The SSL Interplay: Augmentations, Inductive Bias, and Generalization

Vivien Cabannes [1] **Bobak T. Kiani** [2] **Randall Balestriero** [1] **Yann LeCun** [1] **Alberto Bietti** [1]

Abstract

Self-supervised learning (SSL) has emerged as a powerful framework to learn representations from raw data without supervision. Yet in practice, engineers face issues such as instability in tuning optimizers and collapse of representations during training. Such challenges motivate the need for a theory to shed light on the complex interplay between the choice of data augmentation, network architecture, and training algorithm. We study such an interplay with a precise analysis of generalization performance on both pretraining and downstream tasks in a theory friendly setup, and highlight several insights for SSL practitioners that arise from our theory.

1. Introduction

Self-supervised learning (SSL) aims to construct useful representations of data without the need for pre-constructed labels. Due to the recent success and widespread applicability of SSL, established methods for training large neural networks now incorporate pre-training of models in an unsupervised manner over large amounts of data, before fine-tuning/probing them over downstream datasets (Devlin et al., 2019; Chen et al., 2020; Brown et al., 2020; Radford et al., 2021). Self-supervised pretraining generally aims to render the model invariant to certain distorsions/views of the inputs, in order to capture useful features for downstream tasks (e.g., Chen et al., 2020; Caron et al., 2020; Grill et al., 2020; Caron et al., 2021; Bardes et al., 2022). Though very powerful, SSL methods can be challenging to implement properly. They tend to suffer from various practical issues, such as instability and collapse during training and the need to carefully tune parameters related to the architecture, optimization algorithm, representation dimension, and form of augmentations. These different aspects of pretraining can lead to widely different behaviors and representations, as illustrated for instance in Figure 1. These challenges motivate new theoretical insights to better understand why such issues arise and how to better address them.

Our study focuses on the joint-embedding framework and characterizes learned representations for given choices of input distributions, data augmentations, and architecture. To obtain a fine-grained picture, we study linear classes of functions endowed with a reproducing kernel, and analyze a theoretically friendly loss function that models both contrastive and non-contrastive methods. Our work generalizes the discrete data setting of HaoChen et al. (2021) and the finite dimensional setting of Saunshi et al. (2022), encompassing more expressive nonparametric models, potentially with universal approximation properties, and which can capture certain properties of architectures through their limiting kernel limits (Jacot et al., 2018).

Our *contributions* are as follows:

1. We unveil two central integral operators: an "intrinsic" one that depends on the input distribution and choice of augmentations and another capturing the inductive bias associated with the model of computation.
2. We provide new bounds on the downstream generalization error that are sharper than previous work, and which can handle distributions shift between data before and after performing augmentations.
3. We propose new generalization bounds on the pretraining excess risk via tools from convex analysis. This analysis yields novel insights, including an understanding of the benefits of using multiple augmentations per sample (e.g., "multi-crop").
4. We detail several examples where optimal representations are found in closed form, illustrating the role of augmentations, architecture, and regularization in forming representations.
5. We discuss several practical insights for SSL practitioners that emerge from our theory, in particular on how design choices in pretraining may affect downstream performance, and on how to avoid collapse of representations.

Related work. Foundations for theoretically analyzing SSL have emerged in the past few years. Particularly relevant to our work, Balestriero & LeCun (2022); Kiani et al.

[1]Meta AI, New York, NY, USA [2]MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, USA. Correspondence to: Vivien Cabannes <vivc@meta.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

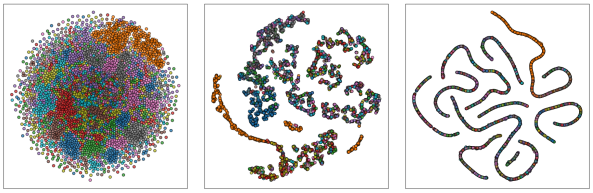

Figure 1. Effect of augmentations and architecture. TSNE of representations learned on MNIST with no augmentations (left) or with rotations and an MLP (middle) or a CNN (right). The representations depend on both the augmentations and the architecture.

(2022) provide theoretically friendly characterizations of many self-supervised learning settings, including closed-form solutions of representations in the kernel setting. For contrastive learning, SSL was first theoretically analyzed by Arora et al. (2019); Tosh et al. (2021a;b); Tian et al. (2021). Notably, HaoChen et al. (2021) recently leveraged tools in spectral graph theory to characterize guarantees on SSL performance under clustering assumptions. These assumptions were deemed impractical by Saunshi et al. (2022), who highlighted the importance of incorporating inductive bias to obtain provable guarantees. This line of work was extended to multi-modal SSL by Lee et al. (2021) where in essence the central symmetric operator T is replaced by a non-symmetric one, and the eigen decomposition is replaced by the singular one. The role of inductive bias has also been scrutinized through analysis of feature learning in training dynamics by Wen & Li (2021) and Tian (2022).

2. Setup

Machine learning streamlines the task of creating algorithms for finding patterns in data. An algorithm is conceptualized as a mapping f from an input $x \in \mathcal{X}$ to an output $y \in \mathcal{Y}$. To construct this mapping $f : \mathcal{X} \to \mathcal{Y}$, one can choose a measure of disagreement $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and minimize the risk

$$\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\rho}[\ell(f(X), Y)], \tag{1}$$

for $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ a distribution on I/O pairs. We denote by $f^* \in \arg\min \mathcal{R}$ an optimal I/O map according to the risk. Mapping raw inputs (e.g., arrays of pixels), to outputs (e.g., classifying an animal in an image), is in general a challenging task. An effective technique consists of first extracting (or engineering) meaningful features $\psi : \mathcal{X} \to \mathbb{R}^k$ from input data before using those features to search f under the form $g \circ \psi$ for $g : \mathbb{R}^k \to \mathcal{Y}$ a simple function.[1]

Though features ψ can be hand-engineered, representation learning aims at improving such design via unsupervised learning procedures. On the one hand, *reconstruction based methods* mask or add noise to inputs via a mapping Mx and aim to reconstruct the original input x from the features $g\circ\psi$ using g, a simple prediction head. Large language models largely rely on this paradigm, usually learning ψ by completing sentences Mx where word tokens are masked (e.g. Devlin et al., 2019). On the other hand, *joint embedding methods* learn ψ by leveraging invariance to small perturbations of the semantic information contained in inputs. This is the paradigm we shall focus on. Recently, joint embedding methods have relied heavily on the concept of data augmentation, such as small rotation, translation, color jittering of images. In particular, contrastive methods learn ψ by enforcing that if two augmentations ξ and ξ' come from the same data point, their representation $\psi(\xi)$ and $\psi(\xi')$ are close; while if they come from different data points, their representation are far away from one another (e.g., Chen et al., 2020). Non-contrastive methods only enforce similarities of augmented datapoints and avoid collapse by enforcing richness of the representation (see, e.g., Bardes et al., 2022). In the following, we focus on a theoretically friendly variant of VICReg (Balestriero & LeCun, 2022) with parameter $\beta > 0$, defined for $\psi : \mathcal{X} \to \mathbb{R}^k$ by

$$\mathcal{L}(\psi) = \beta\, \mathbb{E}_X\, \mathbb{E}_{\xi,\xi'} \left[\|\psi(\xi) - \psi(\xi')\|^2 \,\middle|\, X \right] + \left\| \mathbb{E}_\xi[\psi(\xi)\psi(\xi)^\top] - I \right\|_2^2, \tag{2}$$

where pairs of inputs/augmentations (X, ξ) follow a distribution $\mu \in \Delta_{\mathcal{X}\times\mathcal{X}}$, whose conditional $(\xi \,|\, X)$ arises from the choice of augmentation. The first term in $\mathcal{L}$ enforces invariance of the representation ψ to two augmentations ξ and ξ' of the same input X, while the second term lowers risk of collapse by pushing coordinates $\psi_i : \mathcal{X} \to \mathbb{R}$ of $\psi = (\psi_i)_{i\in[k]}$ to be orthogonal in L^2.

Remark 1 (Contrastive learning with $\mathcal{L}$)**.** *When $\beta = 1$, the population loss $\mathcal{L}$ is equivalent to the spectral contrastive loss studied in HaoChen et al. (2021) as a theoretically friendly proxy for SimCLR (Chen et al., 2020). In other terms, $\mathcal{L}$ analyzes both contrastive and non-contrastive approaches to representation learning.*

Given a representation ψ, one can optimize for f through linear probing by constructing $f = g \circ \psi$ where g is a linear function. f is thereby in the class of functions

$$\mathcal{F} = \left\{ x \mapsto w^\top \psi(x) \,\middle|\, w \in \mathbb{R}^k \right\}. \tag{3}$$

In practice, one might not know the optimal ψ, but can estimate it as $\hat{\psi}$ from empirical data, leading to an estimate $\hat{\mathcal{F}}$ of this class of functions.

Practice	Theory	Quantity
Augmentation	Spectral embedding	T
Architecture	Space of functions	K
Optimization	Regularization	λ
Subtle Interplay		T_λ

Table 1. Analogy between practice and theory that this paper proposes to help disentangle the various phenomena of SSL training.

[1] For convenience, several technicalities, such as measurability, have been deferred to the appendix.

3. Representation learning

In this section, we study the representations induced by pre-training with specific augmentations and inductive biases.

3.1. Closed form solution

Equation (2) admits a closed form solution for ψ upon noting that the invariant part is a quadratic form.

Lemma 2 (Spectral embedding)**.** *There exists a linear positive symmetric operator L in L^2 for which the operator $I - T$ is positive and*

$$\mathbb{E}_X \, \mathbb{E}_{\xi,\xi'} \left[\|\psi(\xi) - \psi(\xi')\|^2 \,\middle|\, X \right] = \sum_{i \in [k]} \psi_i^\top L \psi_i.$$

To be consistent with previous literature, we will rather use $T = I - L/2$, which is also a linear positive symmetric operator, and is defined as, for $\psi_1, \psi_2 \in L^2$

$$\psi_1^\top T \psi_2 = \mathbb{E}_X \, \mathbb{E}_{\xi,\xi'} \left[\psi_1(\xi)^\top \psi_2(\xi) \,\middle|\, X \right]$$

As a consequence, if (λ_i) are the eigenvalues of T and (f_i) are the corresponding eigenvectors, a minimizer of $\mathcal{L}$ is $\psi_i = \sqrt{\mu_i} f_i$ with $\mu_i = 1 - \beta + \beta\lambda_i$.

Lemma 2 is closely tied to the guiding principle in unsupervised learning that a good representation of data should minimize variations over the manifold of the data (Cabannes et al., 2023), and techniques that learn such representations through spectral decomposition of a central operator (see, e.g., Coifman & Lafon, 2006).

3.2. Search within a linear class of functions

In this more technical section, we study solutions of $\mathcal{L}$ for ψ belonging to a linear class of functions. The coordinates of the mapping $\psi : \mathcal{X} \to \mathbb{R}^k$ are typically searched within a space of functions $\Psi \subset \mathbb{R}^{\mathcal{X}}$, leading to $\psi \in \Psi^k$. In our theoretically friendly setup, we assume that Ψ is a linear class of functions endowed with a Hilbertian topology such that the linear evaluations $\psi \mapsto \psi(x)$ are continuous for almost all $x \in \mathcal{X}$. The theory of reproducing kernel Hilbert space (Scholkopf & Smola, 2001) asserts that Ψ can be parameterized by a Hilbert space $\mathcal{H}$ and a mapping $\varphi : \mathcal{X} \to \mathcal{H}$ such that

$$\Psi = \{x \mapsto f_\theta(x) \,|\, f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathcal{H}}, \theta \in \mathcal{H}\}. \quad (4)$$

This generalizes the setting of HaoChen et al. (2021) where $\mathcal{X}$ is assumed to be finite and Ψ is parameterized by $\mathcal{H} = \mathbb{R}^{\mathcal{X}}$ and $\varphi(x) = \delta_x$, as well as the setting of Saunshi et al. (2022) where $\mathcal{H}$ is assumed to be finite dimensional.

To describe architectures such as neural networks with such a linear structure, it is common to linearize those models (e.g. Jacot et al. (2018)) as

$$\psi_\theta(x) = \psi_{\theta_0}(x) + \langle \nabla_{\theta_0} \psi_{\theta_0}(x), \theta - \theta_0 \rangle + o(\|\theta - \theta_0\|),$$

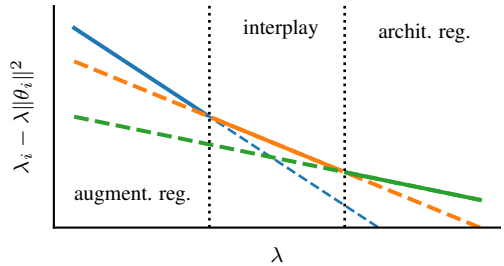

Figure 2. Interplay between T and K as a function of λ. Illustration of Proposition 4 in a setting where $(\lambda_i) = (.9, .75, .5)$ and $(\|\theta_i\|^2) = (.4, .25, .125)$. The plot displays the eigenvalues associated with three different eigenfunctions as a function of λ, β is set to one for convenience. When $\lambda = 0$, the minimizer $\psi_* : \mathcal{X} \to \mathbb{R}$ of (2) is defined through T, here $\varphi_* = f_1$ ($i = 1$, shown in blue), when λ is big $\psi_* = f_3$ (green) mainly depends on K. In the middle, there is an interplay between these two regimes leading to $\psi_* = f_2$ (orange). The three regimes are named the "augmentation", the "architecture" (or VCReg) and the "interplay" regime respectively. This abstract setting can be instantiated with a two-layer ReLU network and cropping as detailed in Figure 8.

where θ are the network parameters, assumed close to their initialization θ_0, and ψ_θ is the neural network. In this case, we may take $\varphi = \nabla_{\theta_0} \psi_{\theta_0}$, which arguably describes some regimes of wide neural networks (Lee et al., 2019).

To minimize $\mathcal{L}$ in practice and improve generalization, a regularization parameter is typically introduced.[2] The following lemma provides a closed form solution of the regularized variant of $\mathcal{L}$.

Lemma 3 (Regularized population loss)**.** *For $\Theta \in \mathbb{R}^k \otimes \mathcal{H}$, and a regularizer $\lambda > 0$, the regularized loss $\mathcal{L}(S\Theta) + \lambda \|\Theta\|_2^2$ can be minimized in closed form with the operator*

$$T_\lambda = (1 - \beta) I + \beta T - \lambda K^{-1}. \quad (5)$$

where $K = SS^\top$ for $S : \mathcal{H} \to L^2(\mu_\Xi); \theta \mapsto f_\theta$ the embedding of $\mathcal{H}$ in L^2. Specifically, if (λ_i) are the (decreasing) eigenvalues of T_λ and (f_i) the corresponding eigenfunctions, a minimizer is given by $\psi_i = \max\{\lambda_i, 0\} f_i$.

3.3. The need for inductive bias

Two different points of view motivate the introduction of the regularizer $\lambda \|\Theta\|$ leading to the operator T_λ.

In the classical viewpoint of statistical learning theory, one would like to retrieve the eigenfunctions of T to minimize $\mathcal{L}$ (Lemma 2). However, when solely accessing finitely many samples of data, eigenfunctions of T should be searched

[2] While we study here the bias of Tikhonov regularization for simplicity, similar studies can be done for early stopped gradient descent or stochastic gradient descent when they are cast as spectral filters, as in Lin et al. (2020), see also literature related to optimization for matrix factorization problems (Chi et al., 2019), which has been applied to SSL by Simon et al. (2023).

within a space of finite capacity (i.e. $\{f \in \Psi \mid \|f\|_\Psi^2 \leq \lambda^{-1}\}$). Though fewer samples are needed for smaller models (e.g. the fewer neurons and layers in a deep network), such small models are unlikely to be expressive enough to represent the ideal solutions. This echoes the classical trade-off between approximation and estimation error. In the case of Laplacians, one can assume that the eigenfunctions of T are smooth thereby belong to a small space of functions that are well-approximated with a finite model of computation. We refer the curious reader to Cabannes et al. (2021a) for results in this vein when $I - T$ is the real Laplacian in L^2.

Another take was suggested by Saunshi et al. (2022), which pointed out that eigenvalues of T can have large multiplicity in realistic situations (in particular in the non-localized augmentations setting of Section 4.2), meaning that the space $\mathcal{F}$ is not uniquely defined from the loss $\mathcal{L}$. As a consequence, defining the optimal solution solely from T is somewhat ill-posed, whereas, when K is properly chosen, T_λ could define a "more principled" representation ψ. Paradoxically, with this viewpoint, *bias could reduce the approximation error*. Figure 3 illustrates such an idea. It leverages the following interpretation of the inductive bias in the friendly setting where T and K commute.

Proposition 4. *If T and K commute, and if (λ_i) are the eigenvalues of T and (f_i) its eigenfunctions, then there exists (θ_i) such that $f_i = f_{\theta_i}$ (4). Moreover, the optimal representation to minimize the regularized loss are the f_i that maximize $\beta\lambda_i - \lambda \|\theta_i\|^2$. In other terms, the regularization biases towards representations that have a small complexity with respect to the model of computation.*

Lemma 3 shows an interesting behavior of the VCReg loss ($\beta = 0$, i.e. VICReg without invariance term). In this setting, the optimal ψ retrieve the largest eigenfunctions of K, recovering kernel PCA. Learning downstream tasks with linear probing of the resulting ψ is equivalent to linear regression with an eigenvalue cut-off, which is a powerful spectral filtering technique (see, e.g. Bun et al., 2017).

4. Illustrative examples

The analysis of this paper relies on two central operators: T, that is "intrinsically" defined from the data distribution and augmentations, and K, which relates to the model of computation (e.g. the network architecture). Once those operators are chosen, Section 5 provides a sharp analysis of convergence and generalization with SSL in the kernel regime. In essence, Assumption 3 requires that the target function (downstream) align well with the learned representation (upstream) when given infinite data. Here, the effect of T and the inductive bias introduced by λK^{-1} on the learned representation can appear abstract. To provide intuition and outline important properties of these operators, this section lays out several concrete examples to help practitioners better understand the role of augmentations and their interplay with the inductive bias of the architecture.

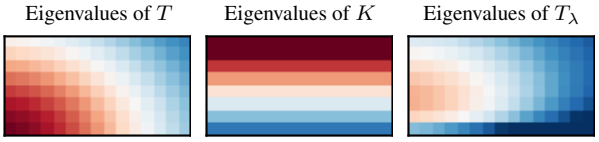

Figure 3. Trade-off on eigenvalues between T and K. Illustration of a harmonic setting where T and K are diagonalized in the same basis. This basis is parametrized by an "invariance score" ($x = m$ in (54)) and a "complexity score" ($y = |S|$ in (54)). The eigenvalues $\lambda_{x,y}(A)$ for $A \in \{T, K, T_\lambda\}$ are represented with colors and displayed in a grid associated with $x \in [15]$ and $y \in [8]$. The sole use of the operator T biases towards invariance (lower x) with high complexity (lower y), while the sole use of K biases toward low complexity. The interplay between the two results in T_λ whose biggest eigenfunctions have high invariance and low complexity, and corresponds to an ideal representation ψ.

Two different perspective have emerged to understand learned representation in SSL. One intuition comes from the spectral clustering literature, and is the object of subsection 4.1. The other intuitive way to understand SSL is based on harmonic analysis, and is the object of subsection 4.2. All in all, this section generalizes previous works by dropping out strong clustering assumptions in the data, showing that what really matter are the eigenfuntions of T, which eventually capture clustering structures when such clustering assumptions are invoked. It further uses harmonic analysis tools to better describe these eigenfunctions as suggested by Saunshi et al. (2022) and detailed in Table 2.

4.1. Low-variation with localized augmentations

When augmentations $(\xi \mid X)$ are localized around the input X, optimizing the loss $\mathcal{L}$ (2) biases towards small gradients of ψ along the directions of augmentations. Formally for $\psi : \mathcal{X} \to \mathbb{R}$, using first order Taylor expansion,

$$\mathbb{E}\left[\|\psi(\xi) - \psi(\xi')\|^2 \,\middle|\, X\right] \simeq \mathbb{E}\left[\langle \nabla\psi(X), \xi - \xi'\rangle^2 \,\middle|\, X\right].$$

Under isotropic augmentations, the objective simplifies as

$$\mathbb{E}\left[\langle \nabla\psi(X), \xi - \xi'\rangle^2 \,\middle|\, X\right] \propto \|\nabla\psi(X)\|^2,$$

which enforces ψ to have small variations on densely populated regions of the input space – reminiscent of popular approaches to tackle representation and semi-supervised learning in the last two decades (van Engelen & Hoos, 2020). More generally, augmentations govern the important directions of invariance for ψ, recovering a finite-differences approach to the Tangent Prop algorithm (Simard et al., 1991).

Low variation methods are particularly useful when data display a clustering structure (c.f. Figure 10 for an illustration with neural networks). If augmentations preserve

the clustering structure, $\mathcal{L}$ is minimized by piecewise constant functions on each cluster, leading to useful features for downstream tasks that involve classifying different clusters (HaoChen et al., 2021; Schiebinger et al., 2015). The inductive biases further deforms those top eigenfunctions to be regular in a sense defined by Ψ (4), e.g., analytic if we use a radial basis function kernel (Sun & Zhou, 2008).

4.2. The role of augmentations

When augmentations are not localized, which is often the case in practice, harmonic analysis provides useful tools to study in-depth the role of augmentations, in particular when data are uniform on the sphere or the hypercube. Our findings on the hypercube are summarized in Table 2. In such a setting, we show that common augmentations enforce smoothness, locality, or invariance to certain symmetries. For example, crops push ψ to focus on details that can appear within the crop size, filtering out long-range interactions between parts of the input that are likely to be spurious features. The following example formalizes this.

Example 1 (Cropping)**.** *Consider the hypercube setting where $\mathcal{X} = \{-1, 1\}^n$ and X is uniformly distributed. A basis of $L^2(\mathcal{X}, \mathbb{R})$ is given by the parity functions $\chi_S : x \mapsto \prod_{i \in S} x_i$ for all the subsets $S \subseteq [n]$. Pre-training via cropping with window sizes $v \times w$ set $T\chi_S = 0$ for all S whose support forms a window larger than the size $v \times w$. For all the other S, $T\chi_S = \lambda_S \chi_S$, where λ_S decreases with the diameter of S. In other terms, pre-training with 2-D cropping eliminates the influence of functions which act globally outside of the cropping window. This, in effect, imparts a locality to the induced representation ψ which is often desirable for generalization.*

This example suggests that the ideal crop size should match the desired scale of details for ψ; e.g., on a dataset with fine-grained details such as iNaturalist, one should reduce the crop window size in comparison to a dataset such as ImageNet. Appendix D discusses further examples of augmentations, such as random noise or translations, and shows how they bias towards smooth or invariant eigenfunctions.

4.3. Interplay with the architecture

While the design of augmentations and architecture can be done separately, changes to the architecture and optimization scheme play an important role in the resulting optimal ψ. Generally, increasing the amount of inductive bias by increasing λ shifts ψ towards smoother functions, in the sense captured by the $\mathcal{H}$ norm, which we illustrate in Figure 4. In practice, the right amount (captured here by the parameter λ) of inductive bias to enforce is often set by a mix of intuition, common knowledge and empirical evidence. For example, Caron et al. (2021) links the inductive bias of early stopping to beneficial outcomes noting that "training longer [...] has been leading to worse performance".

Example 2 (Dot-product kernel)**.** *On the Boolean hypercube setting of Example 1, many linear models (4) take the form $\varphi(x)^\top \varphi(y) = h(\langle x, y \rangle)$ (e.g., the classical NTK linearization of fully connected layer) leading to an integral operator K that is diagonalizable by parity functions. More precisely, there exists $(\nu_i) \in \mathbb{R}^d$ such that $K\chi_S = \nu_{|S|}\chi_S$, where $|S|$ is the cardinality of S and $\nu_{|S|}$ decreases with $|S|$. In the setting of crops, T pushes towards representation on parity functions with small diameter ($\psi = (\chi_S)_S$ for S with small diameters), while the inductive bias acts on the cardinality of the sets S, pushing towards the χ_S that maximize $\nu_{|S|}$. Formal derivations are provided in Appendix D.*

Appendix D provides additional examples. For instance, in the case of translations, there is a similar interplay between a low-degree bias in K versus an invariance bias in T. We also consider convolutional architectures, which can impose locality through constraints on $\mathrm{diam}(S)$, on top of a low-degree bias. Figure 3 shows such trade-offs in eigenvalues, Figure 4 visualizes how this interplay may affect leading eigenfunctions in a spherical setup, and Figure 5 illustrates the resulting effect on different downstream tasks.

5. Convergence analysis

The following section analyzes guarantees on both the pretraining and downstream tasks.[3] For simplicity, we consider the mean-square loss $\ell(y, y') = \|y - y'\|^2$ with $\mathcal{Y} = \mathbb{R}^{d_y}$. The studies of many losses can be reduced to the least-squares case thanks to calibration inequalities (Bartlett et al., 2006) or self-concordance properties (Ostrovskii & Bach, 2018). To precisely study convergence rates, we consider the kernel regime of Section 3.2, where $\mathcal{F}$ is specified through Θ_* of Lemma 3 as

$$\mathcal{F} = \left\{ x \to w^\top \Theta_* \varphi(x) \,\middle|\, w \in \mathbb{R}^k \right\}, \tag{3}$$

and $\hat{\mathcal{F}}$ is defined similarly with $\hat{\Theta}$ as an estimate of Θ_*. In the following, (f_i) denote the eigenfunctions of T_λ ordered by decreasing eigenvalues, and λ is considered to be fixed throughout this section.

5.1. Dealing with distribution shift

Self-supervised learning algorithms often incorporate strong augmentations leading to potentially different marginal distributions over inputs and augmentations. This discrepancy is often overlooked, many theoretical works implicitly assuming $\rho_{\mathcal{X}} = \mu_\Xi$. In practice, the marginal distribution $\rho_{\mathcal{X}}$ of inputs in the downstream task can be meaningfully different from the marginal distribution of augmentations μ_Ξ

[3] The pretraining and downstream tasks refer to minimization of $\mathcal{L}$ and $\mathcal{R}$ respectively.

Augmentation example		Effect of the operator T
Input (no augmentation)		-
Random noise		Attenuate higher order Fourier modes
Cropping		Keep Fourier modes within cropping windows
Translations		Bias towards Fourier modes with cyclic invariance
Flipping		Equate eigenvectors of subsets related by flips

Legend: -1 bit, +1 bit, flipped bit, random bit

Table 2. Effect of common augmentations on the optimal representation ψ through the operator T. Without augmentations, ψ could match any Fourier basis function. Augmentations filter out some of those by attenuating their eigenvalues in T, and the architecture will push ψ to pick some specific frequencies among the remaining ones through the operator K. The table stylizes the effect of usual augmentations on parity functions over bit streams. We refer the reader to Appendix D for further details and derivations.

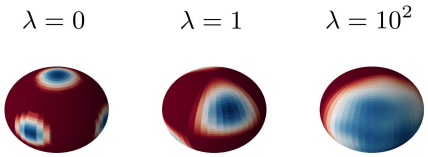

Figure 4. Interplay on the sphere. Level lines of the 7-th eigenfunction of T_λ for three different λ. Augmentations consist of translations of the x, y, z coordinates together with Gaussian perturbations. K is the integral operator associated with the radial basis function. Without regularization (left), the eigenfunction is highly localized at clusters corresponding to the action of the augmentations. Increasing the regularization biases towards smoother harmonic eigenfunctions of K (middle and right).

on which we have imposed orthogonality of the representation ψ in the pretraining task. However, the optimal representation ψ is likely to be invariant to augmentations, meaning that ideally, $\psi(X)$ should have the same distribution when $X \sim \mu_\Xi$ or $X \sim \rho_\mathcal{X}$, which we write formally as $\psi_\# \mu_\Xi = \psi_\# \rho_\mathcal{X}$. Moreover, augmentations are likely to spread the input data distribution, leading to the dominance $\rho_\mathcal{X} \ll \mu_\Xi$. This motivates the following assumptions and definitions.

Assumption 1 (Low expansion)**.** *There exists $c_r > 0$ such that for any function f in the original space of functions Ψ defined in (4),*

$$\|f\|^2_{L^2(\rho_\mathcal{X})} \leq c_r \|f\|^2_{L^2(\mu_\mathcal{X})},$$

Assumption 2. *For any i smaller than the number of positive eigenvalues of T_λ, the projection of the target f^* on f_i in $L^2(\mu_\Xi)$ coincides with the projection on f_i in $L^2(\rho_\mathcal{X})$.*

To make those two concepts more concrete, we provide three examples below.

Example 3. *If $\rho_\mathcal{X}$ has a density against μ_Ξ which is bounded from below by $\delta \in (0, 1]$ on its support, i.e. $\mu_\Xi = \delta\rho_\mathcal{X} + (1-\delta)\mu_\perp$ with $\mu_\perp \in \Delta_\mathcal{X}$, then Assumption 1 is met with $c_r = 1/\delta$.*

Example 4. *Let $\Sigma_\tau = \mathbb{E}_{X\sim\tau}[\varphi(X)\varphi(X)^\top]$ be the covariance matrix of φ under the distribution τ. When there exists c such that $\Sigma_{\rho_\mathcal{X}} \preceq c\Sigma_{\mu_\Xi}$ (i.e $c\Sigma_{\mu_\Xi} - \Sigma_{\rho_\mathcal{X}}$ is positive semi-definite), then Assumption 1 holds with $c_r = c$.*

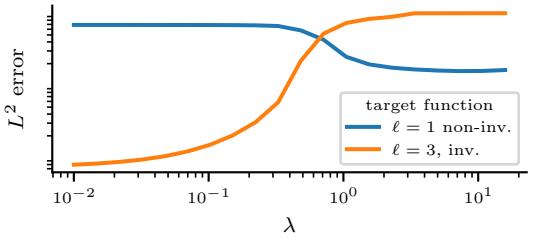

Figure 5. Trade-off on downstream errors. Effect of pretraining regularization λ on the empirical downstream error for two tasks on the sphere $\mathbb{S}^7$. The targets f^*_ℓ are polynomials of degree $\ell \in \{1, 3\}$, with only f^*_3 invariant to translations. K is built from a dot-product kernel that acts as a regularizer on degrees, while T is built from local translations. Designing ψ from T alone ($\lambda = 0$) is helpful to learn globally invariant polynomials in the downstream, while increasing the regularization λ helps to learn polynomials of small degree. Experiment details in Appendix E.2, and Figure 11 showcases a similar behavior for neural networks.

Example 5. *If $\psi_\sharp \mu_\Xi = \psi_\sharp \rho_\mathcal{X}$ holds for the optimal representation $\psi = (f_i)$, with (f_i) the positive eigenfunctions of T_λ, and there exists a measurable function $g : \mathbb{R}^k \to \mathcal{Y}$ such that $f^* = g \circ \psi$, then Assumption 2 is verified.*

In essence, Assumptions 1 and 2 allow for the incorporation of augmented data that does not resemble the original data as long as the model of computation (Assumption 1) and training via the VICReg loss (Assumption 2) do not bias too much towards this aberrant augmented data. Example 3 states that when the augmented data mostly looks like the original samples then one does not have to worry about bias introduced by the model of computation. Example 4 gives a more relaxed guarantee based on second order moments. Finally, Example 5 states that one need not worry about the idiosyncrasies of the augmented data if the learned representations confound augmented data with their original samples.

5.2. Generalization on downstream tasks

This section provides comprehensive generalization guarantees on supervised downstream tasks. The following assumption ensures that the target function $f^*(x) =$

$\mathbb{E}_\rho[Y|X=x]$ of the downstream task is well represented by the pretraining problem.

Assumption 3 (Source condition)**.** *f^* belongs to the positive eigenspace of T_λ, i.e. $f^* \in \mathrm{Span}\{f_i \,|\, \lambda_i > 0\}$.*

Example 6 (Cluster assumption)**.** *If the support of the density μ_Ξ has k connected components, f^* is constant on those clusters, and $\lambda = 0$, then Assumption 3 holds.*

We now give a simplified version of our downstream guarantee. See Theorem 4 in Appendix B for the full statement.

Theorem 1 (Downstream error)**.** *Let $(X_i, Y_i) \sim \rho^{\otimes n}$ be n samples drawn from the distribution for the downstream task and ℓ be the square loss. Define $k_\lambda < +\infty$ as the number of strictly positive eigenvalues of T_λ. Under Assumptions 1, 2, and 3, after a transitory regime, the average excess risk of the optimally-regularized empirical risk minimizer f_n is*

$$\mathbb{E}[\mathcal{R}(f_n) - \mathcal{R}(f^*)] \leq \frac{2k_e\varepsilon^2}{n} + \frac{\log(n)^{1.1}}{n}\|f^*\|_{L^2(\rho)} + c_{f,T_\lambda}^2\left(\mathcal{L}_k(S\hat{\Theta}) - \mathcal{L}_k(S\Theta_*)\right) + c_{f,k}. \quad (6)$$

where ε^2 is the noise level of Y (the supremum of conditional variances), $k_e \leq k$ is the effective dimension of the representation $\psi = \Theta\varphi$ on the downstream task, $c_{f,k} \leq (k_\lambda - k)_+ \|f^*\|^2_{L^2(\rho_\mathcal{X})}$ is a constant relating to the concentration of the energy of f^* the target function on the downstream task with respect to the eigenspaces of T_λ, $c_{f,T_\lambda} \leq \left\|T_\lambda^{-1} f^*\right\|$ is a similar constant taking into account the decay of eigenvalues of T_λ, and the index k in $\mathcal{L}_k$ indicates that we search over $\psi : \mathcal{X} \to \mathbb{R}^k$.

The results of Theorem 1 can be seen as a variance-bias decomposition. A variance term, due to misspecified linear regression, displays rates in $k\log(n)/n$. The $\log(n)$ factor is actually an artefact of our derivations, and could be removed with Theorem 1 of Mourtada et al. (2022). A bias term relates to the approximation error. It captures both the hardness to learn f^* with T_λ through the constants c_{f,T_λ} and $c_{f,k}$, and the error made on the pre-training task through $\mathcal{L} - \mathcal{L}^*$. Note that the proof of Theorem 1 mindfully avoids bounding c_{f,T_λ} by $\left\|T_\lambda^{-1}\right\|_{\mathrm{op}}\|f^*\|$ which would introduce the inverse of the spectral gap of T_λ in the bound, and would not characterize well situation where the target function f^* is actually easy to learn with T_λ. We also remark that for classification tasks, recent work shows that under mild assumptions on $\rho_\mathcal{X}$, and low noise conditions, it should be possible to convert the rates of Theorem 1 into exponentially fast rates for the zero-one loss (Cabannes et al., 2021b). This is particularly the case under the cluster setting studied by HaoChen et al. (2021).[4] The theoretical convergence rates of Theorem 1 are validated experimentally in Figure 6.

[4]See also Rigollet (2007) for fast rates in this setting.

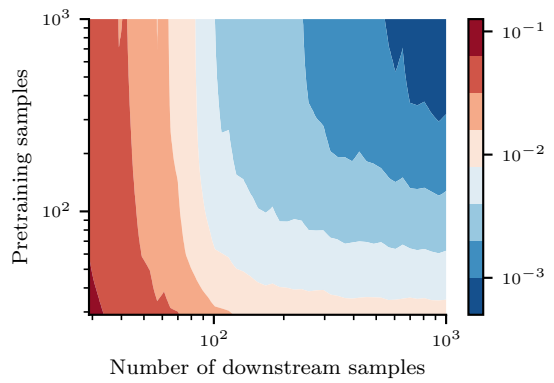

Figure 6. Empirical downstream performance on a simple task (detailed in Appendix D) depends on the number of both downstream samples (x-axis) and pretraining (y-axis) in log-scale. Along the left hand side of the plot, convergence rates of $n_{pre}^{-1/2}$ are observed with respect to the number of pretraining samples (Theorem 3) while along the top, convergence rates of n_{down}^{-1} are observed with respect to the number of downstream samples (Theorem 1). At the bottom, a saturation phenomenon is observed where added downstream samples do not result in noticeable benefits as the excess risk stalls at $\mathcal{R}(\Pi_{\hat{\mathcal{F}}} f^*) - \mathcal{R}(f^*) > 0$.

5.3. Pretraining guarantees

Theorem 1 above states that representations with small pretraining loss can solve downstream tasks that satisfy Assumption 3 but do not address how difficult it is to find such representations. This section aims to bridge that gap. The following theorem details convergence rates of the empirical risk minimizer using Rademacher complexity arguments.

Theorem 2 (Empirical risk minimizer)**.** *Let $\Theta_n \in \mathbb{R}^k \otimes \mathcal{H}$ be the minimizer of the unbiased regularized empirical version of $\mathcal{L}$ based on a dataset $\mathcal{D}_n$. Assume that $\mathcal{D}_n$ is built from n input samples $(X_i) \sim \mu_\mathcal{X}^{\otimes n}$ and m augmentation per samples $(\xi_{ij}) \sim \mu|_{X_i}^{\otimes m}$, then the average excess risk is bounded by*

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{L}(S\Theta_n)] - \mathcal{L}(S\Theta) \leq \frac{12\kappa^2 k}{\lambda\sqrt{n}}\left(1 + \frac{\kappa^2 k}{\lambda}\right), \quad (7)$$

where κ is a bound on $\|\varphi(X)\|$.

Note that the proof of Theorem 2 proceeds with a loose bound on the variance of the empirical risk, which is mainly due to the difficulty in dealing with non-exchangeability of the samples (ξ_{ij}). In essence, the ease of minimizing $\mathcal{L}$ depends on both the variance of $\mathcal{L}$ when estimated with empirical data (or the variance of stochastic gradients when performing SGD), and the size of the space where we aim to find representations $\psi : \mathcal{X} \to \mathbb{R}^k$. With stronger assumptions on the distribution of $\varphi(\xi)$ (e.g., data are clustered, and the law of $(\xi \,|\, X)$ is invariant per cluster), one could show much better behavior of the excess risk with respect to the number of augmentations (e.g., replacing n by the minimum

number of points in one cluster multiplied by the number of views). The following theorem states convergence rates with a stochastic gradient descent algorithm capturing such a potential situation. Proofs and technicalities, based on convex optimization literature, are detailed in Appendix C.

Theorem 3 (Sharper bounds)**.** *There exists an implementable algorithm that guarantees an average excess risk*

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{L}(S\Theta_n)] - \mathcal{L}(S\Theta) \leq 3\kappa^2 c_\lambda c'_\lambda \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_\xi^2}{nm}} + \frac{4\kappa^6 c_\lambda^2}{n} \quad (8)$$

where $c_\lambda = 1 + \kappa^2 k_\lambda/\lambda$, $c'_\lambda = 1 + k_\lambda^2/\lambda^2$, k_λ is the number of positive eigenvalues of T_λ, κ is a bound on $\|\varphi\|$, σ_X relates to the variance of $\mathbb{E}[\psi(\xi) \mid X]$, and σ_ξ relates to the average variance of $(\xi \mid X)$. Moreover, when $K = SS^\top$ or the covariance of the $\varphi(\xi)$ has a finite number of positive eigenvalues (e.g. $\mathcal{X}$ finite or $\mathcal{H}$ finite dimensional), with c_K a constant that relates to the condition number of K, this bound can be tightened to

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{L}(S\Theta_n)] - \mathcal{L}(S\Theta) \leq \frac{4c_K^2 c_\lambda^2}{n}. \quad (9)$$

In the setting studied in HaoChen et al. (2021), we stress that Theorem 3 guarantees convergence rates of $O(n^{-1})$ rather than $O(n^{-1/2})$ on the upstream loss. In effect, we improve the rates of HaoChen et al. (2021, Theorem 4.3) from $n^{-1/2}$ to n^{-1} on both pretraining and downstream tasks.

6. Practical insights

In this section, we relate our theory to commonly observed phenomena when training SSL algorithms and offer best practices informed by our findings.

6.1. The downstream problem

Two regimes should be distinguished for the downstream problem. When few downstream samples are available, *few-shot learning* requires a small effective dimension k_e (6) to lower the estimation error and avoid fitting noise. Limiting k_e (or equivalently the capacity of $\hat{\mathcal{F}}$) can be done either by decreasing the representation dimension k or applying regularization on downstream tasks. This theoretical trade-off between effective dimension and number of downstream examples is illustrated empirically by He & Ozay (2020, Figure 6). On the contrary, when accessing a substantial amount of data for training downstream tasks, one could confidently augment the representation dimension k to decrease the approximation error. This was notably observed on large-scale datasets by Garrido et al. (2022, Figure 1): as k increases, the effective dimension k_e converges to a limit, and the downstream performance keeps increasing until this limit is reached. Remarkably, our theory explains this phenomenon: since k_λ is finite, as k increases, the effective dimension k_e will be bounded by the limiting case where $k = k_\lambda$.[5]

6.2. The pretraining problem

Usefulness of multiple augmentations per sample. Theorem 3 shows how multiple augmentation such as multi-crops can result in faster convergence to an optimal representation ψ. There, the variance of the empirical risk depends on both σ_X due to variation over inputs, and σ_ξ due to variations over resulting views after augmentation. With multiple augmentations per sample, one can reduce the latter variance and improve performance, which was observed with the introduction of multicrops in Caron et al. (2020). However, when the total amount $m \times n$ of pre-processed data is held fixed, it is generally better to process many inputs with two views $m = 2$, rather than a few inputs with many augmentations. This finding matches the empirical observations of Bardes et al. (2022) that if available, fresh samples are always better than more views.

Capacity trouble in pretraining. Theorems 2 and 3 show that, without regularization restricting the capacity of the model of computation, one cannot expect to meaningfully solve the pretraining task. This is captured by the quantity c_λ that goes to infinity as λ goes to zero. Such issues related to the lack of regularization commonly arise in practice. Given $n \times m$ upstream samples (ξ_{ij}), the empirical minimization of VICReg can be implemented by approximating μ with $\sum_{ij} \delta_{(i,\xi_{i,j})}/nm$. In this setting, T is the adjacency matrix of a graph with as many connected components as there are inputs n, as detailed in Appendix E. All the connected components define a maximal eigenvector of the empirical approximation to T, leading to a "collapsed" representation $\psi = \sum_j \delta_{\xi_{ij}}/m$. Regularizing forces the optimizer to search for representation inside the space Ψ which mixes those small clusters letting meaning eigenfunctions emerging (see Figure 7 for an illustration).

6.3. Guidelines for practitioners

Our theoretical study provides several insights that may be useful for SSL practitioners. We highlight a few below.

Avoiding collapse. The common collapse phenomenon, where pretraining ends up fitting noise instead of learning useful features, may be addressed in several ways. Our theory suggests to:

- *Reduce the model capacity*, through *regularization* (e.g., early stopping), or *simpler architectures* (e.g., a shallow CNN instead of an MLP). As a consequence,

[5] Note that without regularization, $(1-\beta)I + \beta T$ is not trace-class so k_e will not converge as k increases.

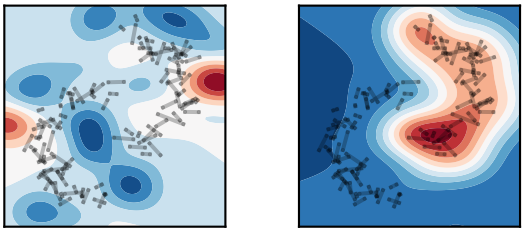

Figure 7. Capacity trouble. Level lines of the top eigenfunction of empirical estimate of T_λ for negligible regularization (left) and small regularization λ (right). Experiments are done with a Gaussian kernel with scale about one tenth of the problem diameter, augmentations are represented as black dots, connected by a line when they provide from the same input X. When λ is negligibly small, capacity troubles arise, infringing the recovery of the cluster structure on the right.

Ψ will have a lower effective dimension, K will encourage "simpler" representations that can be learned with less data, even without any data augmentation.

- *Use stronger augmentations.* T will become more compact, reducing k_λ the dimension of the "positive eigenspace" of T_λ. The ideal ψ will exhibit more structure, thereby its search could be reduced to smaller spaces, making it harder to collapse.

Incorporating priors. Representations are typically used for solving downstream tasks, thus it is crucial to incorporate the right priors during pretraining. Our theory showcases the important role of several factors. *(i) Augmentations* determine the nature of the invariance that is enforced (e.g., low variations, short-range dependencies, translation invariance); affects top eigenfunctions of T. *(ii) Architecture* promotes "simple" representations (e.g., smoothness, locality); affects top eigenfunctions of K. *(iii) Regularization* balances the interplay between augmentations and architecture; affects top eigenfunctions of T_λ. *(iv) Pretraining data* impacts both T and K and their eigenfunctions, e.g., through clustering structure, or natural image statistics.

7. Conclusion

This paper presents a theoretical framework for studying self-supervised learning in the kernel regime. It examines three key operators and their impact on convergence and generalization: T linked with augmentations, K linked with architecture choices, and T_λ resulting from their interplay and tuned by the parameter λ. Our analysis offers useful guarantees and practical guidelines for practitioners to improve the stability and performance of SSL algorithms. Looking beyond the kernel regime, future work can We left for future work the extension of our analysis beyond the kernel setting, in particular to understand non-linear training dynamics in finite width neural network and feature learning capabilities within layers. Moreover, future studies could encompass more techniques that enhance performance in SSL, these include projecting representations before enforcing losses, batching the data, or applying different loss functions.

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Bach, F. *Learning Theory from First Principles*. To appear at MIT press, 2023.

Balestriero, R. and LeCun, Y. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In *NeurIPS*, 2022.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

Bartlett, P. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002.

Bartlett, P., Jordan, M., and Mcauliffe, J. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2006.

Bietti, A. Approximation and learning with deep convolutional models: a kernel perspective. In *International Conference on Learning Representations*, 2022.

Bietti, A. and Mairal, J. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 2019.

Bietti, A., Venturi, L., and Bruna, J. On the sample complexity of learning under invariance and geometric stability. In *Advances in Neural Information Processing Systems*, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in neural information processing systems*, 2020.

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 2015.

Bun, J., Bouchaud, J.-P., and Potters, M. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 2017.

Cabannes, V., Pillaud-Vivien, L., Bach, F., and Rudi, A. Overcoming the curse of dimensionality with Laplacian regularization in semi-supervised learning. In *NeurIPS*, 2021a.

Cabannes, V., Rudi, A., and Bach, F. Fast rates in structured prediction. In *Conference on Learning Theory*, 2021b.

Cabannes, V., Bietti, A., and Balestriero, R. On minimal variations for unsupervised representation learning. *ICASSP*, 2023.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 2007.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Chi, Y., Lu, Y., and Chen, Y. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 2019.

Coifman, R. and Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis*, 2006.

Davis, C. and Kahan, W. The rotation of eigenvectors by a perturbation. *SIAM Journal on Numerical Analysis*, 1970.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Efthimiou, C. and Frye, C. *Spherical harmonics in p dimensions*. World Scientific, 2014.

Favero, A., Cagnetta, F., and Wyart, M. Locality defeats the curse of dimensionality in convolutional teacher-student scenarios. In *Advances in Neural Information Processing Systems*, 2021.

Garrido, Q., Balestriero, R., Najman, L., and Lecun, Y. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. *arXiv preprint arXiv:2210.02885*, 2022.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in neural information processing systems*, 2020.

HaoChen, J., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *NeurIPS*, 2021.

HaoChen, J. Z. and Ma, T. A theoretical study of inductive biases in contrastive learning. *arXiv preprint arXiv:2211.14699*, 2022.

He, B. and Ozay, M. Exploring the gap between collapsed & whitened features in self-supervised learning. In *ICML*, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

Kato, T. *Perturbation Theory for Linear Operators*. Springer, 1995.

Kiani, B. T., Balestriero, R., Chen, Y., Lloyd, S., and LeCun, Y. Joint embedding self-supervised learning in the kernel regime. *arXiv preprint arXiv:2209.14884*, 2022.

Kolmogorov, A. and Tikhomirov, V. ε-entropy and ε-capacity of sets in functional spaces. *Uspekhi Matematicheskikh Nauk*, 1959.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In *NeurIPS*, 2019.

Lee, J., Lei, Q., Saunshi, N., and Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. In *Advances in Neural Information Processing Systems*, 2021.

Lin, J., Rudi, A., Rosasco, L., and Cevher, V. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 2020.

Maurer, A. A vector-contraction inequality for Rademacher complexities. In *Algorithmic Learning Theory*, 2016.

Mei, S., Misiakiewicz, T., and Montanari, A. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, 2021.

Meir, R. and Zhang, T. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 2003.

Micchelli, C., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 2006.

Misiakiewicz, T. and Mei, S. Learning with convolution and pooling operations in kernel methods. In *Advances in Neural Information Processing Systems*, 2022.

Mourtada, J. and Rosasco, L. An elementary analysis of ridge regression with random design. *Comptes Rendus. Mathématique*, 2022.

Mourtada, J., skevičius, T. V., and Zhivotovskiy, N. Distribution-free robust linear regression. *Mathematical Statistics and Learning*, 2022.

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

Ostrovskii, D. and Bach, F. Finite-sample analysis of m-estimators using self-concordance. *Electronic Journal of Statistics*, 2018.

Pillaud-Vivien, L. and Bach, F. Kernelized diffusion maps. *ArXiv*, 2023.

Pinelis, I. and Sakhanenko, A. Remarks on inequalities for large deviation probabilities. *Theory of Probability and Its Applications*, 1986.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Rigollet, P. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 2007.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. Understanding contrastive learning requires incorporating inductive biases. In *ICML*, 2022.

Schiebinger, G., Wainwright, M., and Yu, B. The geometry of kernelized spectral clustering. *The annals of Statistics*, 2015.

Scholkopf, B. and Smola, A. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Simard, P., Victorri, B., LeCun, Y., and Denker, J. Tangent prop - a formalism for specifying selected invariances in an adaptive network. In *NeuRIPS*, 1991.

Simon, J., Knutins, M., Ziyin, L., Geisz, D., Fetterman, A., and Albrecht, J. On the stepwise nature of self-supervised learning. *ArXiv*, 2023.

Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 2007.

Smola, A., Ovári, Z., and Williamson, R. C. Regularization with dot-product kernels. In *Advances in neural information processing systems*, 2000.

Sun, H.-W. and Zhou, D.-X. Reproducing kernel hilbert spaces associated with analytic translation-invariant mercer kernels. *Journal of Fourier Analysis and Applications*, 2008.

Tian, Y. Understanding the role of nonlinearity in training dynamics of contrastive learning. *arXiv preprint arXiv:2206.01342*, 2022.

Tian, Y., Yu, L., Chen, X., and Ganguli, S. Understanding self-supervised learning with dual deep networks, 2021.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, 2021a.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive estimation reveals topic posterior information to linear models. *JMLR*, 2021b.

Tropp, J. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 2015.

van Engelen, J. and Hoos, H. A survey of semi-supervised learning. *Machine Learning*, 2020.

Vitushkin, A. On Hilbert's thirteenth problem. *Proceedings of the USSR Academy of Sciences*, 1954.

Wen, Z. and Li, Y. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, 2021.

Yang, G. and Salman, H. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

A. Mathematical details and simple proofs

A.1. Technical details

Several technicalities were left implicit in the main text, we discuss it now. In particular, we assumed that there exists a minimizer f^* of the risk $\mathcal{R}$ which is true when $\mathcal{Y}$ is finite, or when ℓ is the least-square loss and $(Y \mid X)$ has a second order moment almost everywhere. Moreover, in the proof for least-squares, we will assume for simplicity that $\mathcal{Y} = \mathbb{R}$. The same derivations still holds true when $\mathcal{Y} = \mathbb{R}^k$ although it requires slight precautions such as working in $\mathcal{Y} \otimes \mathcal{H}$ rather than in $\mathcal{H}$ (see Cabannes et al., 2021b, for example).

Integers. The representation dimension k is an integer, and $[k]$ denotes the set $\{1, 2, \cdots, k\}$. For simplicity, we abuse notations and denote by $\mathbb{N}$ the set of strictly positive integers.

Geometry. The product $A \times B$ denotes the set of elements (a, b) with $a \in A$ and $b \in B$. The notation $a^\top$ denotes the adjoint of a which depends on the Hilbertian metric space one consider a to be part of (*e.g.* the adjoint in $L^2(\mu_\Xi)$ is not the same as the adjoint in $L^2(\rho_\mathcal{X})$). The notation $a^\top b$ denotes the scalar product $\langle a, b\rangle$ with the Hilbertian metric space a and b are understood to be part of. The Hilbertian norm on matrices or operators is denoted by $\|\cdot\|_F$ (Frobenius), $\|\cdot\|_2$ or $\|\cdot\|_{HS}$ (Hibert-Schmidt). The operator norm is denoted by $\|\cdot\|_{\text{op}}$. Moreover, the identity is always denoted by I.

Distributions. In order to define probabilities, $\mathcal{X}$ and $\mathcal{Y}$ are assumed to be Polish spacesa endowed with the Borel topologies. We used the simplex notation Δ_A to design the set of probability measure on A, and the tensor notations $\rho^{\otimes n}$ to denotes the measure of n independent random variables all distributed according to ρ. The notation $\varphi_\# \rho$ denote the measure of $\varphi(X)$ when X is distributed according to the measure ρ. The notation $\rho \ll \mu$ means that for any measurable set X the fact that $\mu(X) = 0$ implies that $\rho(X) = 0$. The notation δ_x denotes the Dirac distribution, which satisfies $\langle f, \delta_x\rangle = f(x)$ using the duality bracket between functions and distributions. For any distribution p, the space $L^2(p)$ is made of measurable functions that are square-integrable.

Functions. All functions such as ℓ, f, ψ, φ, and so on, are restricted to be measurable. The notation $\circ$ denote the composition of functions $f \circ g(\cdot) = f(g(\cdot))$. A function $f : \mathcal{X} \to \mathcal{Y}$ is understood as an element of $\mathcal{Y}^\mathcal{X}$, and we use some isomorphism such as $(\mathbb{R}^k)^\mathcal{X} = (\mathbb{R}^\mathcal{X})^k$. We use the notation $\mathbb{R}^k \otimes \mathcal{H}$ to denote linear bounded operator from $\mathcal{H}$ to $\mathbb{R}^k$. This tensor product notation generalizes matrix notations with $\mathbb{R}^k \otimes \mathbb{R}^{d_h} = \mathbb{R}^{k\times d_h}$. In particular,

$$\Psi^k = \left\{x \to \Theta\varphi(x) \,\middle|\, \Theta \in \mathbb{R}^k \otimes \mathcal{H}\right\}.$$

For $\Theta \in \mathbb{R}^k \otimes \mathcal{H}$, one can write Θ in row-style as an element of $\mathcal{H}^k$ as well as its adjoint $\Theta^\top \in \mathcal{H} \otimes \mathbb{R}^k$ in column-style which follows from the fact that $\mathcal{H}^k$ is self-adjoint when endowed with the ℓ^2-product topology.

A.2. Proof of Remark 1

Let us characterize (2) in order to easily implement it with unbiased stochastic gradient. We need to get the expectation outside the norm. This can be done with the following derivations

$$\begin{aligned}
\left\|\mathbb{E}[\psi(\xi)\psi(\xi)^\top] - I\right\|^2 &= \operatorname{Tr}\left((\mathbb{E}[\psi(\xi)\psi(\xi)^\top] - I)(\mathbb{E}[\psi(\xi')\psi(\xi')^\top] - I)\right) \\
&= \mathbb{E}_{\xi,\xi'}\left[\operatorname{Tr}\left(\psi(\xi)\psi(\xi)^\top\psi(\xi')\psi(\xi')^\top\right)\right] - 2\,\mathbb{E}_\xi\left[\operatorname{Tr}\left(\psi(\xi)\psi(\xi)^\top\right)\right] + \operatorname{Tr}(I) \\
&= \mathbb{E}_{\xi,\xi'}\left[(\psi(\xi')^\top\psi(\xi))^2\right] - 2\,\mathbb{E}_\xi\left[\|\psi(\xi)\|^2\right] + k.
\end{aligned}$$

For the first part, we get

$$\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \psi(\xi')\|^2 \,\middle|\, X\right] = 2\,\mathbb{E}_\xi[\|\psi(\xi)\|^2] - 2\,\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top\psi(\xi') \,\middle|\, X\right]$$

As a consequence,

$$\mathcal{L}(\psi;\beta) = 2(\beta - 1)\,\mathbb{E}_\xi[\|\psi(\xi)\|^2] - 2\beta\,\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top\psi(\xi') \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[(\psi(\xi')^\top\psi(\xi))^2\right] + k. \tag{10}$$

In particular, when $\beta = 1$, we retrieve the spectral contrastive loss introduced by HaoChen et al. (2021),

$$\mathcal{L}(\psi) = -2\,\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top\psi(\xi') \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}[(\psi(\xi)^\top\psi(\xi'))^2] + k.$$

A.3. Proof of Lemma 2

First, notice that if we define for $\psi : \mathcal{X} \to \mathbb{R}$, the mapping $\omega(\psi) \to \mathbb{E}_X[\mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \psi(\xi)\|^2 \,\middle|\, X\right]]$, ω is a quadratic form on $L^2(\mu_\Xi)$. As a consequence, it can be represented with a linear self-adjoint operator L on $L^2(\mu_\Xi)$ such that $\omega(\psi) = \langle \psi, L\psi \rangle_{L^2(\mu_\Xi)}$. Because $\omega(\psi) \geq 0$, we have $L \succeq 0$ (with $\succeq$ the Loewner order on symmetric operators, i.e. $A \succeq B$ if $A - B$ is positive). The following lemma show that L is bounded.

Lemma 5. *For any $\psi \in L^2(\mu_\Xi)$, $\omega(\psi) \leq 2\,\|\psi\|^2_{L^2(\mu_\Xi)}$. As a consequence, $L \preceq 2I$.*

Proof. This follows from the fact that

$$\begin{aligned}
\omega(\psi) &= \mathbb{E}_X[\mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \psi(\xi')\|^2 \,\middle|\, X\right]] \\
&= \mathbb{E}_X[\mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \mathbb{E}\left[\psi(\xi) \,|\, X\right] + \mathbb{E}\left[\psi(\xi) \,|\, X\right] - \psi(\xi')\|^2 \,\middle|\, X\right]] \\
&= \mathbb{E}_X[\mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \mathbb{E}\left[\psi(\xi) \,|\, X\right]\|^2 + \|\mathbb{E}\left[\psi(\xi) \,|\, X\right] - \psi(\xi')\|^2 \,\middle|\, X\right]] \\
&= 2\,\mathbb{E}_X[\mathbb{E}_{\xi}\left[\|\psi(\xi) - \mathbb{E}\left[\psi(\xi) \,|\, X\right]\|^2 \,\middle|\, X\right]] \\
&= 2 \min_{\psi_0 : \mathcal{X} \to \mathbb{R}} \mathbb{E}_X[\mathbb{E}_{\xi}\left[\|\psi(\xi) - \psi_0(X)\|^2 \,\middle|\, X\right]] \\
&\leq 2\,\mathbb{E}_X[\mathbb{E}_{\xi}\left[\|\psi(\xi)\|^2 \,\middle|\, X\right]] = 2\,\mathbb{E}_{\xi}\left[\|\psi(\xi)\|^2\right] = 2\,\|\psi\|_{L^2(\mu_\Xi)}
\end{aligned}$$

Hence for any ψ, with the $L^2(\mu_\Xi)$ geometry we have $\psi^\top L \psi \leq \psi^\top \psi$, which implies, since L is self-adjoint, that $\|L\|_{\mathrm{op}} \leq 2$. □

Because $0 \preceq L \preceq 2I$, let us introduce $T = (2I - L)/2$, we have $0 \preceq T \preceq 1$ and, with the $L^2(\mu_\Xi)$ geometry, for $\psi : \mathcal{X} \to \mathbb{R}^k$

$$\mathcal{L}(\psi) = \beta \sum_{i=1}^{k} \psi_i^\top L \psi_i + \left\|\mathbb{E}[\psi(\xi)\psi(\xi')^\top] - I\right\|^2 = 2\beta \sum_{i=1}^{k} \psi_i^\top (I - T)\psi_i + \left\|\mathbb{E}[\psi(\xi)\psi(\xi')^\top] - I\right\|^2.$$

In order to diagonalize all operators without relying on integral formulations of the spectral theorem, we introduce the following mild assumption.

Assumption 4. *Assume that T has a pure point spectrum.*

Example 7. *When the distribution of augmentation have a density p with respect to a any measure and $(x, \xi) \to p\left(\xi \,|\, x\right)/p(\xi)$ is in $L^2(\mu)$, or when $\mathcal{X}$ is finite, T can be shown to be a compact operator, hence to have a pure point spectrum according to the spectral theorem.*

Proof. When $\mathcal{X}$ is finite, the L^2 spaces are finite dimensional, hence locally compact, which implies that all operators are compact. To prove the case with density, let us develop T as an integral operator. We have, in $L^2(\mu_\Xi)$ geometry, for $f : \mathcal{X} \to \mathbb{R}$

$$\begin{aligned}
2f^\top (I - T) f &= \mathbb{E}_X\,\mathbb{E}\left[\|f(\xi) - f(\xi')\|^2 \,\middle|\, X\right] = \mathbb{E}_X\,\mathbb{E}\left[\|f(\xi)\|^2 + \|f(\xi')\|^2 - 2\left\langle f(\xi), f(\xi')\right\rangle \,\middle|\, X\right] \\
&= 2f^\top f - 2\,\mathbb{E}_X\,\mathbb{E}\left[\langle f(\xi), f(\xi')\rangle \,|\, X\right].
\end{aligned}$$

This allow us to identify T with the inner product, we have for $g : \mathcal{X} \to \mathbb{R}$ and p the density of augmentations

$$\begin{aligned}
f^\top T g &= \mathbb{E}_X\,\mathbb{E}\left[\langle f(\xi), g(\xi')\rangle \,|\, X\right] = \int \langle f(\xi), g(\xi')\rangle\, p\left(\xi \,|\, x\right) p\left(\xi' \,|\, x\right) \mathrm{d}\xi'\, \mathrm{d}\xi \mu_{\mathcal{X}}(\mathrm{d}x) \\
&= \int \mu_\Xi(\mathrm{d}\xi) \left\langle f(\xi), \int \mu_\Xi(\mathrm{d}\xi') g(\xi') \frac{\int \mu_{\mathcal{X}}(\mathrm{d}x) p\left(\xi \,|\, x\right) p\left(\xi' \,|\, x\right)}{p(\xi)p(\xi')} \right\rangle.
\end{aligned}$$

As a consequence, one can consider T as the integral operator in $L^2(\mu_\Xi)$ linked with the kernel

$$k(\xi, \xi') = \frac{\int \mu_{\mathcal{X}}(\mathrm{d}x) p\left(\xi \mid x\right) p\left(\xi' \mid x\right)}{p(\xi)p(\xi')}.$$

When this kernel is bounded, or simply when $\xi \to k(\xi, \xi)$ belongs to $L^2(\mu_\Xi)$, T is trace-class hence compact. □

Let us now prove in order to minimize $\mathcal{L}$, one should take the eigenfunctions of the operator $(1-\beta)I + \beta T$ whose corresponding eigenvalues are the biggest positives ones. It can be proven with simple geometry in a somewhat abstract space. To do so, remark that $\psi : \mathcal{X} \to \mathbb{R}^k$ in $L^2(\mu_\Xi, \mathcal{X}, \mathbb{R}^k)$ can be represented as $\tilde{\psi} \in \mathbb{R}^k \otimes L^2(\mu_\Xi, \mathcal{X}, \mathbb{R})$ with the linear map that associates $(\psi_i^\top \varphi)$ to a function $\varphi \in L^2(\mu_\Xi, \mathcal{X}, \mathbb{R})$. Let denote $T_\beta = (1-\beta)I + \beta T$, the upstream loss can be characterized as

$$\begin{aligned}
\mathcal{L}(\psi) &= 2\beta \sum_{i\in[k]} \langle \psi_i, (I-T)\psi_i \rangle + \left\| \mathbb{E}_\xi[\psi(\xi)\psi(\xi)^\top] - I \right\|^2 \\
&= 2\beta \sum_{i\in[k]} \left\langle e_i^\top \psi, (I-T)\psi^\top e_i \right\rangle + \left\| \mathbb{E}_\xi[\sum_{i,j\in[k]} e_i^\top \psi(\xi)\psi(\xi)^\top e_j e_i e_j^\top] - I \right\|^2 \\
&= 2\beta \sum_{i\in[k]} e_i \tilde{\psi}(I-T)\tilde{\psi}^\top e_i + \left\| \sum_{i,j\in[k]} e_i^\top \tilde{\psi}\tilde{\psi}^\top e_j e_i e_j^\top - I \right\|^2 \\
&= 2\beta \operatorname{Tr}\left(\tilde{\psi}(I-T)\tilde{\psi}^\top\right) + \left\| \tilde{\psi}\tilde{\psi}^\top - I \right\|^2 = 2\beta \operatorname{Tr}\left(\tilde{\psi}(I-T)\tilde{\psi}^\top\right) + \operatorname{Tr}\left(\left(\tilde{\psi}\tilde{\psi}^\top - I\right)^2\right). \\
&= \operatorname{Tr}\left(2\beta\tilde{\psi}(I-T)\tilde{\psi}^\top + \tilde{\psi}\tilde{\psi}^\top\tilde{\psi}\tilde{\psi}^\top - 2\tilde{\psi}\tilde{\psi}^\top + I\right). \\
&= \operatorname{Tr}_{L^2(\mu_\Xi)}\left(\tilde{\psi}^\top\tilde{\psi}\tilde{\psi}^\top\tilde{\psi} + (2\beta(I-T) - 2I)\tilde{\psi}^\top\tilde{\psi}\right) + k. \\
&= \operatorname{Tr}_{L^2(\mu_\Xi)}\left(\tilde{\psi}^\top\tilde{\psi}\tilde{\psi}^\top\tilde{\psi} - 2T_\beta\tilde{\psi}^\top\tilde{\psi}\right) + k. = \operatorname{Tr}_{L^2(\mu_\Xi)}\left((\tilde{\psi}^\top\tilde{\psi} - T_\beta)^2 - T_\beta^2\right) + k.
\end{aligned}$$

In order to find the minimizer of $\mathcal{L}$ with this new characterization, slight precautions are needed here since the two operators are not trace-class. The following lemma takes those precautions in order to finish the proof.

Lemma 6. *Let A be a self-adjoint operator on $L^2(\mu_\Xi)$. Assume that there exists c such that $A \preceq cI$ and that A is pure-point spectrum. Then if (λ_i, f_i) denote the eigen-decomposition of A with λ_i in decreasing order, the minimization of $\operatorname{Tr}\left((B-A)^2 - B^2\right)$ under the constraint that B is a self-adjoint positive operator of rank at most k, is reached for $B = \tilde{\psi}^\top\tilde{\psi}$ with $\psi : \mathcal{X} \to \mathbb{R}^k$ such that $\psi_i = \max(0, \lambda_i)^{1/2} f_i$.*

Proof. Let us decompose A into a positive part $A_+ \succeq 0$ and a negative part $A_- \succeq 0$ such that $A = A_+ - A_-$. Using the fact that B is positive self-adjoint, we get

$$\begin{aligned}
\operatorname{Tr}\left((B-A)^2 - A^2\right) &= \operatorname{Tr} B^2 - 2B^{1/2}AB^{1/2} = \operatorname{Tr}\left(B^2 - 2B^{1/2}A_+B^{1/2}\right) + 2\operatorname{Tr}\left(B^{1/2}A_-B^{1/2}\right) \\
&\geq \operatorname{Tr}\left(B^2 - 2B^{1/2}A_+B^{1/2}\right).
\end{aligned}$$

Let us decompose B into k symmetric operators of rank at most one as $B = \sum_{i=1}^k B_i$ such that $B_iB_j = 0$ for any

$i \neq j \in [k]$. Using the different properties of the operators introduced, we proceed with

$$\begin{aligned}
\operatorname{Tr}\left((B-A)^2 - A^2\right) &\geq \sum_{i=1}^{k} \operatorname{Tr}\left(B_i^2\right) - 2\operatorname{Tr}\left(B_i A_+\right) = \sum_{i=1}^{k} \|B_i\|_{\mathrm{op}}^2 - 2\|B_i A_+\|_{\mathrm{op}} \\
&\geq \sum_{i=1}^{k} \|B_i\|_{\mathrm{op}}^2 - 2\|B_i\|_{\mathrm{op}} \|\Pi_{B_i} A_+\|_{\mathrm{op}} \geq \sum_{i=1}^{k} \|B_i\|_{\mathrm{op}}^2 - 2\|B_i\|_{\mathrm{op}} \Big\| \prod_{j<i} (I - \Pi_{B_j}) A_+ \Big\|_{\mathrm{op}} \\
&= \sum_{i=1}^{k} \left(\|B_i\|_{\mathrm{op}} - \Big\| \prod_{j<i} (I - \Pi_{B_j}) A_+ \Big\|_{\mathrm{op}} \right)^2 - \Big\| \prod_{j<i} (I - \Pi_{B_j}) A_+ \Big\|_{\mathrm{op}}^2 \\
&\geq -\sum_{i=1}^{k} \Big\| \prod_{j<i} (I - \Pi_{B_j}) A_+ \Big\|_{\mathrm{op}}^2 \geq -\sum_{i=1}^{k} \sigma_i(A_+)
\end{aligned}$$

where Π_B denote the orthogonal projector on the image of B, and $\sigma_i(A)$ the i-th singular value of A (monotonically ordered with $\sigma_1(A)$ the biggest). The last inequality is due to the Courant-Fisher min-max principle, This inequality can be achieved with Π_{B_i} the projection on the i-th eigenspace of A and $\|B_i\|_{\mathrm{op}} = \sigma_i(A)$. In other terms, B should match the first k positive eigenvalues of A. In the case where A has less than k positive eigenvalues, then B should match all the positive eigenvalues and be null on the range of A_-. □

The following is a direct corollary of the proof above.

Proposition 7 (Uniqueness of minimizers)**.** *The minimizers of $\mathcal{L}$ are unique up to orthogonal transformations and eigenfunction picking. More specifically, if $U \in \mathbb{R}^{k\times k}$ is orthogonal, i.e. $U^\top U = I$, then $\mathcal{L}(\psi) = \mathcal{L}(U\psi)$; and if $\lambda_k = \lambda_{k+1}$, one can choose different eigenfunctions as f_k in the eigen-decomposition (λ_i, f_i) of T_β.*

A.4. Proof of Lemma 3

Let us consider $\psi : \mathcal{X} \to \mathbb{R}^k$ with $\psi_i = f_{\theta_i}$ for $\theta_i \in \mathcal{H}$ and $S : \mathcal{H} \to L^2(\mu_\Xi); \theta \to \theta^\top \varphi(\cdot)$. We can use the tensor notations introduced earlier to parameterized $\psi = S\Theta^\top$ with $\Theta = (\theta_i)_{i\in[k]}$ seen as an element of $\mathbb{R}^k \otimes \mathcal{H}$. The proof of Lemma 3 follows from the fact that

$$\|\Theta\|_2^2 = \left\|\Theta^\top\right\|_2^2 = \left\langle S\Theta^\top, (S^\top S)^{-1} S\Theta^\top \right\rangle_{L^2(\mu_\Xi)} = \sum_{i=1}^{k} S\theta_i^\top K^{-1} S\theta_i = \sum_{i=1}^{k} \psi_i^\top K^{-1} \psi_i.$$

Since $\Psi = S(\mathcal{H}) = \operatorname{im} S = \operatorname{im} K^{1/2} = K^{1/2}(L^2(\mu_\Xi))$, one can consider K^{-1} as the inverse of K such that for $\psi_i \in \ker K$, $\psi_i^\top K^{-1} \psi_i = +\infty$. This is what we implicitly assumed in the main paper, which lead to the (ψ_i) being all in Ψ. Note that in many cases, Ψ is dense in $L^2(\mu_\Xi)$ (Micchelli et al., 2006), and one does not need to take such a precaution since the $\ker K = \{0\}$, and there is only one way to define K^{-1} on $L^2(\mu_\Xi)$.

A.5. Second proof of Lemma 3 with covariance operators

The proof given above of Lemma 3 might seem quite abstract for the reader unfamiliar with reproducing kernel Hilbert space. In this subsection, we provide a somewhat more accessible proof of this Lemma based on covariance operators.

Reusing the unbiased characterization of $\mathcal{L}$ we have

$$\begin{aligned}
\mathcal{L}(\psi;\beta) &= 2(\beta-1)\,\mathbb{E}_\xi[\|\psi(\xi)\|^2] - 2\beta\,\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top\psi(\xi') \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[(\psi(\xi')^\top\psi(\xi))^2\right] + k. \\
&= 2(\beta-1)\operatorname{Tr}\left(\mathbb{E}_\xi[\psi(\xi)\psi(\xi)^\top]\right) - 2\beta\operatorname{Tr}\left(\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\psi(\xi))\psi(\xi')^\top \,\middle|\, X\right]\right) + \operatorname{Tr}\left(\mathbb{E}_\xi\left[\psi(\xi)\psi(\xi)^\top\right]^2\right) + k,
\end{aligned}$$

where the last term provides from the fact that

$$\begin{aligned}
\mathbb{E}_{\xi,\xi'}\left[(\psi(\xi)^\top\psi(\xi'))^2\right] &= \mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top\psi(\xi')\psi(\xi')^\top\psi(\xi)\right] = \mathbb{E}_{\xi,\xi'}\left[\operatorname{Tr}\left(\psi(\xi')\psi(\xi')^\top\psi(\xi)\psi(\xi)^\top\right)\right] \\
&= \operatorname{Tr}\left(\mathbb{E}_\xi[\psi(\xi)\psi(\xi)^\top]\,\mathbb{E}_{\xi'}[\psi(\xi')\psi(\xi')^\top]\right) = \operatorname{Tr}\left(\mathbb{E}_\xi[\psi(\xi)\psi(\xi)^\top]^2\right).
\end{aligned}$$

A.5.1. Operator technicalities

The search for ψ will be done under the form $\Theta\varphi$ for $\Theta \in \mathbb{R}^k \otimes \mathcal{H}$ and $\varphi : \mathcal{X} \to \mathcal{H}$. Let us discuss technicalities related to the infinite dimensional operators that will appear.

Assumption 5. *The Hilbert space $\mathcal{H}$ is separable, and the mapping φ belongs to $L^2(\mu_{\mathcal{X}})$ endowed with Borel topology on both $\mathcal{X}$ and $\mathcal{H}$.*

Remark 8. *The operator $\Sigma = \mathbb{E}_\xi[\varphi(\xi)\varphi(\xi)^\top] \in \mathcal{H} \otimes \mathcal{H}$ is trace-class.*

Proof. This follows from linearity of traces, expectations, together with the fact that $\operatorname{Tr} AB = \operatorname{Tr} BA$,

$$\operatorname{Tr}\Sigma = \mathbb{E}_\xi \operatorname{Tr}\varphi(\xi)\varphi(\xi)^\top = \|\varphi\|^2_{L^2(\mu_{\mathcal{X}})} < +\infty.$$

As a consequence, Σ is compact, hence has a pure point spectrum, and since $\mathcal{H}$ is separable it can be diagonalized with its eigenvectors forming a basis of $\mathcal{H}$. □

We will see later that $\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}$ is indeed isometric to T. Hence, under Assumption 4, $\Sigma^{-1/2}\Sigma_{\mathcal{X}}\Sigma^{-1/2}$ has a pure-point spectrum. However, the following lemma shows that this operator is bounded without using the fact that $T \preceq I$.

Remark 9. *The operator $\Sigma_X = \mathbb{E}_X \mathbb{E}_{\xi,\xi'}\left[\varphi(\xi)\varphi(\xi')^\top \,\middle|\, X\right] \in \mathcal{H} \otimes \mathcal{H}$ verifies $0 \preceq \Sigma_X \preceq \Sigma$ with $\preceq$ the Loewner order ($A \preceq B$ if $B - A$ is semi-definite positive). As a consequence, Σ_X is trace-class and $\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}$ is continuous.*

Proof. This follows from Jensen inequality applies to $A \to AA^\top$, which can be proven using the positivity of covariance operator.

$$0 \preceq \mathbb{E}[(A - \mathbb{E}[A])(A - \mathbb{E}[A])^\top], \qquad \Rightarrow \qquad \mathbb{E}[A]\,\mathbb{E}[A]^\top \preceq \mathbb{E}[AA^\top].$$

As a consequence,

$$\mathbb{E}_{\xi,\xi'}\left[\varphi(\xi)\varphi(\xi')^\top \,\middle|\, X = x\right] \preceq \mathbb{E}_\xi\left[\varphi(\xi)\varphi(\xi)^\top \,\middle|\, X = x\right],$$

which implies that

$$\Sigma_X \preceq \Sigma.$$

As a consequence, $\operatorname{Tr}\Sigma_X \leq \operatorname{Tr}\Sigma < +\infty$ and $\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2} \preceq I$, and $\left\|\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}\right\|_{\mathrm{op}} \leq 1$. The positivity follows from the fact that Σ_X is a covariance operator $\Sigma_X = \mathbb{E}_X\left[\mathbb{E}_\xi\left[\varphi(\xi) \mid X\right]\mathbb{E}_\xi\left[\varphi(\xi) \mid X\right]^\top\right]$. □

A.5.2. Operator formulation

Let us begin by proving a variant of the lemma where everything is expressed in $\mathcal{H}$. We expand later on the isometry between $\mathcal{H}$ and $L^2(\mu_\Xi)$ (due to the isometry between S and $\Sigma^{1/2}$) that allows us to transfer it to the lemma written in the paper.

Lemma 10. *For $(\theta_i) \in \mathcal{H}^k$ and $f_\theta : x \to \langle\varphi(x), \theta\rangle$, and a regularizer $\lambda \in \mathbb{R}$*

$$\mathcal{L}((f_{\theta_i})_{i\in[k]}) + \lambda \sum_{i\in[k]} \|\theta_i\|_2^2 = \operatorname{Tr}\left(\left(\Sigma^{1/2}(\sum_{i\in[k]}\theta_i\theta_i^\top)\Sigma^{1/2} - A)^2 - A^2\right) + k,$$

with A and Σ being operator on $\mathcal{H}$ defined as

$$A = \Sigma^{-1/2}((1-\beta)\Sigma + \beta\Sigma_X - \lambda I)\Sigma^{-1/2}, \quad \Sigma = \mathbb{E}_\xi\left[\varphi(\xi)\varphi(\xi)^\top\right], \quad \Sigma_X = \mathbb{E}_X[\mathbb{E}_{\xi,\xi'}\left[\varphi(\xi)\varphi(\xi')^\top \,\middle|\, X\right]].$$

As a consequence, a minimizer Θ_* of $\mathcal{L}$ is such that Θ_* matches the eigenvalue decomposition of A on positive eigenvalues up to the k-th. Formally, if $A = \sum_{i\in\mathbb{N}} \lambda_i u_i \otimes u_i$ with $u_i \in \mathcal{H}$ and (λ_i) in decreasing order,

$$\Theta_* = (\theta_i)_{i\in[k]}, \text{ with } \theta_i = \sqrt{\max(\lambda_i, 0)}\Sigma^{-1/2}u_i.$$

Moreover, (f_{θ_i}) are orthogonal in $L^2(\mu_\Xi)$, where μ_Ξ denotes the marginal distribution over augmentations.

Proof. Let us now rewrite the different quantities appearing in $\mathcal{L}$ based on the parameterization $\psi = \Theta\varphi$. We have

$$\operatorname{Tr}\left(\mathbb{E}[\psi(\xi)\psi(\xi)^\top]\right) = \operatorname{Tr}\left(\mathbb{E}[\Theta\varphi(\xi)\varphi(\xi)^\top\Theta^\top]\right) = \operatorname{Tr}\left(\Theta\,\mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\Theta^\top\right) = \operatorname{Tr}\left(\Theta\Sigma\Theta^\top\right) = \operatorname{Tr}\left(\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2}\right).$$

The adjoint $\Theta^\top$ is taken with respect to the canonical topology on $\mathcal{H}$ and $\mathbb{R}^k$. Similarly,

$$\operatorname{Tr}\left(\mathbb{E}_X\,\mathbb{E}\left[\psi(\xi)\psi(\xi')^\top \mid X\right]\right) = \operatorname{Tr}\left(\Theta\Sigma_X\Theta^\top\right) = \operatorname{Tr}\left(\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2}\right).$$

For the last term, we get

$$\operatorname{Tr}\left(\mathbb{E}[\psi(\xi)\psi(\xi)^\top]^2\right) = \operatorname{Tr}\left((\Theta\Sigma\Theta^\top)^2\right) = \operatorname{Tr}\left(\Theta\Sigma\Theta^\top\Theta\Sigma\Theta^\top\right) = \operatorname{Tr}\left(\Sigma^{1/2}\Theta^\top\Theta\Sigma\Theta^\top\Theta\Sigma^{1/2}\right)$$

Collecting the different terms, we get

$$\begin{aligned}
&\mathcal{L}(\Theta\varphi) + 2\lambda\operatorname{Tr}(\Theta^\top\Theta) - k \\
&= \operatorname{Tr}\big(2(\beta-1)\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2} - 2\beta\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2} \\
&\qquad\qquad + \Sigma^{1/2}\Theta^\top\Theta\Sigma\Theta^\top\Theta\Sigma^{1/2} + 2\lambda\Sigma^{-1}\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2}\big) \\
&= \operatorname{Tr}\left(\left(\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2} + (\beta-1)I - \beta\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2} + \lambda\Sigma^{-1}\right)^2 - \left((\beta-1)I - \beta\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2} + \lambda\Sigma^{-1}\right)^2\right) \\
&= \operatorname{Tr}\left(\left(\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2} - \Sigma^{-1/2}((1-\beta)\Sigma + \beta\Sigma_X - \lambda)\Sigma^{-1/2}\right)^2 - \left(\Sigma^{-1/2}((1-\beta)\Sigma + \beta\Sigma_X - \lambda)\Sigma^{-1/2}\right)^2\right).
\end{aligned}$$

This proves the first part of the lemma. Remark that the expression of the lemma is slightly different from the generalization to continuous $\mathcal{X}$ suggested by HaoChen et al. (2021) in their Appendix F, that would reuse the work of Schiebinger et al. (2015) considering the covariance operator with feature $\bar{\varphi}(x) = q^{-1/2}(x)\,\mathbb{E}\left[\varphi(\xi) \mid X = x\right]$ where $q : x \to \mathbb{E}_{X\sim\mu_\Xi}[k(x, X)]$ rather than $\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}$.

Finally, let us prove that the f_{θ_i} are orthogonal in L^2, we have

$$\begin{aligned}
\left\langle f_{\theta_i}, f_{\theta_j}\right\rangle_{L^2(\mu_\Xi)} &= \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}\,\mathbb{E}[\left\langle\Sigma^{-1/2}u_i, \varphi(\xi)\right\rangle\left\langle\Sigma^{-1/2}u_i, \varphi(\xi)\right\rangle] \\
&= \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}\,\mathbb{E}[u_i^\top\Sigma^{-1/2}\varphi(\xi)\varphi(\xi)^\top\Sigma^{-1/2}u_j] \\
&= \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}u_i^\top\Sigma^{-1/2}\,\mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\Sigma^{-1/2}u_j \\
&= \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}u_i^\top\Sigma^{-1/2}\Sigma\Sigma^{-1/2}u_j \\
&= \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}u_i^\top u_j = \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}\delta_{ij}.
\end{aligned}$$

This proves the orthogonality of the f_{θ_i} in $L^2(\mu_\Xi)$. □

A.5.3. Isometric formulation

Let us consider $S : \mathcal{H} \to L^2(\mu_\Xi); \theta \to f_\theta$ the embedding of the RKHS in $L^2(\mu_\Xi)$.

Lemma 11. *S is isometric to $\Sigma^{1/2}$, and $K = S^\top S$ is an integral operator that maps $f \in L^2(\mu_\Xi)$ to $Kf \in L^2(\mu_\Xi)$ defined for $\xi \in \mathcal{X}$ as*

$$Kf(\xi) = \mathbb{E}_{\xi'}\left[\varphi(\xi)^\top\varphi(\xi')f(\xi')\right]. \tag{11}$$

Proof. This follows from the fact that both S and $\Sigma^{1/2}$ are a square root of Σ. Indeed, $\Sigma = S^\top S$, since for $\theta \in \mathcal{H}$,

$$\begin{aligned}
\left\langle\theta, S^\top S\theta\right\rangle_{\mathcal{H}} &= \left\langle S\theta, S\theta\right\rangle_{L^2(\mu_\mathcal{X})} = \mathbb{E}_\xi[S\theta(\xi)^2] \\
&= \mathbb{E}_\xi[\langle\theta, \varphi(\xi)\rangle^2] = \mathbb{E}_\xi[\langle\theta, \varphi(\xi)\otimes\varphi(\xi)\theta\rangle] \\
&= \langle\theta, \mathbb{E}[\varphi(\xi)\otimes\varphi(\xi)]\theta\rangle = \langle\theta, \Sigma\theta\rangle.
\end{aligned}$$

As a consequence, S is isometric to $\Sigma^{1/2}$ (if we write the singular value decomposition of S as $USV^\top$, then $\Sigma^{1/2} = USU^\top$). Regarding the part in K, one can check with the same derivation that $S^\top f = \mathbb{E}[f(\xi)\varphi(\xi)] \in \mathcal{H}$ hence the value of $(Kf)(\xi) = (S^\top f)^\top\varphi(\xi) = \mathbb{E}_{\xi'}[f(\xi')\varphi(\xi')^\top\varphi(\xi)]$. □

Using the isometry one can replace $\|S\theta\| = \left\|\Sigma^{1/2}\theta\right\|$ with the Hilbertian norm on $\mathcal{H}$ and $L^2(\mu_\Xi)$, so that for C operating in $\mathcal{H}$, $\operatorname{Tr} SCS^\top = \operatorname{Tr} \Sigma^{1/2} C \Sigma^{1/2}$. Going back to the proof in $\mathcal{H}$, one can replace all the $\Sigma^{1/2}$ by S or its adjoint at the right places to get the following statement.

Lemma 12. *For $\Theta \in \mathbb{R}^k \otimes \mathcal{H}$, and a regularized $\lambda \in \mathbb{R}$*

$$\mathcal{L}(S\Theta) + \lambda \|\Theta\|_2^2 = \operatorname{Tr}((S\Theta^\top \Theta S^\top - T_\lambda)^2 - T_\lambda^2) + k$$

where

$$T = S^{-\top} \Sigma_X S^{-1}, \qquad T_\lambda = (1-\beta)I + \beta T - \lambda K^{-1}, \qquad K = SS^\top,$$

with $S : \mathcal{H} \to L^2(\mu_\Xi); \theta \to f_\theta$ the embedding of $\mathcal{H}$ in $L^2(\mu_\Xi)$, where μ_Ξ denotes the marginal distribution over augmentations. As a consequence, a minimizer Θ_* of $\mathcal{L}; \lambda$ is such that $S\Theta_*^\top$ matches the eigenvalue decomposition of T_λ on positive eigenvalues up to the k-th.

Proof. This lemma follows from the previous discussion. The fact that $S^{-\top}\Sigma S^{-1}$ equates to T on the $L^2(\mu_\Xi)$-closure of Ψ is due to the characterization in Lemma 3. We can nonetheless prove it in a more direct fashion, by adapting Lemma B.9 of Saunshi et al. (2022) to our case. □

A.6. Proof of Proposition 4

Proposition 4 relies on the fact that when two operators commutes, they can be diagonalized in the same basis.

Lemma 13. *When K and T commute, K and T can be diagonalized by the same eigenfunctions (f_i).*

Proof. When the operators commute, if f is an eigenfunction of T with $Tf = \lambda f$, then $TKf = KTf = \lambda K f$. This means that the eigenspace of T, i.e. $\ker(T - \lambda I)$ are stable by K. As a consequence, K can be decomposed with respect to the summation $L^2 = \oplus_{\lambda \in \operatorname{spec}(T)} \ker(T - \lambda I)$. By diagonalizing the restrictions of K on each of those spaces, there exists a basis that diagonalizes both K and T. □

While we did not discuss it in the main text, one should not consider any eigenvalue decomposition of T but only the eigenfunctions that jointly diagonalize T and K. However, note that to find those eigenfunctions, based on Courant-Fisher principle, one can take, recursively on $i \in \mathbb{N}$, $f_i = f_{\theta_i}$ an eigenfunction in $\ker(T - \lambda_i I)$ that maximizes or minimizes $\|\theta_i\|$. Those eigenfunctions (f_i) will diagonalize T_λ, and the optimal representation will pick the ones that maximize $f_i^\top T_\lambda f_i$ as long as this quantity is positive.

If f_i diagonalize K then $f_i \in \operatorname{im} K^{1/2} = \Psi = \operatorname{im} S$, hence there exists a $\theta_i \in \mathcal{H}$ such that $f_i = S\theta_i$. As a consequence, with the $L^2(\mu_\Xi)$ geometry, $f_i^\top K^{-1} f_i = (S\theta_i)^\top (S^\top S)^{-1} S\theta_i = \|\theta_i\|_2^2$. We use this to derive that

$$f_i T_\lambda f_i = (1-\beta) f_i^\top f_i + \beta f_i^\top T f_i - \lambda f_i^\top K^{-1} f_i = 1 - \beta + \beta\lambda_i - \lambda \|\theta_i\|_2^2 .$$

In other term the maximal eigenvalues of T_λ are found by maximizing $\beta\lambda_i - \lambda \|\theta_i\|^2$.

Remark 14. *Recently, HaoChen & Ma (2022) have taken this second perspective on inductive bias perspective by looking at the "barrier" case where one can only match eigenfunctions that belongs to the function space Ψ. In the kernel regime, this is deceptive since, for example, when considering the Gaussian kernel $\varphi(x)^\top \varphi(x') = -\exp(\|x - x'\|^2)$, Ψ is made of analytic functions (Sun & Zhou, 2008), hence cannot parameterize any indicator functions without being one everywhere, therefore their approach would fail to explain how the Gaussian kernel could learn fast under the cluster assumption.*

A.7. Remark about VCReg

When $\mathcal{L} = 0$, finding ψ correspond in finding k functions $(f_{\theta_i})_i$ that are orthogonal in $L^2(\mu_\Xi)$ and maximize $1 - \lambda \|\theta\|^2 = 1 - \lambda f_\theta^\top K^{-1} f_\theta$ before multiplying them by $(1 - \lambda \|\theta_i\|^2)_+$. Using Courant-Fisher min-max principle, the function $(f_{\theta_i})_i$ are given by the k biggest eigenfunctions of K.

A.8. Proof of Example 3

If $\mu_\Xi = \delta\rho_\mathcal{X} + (1-\delta)\mu_\perp$, then for any measurable function f

$$\|f\|_{L^2(\mu_\Xi)}^2 = \delta \int_\mathcal{X} f(x)^2 \rho_\mathcal{X}(\mathrm{d}x) + (1-\delta) \int_\mathcal{X} f(x)^2 \mu_\perp(\mathrm{d}x) \geq \delta \|f(X)\|_{L^2(\rho_\mathcal{X})}^2 .$$

A.9. Proof of Example 4

This follows from the embedding S_ρ and S_μ of $\mathcal{H}$ in $L^2(\rho_\mathcal{X})$ and $L^2(\mu_\Xi)$ respectively. We have seen earlier that $S_\mu^\top S_\mu = \Sigma_{\mu_\Xi}$ and $S_X^\top S_X = \Sigma_{\rho_\mathcal{X}}$. Let $f \in \Psi$, there exists $\theta \in \mathcal{H}$ such that $f = f_\theta$, hence, using the isometry between S and $\Sigma^{1/2}$,

$$\begin{aligned}\|f_\theta\|^2_{L^2(\rho_\mathcal{X})} = \|S_X\theta\|^2_{L^2(\rho_\mathcal{X})} &= \left\|\Sigma^{1/2}_{\rho_\mathcal{X}}\theta\right\|^2_\mathcal{H} = \left\|\Sigma^{1/2}_{\rho_\mathcal{X}}\Sigma^{-1/2}_{\mu_\Xi}\Sigma^{1/2}_{\mu_\Xi}\theta\right\|^2_{L^2(\rho_\mathcal{X})} \\ &\leq \left\|\Sigma^{-1/2}_{\mu_\Xi}\Sigma_{\rho_\mathcal{X}}\Sigma^{-1/2}_{\mu_\Xi}\right\|_{\text{op}} \left\|\Sigma^{-1/2}_{\mu_\Xi}\theta\right\|^2_\mathcal{H} = \left\|\Sigma^{-1/2}_{\mu_\Xi}\Sigma_{\rho_\mathcal{X}}\Sigma^{-1/2}_{\mu_\Xi}\right\|_{\text{op}} \|f_\theta\|^2_{L^2(\mu_\Xi)}\,.\end{aligned}$$

We conclude by using the equivalence the fact that $A \preceq cB$ implies that $B^{-1/2}AB^{-1/2} \preceq c \cdot I$.

A.10. Proof of Example 5

This follows from the definition of the different objects,

$$\Pi^{(\rho_\mathcal{X})}_{f_i}(f) = wf_i, \qquad \text{with} \qquad w = \operatorname*{arg\,min}_{w\in\mathbb{R}} \mathbb{E}_{X\sim\rho_\mathcal{X}}[\|f(X) - wf_i(X)\|^2].$$

We develop this last objective as

$$\begin{aligned}\mathbb{E}_{X\sim\rho_\mathcal{X}}[\|f(X) - wf_i(X)\|^2] &= \mathbb{E}_{X\sim\rho_\mathcal{X}}[\|g(\psi(X)) - wf_i(X)\|^2] = \mathbb{E}_{Z\sim\psi_\#\rho_\mathcal{X}}[\left\|g(Z) - w^\top Z\right\|^2] \\ &= \mathbb{E}_{Z\sim\mu_\Xi}[\left\|f(X) - w^\top f_i(X)\right\|^2].\end{aligned}$$

Hence, the equality of the w and of the projections.

A.11. Proof of Example 6

If μ_Ξ has k connected components, then the indicators of those components will be orthogonal in $L^2(\mu_\Xi)$ while minimizing the invariant term $\mathbb{E}_x\,\mathbb{E}\left[\|\varphi(\xi) - \varphi(\xi')\|^2 \,\middle|\, X\right]$. As a consequence, f^* belongs to the space of the $(f_i)_{i\leq k}$.

B. Control of the downstream convergence

This section is devoted to the proof of Theorem 1. In all the following, k_λ designs the number of positive eigenvalues of T_λ (including multiplicity) as an operator on $L^2(\mu_\Xi)$. We fix $k \leq k_\lambda$, and design by $\mathcal{F}$ the span of the $(f_i)_{i\in[k]}$. In the kernel regime, the space $\mathcal{F}$ can also be written as $\mathcal{F} = \left\{w^\top\Theta_*\varphi \,\middle|\, w \in \mathbb{R}^k\right\}$ for Θ_* the minimizer defined in Lemma 12. We denote by $\hat{\mathcal{F}}$ the space defined similarly from an estimate $\hat{\Theta}$ of Θ_*.

The error on the downstream task could be decomposed into three quantities: the error on the downstream task linked with the capacity of $\hat{\mathcal{F}}$ (12); the error on the upstream task linked to approximation error between $\mathcal{F}$ and $\hat{\mathcal{F}}$ (13), the error due to the fact that the downstream task might not be effectively solved within $\mathcal{F}$ (14).

Lemma 15 (Decomposition intuition)**.** *Let $\mathcal{F}$ and $\hat{\mathcal{F}}$ be two closed convex sets of $L^2(\rho_\mathcal{X})$, and $\Pi_\mathcal{F}$ design the orthogonal projection on the space $\mathcal{F}$ according to $L^2(\rho_\mathcal{X})$ geometry. For any function $f : \mathcal{X} \to \mathcal{Y}$ in $\hat{\mathcal{F}}$, the excess of risk* (1) *can be decomposed as*

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}(f^*) \leq &\left\|f - \Pi_{\hat{\mathcal{F}}}f^*\right\|^2_{L^2(\rho_\mathcal{X})} &(12)\\ &+ 2\left\|(I - \Pi_{\hat{\mathcal{F}}})\Pi_\mathcal{F}f^*\right\|^2_{L^2(\rho_\mathcal{X})} &(13)\\ &+ \|(I - \Pi_\mathcal{F})f^*\|^2_{L^2(\rho_\mathcal{X})}\,, &(14)\end{aligned}$$

Proof. The proof of the lemma follows from classical characterization of the mean square error and a triangular inequality. Introduce the following technical assumption.

Assumption 6. *Assume $(X, Y) \to Y$ to belong to $L^2(\rho)$.*

When $\ell(y, y') = \|y - y'\|^2$, using the fact that $(X, Y) \to Y - \mathbb{E}[Y \mid X]$ is orthogonal to any measurable function that does not depend on Y in $L^2(\rho)$,

$$\mathcal{R}(f) = \mathbb{E}[\|f(X) - Y\|^2] = \mathbb{E}[\|f(X) - \mathbb{E}[Y \mid X]\|^2] + \mathbb{E}[\|\mathbb{E}[Y \mid X] - Y\|^2].$$

As a consequence, $f^*(x) = \mathbb{E}[Y \mid X = x]$ and

$$\mathcal{R}(f) - \mathcal{R}(f^*) = \|f - f^*\|^2_{L^2(\rho_{\mathcal{X}})}.$$

Let us decompose the excess of risk with the orthogonal projection of $\hat{\mathcal{F}}$, we have

$$\mathcal{R}(f) - \mathcal{R}(f^*) = \|f - f^*\|^2_{L^2(\rho_{\mathcal{X}}} = \left\|f - \Pi_{\hat{\mathcal{F}}} f^*\right\|^2_{L^2(\rho_{\mathcal{X}}} + \left\|(I - \Pi_{\hat{\mathcal{F}}}) f^*\right\|^2_{L^2(\rho_{\mathcal{X}}}$$

The second term is worked out as

$$\begin{aligned}
\left\|(I - \Pi_{\hat{\mathcal{F}}}) f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} &= \left\|(I - \Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}} f^* + (I - \Pi_{\hat{\mathcal{F}}})(I - \Pi_{\mathcal{F}}) f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} \\
&\leq 2\left\|(I - \Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}} f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} + \left\|(I - \Pi_{\hat{\mathcal{F}}})(I - \Pi_{\mathcal{F}}) f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} \\
&\leq 2\left\|(I - \Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}} f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} + \|(I - \Pi_{\mathcal{F}}) f^*\|^2_{L^2(\rho_{\mathcal{X}})}
\end{aligned}$$

where the last inequality is due to the fact that projections contract distances. □

For linear probing (3), when the downstream task is learned with n data points and a noise level ε, (12) is expected to behave as $\varepsilon^2 k/n$ (Mourtada & Rosasco, 2022). In this linear setting, (13) should be seen as a measure of angle between $\mathcal{F}$ and $\hat{\mathcal{F}}$ seen through the eyes of f^* (Davis & Kahan, 1970; Kato, 1995).

B.1. Controlling (12)

The downstream task error relates to the generalization error of mis-specified linear model. To bound it, we will use the convergence rates analysis through concentration of integral operators of Smale & Zhou (2007) and Caponnetto & De Vito (2007). It requires reworking slightly the previous decomposition.

Lemma 16 (Warm-up)**.** *Let $\hat{\mathcal{F}}$ be the span of the $(\psi_i)_{i\in[k]}$, with $S_\psi : \mathbb{R}^k \to L^2$ defined as $S_\psi w = w^\top \psi$, then*

$$\Pi_{\hat{\mathcal{F}}} f^* = S_\psi \, \mathbb{E}[\psi(X)\psi(X)^\top]^{-1} \, \mathbb{E}[Y\psi(X)]. \tag{15}$$

Based on data (X_i, Y_i), one can define the empirical risk minimizer $f_n = S_\psi w_n$, where w_n is the minimizer of

$$w_n \in \operatorname*{arg\,min}_{w\in\mathbb{R}^k} \sum_{i=1}^n \left\|w^\top \psi(X_i) - Y_i\right\|^2 = [\frac{1}{n}\sum_{i=1}^n \varphi(X_i)\varphi(X_i)^\top]^{-1} \frac{1}{n}\sum_{j=1}^n Y_i \varphi(X_i). \tag{16}$$

Proof. The two formula can be proven at once by remarking that if $\Pi_{\hat{\mathcal{F}}} f^*$ is defined as $S_\psi w$ for w minimizing

$$\mathbb{E}[\left\|w^\top \varphi(X) - Y\right\|^2] = w^\top \, \mathbb{E}[\varphi(X)\varphi(X)^\top] w - 2w^\top \, \mathbb{E}[Y\varphi(X)] + \mathbb{E}[\|Y\|^2].$$

Minimizing this quadratic form leads to the first results. The second result is proven in the same way after substituting the distribution over (X, Y) by the empirical one $n^{-1}\sum_{i\in[n]} \delta_{(X_i, Y_i)}$. □

As a consequence of this warm-up lemma, let us introduce some notations, for $\psi : \mathcal{X} \to \mathbb{R}^k$ and some data (X_i), define

$$S_\psi : \mathbb{R}^k \to L^2(\rho_{\mathcal{X}}); w \to w^\top \psi, \qquad \hat{S}_\psi : \mathbb{R}^k \to \ell^2(n); w \to (w^\top \psi(X_i))_{i\in[n]}, \tag{17}$$

where $\ell^2(n)$ is endowed with normalized (i.e. probability-like) scalar product $\langle a, b\rangle = n^{-1}\sum_{i\in[n]} a_i b_i$. Similarly to Lemma 11, one can show that the adjoint of S_ψ and $\hat{S}_\psi$, and the covariance operators are

$$S_\psi : L^2(\rho_{\mathcal{X}}) \to \mathbb{R}^k; f \to \mathbb{E}_{\rho_{\mathcal{X}}}[f(X)\psi(X)], \qquad \hat{S}_\psi : \ell^2(n) \to \mathbb{R}^k; (Y_i)_{i\in n} \to \frac{1}{n}\sum_{i\in[n]} Y_i \psi(X_i).$$

$$\Sigma_\psi = S_\psi S_\psi^\top = \mathbb{E}_{\rho_{\mathcal{X}}}[\psi(X)\psi(X)^\top], \qquad \hat{\Sigma}_\psi = \hat{S}_\psi \hat{S}_\psi^\top = \frac{1}{n}[\psi(X)\psi(X)^\top], \tag{18}$$

In this subsection, we will only consider S and Σ associated with ψ and we remove the indices for convenience. To simplify notation when $f \in L^2(\rho_{\mathcal{X}})$ we will write $\hat{S}^\top f$ for $\hat{S}^\top (f(X_i))_{i\in[n]}$.

Assumption 7 (Homoskedastic noise)**.** *There exists $\varepsilon > 0$ such that for $\rho_{\mathcal{X}}$-almost all x, the variance of $(Y \mid X = x)$ is bounded by ε^2.*

Lemma 17 (Bias-Variance decomposition)**.** *Based on data (X_i, Y_i), one can define the regularized empirical risk minimizer $f_n = S_\psi w_n$ with a regularization parameter $\gamma > 0$ as*

$$w_n \in \operatorname*{arg\,min}_{w\in\mathbb{R}^k} \sum_{i=1}^n \left\| w^\top \psi(X_i) - Y_i \right\|^2 + \gamma \left\| w \right\|^2 . \tag{19}$$

When doing so, under Assumption 7, the average excess of risk can be decomposed as, with $M = \sup \|\psi(X)\|$,

$$\begin{aligned}\mathbb{E}_{(X_i,Y_i)}[\left\| f_n - \Pi_{\hat{\mathcal{F}}} f^* \right\|^2_{L^2(\rho_{\mathcal{X}})}] \leq \frac{\varepsilon^2}{n}\left(1 + \frac{M^2}{\gamma n}\right) \operatorname{Tr}\left((\Sigma+\gamma)^{-1}\Sigma\right) \\ + 2\gamma \left(1 + \frac{M^2}{\gamma n}\right)^2 \left\langle \Pi_{\hat{\mathcal{F}}} f^*, \Sigma(\Sigma+\gamma)^{-1}\Pi_{\hat{\mathcal{F}}} f^* \right\rangle_{L^2(\rho_{\mathcal{X}})} \\ + 2\,\mathbb{E}_{(X_i)} \left\| S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top (I - \Pi_{\hat{\mathcal{F}}}) f^* \right\|^2_{L^2(\rho_{\mathcal{X}})} .\end{aligned} \tag{20}$$

Proof. Retaking the warm-up lemma, one can show that

$$w_n = (\hat{\Sigma}+\gamma)^{-1}\hat{S}(Y_i)_{i\in[n]}.$$

As a consequence, using the usual bias-variance decomposition, and the fact that $f^* = \mathbb{E}_\rho\left[Y \mid X = \cdot\right]$, we develop

$$\begin{aligned}&\mathbb{E}_{(Y_i \mid X=X_i)}[\left\| f_n - \Pi_{\hat{\mathcal{F}}} f^* \right\|^2] = \mathbb{E}_{(Y_i \mid X=X_i)}[\left\| S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top (Y_i)_{i\in[n]} - \Pi_{\hat{\mathcal{F}}} f^* \right\|^2] \\ &= \mathbb{E}_{(Y_i \mid X=X_i)}\left[\left\| S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top (Y_i - \mathbb{E}\left[Y \mid X = X_i\right])_{i\in[n]} \right\|^2\right] + \left\| S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top f^* - \Pi_{\hat{\mathcal{F}}} f^* \right\|^2 .\end{aligned}$$

The first term can be worked out with Mourtada & Rosasco (2022) techniques as

$$\mathbb{E}_{(X_i,Y_i)}\left[\left\| S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top (Y_i - \mathbb{E}\left[Y \mid X = X_i\right])_{i\in[n]} \right\|^2\right] \leq \frac{\varepsilon^2}{n}\left(1 + \frac{R^2}{\gamma n}\right) \operatorname{Tr}\left((\Sigma+\gamma)^{-1}\Sigma\right)$$

under the assumption that the variance of $(Y \mid X)$ is bounded by ε^2.

We work out the second term with

$$\left\| S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top f^* - \Pi_{\hat{\mathcal{F}}} f^* \right\| \leq \left\| S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top (f^* - \Pi_{\hat{\mathcal{F}}} f^*) \right\| + \left\| S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top \Pi_{\hat{\mathcal{F}}} f^* - \Pi_{\hat{\mathcal{F}}} f^* \right\| .$$

Once again, the last part can be worked out with techniques of Mourtada & Rosasco (2022) to get

$$\mathbb{E}_{(X_i,Y_i)}[\left\| S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top \Pi_{\hat{\mathcal{F}}} f^* - \Pi_{\hat{\mathcal{F}}} f^* \right\|^2] \leq \gamma \left(1 + \frac{R^2}{\gamma n}\right)^2 \left\langle \Pi_{\hat{\mathcal{F}}} f^*, \Sigma(\Sigma+\gamma)^{-1}\Pi_{\hat{\mathcal{F}}} f^* \right\rangle_{L^2(\rho_{\mathcal{X}})} .$$

This provides the decomposition of the lemma. □

Let us work out the last term in (20).

Lemma 18. *For $t = \left\| (\Sigma+\gamma)^{-1/2}(\Sigma - \hat{\Sigma})(\Sigma+\gamma) \right\|_{\mathrm{op}}$ and M such that $\|\psi(X)\| \leq M$ almost everywhere,*

$$\left\| S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top (I - \Pi_{\hat{\mathcal{F}}}) f^* \right\|_{L^2(\rho_{\mathcal{X}})} \leq \min\left\{\frac{1}{1-t}, 1 + t\cdot\frac{M^2+\gamma}{\gamma}\right\} \left\| \Sigma_\gamma^{-1/2}\hat{S}^\top (I - \Pi_{\hat{\mathcal{F}}}) f^* \right\| . \tag{21}$$

Proof. Let us set $f = (I - \Pi_{\hat{\mathcal{F}}})f^*$ and $A_\gamma = A + \gamma I$ for simplicity. Remark that f is orthogonal to the image of S, hence $S^\top f = 0$. We decompose the last quantity with

$$\begin{aligned}
\hat{\Sigma}_\gamma^{-1}\hat{S}^\top f &= (\hat{\Sigma}_\gamma^{-1}\hat{-}\Sigma_\gamma^{-1})\hat{S}^\top f + (\Sigma_\gamma)^{-1}\hat{S}^\top f \\
&= \hat{\Sigma}_\gamma^{-1}\hat{(}\Sigma_\gamma - \hat{\Sigma}_\gamma)\Sigma_\gamma^{-1}S^\top f + \Sigma_\gamma^{-1}\hat{S}^\top f \\
&= \hat{\Sigma}_\gamma^{-1}\hat{(}\Sigma - \hat{\Sigma})\Sigma_\gamma^{-1}\hat{S}^\top f + \Sigma_\gamma^{-1}\hat{S}^\top f \\
&= \Sigma_\gamma^{-1/2}\left(\Sigma_\gamma^{1/2}\hat{\Sigma}_\gamma^{-1}\Sigma_\gamma^{1/2}\Sigma_\gamma^{-1/2}\hat{(}\Sigma - \hat{\Sigma})\Sigma_\gamma^{-1/2} + I\right)\Sigma_\gamma^{-1/2}\hat{S}^\top f
\end{aligned}$$

Using the fact that S is isometric to $\Sigma^{1/2}$ which itself if smaller than $\Sigma_\gamma^{1/2}$ (with the Loewner order), we have

$$\left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top(I - \Pi_{\hat{\mathcal{F}}})f^*\right\|_{L^2(\rho_{\mathcal{X}})} \leq \left(1 + \left\|\Sigma_\gamma^{1/2}\hat{\Sigma}_\gamma^{-1}\Sigma_\gamma^{1/2}\right\|_{\text{op}}\left\|\Sigma_\gamma^{-1/2}\hat{(}\Sigma - \hat{\Sigma})\Sigma_\gamma^{-1/2}\right\|_{\text{op}}\right)\left\|\Sigma_\gamma^{-1/2}\hat{S}^\top f\right\|$$

We know that

$$\left\|\Sigma_\gamma^{1/2}\hat{\Sigma}_\gamma^{-1}\Sigma_\gamma^{1/2}\right\|_{\text{op}} \leq \gamma^{-1}(\|\Sigma\|_{\text{op}} + \gamma) \leq \gamma^{-1}(\sup_{x\in\text{supp}\,\rho_{\mathcal{X}}}\|\psi(x)\|^2 + \gamma)$$

We also have that for A and $\hat{A}$ self adjoint and any $t > 0$, the sequence of implications

$$\begin{aligned}
\left\|A^{-1/2}(A - \hat{A})A^{-1/2}\right\|_{\text{op}} \leq t \quad &\Leftrightarrow \quad -tI \preceq A^{-1/2}(\hat{A} - A)A^{-1/2} \preceq tI \\
&\Leftrightarrow \quad -tA \preceq \hat{A} - A \preceq tA \\
&\Leftrightarrow \quad (1-t)A \preceq \hat{A} \preceq (1+t)A \\
&\Leftrightarrow \quad (1+t)^{-1}A^{-1} \preceq \hat{A}^{-1} \preceq (1-t)^{-1}A^{-1} \\
&\Leftrightarrow \quad (1+t)^{-1} \preceq A^{1/2}\hat{A}^{-1}A^{1/2} \preceq (1-t)^{-1}.
\end{aligned}$$

Combining the different results leads to the lemma. □

Probabilistic arguments will show that t, as well as $\left\|(\Sigma + \gamma^{-1/2}\hat{S}^\top(I - \Pi_{\hat{\mathcal{F}}})f^*\right\|$, vanishes to zero in $n^{-1/2}$. We will use Bernstein concentration inequality.

Lemma 19 (Bernstein concentration inequalities)**.** *Let denote by $\mathcal{A}$ a Hilbert space and by $(Z_i)_{i\in[n]}$ a sequence of independent random vectors on $\mathcal{A}$ such that $\mathbb{E}[Z_i] = 0$, and such that there exists two positive constants M and σ such that for all $m > 2$*

$$\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|Z_i\|^m] \leq m!\sigma^2 M^{m-2}/2$$

For any $t > 0$,

$$\mathbb{P}(\|\frac{1}{n}\sum_{i=1}^n Z_i\| \geq t) \leq 2\exp\left(\frac{-nt^2}{2\sigma^2 + 2tM}\right).$$

In particular when the (Z_i) are bounded by $3M$, and $\sigma^2 = n^{-1}\sum_{i\in n}\mathbb{E}[Z_i^2]$, the condition holds. When, instead, Z_i are symmetric matrices in $\mathbb{R}^{k\times k}$ and $\|\cdot\|$ is the operator norm, the same bound holds with $k\exp(\cdots)$ instead of $2\exp(\cdots)$ on the right-hand side, where $\sigma^2 = \left\|n^{-1}\sum_{i\in[n]}\mathbb{E}[Z_i^2]\right\|$.

Proof. See Corollary 1 in Pinelis & Sakhanenko (1986) for the first part, and Tropp (2015) for the matrix version. □

Lemma 20. *For any $t > 0$, the vector part in last term of the bias decomposition (20) can be controlled with*

$$\mathbb{P}\left(\left\|\Sigma_\gamma^{-1/2}\hat{S}^\top(I - \Pi_{\hat{\mathcal{F}}})f^*\right\| \geq t\right) \leq 2\exp\left(\frac{-nt^2}{a(b + 2M\gamma^{-1/2}t/3)}\right) \tag{22}$$

where $b = 2\operatorname{Tr}(\Sigma+\gamma)^{-1}\Sigma$, $M = \sup\|\psi(X)\|$ *and* $a = \|f^*\|_{L^\infty} + M\|f^*\|_{L^2}$. *Moreover, this vector part is bounded by* $\gamma^{-1}a^2M^2$. *The matrix part in the last term of* (20) *is controlled with*

$$\mathbb{P}\left(\left\|\Sigma_\gamma^{-1/2}(\hat\Sigma-\Sigma)\Sigma_\gamma^{-1/2}\right\|_{\mathrm{op}} \geq t\right) \leq k\exp\left(\frac{-nt^2}{2M^2\gamma^{-1}(1+t/3)}\right) \tag{23}$$

Moreover, this matrix part is bounded by $\gamma^{-2}M^4$.

Proof. Let us introduce

$$Z_i = (I-\Pi_{\hat{\mathcal{F}}})f^*(X_i)(\Sigma+\gamma)^{-1/2}\psi(X_i) \in \mathbb{R}^k. \tag{24}$$

One can check that

$$\frac{1}{n}\sum_{i\in[n]} Z_i = (\Sigma+\gamma)^{-1/2}\frac{1}{n}\sum_{i\in[n]}(I-\Pi_{\hat{\mathcal{F}}})f^*(X_i)\psi(X_i) = (\Sigma+\gamma)^{-1/2}\hat{S}^\top(I-\Pi_{\hat{\mathcal{F}}})f^*,$$

as well as, since $\operatorname{im} S = \hat{\mathcal{F}}$

$$\mathbb{E}[Z_i] = S^\top(I-\Pi_{\hat{\mathcal{F}}})f^* = 0.$$

Moreover,

$$\|Z_i\| = \left\|(\Sigma+\gamma)^{-1/2}\psi(X_i)\right\| \left|(I-\Pi_{\hat{\mathcal{F}}})f^*(X_i)\right| \leq \gamma^{-1/2}M(\|f^*\|_{L^\infty} + M\|f^*\|_{L^2}).$$

where $R = \sup_X\|\psi(X)\|$ and we have used the fact that

$$\begin{aligned}\left|(I-\Pi_{\hat{\mathcal{F}}})f^*(X_i)\right| &\leq |f^*(X_i)| + \left|\Pi_{\hat{\mathcal{F}}}f^*(X_i)\right| = |f^*(X_i)| + \left|\left\langle SS^{-1}\Pi_{\hat{\mathcal{F}}}f^*, \varphi(X)\right\rangle\right| \\ &\leq |f^*(X_i)| + \left\|SS^{-1}\right\|_{\mathrm{op}}\left\|\Pi_{\hat{\mathcal{F}}}f^*\right\|_{L^2}\|\varphi(X_i)\| \leq \|f^*\|_{L^\infty} + M\|f^*\|_{L^2}.\end{aligned}$$

Finally, we have

$$\begin{aligned}\mathbb{E}[\|Z_i\|^2] &= \mathbb{E}\left[\left\|(\Sigma+\gamma)^{-1/2}\psi(X_i)\right\|^2\left|(I-\Pi_{\hat{\mathcal{F}}})f^*(X_i)\right|^2\right] \\ &\leq \mathbb{E}\left[\left\|(\Sigma+\gamma)^{-1/2}\psi(X_i)\right\|^2\right](\|f^*\|_{L^\infty} + M\|f^*\|_{L^2}) \\ &= \operatorname{Tr}\left((\Sigma+\gamma)^{-1}\Sigma\right)(\|f^*\|_{L^\infty} + M\|f^*\|_{L^2}).\end{aligned}$$

Using Bernstein inequality leads to the control on the vector term.

For the matrix term, let us introduce

$$Z_i = U_iU_i^\top - \mathbb{E}[U_iU_i^\top], \qquad U_i = (\Sigma+\gamma)^{-1/2}\varphi(X_i).$$

We have $(\Sigma+\gamma)^{-1/2}(\hat\Sigma-\Sigma)(\Sigma+\gamma)^{-1/2} = \frac{1}{n}\sum_{i\in[n]} Z_i$, and

$$\sup\|Z\| \leq \sup\|U\|^2 \leq \gamma^{-1}M^2.$$

Finally, using the fact that $\|U_i\|^2 \preceq U_i\sup\|U_i\|$, with the variational definition of the mean, with the infimum taken with respect to the Loewner order

$$\mathbb{E}[Z_i^2] = \inf_a \mathbb{E}[(Z_i-a)^2] \preceq \mathbb{E}[(U_iU_i^\top)^2] \preceq \sup\|U\|^2\,\mathbb{E}[U_i^\top U_i] = \sup\|U\|^2(\Sigma+\gamma)^{-1}\Sigma \preceq \sup\|U\|^2 I$$

Applying the matrix version of Bernstein inequality leads to the lemma. □

We now turn the deviation inequalities of the last lemma into a bound on the average.

Lemma 21. *Retaking the notation of the previous lemma.*

$$\begin{aligned}\mathbb{E}_{(X_i)}\left\|S(\hat\Sigma+\gamma)^{-1}\hat{S}^\top(I-\Pi_{\hat{\mathcal{F}}})f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} &\leq k\exp\left(\frac{-3n\gamma}{(3+\sqrt{2})M^2}\right)(\gamma^{-4}M^6a^2(M^2+2\gamma))^2 \\ &\quad + \frac{16ab}{n} + \frac{512a^2M^2}{9\gamma n^2}.\end{aligned}$$

Proof. In essence, we have two random variables, $X = \left\|\Sigma_\gamma^{-1/2}(\hat{\Sigma}-\Sigma)\Sigma_\gamma^{-1/2}\right\|_{\text{op}}^2$, and $Y = \left\|\Sigma_\gamma^{-1/2}\hat{S}^\top(I-\Pi_{\hat{\mathcal{F}}})f^*\right\|^2$ the vector one. We proceed with computation using the fact that for X positive $\mathbb{E}[X] = \int_{t>0}\mathbb{P}(X>t)\,\mathrm{d}t$ and that $ab>t$, implies for any s that $a>1+s$ or $b>t/(1+s)$,

$$\begin{aligned}
&\mathbb{E}[\min\left\{\frac{1}{1-X}, 1+\frac{(M^2+\gamma)X}{\gamma}\right\}^2 Y^2] \\
&= \int_{t\in(0,\sup(1+\gamma^{-1}(M^2+\gamma)X)^2Y^2)} \mathbb{P}\left(\min\left\{\frac{1}{1-X}, 1+\frac{(M^2+\gamma)X}{\gamma}\right\}^2 Y^2 > t\right)\mathrm{d}t \\
&\leq \int \inf_s \mathbb{P}\left(\min\left\{\frac{1}{1-X}, 1+\frac{(M^2+\gamma)X}{\gamma}\right\}^2 > 1+s\right) + \mathbb{P}(Y^2 > t/(1+s))\,\mathrm{d}t.
\end{aligned}$$

Rather than solving this in closed form, we will proceed with a much simpler bound that consists in taking $s=1$ without any optimization. It gives the much simpler formula

$$\mathbb{E}[\min\left\{\frac{1}{1-X}, 1+\frac{(M^2+\gamma)X}{\gamma}\right\}^2 Y^2] \leq \mathbb{P}(\frac{1}{(1-X)^2} > 2)\sup(1+\gamma^{-1}(M^2+\gamma)X)^2Y^2 + 2\,\mathbb{E}[Y^2].$$

For Y we can use the same technique as before, using that $\exp(-(a+b)^{-1}) \leq \exp(-\max(2a,2b)^{-1}) \leq \exp(-(2a)^{-1}) + \exp(-(2b)^{-1})$, we get

$$\begin{aligned}
\mathbb{E}[Y^2] &= \int_{t>0}\mathbb{P}(Y^2>t)\,\mathrm{d}t \leq \int_{t>0} 2\exp\left(-\frac{-nt}{a(b+2M\gamma^{-1/2}t^{1/2}/3)}\right)\mathrm{d}t \\
&\leq 4\int_{t>0}\exp\left(-\frac{-nt}{2ab}\right) + \exp\left(-\frac{-nt^{1/2}}{4aM\gamma^{-1/2}/3)}\right)\mathrm{d}t \\
&= 8abn^{-1} + 256a^2M^2\gamma^{-1}n^{-2}/9
\end{aligned}$$

We conclude the lemma with the previous one. □

Let us now simplify the constant that appear in the bound derived so far.

Lemma 22 (Simplifying constants)**.** *The constant is the previous bound can be worked out as*

$$\mathrm{Tr}\left(\Sigma(\Sigma+\gamma)^{-1}\right) \leq k, \qquad M \leq \lambda^{-1}k\sup\|\varphi\|, \qquad \left\langle \Pi_{\hat{\mathcal{F}}}f^*, \Sigma(\Sigma+\gamma)^{-1}\Pi_{\hat{\mathcal{F}}}f^*\right\rangle_{L^2(\rho_{\mathcal{X}})} \leq \|f^*\|_{L^2(\rho_{\mathcal{X}})}.$$

We also have

$$\|f^*\|_{L^2(\rho_{\mathcal{X}})} \leq \|f^*\|_{L^\infty(\rho_{\mathcal{X}})} \leq \sigma, \qquad \varepsilon^2 \leq \sigma^2, \quad \sigma^2 = \sup_x \mathbb{E}\left[Y^2 \,\middle|\, X=x\right]$$

As a consequence, the constant a appearing earlier is smaller than $(1+M)\sigma$.

Proof. The first bound is a direct application of the fact that $\Sigma \preceq \Sigma+\lambda$, hence $\mathrm{Tr}((\Sigma+\gamma)^{-1}\gamma) \leq \mathrm{Tr}(I) = k$. The second bound is due to the fact that $\psi = \hat{\Theta}\varphi$, hence $\|\psi\| \leq \left\|\hat{\Theta}\right\|_{\text{op}}\|\varphi\| \leq \left\|\hat{\Theta}\right\|_F\|\varphi\|$. In the meantime, if $\hat{\Theta}$ was regularized

$$\lambda\left\|\hat{\Theta}\right\|_F^2 \leq \hat{\mathcal{L}}(\Theta) + \lambda\left\|\hat{\Theta}\right\|^2 \leq \hat{\mathcal{L}}(0) = k.$$

For the part in f^*, we have that

$$\left\|\Sigma^{1/2}(\Sigma+\gamma)^{-1/2}\Pi_{\hat{\mathcal{F}}}f^*\right\| \leq \left\|\Pi_{\hat{\mathcal{F}}}f^*\right\| \leq \|f^*\|.$$

Finally, the last equality is due to the fact that $f^*(x)$ is the mean of Y conditionally to $X=x$,

$$\|f^*(X)\| = \|\mathbb{E}\left[Y \mid X\right]\| \leq \mathbb{E}\left[Y^2 \,\middle|\, X\right]^{1/2} \leq \sigma.$$

This ends the lemma □

Lemma 23. *Under Assumption 7, when $\gamma = M^2 \log(n)^{1+\delta} n^{-1}$, with $\delta > 0$, there exists a $N > 0$ such that for any $n > N$, the excess of risk of the regularized empirical risk (19) minimizer reads*

$$\mathbb{E}_{(X_i,Y_i)}[\mathcal{R}(f_n) - \mathcal{R}(f^*)] \leq \frac{2k_e\varepsilon^2}{n} + \frac{8M^2\log(n)^{1+\delta}}{n}\|f^*\|_{L^2(\rho_{\mathcal{X}})} + \frac{64ka}{n} + 2\left\|I - \Pi_{\hat{\mathcal{F}}}\Pi_{\mathcal{F}}f^*\right\|^2 + \|I - \Pi_{\mathcal{F}}f^*\|^2 \quad (25)$$

where $k_e = \operatorname{Tr}\left(\Sigma(\Sigma+\gamma I)^{-1}\right) \leq k$ is the effective dimension, $a = \left\|I - \Pi_{\hat{\mathcal{F}}}f^*\right\|_{L^\infty} \leq \|f^*\|_{L^\infty} + M\|f\|_{L^2}$, and $M = \sup\|\psi\| \leq k\lambda^{-1}\sup\|\varphi\|$.

Proof. When $\gamma = c\log(n)^{1+\delta}n^{-1}$ the excess of risk reads

$$\begin{aligned}\mathbb{E}_{(X_i,Y_i)}[\mathcal{R}(f_n) - \mathcal{R}(f^*)] \leq \frac{k_e\varepsilon^2}{n}\left(1 + \frac{M^2}{c\log(n)}\right) + \frac{2c\log(n)^{1+\delta}}{n}\left(1 + \frac{M^2}{c\log(n)}\right)^2\|f^*\|_{L^2(\rho_{\mathcal{X}})} + \frac{64ka}{n} \\ + \frac{114a^2M^2}{9cn\log(n)} + O(\exp(-\log(n)^{1+\delta/2})) + 2\left\|I - \Pi_{\hat{\mathcal{F}}}\Pi_{\mathcal{F}}f^*\right\|^2 + \|I - \Pi_{\mathcal{F}}f^*\|^2\end{aligned}$$

Taking $c = M^2$ leads to the lemma. □

B.2. Controlling (13)

An ideal control of (13) would leverage closed form solutions to both the population and empirical risk and use concentration inequalities on integral operators, as in Cabannes et al. (2021b); Pillaud-Vivien & Bach (2023). Yet, those proof proceed with the estimation of the smallest eigenfunctions of $(\hat{\Sigma}_X + \lambda)^{-1/2}\Sigma(\hat{\Sigma}_X + \lambda)^{-1/2}$, rather than the biggest of $\Sigma^{-1/2}(\Sigma_X - \lambda)\Sigma^{-1/2}$. In this proof, we will rather utilize derivations based on empirical processes concentration, together with the following "transfer bound".

Lemma 24 (Transfer bound)**.** *For $\hat{\Theta} \in \mathbb{R}^k \otimes \mathcal{H}$, and $\hat{\mathcal{F}} = \left\{x \to w^\top\hat{\Theta}\varphi(x) \,\middle|\, w \in \mathbb{R}^k\right\}$,*

$$\sum_{i\in[k]}\lambda_i^2\left\|(\Pi_{\mathcal{F}}^{(\mu_\Xi)} - \Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})f_i\right\|_{L^2(\mu_\Xi)}^2 - \sum_{k<i\leq k_\lambda}\lambda_i^2\left\|\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)}f_i\right\|_{L^2(\mu_\Xi)}^2 \leq \mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta_*;\lambda), \quad (26)$$

where $\Pi_{\mathcal{F}}^{(\tau)}$ is the projection orthogonal on $\mathcal{F}$ in $L^2(\tau)$.

Proof. For simplicity, let us remove the dependency to μ_Ξ in the proof. Let us introduce $C = S\hat{\Theta}\hat{\Theta}^\top S^\top$, C is a positive operator of rank k in L^2, let us write it as $C = \sum_{i\in[k]}\mu_i g_i g_i^\top$ with $\mu_i \geq 0$.

$$\mathcal{L}(\hat{\Theta};\lambda) - k = \operatorname{Tr}\left((C - T_\lambda)^2 - T_\lambda^2\right) = \operatorname{Tr}\left(C^2 - 2C^{1/2}T_\lambda C^{1/2}\right).$$

Let us decompose $T_\lambda = T_+ - T_-$ where T_+ and T_- are positive. Since $T_\lambda \preceq T_+$, $-C^{1/2}T_\lambda C^{1/2} \succeq C^{1/2}T_+C^{1/2}$, hence

$$\mathcal{L}(\hat{\Theta};\lambda) - k \geq \operatorname{Tr}\left(C^2 - 2C^{1/2}T_+C^{1/2}\right) = \sum_{i\leq k}\mu_i^2 - 2\mu_i\left\|T_+^{1/2}g_i\right\|^2.$$

Minimizing this quantity with respect to μ_i, leads to

$$\mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta;\lambda) \geq \sum_{i\leq k}\lambda_i^2 - \sum_{i\leq k}\left\|T_+^{1/2}g_i\right\|^4.$$

Let us know introduce (f_i) the eigenfunctions of T_λ. With $U = (\langle g_i, f_j\rangle^2)_{ij} \in \mathbb{R}^{k\times k_\lambda}$ and $\lambda = (\lambda_i) \in \mathbb{R}^{k_\lambda}$, we have

$$\sum_{i\leq k}\left\|T_+^{1/2}g_i\right\|^4 = \sum_{i\leq k}(g_i^\top T_+ g_i)^2 = \sum_{i\leq k}\left(\sum_{j\leq k_\lambda}\lambda_j\langle g_i, f_j\rangle^2\right)^2 = \sum_{j,m\leq k_\lambda}\lambda_j\lambda_m\sum_{i\leq k}\langle g_i, f_j\rangle^2\langle g_i, f_m\rangle^2 = \lambda^\top U^\top U\lambda.$$

Note that U is at most doubly stochastic since both (g_i) and (f_i) are orthonormal families, thus $\|U\| \leq 1$, and $U^\top U \preceq I$. If one replace the f_i by $f_i / \left\|\Pi_{\hat{\mathcal{F}}}\right\|$ in the definition of U that would become $\tilde{U} = \operatorname{diag}((\left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2)_{i \leq k_\lambda})^{-1} U$, $\tilde{U}$ is still right stochastic. Hence

$$U^\top U \preceq \operatorname{diag}(\left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2_{i \leq k_\lambda})^2 \operatorname{diag}(\left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2_{i \leq k_\lambda}).$$

It follows that

$$\sum_{i \leq k} \left\|T_+^{1/2} g_i\right\|^4 \leq \lambda^\top U^\top U \lambda = \sum_{i \leq k_\lambda} \lambda_i^2 \left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2.$$

This allows to simplify the lower bound as

$$\begin{aligned}
\mathcal{L}(\hat{\Theta}; \lambda) - \mathcal{L}(\Theta; \lambda) &\geq \sum_{i \leq k} \lambda_i^2 \left(f_i^\top f_i - f_i^\top \Pi_{\hat{\mathcal{F}}} f_i\right) - \sum_{k < i \leq k_\lambda} \lambda_i^2 f_i^\top \Pi_{\hat{\mathcal{F}}} f_i \\
&= \sum_{i \leq k} \lambda_i^2 \left\langle f_i, (I - \Pi_{\hat{\mathcal{F}}}) f_i \right\rangle - \sum_{k < i \leq k_\lambda} \lambda_i^2 \left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2 \\
&= \sum_{i \leq k} \lambda_i^2 \left\|(\Pi_{\mathcal{F}} - \Pi_{\hat{\mathcal{F}}}) f_i\right\|^2 - \sum_{k < i \leq k_\lambda} \lambda_i^2 \left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2.
\end{aligned}$$

This ends the proof of our transfer bound. □

The left-hand side in Lemma 24 is to be linked with the desired control of (13). In order to deal more finely with distribution-shift, we introduce the following generic variant of Assumptions 1 and 2.

Assumption 8 (Low expansion)**.** *Assume that for any function of the original space of functions $f \in \Psi$ (4),*

$$\|f\|_{L^2(\rho_{\mathcal{X}})} \leq \zeta\left(\|f\|_{L^2(\mu_{\mathcal{X}})}\right),$$

with $\zeta : \mathbb{R} \to \mathbb{R}$ continuous, increasing and $\zeta(0) = 0$.

Definition 25 (Distribution ε-robustness)**.** *A close convex set of functions $\mathcal{F}$ will be said to be ε-robust to distribution shift conditionally to the function f if*

$$\left\|\Pi_{\mathcal{F}}^{(\rho_{\mathcal{X}})} f - \Pi_{\mathcal{F}}^{(\mu_{\Xi})} f\right\|_{L^2(\rho_{\mathcal{X}})} \leq \varepsilon \|f\|_{L^2(\rho_{\mathcal{X}})},$$

where $\Pi_{\mathcal{F}}^{(\tau)}$ is the projection orthogonal on $\mathcal{F}$ in $L^2(\tau)$.

Assumption 9. *There exists a profile $\sigma : \mathbb{R}^2 \to \mathbb{R}$ increasing and bounded such that for any $k \in \mathbb{N}$, $\operatorname{Span}\{f_i\}_{i \in [k]}$ is $\sigma(k)$-robust to f^*.*

Lemma 26 (Decomposition)**.** *Under Assumptions 8 and 9, with $\mathcal{F}_l$ the span of the $(f_i)_{i \in [l]}$*

$$\left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})}) \Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} f^*\right\|_{L^2(\rho_{\mathcal{X}})} \leq \sigma(l) + \zeta\left(\sum_{i \leq l} \left|\langle f^*, f_i \rangle_{L^2(\mu_{\Xi})}\right| \left\|(\Pi_{\mathcal{F}_l}^{(\mu_{\Xi})} - \Pi_{\hat{\mathcal{F}}}^{(\mu_{\Xi})}) f_i\right\|_{L^2(\mu_{\Xi})}\right). \tag{27}$$

Proof. Using the fact that $I - \Pi$ is a projection when Π is a projection, and that projections contract distance, we get

$$\begin{aligned}
\left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})}) \Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} f^*\right\|_{L^2(\rho_{\mathcal{X}})} &\leq \left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})})(\Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} - \Pi_{\mathcal{F}_l}^{(\mu_{\Xi})}) f^*\right\|_{L^2(\rho_{\mathcal{X}})} + \left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})}) \Pi_{\mathcal{F}_l}^{(\mu_{\Xi})} f^*\right\|_{L^2(\rho_{\mathcal{X}})} \\
&\leq \left\|(\Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} - \Pi_{\mathcal{F}_l}^{(\mu_{\Xi})}) f^*\right\|_{L^2(\rho_{\mathcal{X}})} + \left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})}) \Pi_{\mathcal{F}_l}^{(\mu_{\Xi})} f^*\right\|_{L^2(\rho_{\mathcal{X}})}.
\end{aligned}$$

Under Assumption 9, the first term in the right-hand side of the previous equation is bounded by $\sigma(l)$. Regarding the second term, under Assumption 8, for $f \in \Psi$ and $f' \in \hat{\mathcal{F}} \subset \Psi$, we have

$$\left\|(I - \Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})}) f\right\|_{L^2(\rho_{\mathcal{X}})} \leq \|f - f'\|_{L^2(\rho_{\mathcal{X}})} \leq \zeta\left(\|(f - f'\|_{L^2(\mu_{\Xi})}\right).$$

Taking the minimum on the right-hand side and using the fact that ζ is increasing leads to

$$\left\|(I-\Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})})f\right\|_{L^2(\rho_{\mathcal{X}})} \leq \zeta\left(\left\|(I-\Pi_{\mathcal{F}_l}^{(\mu_\Xi)})f\right\|_{L^2(\mu_\Xi)}\right).$$

Applied to $\Pi_{\mathcal{F}_l}^{(\mu_\Xi)} f^*$, this leads to

$$\left\|(I-\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})\Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} f^*\right\|_{L^2(\rho_{\mathcal{X}})} \leq \zeta\left(\left\|(I-\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})\Pi_{\mathcal{F}_l}^{(\mu_\Xi)} f^*\right\|_{L^2(\mu_\Xi)}\right).$$

We are done with all the quantities that relate to the distribution shift. Under Assumption 3, we have

$$\begin{aligned}\left\|(I-\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})\Pi_{\mathcal{F}_l}^{(\mu_\Xi)} f^*\right\|_{L^2(\mu_\Xi)} &= \left\|\sum_{i\leq l}\langle f^*, f_i\rangle_{\mu_\Xi}(I-\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})f_i\right\|_{L^2(\mu_\Xi)} \\ &\leq \sum_{i\leq l}\left|\langle f^*, f_i\rangle_{\mu_\Xi}\right|\left\|(I-\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})f_i\right\|_{L^2(\mu_\Xi)}.\end{aligned}$$

Collecting the previous equations leads to the lemma. □

To get a finer control of (13), remark that the left-hand side of (26) has some additional constraints that can help us to tighten our bound. For simplicity, we will remove all the dependency to μ_Ξ in the following. In essence, we want to lower bound the λ_i^2 and to upper bound the $\left\|(\Pi_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})f_i\right\|$. The next lemma adds a constraint the maximal error one can make on (13) under a constraint on $\mathcal{L}(\Theta;\lambda)$.

Lemma 27. *When $\mathcal{F}$ is of dimension k and $\hat{\mathcal{F}}$ is of dimension k' we have*

$$\sum_{i\leq k}\left\|(\Pi_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})f_i\right\|^2 = k-k'+\sum_{i>k}\left\|\Pi_{\hat{\mathcal{F}}}f_i\right\|^2 \leq k. \tag{28}$$

Proof. Let us consider two projection U and V onto the span of $(u_i)_{i\in[k]}$ and $(v_i)_{i\in[k]}$ with $(u_i)_{i\in\mathbb{N}}$ and $(v_i)_{i\in\mathbb{N}}$ two orthonormal basis of the ambient space. We have, with Hilbert-Schmidt norm everywhere,

$$\|U(I-V)\|^2 = \|U\|^2 - \|UV\|^2 = k - \|UV\|^2 = k - \left\|(UV)^\top\right\|^2 = k-k'+k'-\|VU\|^2 = k-k'+\|V(I-U)\|^2.$$

Based on invariant of the Hilbert-Schmidt norm to adjoint, and the fact that projection are self-adjoint, we have

$$\|U(I-V)\|^2 = \|(I-V)U\|^2 = k-k'+\|V(I-U)\|^2 = k-k'+\|(I-U)V\|^2.$$

Finally, we also know that since projection contracts distances $\|(I-V)U\|^2 \leq \|U\|^2 = k$. The claim of the lemma consists in writing explicitly

$$\begin{aligned}\left\|(I-\Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}}\right\|^2 &= \left\|(\Pi_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}}\right\|^2 = \sum_{i\leq k}\left\|(\Pi_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})f_i\right\|^2 \\ &= k-k'+\left\|\Pi_{\hat{\mathcal{F}}}(I-\Pi_{\mathcal{F}})\right\|^2 = k-k'+\sum_{i>k}\left\|\Pi_{\hat{\mathcal{F}}}f_i\right\|^2 \leq k.\end{aligned}$$

This is lead to the statement of the lemma. □

Given a control on (2), finding an upper bound on (13) reduces to a purely algebraic one. In order to find the worse value that $\sum_{i\leq k}|\langle f^*, f_i\rangle|\left\|(\Pi_{\mathcal{F}}^{(\mu_\Xi)}-\Pi_{\hat{\mathcal{F}}})f_i\right\|_{L^2(\mu_\Xi)}$ can take, let us introduce

$$x_i = \left\|(\Pi_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})f_i\right\|, \quad c_i = |\langle f^*, f_i\rangle|. \tag{29}$$

The previous results lead to the following maximization problem in order to find the worse value of (13),

$$\max_x \sum_{i\leq k} c_i x_i \tag{30}$$

$$\text{subject to} \quad \sum_{i\leq k} \lambda_i^2 x_i^2 - \sum_{k<i\leq k_\lambda} \lambda_i^2 x_i^2 \leq \varepsilon \tag{Lemma 24}$$

$$\sum_{i\leq k} x_i^2 = k - k' + \sum_{k<i\leq k_\lambda} x_i^2 \leq k \tag{Lemma 27}$$

B.3. Keeping it simple and concluding after controlling (14)

Solving smartly the algebraic problem above to get the best bound on (13) requires distinguishing between many cases. While it might be relevant to distinguish those different cases and show different convergence regimes, this subsection proceed in a simpler way, although less tight. In particular, we can simplify the problem with respect to the $(x_i)_{ik}$, using the fact that $k' \leq k$ (it is minimum between the number of positive eigenvalues of $\hat{T}_\lambda$ based on samples and k), it leads to $x_{k+1}^2 = \sum_{i\leq k} x_k^2$ and $x_{k+1+j}^2 = 0$, (30) becomes

$$\max_x \sum_{i\leq k} c_i x_i \tag{30}$$

$$\text{subject to} \quad \sum_{i\leq k} (\lambda_i^2 - \lambda_{k+1}^2) x_i^2 \leq \varepsilon$$

In general, one could refine this formulation by introducing a probability argument that tells us how much one can expect the error between $\Pi_{\hat{\mathcal{F}}}$ and $\Pi_{\mathcal{F}}$ to concentrates on the eigenspace linked to the smallest eigenvalue of T_λ^2. The problem shows two behaviors, if the c_i decrease faster than the λ_i than we want to charge the energy of $(x_i)_{i\leq k}$ on the smallest indices. Otherwise, we want to charge the $(x_i)_{i\leq k}$ on the biggest indices.

To keep it simple, we will optimize $\mathcal{L}$ without any rank restriction first, which allow considering $\lambda_{k_\lambda+1} = 0$, before thresholding the rank to get to a space of dimension k.

Lemma 28. *Under Assumptions 8 and 9, with $\mathcal{F}_l$ the span of the first l eigenfunctions of T_λ,*

$$\begin{aligned}&\left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_\mathcal{X})}) f^*\right\|_{L^2(\rho_\mathcal{X})}^2 \\ &\quad \leq \inf_{l\leq k} \left\|(I - \Pi_{\mathcal{F}_l}^{(\rho_\mathcal{X})}) f^*\right\|_{L^2(\rho_\mathcal{X})}^2 + 4\sigma(l)^2 + 4\zeta^2 \left(\left\|\tilde{T_\lambda}^{-1} \Pi_{\mathcal{F}_l}^{(\mu_\Xi)} f^*\right\|_{L^2(\mu_\Xi)} \left(\mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta;\lambda)\right)^{1/2} \right).\end{aligned} \tag{31}$$

where $\tilde{T}_\lambda = \sum_{i\in[k]} (\lambda_i^2 - \lambda_{k+1}^2)^{1/2} f_i f_i^\top$. Moreover, when the search for $\hat{\mathcal{F}}$ is done without rank restriction on Θ, before thresholding to get reduce $\hat{\mathcal{F}}$ to a space of dimension k, under the strong Assumptions 1 and 2, as well as Assumption 3

$$\left\|(I - \Pi_{\hat{\mathcal{F}}_k}) f^*\right\|^2 \leq |k - k_\lambda| \left\|f^*\right\|_{L^2(\rho_\mathcal{X})}^2 + 2c_r \left\|T_\lambda^{-1} f^*\right\|_{L^2(\mu_\Xi)}^2 \left\{\mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta_*;\lambda)\right\}. \tag{32}$$

Proof. Keeping the algebraic notation above, this comes from a simple application of Cauchy-Schwarz, for $(a_i) \in \mathbb{R}^k$

$$\sum_{i\leq l} c_i x_i = \sum_{i\in[l]} \frac{c_i}{a_i} a_i x_i \leq \left(\sum_{i\leq[l]} \frac{c_i^2}{a_i^2}\right)^{1/2} \left(\sum_{i\in[l]} a_i^2 x_i^2\right)^{1/2}.$$

When applies to the quantities in (29) and $a_i = \lambda_i^2 - \lambda_{k+1}^2$ and $l \leq k$, the previous lemma leads to

$$\begin{aligned}
\left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})})\Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} f^*\right\|_{L^2(\rho_{\mathcal{X}})} &\leq \sigma(l) + \zeta\left(\sum_{i\leq l}\left|\langle f^*, f_i\rangle_{L^2(\mu_\Xi)}\right| \left\|(\Pi_{\mathcal{F}_l}^{(\mu_\Xi)} - \Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})f_i\right\|_{L^2(\mu_\Xi)}\right) \\
&\leq \sigma(l) + \zeta\left(\sum_{i\leq l} c_i x_i\right) \leq \sigma(l) + \zeta\left(\left(\sum_{i\leq l}\frac{c_i^2}{a_i^2}\right)^{1/2}\left(\sum_{i\in[l]} a_i^2 x_i^2\right)^{1/2}\right) \\
&\leq \sigma(l) + \zeta\left(\left(\sum_{i\leq l}\frac{c_i^2}{a_i^2}\right)^{1/2}\left(\mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta;\lambda)\right)^{1/2}\right).
\end{aligned}$$

We conclude by remarking that $\sum_{i\leq l}\frac{c_i^2}{a_i^2} = \left\|\tilde{T}_\lambda^{-1}\Pi_{\mathcal{F}_l^{(\mu_\Xi)}} f^*\right\|_{L^2(\mu_\Xi)}$.

For the second part, set $\hat{\mathcal{F}}_k$ the k first eigenfunctions to the all the one retrieve with the empirical minimization of $\mathcal{L}$, and $\mathcal{F}$ to be the span of all the eigenfunctions linked with positive eigenvalues of T_λ. Let us rework the decomposition of the excess of risk, we have

$$\begin{aligned}
\left\|(I - \Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2 &= \left\|\Pi_{\hat{\mathcal{F}}_{k_\lambda}}(I - \Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2 + \left\|(I - \Pi_{\hat{\mathcal{F}}_{k_\lambda}})(I - \Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2 \\
&= \left\|(\Pi_{\hat{\mathcal{F}}_{k_\lambda}} - \Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2 + \left\|(I - \Pi_{\hat{\mathcal{F}}_{k_\lambda}})f^*\right\|^2 \\
&\leq \left\|(\Pi_{\hat{\mathcal{F}}_{k_\lambda}} - \Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2 + 2\left\|(I - \Pi_{\hat{\mathcal{F}}_{k_\lambda}})\Pi_{\mathcal{F}} f^*\right\|^2 + \|(I - \Pi_{\mathcal{F}})f^*\|^2 \\
&\leq |k - k_\lambda|\,\|f^*\|^2 + 2\left\|(I - \Pi_{\hat{\mathcal{F}}_{k_\lambda}})\Pi_{\mathcal{F}} f^*\right\|^2.
\end{aligned}$$

The last bound begin due to Assumption 3, as well as the lax bounding that on the operator norm of two projections. When one could remove the $k - k_\lambda$ we let it as we expect the quantity to behave it this way, with a constant similar to $\|f^*\|^2/k_\lambda$ instead of $\|f^*\|^2$. □

We can now state the master theorem.

Theorem 4. *Under Assumptions 3, 7, 8 and 9, there exists a regularizer γ such that the regularized empirical risk minimizer verifies that: for any $\delta > 0$, there exists an $N_\delta > 0$ such that for any $n > N_\delta$, the excess of risk of the regularized empirical risk (19) minimizer reads*

$$\begin{aligned}
\mathcal{R}(f) - \mathcal{R}(f^*) \leq{}& \frac{2k_e\varepsilon^2}{n} + \frac{8M^2\log(n)^{1+\delta}}{n}\|f^*\|_{L^2(\rho_{\mathcal{X}})} + \frac{64ka}{n} \\
&+ \inf_{l\leq k}\left\|(\Pi_{\mathcal{F}_{k_\lambda}} - \Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})})f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} + 4\sigma(l)^2 + 4\zeta^2\left(\left\|\tilde{T_\lambda}^{-1}\Pi_{\mathcal{F}_l}^{(\mu_\Xi)} f^*\right\|_{L^2(\mu_\Xi)}\left(\mathcal{L}_k(\hat{\Theta};\lambda) - \mathcal{L}_k(\Theta;\lambda)\right)^{1/2}\right). \quad (33)
\end{aligned}$$

where $\mathcal{F}_l$ the span of l-th first eigenfunction of T_λ, k_λ the number of strictly positive eigenfunctions of T_λ, $k_e \leq k$ is the effective dimension of ψ in $L^2(\rho_{\mathcal{X}})$, $a = \left\|I - \Pi_{\hat{\mathcal{F}}} f^*\right\|_{L^\infty} \leq \|f^*\|_{L^\infty} + M\|f\|_{L^2}$, $M = \sup\|\psi\| \leq k\lambda^{-1}\sup\|\varphi\|$, and $\tilde{T}_\lambda = \sum_{i\in[k]}(\lambda_i^2 - \lambda_{k+1}^2)^{1/2} f_i f_i^\top$. Moreover, under the sole Assumptions 1 and 2, we have the simpler bound

$$\begin{aligned}
\mathcal{R}(f) - \mathcal{R}(f^*) \leq{}& \frac{2k_e\varepsilon^2}{n} + \frac{8M^2\log(n)^{1+\delta}}{n}\|f^*\|_{L^2(\rho_{\mathcal{X}})} + \frac{64ka}{n} + \max(k - k_\lambda, 0)\|f^*\|^2_{L^2(\rho_{\mathcal{X}})} \\
&+ 2c_r\left\|T_\lambda^{-1}\Pi_{\mathcal{F}_\lambda} f^*\right\|^2_{L^2(\mu_\Xi)}\left\{\mathcal{L}_{k_\lambda}(\hat{\Theta};\lambda) - \mathcal{L}_{k_\lambda}(\Theta_*;\lambda)\right\} + \|(I - \Pi_{\mathcal{F}_\lambda})f^*\|_{L^2(\mu_\Xi)}
\end{aligned}$$

Where $\hat{\Theta}$ is understood as belonging to $\mathbb{R}^{k_\lambda} \otimes \mathcal{H}$ in this last expression and $\mathcal{F}_\lambda$ the eigenspace linked with positive eigenvalues of T_λ.

B.4. Discussion

B.4.1. Finite number of positive eigenvalues

The following result relates the eigenvalues of T_λ with those of K. It notably proves that k_λ is finite when K is trace-class, which is one claim of Theorem 1.

Lemma 29 (Relating capacity between K and T_λ)**.** *If (μ_i) are the eigenvalues of K, then the number of eigenvalues of T_λ that are bigger than $t \in \mathbb{R}$ is smaller than the cardinality of $\{i \,|\, \mu_i > \lambda/(1-t)\}$. Moreover, if there exists $q > 0$ such that $\mathrm{Tr}\left(K^{1/q}\right) < +\infty$, then there exists a c_q such that if (μ_i) are the eigenvalues of K, we have $\mu_i \leq c_q i^{-q}$. As a consequence, in this setting, for any $t \in \mathbb{R}$, the number of eigenvalues of T_λ that is bigger than t is smaller than $(c_q(1-t)/\lambda)^{1/q}$.*

Proof. Let us consider the set of eigenvectors (f_i) whose eigenvalues are bigger than t. Consider the span of this space, we want to quantify its dimension. We know that all unitary vectors in this span satisfies

$$t \leq x^\top T_\lambda x \leq x^\top T x - \lambda x^\top K^{-1} x \leq 1 - \lambda x^\top K^{-1} x$$

Hence

$$x^\top K^{-1} x \leq \frac{1-t}{\lambda}$$

This means that this span does not intersect the span of φ_i for φ_i the eigenvectors of K^{-1} such that the eigenvalues are bigger than $\lambda/(1-t)$. In other terms, this linear space does not intersect a linear space of co-dimension d where d is the cardinality mentioned in the lemma statement. Let us denote by U the space we are interested in and by V the space it does not intersect beside in the origin, and by E the ambient space Since $U \cap V = \{0\}$, the quotient $(U+V)/V$ is isomorphic to U, hence

$$\dim(U) = \dim\left(\frac{U+V}{V}\right) \leq \dim\left(\frac{E}{V}\right) = \mathrm{codim}(V) = d.$$

This concludes the proof of the first part of the lemma.

The second claim follows from the fact that $\mu_i^{1/q}$ are summable and decreasing, hence the sequence $S_n = \sum_{i \leq n} \mu_i^{1/q}$ is a Cauchy sequence. As a consequence, there exists $N \in \mathbb{N}$, such that for any $s > N/2$, we have

$$s\mu_{2s}^{1/q} \leq S_{2s} - S_s \leq 1/2.$$

Hence, for all $s \geq N$, we have $\mu_s \leq s^{-1/q}$, hence $\mu_s / s^{-1/q}$ is bounded. Denoting by c_q the maximum, leads to the first result. The final statement is a consequence of the fact that $c_q i^{-q} > \lambda/(1-t)$ implies $i < (c_q(1-t)/\lambda)^{1/q}$. □

Example 8. *When considering the radial basis function kernel $\varphi(x)^\top \varphi(x') = \exp(-\|x - x'\|^2)$, Ψ is the space of analytical functions (Sun & Zhou, 2008), which is known to be small compared to L^2 spaces (Kolmogorov & Tikhomirov, 1959). As a consequence, one can think as $q = +\infty$ in the previous lemma. More in general, when φ is bounded, K is trace-class and one can take $q = 1$.*

Proof. The capacity of K is relates to the capacity of $K(\left\{ f \,\middle|\, \|f\|_{L^2(\mu_\equiv)} \leq 1 \right\})$, which itself relates to the capacity of $\Psi = \mathrm{im}\, K^{1/2}$. This explains why q can be taken, in essence, as arbitrarily big (Bach, 2023).

When φ is bounded, the following

$$\begin{aligned}\mathrm{Tr}\,(K) &= \mathrm{Tr}\left(SS^\top\right) = \mathrm{Tr}\left(S^\top S\right) = \mathrm{Tr}\left(\mathbb{E}[\varphi(X)\varphi(X)^\top]\right) = \mathbb{E}[\mathrm{Tr}\left(\varphi(X)\varphi(X)^\top\right)] \\ &= \mathbb{E}[\varphi(X)^\top \varphi(X)] = \mathbb{E}[\|\varphi(X)\|^2] < +\infty,\end{aligned}$$

proves that K is trace class. □

B.4.2. Derivation for vanishing bias

In the main text, we have assumed that T_λ was the right operator to define the solution of the representation learning (which explains Assumption 3). This might offend the purist as it would be nicer to define a principled solution that does not depend on the choice of the architecture (yet that might be easier to approximate with some architecture than others). This suggests studying the behavior of the last expression in Theorem 4 when λ goes to zero.

We let for future work a more precise study of the inductive bias in this vanishing setting: in essence, the choice of architecture Ψ perturbs T by λK^{-1} to make it T_λ, and ideally, we would like to quantify the speed at which T_λ converges to T when seen through the eyes of f^* as we decrease the regularization parameter. In the kernel regime, it could be characterized by perturbation theory (Kato, 1995), and refinement of Davis-Kahan theorem (Davis & Kahan, 1970) taking into account Assumption 3. Moreover, when K and T commute, the interplay can be studied in a more direct fashion thanks to Proposition 4.

C. Control of the upstream excess of risk

In order to control the excess of risk, one can use technique steaming from optimization as well as technique steaming from classical statistical learning.

C.1. Rademacher complexity

First, let us remark that $\mathcal{L}$ is a quadratic function when parameterized with $\Lambda = \Theta^\top \Theta \in \mathcal{H} \otimes \mathcal{H}$.

Lemma 30. *Let $\Theta \in \mathbb{R}^k \otimes \mathcal{H}$, denote $\Lambda = \Theta^\top \Theta \in \mathcal{H} \otimes \mathcal{H}$*

$$\mathcal{L}(S\Theta) = 2(\beta - 1)\, \mathbb{E}_\xi[\langle \Lambda, \varphi(\xi)\varphi(\xi)^\top \rangle] - 2\beta\, \mathbb{E}_X\, \mathbb{E}_{\xi,\xi'}\left[\langle \Lambda, \varphi(\xi')\varphi(\xi)^\top \rangle \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \rangle^2\right] + k. \quad (34)$$

Moreover, the regularization reads $\lambda \|\Theta\|^2 = \lambda \operatorname{Tr} \Lambda = \lambda \langle \Lambda, I \rangle$.

Proof. Consider $\psi = \Theta\varphi$, we have

$$\begin{aligned}
\mathcal{L}(\psi; \beta) &= 2(\beta - 1)\, \mathbb{E}_\xi[\psi(\xi)^\top \psi(\xi)] - 2\beta\, \mathbb{E}_X\, \mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top \psi(\xi') \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[(\psi(\xi')^\top \psi(\xi))^2\right] + k. \\
&= 2(\beta - 1)\, \mathbb{E}_\xi[\varphi(\xi)^\top \Lambda \varphi(\xi)] - 2\beta\, \mathbb{E}_X\, \mathbb{E}_{\xi,\xi'}\left[\varphi(\xi)^\top \Lambda \varphi(\xi') \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[(\varphi(\xi')^\top \Lambda \varphi(\xi))^2\right] + k. \\
&= 2(\beta - 1)\, \mathbb{E}_\xi[\operatorname{Tr}\left(\Lambda \varphi(\xi)\varphi(\xi)^\top\right)] - 2\beta\, \mathbb{E}_X\, \mathbb{E}_{\xi,\xi'}\left[\operatorname{Tr}\left(\Lambda \varphi(\xi')\varphi(\xi)^\top\right) \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[\operatorname{Tr}\left(\Lambda \varphi(\xi)\varphi(\xi')^\top\right)^2\right] + k.
\end{aligned}$$

The lemma follows from the characterization of the Hilbert-Schmidt geometry with the trace, the fact that Λ is self-adjoint, and that the regularization reads $\|\Theta\|^2 = \operatorname{Tr} \Theta^\top \Theta$. □

Let us recall three useful facts from the statistical learning literature.

Lemma 31. *Let $\mathcal{R}(\zeta) = \mathbb{E}_Z[\ell(\zeta, Z)]$, ζ^* be the minimizer of $\mathcal{L}$ inside a domain for ζ, and ζ_n be the minimizer of $\mathcal{R}_{(Z_i)}(\zeta) = \frac{1}{n}\sum_{i\in[n]} \ell(\zeta, Z_i)$ based on exchangeable data Z_i such that $\mathbb{E}_{(Z_i)}[\mathcal{R}_{(Z_i)}] = \mathcal{R}$. The average excess of risk of ζ_n is bounded by Rademacher complexity as*

$$\mathcal{R}(\zeta_n) - \mathcal{R}(\zeta_*) \leq 4\, \mathbb{E}_{(Z_i),(\sigma_i)}\left[\sup_\zeta \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(\zeta, Z_i)\right] \quad (35)$$

where σ_i are i.i.d variables taking values one and minus one with probability one half.

Proof. The proof is a classical result from learning theory (Bartlett & Mendelson, 2002), its proof consists in introducing both the empirical risk of ζ_n and ζ, and bounding the difference between the empirical and population of ζ by the supremum of this deviation over the entire domain of ζ. This is followed by the replacement of the population risk by the average empirical one, and a symmetrization trick that introduce the variable (σ_i) based on exchangeability of the (Z_i). □

Lemma 32. *For linear model, the Rademacher complexity can be bounded as*

$$\mathbb{E}_{(Z_i),(\sigma_i)}\left[\sup_{\|\zeta\|\leq M} \frac{1}{n}\sum_{i=1}^n \sigma_i \langle Z_i, \zeta \rangle\right] \leq \frac{M}{\sqrt{n}}\sqrt{\mathbb{E}[\|Z\|^2]}. \quad (36)$$

Proof. This is a classical result on Rademacher complexity of ball constraints predictors (Bartlett & Mendelson, 2002). □

Lemma 33. *Moreover, when $h : \mathbb{R} \to \mathbb{R}$ is Lipschitz, the following contraction principle holds*

$$\mathbb{E}[\sup_f \frac{1}{n}\sum_{i=1}^n \sigma_i h(f(Z_i))] \leq \sup \|\mathrm{d}h(x)\| \, \mathbb{E}[\sup_f \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)]$$

Proof. This follows from contraction of space capacity by Lipschitz functions (Vitushkin, 1954), see Meir & Zhang (2003) for a proof in the context of machine learning. □

We can now state the convergence property based on Rademacher complexity.

Lemma 34. *Let $\Theta_n \in \mathbb{R}^k \otimes \mathcal{H}$ be the minimizer of the unbiased regularized empirical version of $\mathcal{L}$ based on a dataset $\mathcal{D}_n$. Assume that $\mathcal{D}_n$ is built from n input samples (X_i) and m augmentation per samples (ξ_{ij}), then the average excess of risk is bounded by*

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{L}(S\Theta_n)] - \mathcal{L}(S\Theta) \leq \frac{8\kappa^2 \sup \|\Lambda\|_{HS}}{\sqrt{n}} \left(\frac{m+1+\beta}{m} + \frac{\sqrt{2}\kappa^2 \sup \|\Lambda\|_{HS} (m^2+1)}{m^2} \right), \tag{37}$$

where κ is a bound on $\|\varphi(X)\|$.

Proof. Following the previous lemmas on Rademacher complexity we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}[\mathcal{L}(S\Theta_n);\lambda] - \mathcal{L}(S\Theta;\lambda) &\leq 8\,\mathbb{E}_{\mathcal{D}_n,\sigma}\left[\sup_\Lambda \frac{1-\beta}{n}\sum_{i\in[n]}\sigma_i \left\langle \Lambda, \frac{1}{m}\sum_{j\in[m]} \varphi(\xi_{ij})\varphi(\xi_{ij})^\top \right\rangle\right] \\
&\quad + 8\,\mathbb{E}_{\mathcal{D}_n,\sigma}\left[\sup_\Lambda \frac{\beta}{n}\sum_{i\in[n]}\sigma_i \left\langle \Lambda, \frac{2}{m}\sum_{j\in[m/2];j+k-1=m} \varphi(\xi_{ij})\varphi(\xi_{ik})^\top \right\rangle\right] \\
&\quad + 4\,\mathbb{E}_{\mathcal{D}_n,\sigma}\left[\frac{2}{n}\sum_{i\in[n/2];i+j-1=n}\sigma_i \frac{1}{m^2}\sum_{k,l\in[m]} \left\langle \Lambda, \varphi(\xi_{ik})\varphi(\xi_{jk})^\top \right\rangle^2\right] \\
&\leq \frac{8\sup\|\Lambda\|_{HS}}{\sqrt{n}}\left((1-\beta)\,\mathbb{E}_X\left[\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \varphi(\xi_i)\varphi(\xi_i)^\top\right\|_{HS}^2 \middle| X\right]\right]^{1/2}\right) \\
&\quad + \frac{8\sup\|\Lambda\|_{HS}}{\sqrt{n}}\left(\beta\,\mathbb{E}_X\left[\mathbb{E}\left[\left\|\frac{2}{m}\sum_{i,j=1}^{m/2} \varphi(\xi_i)\varphi(\xi_j)^\top\right\|_{HS}^2 \middle| X\right]\right]^{1/2}\right) \\
&\quad + \frac{8\sup\|\Lambda\|_{HS}}{\sqrt{n}}\left(\sqrt{2}\sup\left|\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top\right\rangle\right| \mathbb{E}\left[\left\|\frac{1}{m^2}\sum_{i,j=1}^m \varphi(\xi_i)\varphi(\xi_j)^\top\right\|_{HS}^2\right]^{1/2}\right).
\end{aligned}$$

To work out those terms, remark that if (Z_i) are i.i.d. variables,

$$\mathbb{E}[\left\|\frac{1}{p}\sum_{i\in[p]} Z_i\right\|^2] = \mathbb{E}[\left\|\frac{1}{p}\sum_{i\in[p]} Z_i - \mathbb{E}[Z]\right\|^2] + \|\mathbb{E}[Z]\|^2 = \frac{1}{p}\,\mathbb{E}[\|Z-\mathbb{E}[Z]\|^2] + \|\mathbb{E}[Z]\|^2.$$

While one could work out each term, the lemma consists in simply bounding φ by κ, hence all the mean and standard deviation one can obtain with expression of φ by κ. □

The expression in the main text is due to the following lemma.

Lemma 35. *When minimizing a regularized risk, one can reduce the search of Θ under the constraint $\|\Lambda\|_{HS} \leq \lambda^{-1}k$.*

Proof. When regularizing we have

$$\|\Lambda\|_{HS} = \left\|\Theta^\top\Theta\right\|_{HS} \leq \|\Theta\|_{\mathrm{op}} \|\Theta\|_{HS} \leq \|\Theta\|_{HS}^2 ,$$

and for minimizer of the empirical or population risk

$$\lambda \|\Theta\|^2 \leq \mathcal{L}(S\Theta) + \lambda \|\Theta\|^2 \leq \mathcal{L}(0) = k,$$

which explains the statement of the lemma. □

The attentive reader would remark that compared to the bound of HaoChen et al. (2021) we gain a factor $k^{-1/2}$. Indeed, this factor could be recovered in HaoChen et al. (2021) by using the techniques of Maurer (2016) rather than a trivial bound on Rademacher complexity of vector-valued function spaces in $k \max_{i\in[k]} \hat{\mathcal{R}}(\mathcal{F}_i)$ with HaoChen et al. (2021) notations.

C.2. Convex optimization

When a least-square problem benefits from additional structure, such as smoothness or strong convexity, results from convex optimization could lead to improvement over the usual convergence rates in $n^{-1/2}$. Recall basic results from convex optimization.

Lemma 36. *Let $\mathcal{L}(\Theta) = \mathbb{E}_Z[\ell(\Theta, Z)]$ be a convex function optimized over a convex domain. Given n samples (Z_i), (unbiased) stochastic gradient descent with final averaging can achieve an excess of risk*

$$\mathbb{E}_{(Z_i)} \mathcal{L}(\hat{\Theta}) - \mathcal{L}(\Theta_*) \leq \sqrt{2} M V n^{-1/2} \tag{38}$$

with $M^2 = \|\Theta_* - \Theta_0\|^2$ and $V^2 = \mathbb{E}[\|\nabla_\Theta \ell(\Theta, Z_i)\|^2]$. Moreover, if $\mathcal{L}$ is α-smooth, then it can achieve

$$\mathbb{E}_{(Z_i)} \mathcal{L}(\hat{\Theta}) - \mathcal{L}(\Theta_*) \leq \sqrt{2} M \sigma n^{-1/2} + \alpha M^2 n^{-1} \tag{39}$$

where $\sigma^2 = \mathbb{E}[\|\nabla\mathcal{L} - \nabla\ell\|^2]$. Finally, when $\mathcal{L}$ is α-strongly convex, it achieves

$$\mathbb{E}_{(Z_i)} \mathcal{L}(\hat{\Theta}) - \mathcal{L}(\Theta_*) \leq \frac{2V^2}{\alpha(n+1)}. \tag{40}$$

As a consequence, given n data samples, there exists an empirical estimate of $\hat{\Theta}$ that guarantee those generalization bounds.

Proof. This lemma is a direct consequence of Theorems 6.1, 6.2 and 6.3 of Bubeck (2015). □

It should be noted that when parameterized with $\Lambda = \Theta^\top\Theta$, $\mathcal{L}$ is a quadratic form as stated by Lemma 30, yet it is minimized over a non-convex domain, the domain of symmetric operator of rank k. We will relax this constraint and consider the harder problem of optimizing over Λ in the set of self-adjoint positive operators. This is justified by the fact that Theorem 4 provides guarantee on the downstream task, even when one relaxes the rank constraint on Λ.

To benefit from Lemma 36, one should consider an unbiased expression of $\mathcal{L}$. Consider the minibatch scheme that consist in sampling two inputs X_1, X_2, and m augmentations ξ_{ij} for each X_i, formally

$$X_i \sim \mu_{\mathcal{X}}^{\otimes 2}, \qquad \xi_{ij} \sim \mu|_{X_i}^{\otimes m}. \tag{41}$$

Here $\mu_{\mathcal{X}}$ denotes the marginal of μ with respect to X, which is likely to match $\rho_{\mathcal{X}}$, and $\mu|_X$ denotes the distribution of Ξ conditionally to X.

Lemma 37. *An unbiased formulation of $\mathcal{L}$ is based on ℓ defined as*

$$\begin{aligned}\nabla_\Lambda \ell(S\Theta; \lambda) = \frac{2(\beta-1)}{m} \sum_{j\in[m]} \varphi(\xi_{1j})\varphi(\xi_{1j})^\top - \frac{2\beta}{m(m-1)} \sum_{1\leq j\neq k\leq m} \varphi(\xi_{1j})\varphi(\xi_{1k})^\top \\ + \frac{1}{m^2} \sum_{i,i'=1}^{2} \sum_{j,k=1}^{m} \left\langle \Lambda, \varphi(\xi_{ij})\varphi(\xi_{i'k})^\top \right\rangle \varphi(\xi_{ij})\varphi(\xi_{i'k})^\top .\end{aligned} \tag{42}$$

Moreover, when $\mathcal{L}$ is regularized, one has to add $+\lambda I$ to get a gradient on the regularized risk.

Proof. This formula follows from Lemma 30. □

In order to bound the norm squared of the gradient, one can use the following lemma.

Lemma 38. *For ℓ given in (42), bounds on the gradient norm and its variance are*

$$\|\nabla_\Lambda \ell\| \leq 2\kappa^2 + \kappa^4 \sup \|\Lambda\|, \qquad \text{and} \qquad \mathbb{E}[\|\nabla_\Lambda \ell - \nabla \mathcal{L}\|^2] \leq (\sigma_X^2 + m^{-1}\sigma_\xi^2)(1 + \sup \|\Lambda\|^2), \tag{43}$$

where σ_X relates to the variance of $\mathbb{E}[\psi(\xi) \mid X]$ and σ_ξ relates to the average variance of $(\xi \mid X)$.

Proof. Let us decompose $\nabla \ell$ into three terms $\nabla \ell = a + b + c$ as appearing in (42), we have

$$\begin{aligned}
&\|a\| \leq 2(1-\beta) \left\|\varphi(\xi)\varphi(\xi)^\top\right\| \leq 2(1-\beta)\kappa^2 \\
&\|b\| \leq 2\beta \left\|\varphi(\xi)\varphi(\xi')^\top\right\|^2 \leq \beta\kappa^2 \\
&\|c\| \leq \left\|\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \right\rangle \varphi(\xi)\varphi(\xi')^\top\right\| \leq \sup \|\Lambda\| \, \kappa^4.
\end{aligned}$$

To bound the variance, one can proceed with

$$\mathbb{E}\|\nabla \ell - \nabla \mathcal{L}\|^2 \leq 3\,\mathbb{E}\|a - \mathbb{E}[a]\|^2 + 3\,\mathbb{E}\|b - \mathbb{E}[b]\|^2 + 3\,\mathbb{E}\|c - \mathbb{E}[c]\|^2 .$$

Let us begin with the part in a,

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i\in[m]} \varphi(\xi_{1i})\varphi(\xi_{1i})^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i\in[m]} \varphi(\xi_{1i})\varphi(\xi_{1i})^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \,\middle|\, X = X_1]\right\|^2\right] \\
&\quad + \mathbb{E}\left[\left\|\mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \mid X = X_1] - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\right\|^2\right] \\
&= \frac{1}{m}\,\mathbb{E}_X\,\mathbb{E}_\xi\left[\left\|\varphi(\xi)\varphi(\xi)^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \mid X]\right\|^2 \,\middle|\, X\right] + \mathbb{E}\left[\left\|\mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \mid X] - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\right\|^2\right].
\end{aligned}$$

Similarly, the part in b can be expressed as

$$\begin{aligned}
\mathbb{E}[\|b - \mathbb{E}[b]\|^2] = 2\beta\left(\frac{2}{m}\,\mathbb{E}_X\,\mathbb{E}_\xi\left[\left\|\varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \mid X]\right\|^2 \,\middle|\, X\right]\right) \\
+ 2\beta\,\mathbb{E}\left[\left\|\mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \mid X] - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top]\right\|^2\right].
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}[\|c - \mathbb{E}[c]\|^2] &= \frac{1}{m^2}\,\mathbb{E}_X\,\mathbb{E}_\xi\left[\left\|\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top\right\rangle \varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top\right\rangle \varphi(\xi)\varphi(\xi')^\top \mid X, X']\right\|^2 \,\middle|\, X, X'\right] \\
&\quad + \mathbb{E}\left[\left\|\mathbb{E}\left[\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top\right\rangle \varphi(\xi)\varphi(\xi')^\top \mid X, X'\right] - \mathbb{E}[\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top\right\rangle \varphi(\xi)\varphi(\xi')^\top]\right\|^2\right] \\
&= \frac{1}{m^2}\,\mathbb{E}_X\,\mathbb{E}_\xi\left[\left\|\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X']\right\rangle\right\|^2 \,\middle|\, X, X'\right] \\
&\quad + \mathbb{E}\left[\left\|\left\langle \Lambda, \mathbb{E}\left[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X'\right] - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top]\right\rangle\right\|^2\right]. \\
&\leq \frac{1}{m^2}\,\|\Lambda\|^2\,\mathbb{E}_X\,\mathbb{E}_\xi\left[\left\|\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X']\right\|^2 \,\middle|\, X, X'\right] \\
&\quad + \|\Lambda\|^2\,\mathbb{E}\left[\left\|\mathbb{E}\left[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X'\right] - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top]\right\|^2\right].
\end{aligned}$$

As a consequence, we get

$$\mathbb{E}\|\nabla \ell - \nabla \mathcal{L}\|^2 \leq 3\left(2(1-\beta)\left(\frac{\sigma_{\xi,1}^2}{m} + \sigma_{X,1}^2\right) + 2\beta\left(\frac{2\sigma_{\xi,2}^2}{m} + \sigma_{X,2}^2\right) + \sup\|\Lambda\|^2\left(\frac{\sigma_{\xi,3}^2}{m^2} + \sigma_{X,3}^2\right)\right)$$

where

$$\sigma_{\xi,1}^2 = \mathbb{E}_X \mathbb{E}_\xi \left[\left\| \varphi(\xi)\varphi(\xi)^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \mid X] \right\|^2 \Big| X \right]$$
$$\sigma_{X,1}^2 = \mathbb{E} \left[\left\| \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \mid X] - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top] \right\|^2 \right]$$
$$\sigma_{\xi,2}^2 = \mathbb{E}_X \mathbb{E}_\xi \left[\left\| \varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \mid X] \right\|^2 \Big| X \right]$$
$$\sigma_{X,2}^2 = \mathbb{E} \left[\left\| \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \mid X] - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top] \right\|^2 \right]$$
$$\sigma_{\xi,3}^2 = \mathbb{E}_X \mathbb{E}_\xi \left[\left\| \varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X'] \right\|^2 \Big| X, X' \right]$$
$$\sigma_{X,3}^2 = \mathbb{E} \left[\left\| \mathbb{E}\left[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X' \right] - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top] \right\|^2 \right].$$

Using the fact that $m^2 \geq m$ and choosing the right σ_X and σ_ξ leads to the lemma. □

The following lemma states the convexity properties of $\mathcal{L}$

Lemma 39. *As a function of Λ, the objective $\mathcal{L}$ is α-smooth with $\alpha = \kappa^4$, where κ is a bound on $\|\varphi\|$. Moreover, when $\mathcal{X}$ is finite, it is α'-strongly, with α' being the square of eigen gap of $K = SS^\top$.*

Proof. This is a consequence of Lemma 30, $\mathcal{L}$ is a quadratic function, with the quadratic part being

$$\mathbb{E}[\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \rangle^2] = \langle \Lambda, \mathbb{E}[\varphi(\xi')\varphi(\xi')^\top] \otimes \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\Lambda \rangle = \langle \Lambda, \Sigma \otimes \Sigma\Lambda \rangle .$$

In other terms, the hessian of $\mathcal{L}$ is $\Sigma \otimes \Sigma \in \mathcal{H}^{\otimes 2} \otimes \mathcal{H}^{\otimes 2}$. As a consequence, $\Sigma \otimes \Sigma \preceq \|\Sigma \otimes \Sigma\|_{\mathrm{op}} I = \|\Sigma\|_{\mathrm{op}}^2 I \preceq \kappa^4 I$. Similarly, $\Sigma \otimes \Sigma \succeq \left\|\Sigma^{-1}\right\|_{\mathrm{op}}^{-2} I = \gamma_\xi^2 I$, where γ_ξ is the eigen gap of Σ, hence of K. □

There are few remaining difficulties that must be addressed before concluding. First, although the identity is not Hilbert-Schmidt, it should be noted that the term in λ will only contract distances in the stochastic gradient descent. As a consequence, optimizing the regularized risk will only contract the descent trajectory (to prove it formally one could go back to the proofs of Bubeck (2015)). Finally, we have described a descent in the space of self-adjoint positive operators, without incorporating any constraints on the rank of Λ. Notice that based on Lemma 35, on can restrict the search of Λ to inside the domain $\|\Lambda\| \leq k_\lambda/\lambda$. Finally, if Λ minimizes the loss $\mathcal{L}$, then one can show that thresholding its eigenvalues to make it of dimension at most k can only increase the loss $\mathcal{L}$ by a bounded multiplicative factor. We note that without explicit regularization, the previously described stochastic gradient descent algorithm with early stopping has a regularization effect that could be studied in the spectral filtering framework of Lin et al. (2020).

D. Examples

This section is devoted to illustrate what T and K are under simple distributions thanks to harmonic analysis techniques.

D.1. Harmonics analysis on the sequence of bits, a.k.a. the Boolean hypercube

A fine-grained analysis of the role of classical augmentations can be derived in settings that allow précise derivations. We shall focus on invariant data distribution such as the uniform distribution, and augmentations consisting of permutations or perturbations of coordinates that left this distribution invariant. While such distributions may lack structure present in real data, they allow for a precise study of the effect of certain architectures and augmentations, which may also partly apply to more realistic data. The study involves the construction of appropriate L^2 bases that ease the study of the effect of both the kernel operator K and the smoothing operator T defined from augmentations. These are closely related to the study of invariant kernels (see, e.g., Bietti et al., 2021; Bietti, 2022; Mei et al., 2021; Misiakiewicz & Mei, 2022).

Will focus here on the data that are n-bit inputs on the Boolean cube $\mathcal{X} = \{-1, +1\}^d$ with uniform distribution. To be able to use the harmonic analysis tools to their fullest, we assume that inputs are sampled from the uniform distribution on $\mathcal{X}$. In this setting, the space of function $L^2(\mathcal{X}) = L^2(\mathcal{X}, \mathbb{R}, \mu_\mathcal{X})$ is defined through the usual scalar product, for $f, g : \mathcal{X} \to \mathbb{R}$,

$$\langle f, g \rangle = \mathbb{E}_{x \sim \tau}[f(x)g(x)] = \frac{1}{2^d} \sum_{x \in \mathcal{X}} f(x)g(x).$$

D.2. The role of augmentation

Let us know analyze the role of augmentation in the definition of T on the Boolean cube. For simplicity and ease of notation, we assume indexing of the bits is taken $\bmod d$, e.g., $x_{-1} = x_d$.

D.2.1. Study through parity functions

Parity functions. A useful basis in this space are the parity functions, which can be seen as Fourier functions in this L^2-space (O'Donnell, 2014). They are defined for each subset $S \subseteq [d]$ as counting the parity of x within this set

$$\chi_S(x) = \prod_{i \in S} x_i. \tag{44}$$

Lemma 40. *The parity functions χ_S form an orthonormal basis of $L^2(\mathcal{X})$.*

Proof. It is straightforwards to check that $\langle \chi_S, \chi_S \rangle = 1$. If $S \neq S'$, then w.l.o.g. there is an $i \in S \setminus S'$, and we have

$$\langle \chi_S, \chi_{S'} \rangle = \mathbb{E}_x[x_i \chi_{S \setminus \{i\}}(x) \chi_{S'}(x)] = \mathbb{E}_{x_i}[x_i]\, \mathbb{E}_{x_{-i}}[\chi_{S \setminus \{i\}}(x) \chi_{S'}(x)] = 0.$$

This proves orthogonality of this basis. □

Let us begin with augmentations that are easily to study with the parity basis.

Proposition 41 (Random noise)**.** *Consider the flip of each bit of x with probability equal to p formally via the operation*

$$B_y^p(x) = x \odot y, \quad y \sim \mathrm{Ber}(\{-1, +1\}, p)^{\otimes d}, \tag{45}$$

where the operation $x \odot y$ applies pointwise multiplication and the distribution $\mathrm{Ber}(\{-1, +1\}, p)$ returns the value -1 with probability p and $+1$ with probability $1 - p$. Under the augmentations $\xi = X \otimes y$, T is diagonalized in the parity basis with

$$T\chi_S = |1 - 2p|^{|S|}\, \chi_S. \tag{46}$$

In other terms, T applies a factor $|1 - 2p|^{|S|}$ to reduce the effect of higher order Fourier functions.

Proof. Recall the formula $g^\top T f = \mathbb{E}_X\, \mathbb{E}_{\xi, \xi'} [\langle f(\xi), g(\xi') \rangle \mid X]$. As a consequence, with y, y' denoting the noise strings (each bit equal to -1 with probability p) and $S \triangle S' = (S \cup S')/(S \cap S')$,

$$\begin{aligned}
\chi_S^\top T \chi_{S'} &= \mathbb{E}_X[\mathbb{E}_{y,y'}[\chi_S(X \odot y)\chi_{S'}(X \odot y')]] = \mathbb{E}_X\left[\mathbb{E}_{y,y'}\left[\prod_{i \in S} X_i y_i \prod_{j \in S'} X_j y'_j\right]\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_y\left[\prod_{i \in S \triangle S'} X_i y_i\right]\right] \mathbb{E}_{y,y'}\left[\prod_{i \in S \cap S'} y_i y'_i\right] \\
&= \mathbb{E}[\chi_{S \triangle S'}(X)] \cdot |1 - 2p|^{|S \triangle S'|} |1 - 2p|^{2|S \cap S'|} = |1 - 2p|^{|S|} \delta_{S,S'}.
\end{aligned}$$

Therefore, in the case of bit-flip augmentations, T is diagonalized in the parity basis. □

Proposition 42 (Cropping/Masking)**.** *Consider the cropping operation within a window of size w, formally defined as*

$$[M_a^w(x)]_i = \begin{cases} x_i & \text{if } i \in [a, a+w) \\ \mathrm{Ber}(\{-1, +1\}, 0.5) & \text{otherwise} \end{cases}, \qquad a \sim \mathcal{U}([d]), \tag{47}$$

where $[a, a+w) = \{a, a+1, \ldots, a+w-1\}$, a is drawn from the uniform distribution over $[d]$, and the distribution $\mathrm{Ber}(\{-1, +1\}, 0.5)$ returns a random bit with equal probability for $+1$ and -1 thus effectively masking the values outside of the window in $[a, a+w)$. Under the augmentations $\xi = M_a^w(X)$, T is diagonalized in the parity basis with

$$T\chi_S = \frac{\max\{1 + w - \mathrm{diam}(S), 0\}^2}{d^2} \cdot \chi_S \qquad \text{with} \qquad \mathrm{diam}(S) = \min\{v \mid v, a \in [d]; S \subseteq [a, a+v)\}. \tag{48}$$

In other terms, the action of cropping effectively removes any dependence on the kernel with parity functions of high order whose support falls outside the windows of size w.

Proof. In this setting,

$$\begin{aligned}
\chi_S^\top T \chi_{S'} &= \mathbb{E}_X[\mathbb{E}_{a,b}[\chi_S(M_a^w(X))\chi_{S'}(M_b^w(X))]] \\
&= \frac{1}{d^2}\sum_{a,b=1}^{d} \mathbb{E}_{X,\nu,\nu'}\left[\prod_{i\in S\cap[a,a+w)} x_i \prod_{i'\in S\setminus[a,a+w)} \nu_{i'} \prod_{j\in S'\cap[b,b+w)} x_j \prod_{j'\in S'\setminus[b,b+w)} \nu'_{j'}\right] \\
&= \frac{1}{d^2}\sum_{a,b=1}^{d} \mathbf{1}_{S\subseteq[a,a+w)}\, \mathbf{1}_{S'\subseteq[b,b+w)}\, \mathbb{E}_X[\chi_S(X)\chi_{S'}(X)] = \frac{1}{d^2}\sum_{a,b=1}^{d} \mathbf{1}_{S\subseteq[a,a+w)}\, \mathbf{1}_{S'\subseteq[b,b+w)}\, \delta_{S,S'} \\
&= \left(\frac{1}{d}\sum_{a=1}^{d} \mathbf{1}_{S\subseteq[a,a+w)}\right)^2 \delta_{S,S'}.
\end{aligned}$$

The count of the sum relates to the diameter of S. □

Proposition 43 (2D Cropping)**.** *Consider that 2D setting $\mathcal{X} = \{-1,+1\}^{m\times d}$ where inputs are organized into an $m \times d$ grid. Consider the cropping operation to a window of size $v \times w$, formally*

$$[M_{a,b}^{v\times w}(x)]_{i+jm} = \begin{cases} x_{i+jm} & \text{if } i \in [a, a+v), j \in [b, b+w) \\ \mathrm{Ber}(\{-1,+1\}, 0.5) & \text{otherwise} \end{cases}, \qquad (a,b) \sim \mathcal{U}\left([m]\times[d]\right). \tag{49}$$

Under the augmentation $\xi = M_{a,b}^{v\times w}(X)$, T is diagonalizable in the parity basis and

$$T\chi_S = \frac{1}{m^2 d^2}\left(1 + v - \mathrm{diam}_{e_1} S\right)_+^2 \cdot \left(1 + w - \mathrm{diam}_{e_2} S\right)_+^2 \chi_S, \tag{50}$$

where $\mathrm{diam}_{e_1} S$ is the diameter of S projected onto the first dimension.

Proof. This follows from the proof of the 1D case. □

Proposition 44 (Flipping)**.** *Consider the operator which, with probability p, flip the indices into reverse order, formally*

$$[R(x)]_i = x_{-i}. \tag{51}$$

Under the augmentation $\xi = R(X)$,

$$T = (1 - 2p + 2p^2)I + 2p(1-p)J, \tag{52}$$

where J is the involution that matches any set S to its mirror $\tilde{S} = \{-i \,|\, i \in S\}$. In this setting, T is diagonalized by the $\sqrt{2}(\chi_S + \chi_{\tilde{S}})$ and $\sqrt{2})(\chi_S - \chi_{\tilde{S}})$ for $S \subseteq [d]$.

Proof. In this setting,

$$\begin{aligned}
\chi_S^\top T \chi_{S'} &= ((1-p)^2 + p^2)\,\mathbb{E}_X\left[\chi_S(X)\chi_{S'}(X)\right] + 2p(1-p)\,\mathbb{E}_X\left[\chi_{\tilde{S}}(X)\chi_{S'}(X)\right] \\
&= (1 - 2p + 2p^2)\delta_{S,S'} + 2p(1-p)\delta_{\tilde{S},S'},
\end{aligned}$$

which explain the lemma. □

Remark 45. *Up to now, we have studied all the operators in the space $L^2(\mathcal{X}, \mathbb{R}, \mu_X)$ while the main text considered those operators in $L^2(\mathcal{X}, \mathbb{R}, \mu_\Xi)$, this is justified by the fact that all transformations studied earlier let invariant the uniform distribution, hence*

$$L^2(\mu_\mathcal{X}) = L^2(\mu_\Xi). \tag{53}$$

D.2.2. Study of translations through cyclic parities

In order to study augmentations that consist of permutations, and more specifically translations, the parity basis is not adapted to diagonalize T. Instead we define below a different basis that incorporates cyclic symmetries (Misiakiewicz & Mei, 2022). We note that a similar study may be carried on other distributions, e.g., uniform on the sphere, product of spheres, or torus (Bietti et al., 2021; Bietti, 2022; Favero et al., 2021; Mei et al., 2021).

Cyclic parity functions. The functions χ_S are polynomials that can be grouped by their degree $\ell = |S|$ into spaces $V_{d,\ell}$, whose direct sum yields the full $L^2(\mathcal{X})$ space, with

$$\dim V_{d,\ell} = |\{S \subseteq [d], |S| = \ell\}| = \binom{d}{\ell}.$$

Those different spaces can be further decomposed into orbits under the action of a group. In particular for the group of permutations $G = \mathfrak{S}_d$, we define the action $\mathcal{A} : G \times \mathcal{X} \to \mathcal{X}$ denoted $\mathcal{A}(a, x) = a \cdot x$ as

$$(a \cdot x)_i = x_{a^{-1}(i)}.$$

To give a concrete example of the study of augmentations through harmonic analysis, let us focus more specifically on the action of translation, which form a sub-group of permutations. For simplicity, we will denote this group $[d]$ which is understood as $\mathbb{Z}/d\mathbb{Z}$, acting on $\mathcal{X}$ as

$$(a \cdot x)_i = x_{i-a}$$

where $i - a$ being understood modulo d. Define the orbits of this action as $\{S + a \,|\, a \in [d]\}$ for $S \subseteq [d]$. On those different orbits, one can define the following "cyclic parities" $\psi_{m,S} : \mathcal{X} \to \mathbb{C}$:

$$\psi_{m,S} = \frac{1}{\sqrt{k_S}} \sum_{k \in [k_S]} e^{2i\pi km/k_S} \chi_{S+k} = \frac{\sqrt{k_S}}{d} \sum_{k \in [d]} e^{2i\pi k \frac{md}{k_S}/d} \chi_{S+k} \qquad \text{where} \qquad k_S = |\mathrm{orb}(S)|\,, \tag{54}$$

where $m \in [k_S]$ and S is taken as a representant of an orbit.

Lemma 46. *The cyclic parities $(\psi_{m,S})$, for $m \in [k_S]$ and S is a set of representers of each orbit of the translations action, form an orthogonal basis of $L^2(\mathcal{X}, \mathbb{C}, \mu)$ where μ is the uniform measure on $\mathcal{X}$. Moreover, they diagonalize the operators $A : L^2 \to L^2$ defined as $Af(x) = f(a \cdot x)$ for any $a \in [d]$.*

Proof. The first part follows from the fact that $L^2(\mathcal{X})$ can be decomposed into the direct sum linked with the $V_{d,\ell}$ for $\ell \in [0, d]$, that each subspace can be decomposed into the orbits of the action translation $\mathrm{orb}(S) = \{S + a \,|\, a \in [d]\}$ (note that translation do not change the cardinals of the sets S). Those latter spaces can be parameterized through the discrete Fourier transform, yielding the $\psi_{m,S}$.

A natural way to "find" those basis is when trying to diagonalize an operator T such that $(\chi_S^\top T \chi_{S'})_{S,S' \subseteq [d]}$ that is block diagonal, where each block corresponds to a circulant matrix on an orbit, which can be diagonalized with the discrete Fourier transform. This is especially the case for operator of the lemma

$$A\chi_S = \chi_{S+a}$$

The above is only nonzero when $[d] \cdot S$ intersects $[d] \cdot S'$, which implies $\mathrm{orb}(S) = \mathrm{orb}(S')$ thereby constructing a block diagonal structure. Indexing the elements of the i-th block by $S_{i,k} = S_i + k$ for $k \in [d]$, we have

$$\chi_{S_{i,k}}^\top A \chi_{S_{i,k'}} = \mathbf{1}_{S_{i,k} = S_{i,k'} + a}, = \mathbf{1}_{S_i + k = S_i + k' + a}, = \mathbf{1}_{k - k' = a},$$

which only depends on the value of $(k - k')$. Therefore, each block above is a circulant matrix which is diagonalized by the discrete Fourier transform. The eigenvectors of this matrix are

$$v_m = \frac{1}{\sqrt{k_S}} \sum_{k \in [k_S]} e^{2i\pi km/k_S} e_k, \qquad \text{where} \qquad k_S = |\mathrm{orb}(S)|\,,$$

for $m \in [k_S]$ and the corresponding eigenvalues read

$$\mu_m = \sum_{k \in [k_S]} c_{k_S - k} \exp\left(\frac{2i\pi km}{k_S}\right),$$

where

$$c_i = \mathbf{1}_{i=a},$$

Using the fact that we wrote those matrices for $e_i \simeq \chi_{S+i}$ yields the lemma. □

The study of the operator T can be simplified thanks to its square root $A : L^2(\mu_\Xi) \mapsto L^2(\mu_\mathcal{X})$ formally defined by

$$Af(x) = \mathbb{E}_\xi \left[f(\xi) \,|\, X = x \right] \tag{55}$$

and verifying

$$\langle f, Tg \rangle_{L^2(\mu_\Xi)} = \mathbb{E}_X \, \mathbb{E}_{\xi,\xi'}[f(\xi)g(\xi')|X] = \langle Af, Ag \rangle_{L^2(\mu_\mathcal{X})}. \tag{56}$$

This decomposition will be particularly useful, when $\mu_\mathcal{X}$ is invariant under the action of permutations, which implies $\mu_\mathcal{X} = \mu_\Xi =: \mu$.

Lemma 47. *In the uniform Boolean setting, when augmentations are defined as $\xi = a \cdot X$ where a is a permutation sampled from the probability distribution $p \in \Delta_{\mathfrak{S}_d}$,*

$$Tf(x) = \sum_{a,b \in \mathfrak{S}_d} p(a)p(b) f((a^{-1}b) \cdot x).$$

Proof. The square root of T is defined as $Af(x) = \sum_{a \in \mathfrak{S}_d} p(a) f(a \cdot x)$. Let us focus on the case where $p(b) = \delta_{a=b}$, using the fact that $\mu_\mathcal{X}$ is the uniform measure, hence is left invariant by translation, we compute the adjoint of A with

$$\begin{aligned}
\langle Af, g \rangle_{L^2(\mu_\mathcal{X})} &= \frac{1}{2^d} \sum_{x \in \mathcal{X}} Af(x)g(x) = \frac{1}{2^d} \sum_{x \in \mathcal{X}} p(a) f(a \cdot x) g(x) \\
&= \frac{1}{2^d} \sum_{x \in \mathcal{X}} p(a) f(a \cdot x) g(a^{-1} \cdot a \cdot x) = \frac{1}{2^d} \sum_{x \in \mathcal{X}} p(a) f(x) g(a^{-1} \cdot x) \\
&= \left\langle f, x \mapsto p(a) g(a^{-1} \cdot x) \right\rangle_{L^2(\mu_\mathcal{X})}. \\
&= \left\langle f, A^\top g \right\rangle_{L^2(\mu_\Xi)}.
\end{aligned}$$

In the general case, we get by linearity,

$$A^\top f(x) = \sum_{a \in \mathfrak{S}_d} p(a) f(a^{-1} \cdot x).$$

Computing $T = A^\top A$ leads to the result. Remark that if we further assume that p is symmetric (i.e., $p(a) = p(a^{-1})$), then we have $A^\top = A$, so that $T = A^2$. □

This allows us to characterize more finely the effect of translation on the operator T.

Proposition 48 (Translations)**.** *Consider the translation operator defined formally as*

$$[T_a(x)]_i = x_{i-a}, \qquad a \sim p \in \Delta_{[d]} \tag{57}$$

Under the augmentation $\xi = T_a(X)$, T is diagonalized in $\mathbb{C}$ by the cyclic parity functions (54).

$$T\psi_{m,S} = \frac{d^2}{k_S^2} \left| \hat{p}\left(\frac{md}{k_S}\right) \right|^2 \psi_{m,S}, \tag{58}$$

where $\hat{p}$ is the Fourier transform of p, defined for $\omega \in [d]$ by

$$\hat{p}(\omega) = \sum_{a \in [d]} p(a) \exp\left(\frac{-2i\pi a\omega}{d}\right) \tag{59}$$

Proof. In the case of translation, we have

$$Af(x) = \sum_{a \in [d]} p(a) f(a \cdot x) = \sum_{a \in [d]} p(a) A_a f(x),$$

were A_k be the operator that associate f to $x \to f(k \cdot x)$, it is a translation operator and retaking the proof of Lemma 46, $A_k \psi_{m,S} = e^{-2i\pi km/k_S} \psi_{m,S}$. This leads to

$$A\psi_{m,S} = \sum_{a \in [k_S]} p(a) \exp\left(\frac{-2i\pi am}{k_S}\right) \psi_{m,S} = \sum_{a \in [d]} \frac{d}{k_S} \cdot p(a) \exp\left(\frac{-2i\pi am}{k_S}\right) \psi_{m,S} = \frac{d}{k_S} \cdot \hat{p}\left(\frac{md}{k_S}\right) \psi_{m,S},$$

and

$$T = A^*A = \sum_{m,S} \frac{d^2}{k_S^2} \left| \hat{p}\left(\frac{md}{k_S}\right) \right|^2 \psi_{m,S}\psi_{m,S}^*.$$

This proves the lemma. □

We now show how different sampling distributions over translations induce varying smoothing effects in the operator T.

Example 9 (Smoothing effect of translations)**.** *To see the effect of augmentation strength, consider a distribution p over translations that takes the form $p(a) = \omega p_0(\omega a)$, where p_0 is a localized window shape (e.g., uniform or Gaussian) that sums to 1. Here $\omega \approx 1/\Delta$ is inversely related to the window size Δ, which controls the "strength" or range of augmentations. Then we have*

$$|\hat{p}(m)|^2 = |\hat{p}_0(m/\omega)|^2 \approx |\hat{p}_0(\Delta m)|^2.$$

Here, the squared Fourier coefficients $|\hat{p}_0(m)|^2$ typically decay with the frequency m, which shows that T has a smoothing effect that penalizes eigenfunctions $\psi_{m,S}$ with larger m, i.e., those which oscillate more quickly. The above formula also highlights that the increasing the augmentation strength Δ will lead to faster decay with m, while leaving the translation-invariant eigenfunctions ($m = 0$) unaffected.

D.3. The role of architectures

A particularly useful feature space φ to define the linear class of functions Ψ is the set:

$$\varphi : \mathcal{X} \mapsto \mathbb{R}^{2^d}; x \mapsto (e_S \chi_S(x))_{S \subseteq [d]}. \tag{60}$$

for any sequences $(e_S) \in \mathbb{R}^{2^d}$. Linear model of this form can be diagonalized in the parity basis, which allows one to effectively study the interplay between the role of augmentation and the role of the architecture.

Lemma 49. *For any linear model defined through the features φ in (60), the integral operator $K : L^2(\mathcal{X}) \mapsto L^2(\mathcal{X})$ is diagonalized in the parity basis,*

$$K\chi_S = e_S^2 \chi_S. \tag{61}$$

Proof. This follows from the fact that $Kf(x) = d^{-1} \sum_{x' \in [d]} k(x, x') f(x')$ where $k(x, x') = \varphi(x)^\top \varphi(x')$. □

Among those classes of functions are dot-product kernel that verifies $k(x, y) := \varphi(x)^\top \varphi(y) = \tilde{h}(\|x - y\|^2) = h(x^\top y)$. Once again, those kernels are particularly well adapted to the Fourier geometry of the Boolean hypercube.

Lemma 50 (Spectral decomposition of dot-product kernel)**.** *Any dot-product kernel is diagonalizable in the parity basis. Specifically, there exists $(\nu_i)_{i \in [0,d]} \in \mathbb{R}^{d+1}$ such that, when $\mu_{\mathcal{X}}$ is the uniform distribution on the hypercube,*

$$K\chi_S = \nu_{|S|}\chi_S. \tag{62}$$

Proof. One can check that $x^\top y = d - 2k$ for k the number of bits that differs in x and y. Define Q_ℓ the degree-ℓ averaged polynomials of degree ℓ as

$$\sum_{S \subseteq [d], |S| = \ell} \chi_S(x)\chi_S(y) = \binom{d}{\ell} Q_{\ell,d}(\langle x, y \rangle), \tag{63}$$

for any Boolean strings x and y. The $Q_{\ell,d}$ are well defined since the left-hand side is translation invariant. Moreover, leveraging the orthogonality of the χ_S, one can show that the $(Q_{\ell,d})_{\ell \in [0,d]}$ form a basis of functions on $\{d - 2k \mid k \in [0, d]\}$. More exactly, the $m \to \binom{d}{\ell}^{-1/2} Q_{\ell,d}(m)$ are orthonormal basis of the L^2 space endowed with τ the pushforwards measure of the uniform distribution on $\mathcal{X}$ through the mapping $x \to \langle x, y \rangle$ for any fixed y, and the dimensions match. As a consequence, there exists ν_ℓ such that

$$h(\langle x, y \rangle) = \sum_{\ell \in [0,d]} \nu_\ell \binom{d}{\ell} Q_{\ell,d}(\langle x, y \rangle)$$

where ν_ℓ can be found by computing the scalar product between h and Q_ℓ in $L^2(\tau)$.

$$\nu_\ell = \langle h, Q_\ell \rangle_{L^2(\tau)} . \tag{64}$$

Finally, using the fact that, in the uniform setting, $Kf(x) = d^{-1} \sum_{x' \in [d]} k(x, x') f(x')$ where $k(x, x') = \varphi(x)^\top \varphi(x')$, we have

$$K\chi_S(x) = \mathbb{E}_Y[h(\langle x, Y \rangle)\chi_S(Y)] = \sum_\ell \nu_\ell \sum_{S' \subset [d], |S'|=\ell} \chi_{S'}(x)\, \mathbb{E}[\chi_{S'}(Y)\chi_S(Y)] = \nu_\ell \chi_S(x).$$

This ends the proof of this lemma. □

Lemma 49 can also be shown on the sphere. Its proof showcase the Q_ℓ which act as normalized Legendre (or Gegenbauer) polynomials. See, *e.g.* Smola et al. (2000); Bietti et al. (2021); Mei et al. (2021) for details. Note that for common kernel functions on the sphere, such as the ones appearing in the NTK, the ν_k decay polynomially with k (Bach, 2017; Bietti & Mairal, 2019).

The features (60) are rich enough to describe the neural tangent kernels of simple architectures with fully connected or convolutional layers. First, we describe the general form of such NTKs as below.

Proposition 51 (Linearization of simple network)**.** *Define a simple neural architecture as*

$$f(x) = \sqrt{\frac{\Delta}{N\omega d}} \sum_{i \in [N]} \sum_{k \in [d/\Delta]} a_{ik} \sum_{s \in [\omega]} \sigma\left(\left\langle w_i, x^{(q)}_{(k\Delta+s)} \right\rangle\right), \tag{65}$$

where $x^{(q)}_{(k)} = (x_k, x_{k+1}, \cdots, x_{k+q-1})$ is a local patch of size q (with indices being defined modulo d), w_i the weights initialized from a rotation-invariant distribution $\mathcal{W}$, $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is an activation function, $\omega \in \mathbb{N}$ is the size of the average pooling window, $\Delta \in \mathbb{N}$ is the pooling window, and N is the channel number. The linearization of this network near initialization yields the kernel

$$k(x, x') = \varphi(x)^\top \varphi(x') = \frac{\Delta}{d\omega} \sum_{k \in [d/\Delta]} \sum_{s, s' \in [\omega]} h\left(\left\langle x^{(q)}_{(k\Delta+s)}, y^{(q)}_{(k\Delta+s')} \right\rangle / q\right) \tag{66}$$

where

$$h(\langle u, v \rangle / q) = \mathbb{E}_{w \sim \mathcal{W}}\left[\sigma(\langle u, w \rangle / \sqrt{q})\sigma(\langle v, w \rangle / \sqrt{q}) + \sigma'(\langle u, v \rangle / \sqrt{q})\sigma'(\langle u, v \rangle / \sqrt{q}) \cdot \langle u, v \rangle / q\right]. \tag{67}$$

Proof. Such a linearization can be found, e.g., in Proposition 3 of Misiakiewicz & Mei (2022). □

Proposition 52 (Linearization of a fully connected network)**.** *A one hidden layer fully connected layer*

$$f_{FC}(x) = \frac{1}{\sqrt{N}} \sum_{i \in [N]} a_{ik} \sigma(w_i^\top x),$$

can be linearized as a dot-product kernel with $k_{FC}(x, y) = h(x^\top y/d)$ for h defined in (67). Moreover, the resulting integral operator K_{FC} is diagonalized in the parity basis as

$$K_{FC}\chi_S = \nu_h(d, |S|)\chi_S,$$

where the coefficients are given by $\nu_h(d, \ell) = \langle h, Q_\ell \rangle_{L^2(\tau)}$ as in (64).

Note that eigenvalues $\nu_h(d, \ell)$ are non-increasing with ℓ, and for fixed ℓ and large d they satisfy $\nu_h(d, \ell) = \Theta_d(d^{-\ell})$. More generally, it can be shown that $\lim_{d \to \infty} d^k \nu_h(d, \ell) = \frac{d^k}{dt^k} h(t)\big|_{t=0}$.

Proof. The first part is a direct consequence of the prior proposition with $\omega = 1$ and $q = \Delta = d$. The second part is due to Lemma 50, and (63). For the statements on eigenvalues, see (Yang & Salman, 2019). □

Proposition 53 (Linearization of a convolutional network)**.** *A convolutional layer followed by a fully connected layer*

$$f_{CNN}(x) = \frac{1}{\sqrt{Nd}} \sum_{i\in[N]} \sum_{k\in[d]} a_{ik}\sigma\left(w_i^\top x_{(k)}^{(q)}\right),$$

can be linearized with the h of (67) as

$$k_{CNN}(x,y) = \frac{1}{d}\sum_{k\in[d]} h\left(\left\langle x_{(k)}^{(q)}, y_{(k)}^{(q)}\right\rangle / q\right).$$

In the Boolean setting, the resulting integral operator K_{CNN} is diagonalized in both the parity and the cyclic basis as

$$K_{CNN}\psi_{m,S} = \begin{cases} \nu_h(q,|S|)\frac{(q+1-\mathrm{diam}(S))_+}{q}\psi_{m,S}, & \text{if } \mathrm{diam}(S)\le q, \\ 0 & \text{otherwise.} \end{cases}$$

where $\nu_h(q,\ell)$ are defined by Proposition 52.

Proof. The first part corresponds to the case $\omega = \Delta = 1$. The second part is due to the expansion of h over the $\mathbb{Q}_\ell$ basis, which leads to (see Eq. (30) in Misiakiewicz & Mei (2022) for details)

$$k_{CNN}(x,y) = \sum_{S\subseteq[d],\mathrm{diam}(S)\le q} q^{-1}\nu_h(q,|S|)(q+1-\mathrm{diam}(S))_+\chi_S(x)\chi_S(y).$$

The fact that K let the $\{S \,|\, |S| = a, \mathrm{diam}(S) = b\}$ invariant, since the eigenvalues only depends on $|S|$ and $\mathrm{diam}(S)$, allows to change from the parity basis to the cyclic basis. □

When pooling is included in the kernel and $\omega > 1$ in (65), then the architecture enforces local translation invariance. As a simple example, consider the setting of global average pooling $\omega = d$ where strict invariance to translations is enforced and parity functions are projected onto their sum of elements of the orbit to form the eigenbasis. In this case, K is no longer diagonalized in the parity basis, but it is diagonal in the basis of cyclic parities.

D.4. Interplay between augmentations and architecture

In the uniform Boolean setting, the interplay between augmentations and architecture is made easy by the fact that many operators K and T commutes.

Lemma 54. *The operator K associated with a dot-product kernel in the uniform Boolean setting commutes with all the operators T that can be built from bitwise noise, cropping, translations or index flip.*

Proof. In the case of a dot-product kernel in the uniform setting, the spaces $V_{d,\ell}$ are eigenspaces of K. Those spaces are left invariant by all the T defined through usual augmentations, since translations and index-flip operations preserve the cardinality of subsets. As a consequence, K and T can be diagonalized in the same basis, hence they commute. □

As a consequence of the previous lemma, the integral operator K associated with the linear model of fully connected layer commute with all the operators T defined for usual augmentations. It is also the case for the convolutional layer with T deriving from random noise, cropping, or translation.[6] As a consequence, the interplay between the architecture and the augmentations can be studied easily thanks to Proposition 4.

Example 10 (Interplay between FC kernel and translation augmentations)**.** *Recall from Example 9 that when sampling translations from a localized window, the eigenvalues of T are of the form $|\hat{p}(m)|^2$ and typically decay with the frequency index m in $\psi_{m,S} = \frac{1}{\sqrt{k_S}}\sum_{k\in[k_S]} e^{2i\pi km/k_S}\chi_{S+k}$ for any set S with no periodicity. In contrast, the eigenvalues $\nu_h(d,|S|)$ of K for eigenfunctions $\psi_{m,S}$ decay as $\Theta_d(d^{-|S|})$, independently of m. Regularization with parameter λ thus shrinks the eigenvalues to $|\hat{p}(m)|^2 - \lambda\nu_h(d,|S|)^{-1}$ after pre-training. This most notably eliminates contributions from eigenfunctions $\psi_{m,S}$ where m is small (i.e., near-invariant) but $|S|$ is large. See Figures 3 and 5 for an illustration.*

[6] Since it lets invariant the orbit of translation.

Example 11 (Interplay between kernel for CNN and translation augmentations)**.** *Consider the setting as before in Example 10 with translations sampled from a localized window. For a single layer CNN with patch width q, eigenfunctions correspond to parity functions χ_S, or cyclic parities $\psi_{m,S}$ where $\mathrm{diam}(S) \leq q$ with corresponding eigenvalue $\nu_h(q,\ell)\frac{q+1-\mathrm{diam}(S)}{d}$. Here, the eigenfunctions $\psi_{m,S}$ of T for S with diameter larger than q are completely eliminated, regardless of the regularization strength λ, . For eigenfunctions $\psi_{m,S}$ where $\mathrm{diam}(S) \leq q$, the CNN shrinks the contribution to $|\hat{p}(m)|^2 - \lambda(\nu_h(q,\ell)\frac{q+1-\mathrm{diam}(S)}{d})^{-1}$, which shrinks more when $\mathrm{diam}(S)$ is larger.*

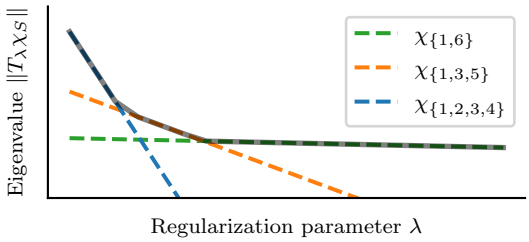

Figure 8. Illustration of the interplay between T and K as a function of λ where K is the NTK of a 2-layer ReLU network and T performs crops of window size 8 on 12-bit inputs. Here we plot eigenvalues of three different parity functions in the eigenbasis of both operators. Parity functions which large diameters have smaller eigenvalues for T (here, the parity function with largest diameter is $\chi_{\{1,6\}}(X) = X_1X_6$). Eigenvalues of K, in contrast, bias towards parities supported over fewer bits. Therefore, small regularization biases towards parities with small diameter whereas added regularization penalizes parities with high cardinality.

D.5. Remark on the Sphere setup

In experiments, we also consider a setup with uniform data on the sphere $\mathcal{X} = \mathbb{S}^{d-1}$, with augmentations consisting of permutations, and a dot-product kernel $\varphi(x)^\top\varphi(y) = h(x^\top y)$. A natural choice of basis functions for $L^2(\mathcal{X})$ in this case are *spherical harmonics* (Efthimiou & Frye, 2014). These consist of homogeneous harmonic polynomials, and similar to the parity case, these can be grouped by degree, leading to orthogonal spaces $V_{d,\ell}$ of spherical harmonics of any degree $\ell \geq 0$, with

$$N(d,\ell) := \dim V_{d,\ell} = \frac{2\ell+d-2}{\ell}\binom{\ell+d-3}{d-2}.$$

It is well-known that for dot-product kernels, K is diagonal in such a basis (Smola et al., 2000; Bach, 2017), with decaying eigenvalues that only depend on the degree ℓ. These are given analogously to the hypercube setting by

$$\nu_h(d,\ell) = \mathbb{E}_{t\sim\tau}[h(t)Q_{\ell,d}(t)],$$

where $Q_{\ell,d}$ are now Legendre (Gegenbauer) polynomials of degree ℓ orthogonal w.r.t. a different measure $d\tau(t) = (1-t^2)^{\frac{d-3}{2}}dt$ over $[-1,1]$.

Since the spaces $V_{d,\ell}$ are left stable by the operator $T = A^\top A$, it is possible to show that there exists a choice of spherical harmonics that also diagonalizes T (see, e.g., Bietti et al., 2021, Lemma 12). We may then see the eigenvalues $\lambda_{\ell,j}$ of T in this basis as capturing the invariance of the corresponding harmonic $Y_{\ell,j}$, in particular $Y_{\ell,j}$ is invariant to all augmentations when $\lambda_{\ell,j} = 1$, and non-invariant or only partially invariant when $\lambda_{\ell,j} < 1$.

Ordering $\lambda_{k,j}$ at fixed k by decreasing j, the interplay between T and K then resembles the one described, e.g., in Figure 3.

E. Experiments

E.1. Implementation details

Previously, we extensively studied the embedding of $\mathcal{H}$ in L^2 defined as $S : \mathcal{H} \mapsto L^2; \theta \to \varphi(\cdot)^\top\theta$. Given samples $(\xi_{ij})_{i\leq n, j\leq m}$, all the action on $\mathcal{H}$ can be reduced to the span of the $\varphi(\xi_{ij})$ (which is known as the representer theorem), and S can be reduced to the embedding $\hat{S} : \mathcal{H} \mapsto \mathbb{R}^{nm}; \theta \to (\frac{1}{nm}\varphi(X_{ij})^\top\theta)_{ij}$. This leads to the implementation

$$\hat{T_\lambda} = \beta + (1-\beta)\hat{T} - \lambda\hat{K}.$$

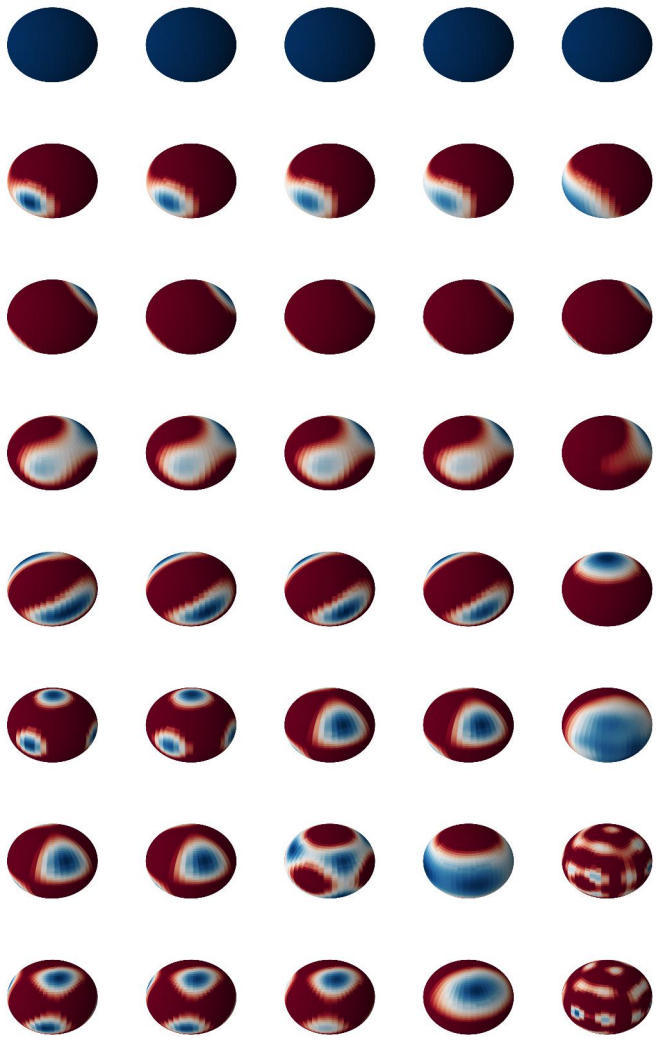

Figure 9. Extending Figure 4. The i-th row representing the i-th eigenfunctions of T_λ (ordered by decreasing eigenvalues). Regularization λ increases over the columns as $\lambda \in \{0, .1, 1, 10, 100\}$. Small λ biases towards functions invariant to the translation augmentation chosen here whereas large λ biases towards smoother functions on the sphere corresponding to low order spherical harmonics in this setting. The last two on the right are artifacts of the instability of the pseudo-inverse for K (leading to the implementation $\varphi K^{-1}\varphi = 0$ while we have defined $\varphi^\top K^{-1}\varphi = +\infty$ when $K\varphi = 0$).

Where, $\hat{T} \in \mathbb{R}^{nm}$ is the matrix equal to the following where we index elements in $\mathbb{R}^{nm}$ by ij with $i \in [n]$ and $j \in [m]$,

$$\hat{T} = I + \sum_{ijk} e_{ij} e_{ik}^\top,$$

and K is the Gram matrix defined as

$$nm \cdot e_{ij}^\top K e_{kl} = k(\xi_{ij}, \xi_{kl}) = \varphi(\xi_{ij})^\top \varphi(\xi_{kl}).$$

Note the the matrix $\hat{T} - I$ can be seen as the adjacency matrix of the graph that connects augmentations if and only if they come from the same input. Equivalently, $\hat{T}$ can be seen as a Laplacian matrix. An eigenvector of $\hat{T}_\lambda$ in $\mathbb{R}^{nm}$ is projected back onto L^2 thanks to $S\hat{S}^{-1} = S\hat{S}(\hat{S}\hat{S}^\top)^{-1} = K_x^\top K^{-1}$ where

$$nmK_x = \left(\varphi(x)^\top \varphi(\xi_{ij})\right)_{ij} \in \mathbb{R}^{nm}.$$

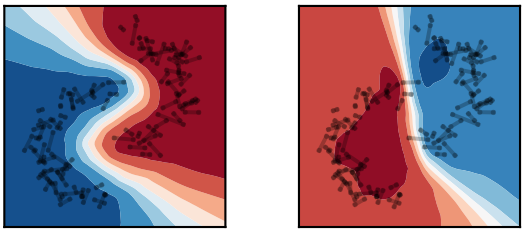

Figure 10. VCReg with Neural networks. Contour plots of the minimizer $\psi : \mathcal{X} \to \mathbb{R}$ of $\mathcal{L}$ for $\beta = 1$ (left) and $\beta = 0$ (right) with a two layer fully connected neural network when $k = 1$, $\mathcal{X} = \mathbb{R}^2$, X is distributed according to a half-moon structure and $\xi = X + \varepsilon$ for a small noise ε. Augmentations are represented as black dots, connected by a line when they provide from the same input X.

E.2. Experiment details for Figure 5

We consider data uniformly distributed on the sphere $\mathbb{S}^{d-1}$ with $d = 8$, augmentations consisting of cyclic shifts of $\{-1, 0, 1\}$, and a dot-product kernel of the form $k(x, y) = (1 + x^\top y)\kappa(x^\top y)$, with $\kappa(u) = 1 - \arccos(u)/\pi$.

The target functions f_ℓ^* are given by:

$$f_1^*(x) = \frac{1}{3} \sum_{j=1}^{3} Q_{1,d}(x_j)$$
$$f_3^*(x) = \frac{1}{d} \sum_{j=1}^{d} Q_{3,d}(x_j),$$

where $Q_{\ell,d}$ are the Gegenbauer polynomials introduced in Appendix D.5. Note that f_3^* is a cyclic-invariant spherical harmonic of degree 3, while f_1^* is a non-invariant spherical harmonic of degree 1 (though is has some local shift stability). Labels on the downstream tasks are generated from the f_ℓ^* without noise.

Figure 5 shows the downstream relative excess risk $\|\hat{f}_n - f_\ell^*\|_{L^2}^2 / \|f_\ell^*\|_{L^2}^2$, approximated over 1500 test datapoints, as a function of the regularization parameter λ used in pretraining. We use the same $n = 300$ samples for pretraining and downstream linear prediction. Pretraining uses all 3 augmentations for each sample, with a representation dimension $k = 20$. The downstream problem is solved with kernel ridge regression using the induced kernel from pretraining, and the ridge parameter is tuned on test samples to avoid dealing with model selection issues.

E.3. Experiment details for Figure 6

Figure 6 considers a classification problem involving four classes with a pretraining task specifically constructed to design a representation $\psi : \mathcal{X} \to \mathbb{R}^k$ for $k = 4$ that solves this particular classification problem. The dataset we consider is the halfmoon dataset, where $X = Z + \mathbf{1}_{\langle Z, e_1 \rangle > 0} e_2 + U$, $Z \sim \mathcal{U}\left(\mathbb{S}^2\right)$, and $U \sim \mathcal{N}(0, \sigma^2 I)$ for $\sigma = 0.1$. Augmentations apply Gaussian noise, $\xi = X + V$ for $V \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.1$. This setting corresponds to that with a Laplacian where

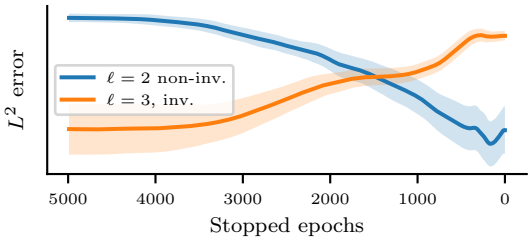

Figure 11. Behavior of Figure 5 with a neural network. The regularization parameter λ is replaced by early stopping of SGD. We consider a neural network with two hidden layers, both made of 200 neurons. Optimization was performed with gradient descent with a constant step-size. Randomness due to weights initialization is averaged over 100 trials, the standard deviation being shown on the Figure.

$\mathcal{L}(\psi) \simeq \|\nabla\psi\|^2_{L^2(\rho_{\mathcal{X}})}$. As a consequence, the ideal ψ will correspond to the top eigenvalues of the Laplacian. I.e., the first two span the constant functions on both moons, the next two are waves with a single oscillation on a given moon, etc. In essence, one can view the harmonics on $L^2([0,1])$ as $x \to \cos(2i\pi\omega x + \chi)$ for $\chi \in \{0, \pi/2\}$ and $\omega \in \mathbb{N}$, deforming the segment $[0,1]$ to match one moon, and duplicating this basis on the other moons. In this setting, eigenfunctions are not analytic, since analytic functions cannot be dissociated on two different manifold (e.g., a locally constant analytic function is globally constant). As a consequence, searching for the eigenfunction with the radial basis function kernel (Ψ only contains analytical function in this case (Sun & Zhou, 2008)) requires proper tuning of the regularization parameter as a function of the number of samples. This explains our choice of the exponential kernel in this experiment, which corresponds to $\varphi(x)^\top\varphi(y) = \exp(-\|x-y\|/\sigma)$ and is associated with the looser Sobolev space that is still a Reproducing kernel Hilbert space (in $\mathbb{R}^2$, this is H^1). This improves the learning of the top eigenfunctions of T without varying λ, better illustrating the convergence rates of Theorems 1, 2 and 3.

In our experiments, we fixed $\lambda = 10^{-3}$ and the scale of the exponential kernel σ to be about one fifth of the problem diameter. We plot the eigenfunctions of T derived empirically with $n_{pre} = 2000$ samples in Figure 13. The classification tasks aims to learn the four classes described on the left of Figure 12. Class labels include some noise as indicated by the level lines of the conditional probability of Y as a function of X shown in the middle of Figure 12. A training set example is shown on the right of this figure with $n_{down} = 100$. In the experiments we fix $k = 5$, which ensures that there is strong correlation in performance between the pretraining and downstream tasks. The downstream task is optimized with a least-squares surrogate: we learn $g : \mathcal{X} \to \mathbb{R}^4$ that minimizes the least-square error $\mathbb{E}[\|g(X) - e_Y\|^2]$ before decoding it as $f(X) = \arg\max_{i\in[4]} g_i(X)$ to get an estimate of the ideal mapping $f^* : \mathcal{X} \to \mathcal{Y}$. We report the downstream generalization error on both the least-squares (surrogate) loss and the 0-1 loss on Figure 14. This error is computed as the average over 100 trials on the pretraining task and 200 trials on the downstream task.

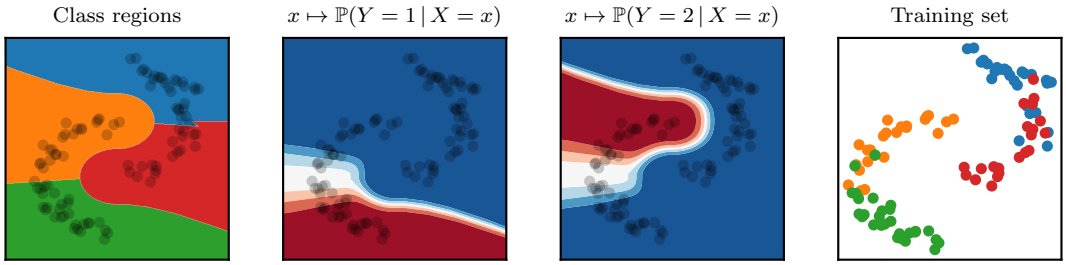

Figure 12. Setting of Figure 6. The downstream task consists in learning four classes in $\mathcal{X} = \mathbb{R}^2$ with are represented on the left. Those classes are generated with noise. The level lines of the conditional distribution of Y given X are represented on the middle for the left moons; the right moon follows the same structure. A training set example is on the right.

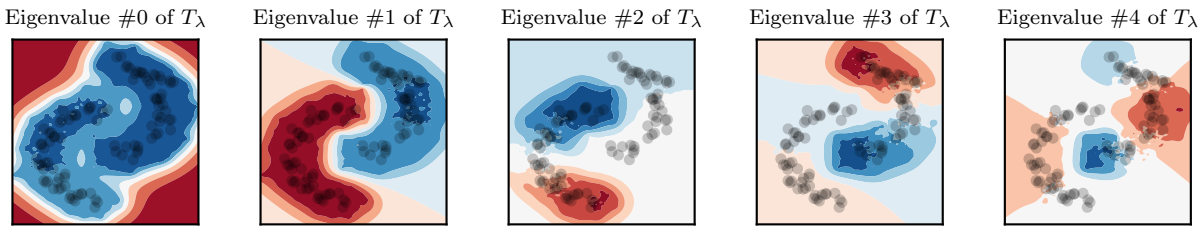

Figure 13. Eigenvalues of T_λ estimated empirically with 2000 pretraining samples on the problem that yield the empirical rates displayed on Figure 6.

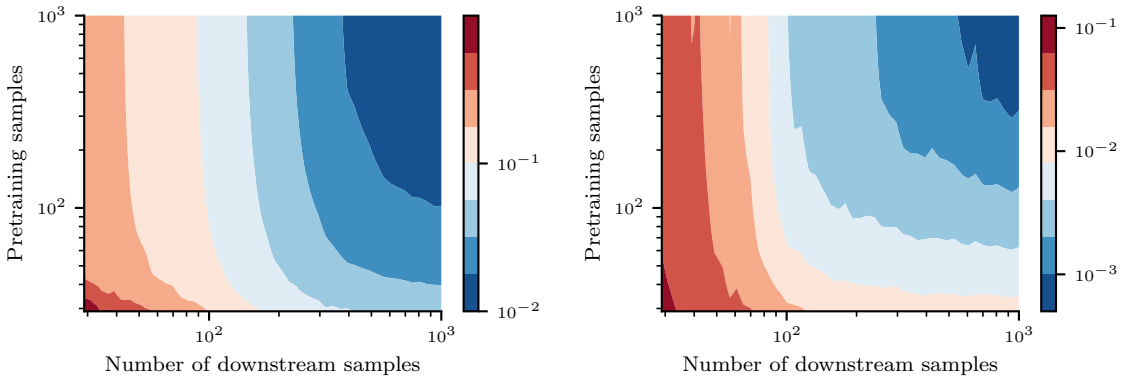

Figure 14. Averaged downstream error computed over 100 trials on the pretraining task and 200 trials on the downstream task, for both the least-squares loss (right) and the 0-1 loss (left).