keep $\boldsymbol{v}$ unchanged. It is not hard to see that this subspace has dimension one, and coincides with the span of the function $Q_\ell^{(d)}(\langle \boldsymbol{v}, \cdot\rangle)$.

We will use the following properties of Gegenbauer polynomials

1. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$\langle Q_j^{(d)}(\langle \boldsymbol{x}, \cdot\rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot\rangle)\rangle_{L^2} = \frac{1}{B(d;k)}\delta_{jk} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle). \tag{30}$$

2. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) = \frac{1}{B(d;k)} \sum_{i=1}^{B(d;k)} Y_{ki}^{(d)}(\boldsymbol{x}) Y_{ki}^{(d)}(\boldsymbol{y}). \tag{31}$$

These properties imply that —up to a constant— $Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)$ is a representation of the projector onto the subspace of degree -k spherical harmonics

$$(\overline{\mathsf{P}}_k f)(\boldsymbol{x}) = B(d;k) \int_{\mathbb{S}^{d-1}(\sqrt{d})} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\, f(\boldsymbol{y})\, \tau_d(\mathrm{d}\boldsymbol{y})\,. \tag{32}$$

For a function $\sigma \in L^2([-\sqrt{d}, \sqrt{d}], \tau_d^1)$ (where τ_d^1 is the distribution of $\langle \boldsymbol{e}_1, \boldsymbol{x}\rangle$ when $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$), denoting its spherical harmonics coefficients $\xi_{d,k}(\sigma)$ to be

$$\xi_{d,k}(\sigma) = \int_{[-\sqrt{d}, \sqrt{d}]} \sigma(x) Q_k^{(d)}(\sqrt{d}x)\tau_d^1(\mathrm{d}x), \tag{33}$$

then we have the following equation holds in $L^2([-\sqrt{d}, \sqrt{d}], \tau_d^1)$ sense

$$\sigma(x) = \sum_{k=0}^{\infty} \xi_{d,k}(\sigma) B(d;k) Q_k^{(d)}(\sqrt{d}x).$$

For any rotationally invariant kernel $H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = h_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d)$, with $h_d(\sqrt{d}\,\cdot) \in L^2([-\sqrt{d}, \sqrt{d}], \tau_d^1)$, we can associate a self adjoint operator $\mathcal{H}_d : L^2(\mathbb{S}^{d-1}(\sqrt{d})) \to L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ via

$$\mathcal{H}_d f(\boldsymbol{x}) \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} h_d(\langle \boldsymbol{x}, \boldsymbol{x}_1\rangle/d)\, f(\boldsymbol{x}_1)\, \tau_d(\mathrm{d}\boldsymbol{x}_1)\,. \tag{34}$$

By rotational invariance, the space V_k of homogeneous polynomials of degree k is an eigenspace of $\mathcal{H}_d$, and we will denote the corresponding eigenvalue by $\xi_{d,k}(h_d)$. In other words $\mathcal{H}_d f(\boldsymbol{x}) \equiv \sum_{k=0}^{\infty} \xi_{d,k}(h_d)\overline{\mathsf{P}}_k f$. The eigenvalues can be computed via

$$\xi_{d,k}(h_d) = \int_{[-\sqrt{d}, \sqrt{d}]} h_d\big(x/\sqrt{d}\big) Q_k^{(d)}(\sqrt{d}x)\tau_d^1(\mathrm{d}x)\,. \tag{35}$$

G.4 HERMITE POLYNOMIALS

The Hermite polynomials $\{\mathrm{He}_k\}_{k\ge 0}$ form an orthogonal basis of $L^2(\mathbb{R},\gamma)$, where $\gamma(\mathrm{d}x) = e^{-x^2/2}\mathrm{d}x/\sqrt{2\pi}$ is the standard Gaussian measure, and He_k has degree k. We will follow the classical normalization (here and below, expectation is with respect to $G\sim\mathcal{N}(0,1)$):

$$\mathbb{E}\big\{\mathrm{He}_j(G)\,\mathrm{He}_k(G)\big\} = k!\,\delta_{jk}\,. \tag{36}$$

As a consequence, for any function $g\in L^2(\mathbb{R},\gamma)$, we have the decomposition

$$g(x) = \sum_{k=0}^{\infty}\frac{\mu_k(g)}{k!}\,\mathrm{He}_k(x)\,,\qquad \mu_k(g)\equiv\mathbb{E}\big\{g(G)\,\mathrm{He}_k(G)\big\}\,. \tag{37}$$

The Hermite polynomials can be obtained as high-dimensional limits of the Gegenbauer polynomials introduced in the previous section. Indeed, the Gegenbauer polynomials (up to a $\sqrt{d}$ scaling in domain) are constructed by Gram-Schmidt orthogonalization of the monomials $\{x^k\}_{k\ge 0}$ with respect to the measure $\tilde{\tau}_d^1$, while Hermite polynomial are obtained by Gram-Schmidt orthogonalization with respect to γ. Since $\tilde{\tau}_d^1\Rightarrow\gamma$ (here $\Rightarrow$ denotes weak convergence), it is immediate to show that, for any fixed integer k,

$$\lim_{d\to\infty}\mathrm{Coeff}\{Q_k^{(d)}(\sqrt{d}x)\,B(\mathbb{S}^{d-1};k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\,\mathrm{He}_k(x)\right\}\,. \tag{38}$$

Here and below, for P a polynomial, $\mathrm{Coeff}\{P(x)\}$ is the vector of the coefficients of P. As a consequence, for any fixed integer k, we have

$$\mu_k(\sigma) = \lim_{d\to\infty}\xi_{d,k}(\sigma)(B(d;k)k!)^{1/2}, \tag{39}$$

where $\mu_k(\sigma)$ and $\xi_{d,k}(\sigma)$ are given in Eq. (37) and (33).

G.5 THE INVARIANT FUNCTION CLASS AND THE SYMMETRIZATION OPERATOR

Let $\mathcal{G}_d$ be a group that is isomorphic to a subgroup of $\mathcal{O}(d)$, the orthogonal group in d dimension. That means, each element of $\mathcal{G}_d$ can be identified with a matrix in $\mathcal{O}(d)\subseteq\mathbb{R}^{d\times d}$, and the group addition operation in $\mathcal{G}_d$ can be regarded as matrix multiplications in $\mathcal{O}(d)$. For any $\boldsymbol{x}\in\mathbb{S}^{d-1}(\sqrt{d})$ and $g\in\mathcal{G}_d$, we define group action $g\cdot\boldsymbol{x}$ to be the multiplication of matrix representation of g with the vector $\boldsymbol{x}$. We equip $\mathcal{G}_d$ with a probability measure π_d, which is the uniform probability measure on $\mathcal{G}_d$. More specifically, the Borel sigma algebra on $\mathcal{G}_d$ is defined as the Borel sigma algebra of $\mathcal{O}(d)$ restricted on $\mathcal{G}_d$. The uniform probability measure π_d satisfies the property that, for any Borel-measurable set $B\subseteq\mathcal{G}_d$ and any $g\in\mathcal{G}_d$, we have

$$\pi_d(B) = \pi_d(gB).$$

Let $L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ be the class of L^2 functions on $\mathbb{S}^{d-1}(\sqrt{d})$ equipped with uniform probability measure $\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. We define the invariant function class to be

$$L^2(\mathbb{S}^{d-1}(\sqrt{d}),\mathcal{G}_d) = \Big\{f\in L^2(\mathbb{S}^{d-1}(\sqrt{d})) : f(\boldsymbol{x}) = f(g\cdot\boldsymbol{x}),\ \ \forall\boldsymbol{x}\in\mathbb{S}^{d-1}(\sqrt{d}),\ \ \forall g\in\mathcal{G}_d\Big\}.$$

We define the symmetrization operator $\mathcal{S}: L^2(\mathbb{S}^{d-1}(\sqrt{d}))\to L^2(\mathbb{S}^{d-1}(\sqrt{d}),\mathcal{G}_d)$ to be

$$(\mathcal{S}f)(\boldsymbol{x}) = \int_{\mathcal{G}_d} f(g\cdot\boldsymbol{x})\pi_d(\mathrm{d}g).$$

G.6 ORTHOGONAL POLYNOMIALS ON INVARIANT FUNCTION CLASS

We define $V_{d,\le k}\subseteq L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ to be the subspace spanned by all the degree ℓ polynomials, $V_{d,>k}\equiv V_{d,\le k}^{\perp}\subseteq L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ to be the orthogonal complement of $V_{d,\le k}$, and $V_{d,k} = V_{d,\le k}\cap V_{d,\le k-1}^{\perp}$. In words, $V_{d,k}$ contains all degree k polynomials that orthogonal to all polynomials of degree at most $k-1$. We further define $V_{d,<k} = V_{d,\le k-1}$ and $V_{d,\ge k} = V_{d,>k-1}$.

Let $\overline{\mathsf{P}}_{\leq \ell}$ to be the projection operator on $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mathrm{Unif})$ that project a function onto $V_{d,\leq \ell}$, the space spanned by all the degree ℓ polynomials. Then it is easy to see that $\overline{\mathsf{P}}_{\leq \ell}$ and $\mathcal{S}$ operator commute. This means, for any $f \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$, we have

$$\overline{\mathsf{P}}_{\leq \ell}[\mathcal{S}(f)] = \mathcal{S}[\overline{\mathsf{P}}_{\leq \ell}(f)].$$

Similarly, we can define $\overline{\mathsf{P}}_{\ell}$, $\overline{\mathsf{P}}_{<\ell}$, $\overline{\mathsf{P}}_{>\ell}$, $\overline{\mathsf{P}}_{\geq \ell}$, which commute with $\mathcal{S}$. We denote $V_{d,\ell}(\mathcal{G}_d) \equiv \mathcal{P}_\ell(\mathbb{S}^{d-1}(\sqrt{d}), \mathcal{G}_d)$ to be the space of polynomials in the images of $\overline{\mathsf{P}}_\ell \mathcal{S}$. Then we have

$$\mathcal{P}_\ell(\mathcal{A}_d, \mathcal{G}_d) = \overline{\mathsf{P}}_\ell(L^2(\mathbb{S}^{d-1}(\sqrt{d}), \mathcal{G}_d)) = \mathcal{S}[\overline{\mathsf{P}}_\ell(L^2(\mathcal{A}_d))].$$

We denote $D(d;k) = D(\mathbb{S}^{d-1}(\sqrt{d}); \mathcal{G}_d; k) \equiv \dim(\mathcal{P}_k(\mathbb{S}^{d-1}(\sqrt{d}), \mathcal{G}_d))$ to be the dimension of $\mathcal{P}_k(\mathbb{S}^{d-1}(\sqrt{d}), \mathcal{G}_d)$. We denote $\{\overline{Y}_{kl}^{(d)}\}_{l \in [D(\mathbb{S}^{d-1}(\sqrt{d});k)]}$ to be a set of orthonormal polynomial basis in $\mathcal{P}_k(\mathbb{S}^{d-1}(\sqrt{d}), \mathcal{G}_d)$. That means

$$\mathbb{E}_{\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\overline{Y}_{k_1 l_1}^{(d)}(\boldsymbol{x}) \overline{Y}_{k_2 l_2}^{(d)}(\boldsymbol{x})] = \mathbf{1}\{k_1 = k_2, l_1 = l_2\},$$

and

$$\overline{Y}_{kl}^{(d)}(\boldsymbol{x}) = \overline{Y}_{kl}^{(d)}(g \cdot \boldsymbol{x}), \ \ \forall \boldsymbol{x} \in \mathbb{S}^{d-1}(\sqrt{d}), \ \ \forall g \in \mathcal{G}_d.$$