Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2008-030
CBCL-272

April 4, 2008

On a model of visual cortex: learning invariance and selectivity

Andrea Caponnetto,, Tomaso Poggio and , and Steve Smale

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

CBCL Paper
April 4, 2008

On a model of visual cortex: learning invariance and selectivity from image sequences

Andrea Caponnetto[◇], Tomaso Poggio[†] and Steve Smale[‡]

Department of Mathematics, City University of Hong Kong[◇]
CBCL, McGovern Institute, Artificial Intelligence Lab, BCS, MIT[†]
Toyota Technological Institute at Chicago and University of California, Berkeley[‡]

Abstract

In this paper we present a class of algorithms for similarity learning on spaces of images. The general framework that we introduce is motivated by some well-known hierarchical pre-processing architectures for object recognition which have been developed during the last decade, and which have been in some cases inspired by functional models of the ventral stream of the visual cortex. These architectures are characterized by the construction of a hierarchy of "local" feature representations of the visual stimulus. We show that our framework includes some well-known techniques, and that it is suitable for the analysis of dynamic visual stimuli, presenting a quantitative error analysis in this setting.

1 Introduction

During the last decade a great number of techniques have been proposed to learn similarity measures. Instances include techniques which utilize clouds of unlabelled input samples [24] [9] [4] [5], and techniques utilizing various kinds of additional side information [6], as homonymous and heteronymous example pairs [28] [3] [23] [7] [15] [20] [25], or invariances in pattern recognition [11] [18] [19] [1] [26].

Some of this algorithms have been designed on the basis of physiological and psychophysical evidence, trying to model the functional structure of primary visual cortex. In this paper we mainly refer to the algorithms of Serre et al. [21] and of Mutch and Lowe [17], which in turn extend the model of Riesenhuber and Poggio [19]. These are some of the most recent models which attempt to describe in a quantitative way information processing in the ventral stream of the visual cortex, and which include also convolutional networks [13] and Neocognitrons [11].

All these models, starting from a image layer, successively perform the computation of the "neural responses" in higher layers of the network, alternating layers of "**S**" units and "**C**" units. This alternating structure is analogous to the V1 simple and complex cells discovered by Hubel and Wiesel in the late sixties [12]. Broadly speaking, the function of "**S**" units is to increase selectivity relative to relevant variations of the input stimulus, while "**C**" units increase the invariance with respect to translations and scalings.

More formally, the response Y of a simple "**S**" unit receiving the pattern of "synaptic inputs" $(X_1, X_2, \dots)$ from the previous layer is given by

$$Y = \exp\left(-\lambda \sum_j (W_j - X_j)^2\right), \tag{1}$$

where λ defines the sharpness of the tuning around the preferred stimulus of the unit corresponding to the weight vector $W = (W_1, W_2, \dots)$.

Conversely, the "**C**" units are responsible for the pooling operation. That is, the response Y of a complex unit corresponds to the strongest input X_j from its afferents in the previous "**C**" layer

$$Y = \max_j X_j. \tag{2}$$

The overall goal of the model is to increase feature invariance while maintaining specificity using a multi-layer hierarchical architecture. Most notably, models originally motivated by physiological and psychophysical evidence have been proven extremely effective in pattern recognition tasks and in specific contexts comparable to state-of-the-art algorithms [22] [21].

In this paper we present a class of hierarchical algorithms for learning similarities and invariances on spaces of images which, to some extent, generalizes the type of algorithm described above.

In Section 2 we formally introduce our framework. In Section 3 we show how to draw a parallel between the framework and the models of the ventral stream. In Section 4 we develop an error analysis in presence of dynamic visual stimuli, and finally in Section 5 we describe in detail how to implement an algorithm using samples from streams of images. All the proofs of the presented results are collected in the Appendix.

2 The hierarchy of patches and the local feature mappings

The type of multi-layer architecture that we are going to describe is aimed at associating to an "image" $f \in \mathrm{Im}(R)$ on the "retina" $R \subset \mathbb{R}^2$, an ensemble of "local" feature representations $\phi_v(f)$. The "local" representation $\phi_v(f)$ takes value on the separable Hilbert space $\mathcal{H}_v$, and encodes information relative to the properties of f over a small patch, or "receptive field", v of R. The patches, which we assume to be disks in R, are organized in layers according to their size. The lower layer $V(0)$ is a (finite) collection of disks in R of radius σ_0, the upper layer $V(1)$ a (finite) collection of disks of radius $\sigma_1 > \sigma_0$ and so on up to the uppermost layer V(K) containing one or more disks of radius σ_K, with

$$\sigma_0 < \sigma_1 < \cdots < \sigma_{K-1} < \sigma_K.$$

The layers of patches $V(0), V(1), \ldots, V(K)$ are equipped with a natural tree structure. We say that a patch v in $V(j)$ is a child of w in $V(j+1)$, and we write $v \in \mathrm{Ch}(w)$, whenever $v \subset w$. In the following we always assume that $\mathrm{Ch}(w)$ is nonempty for every $w \in V(j)$ and $j > 0$.

For sake of simplicity we also assume that the patches are evenly distributed, in the sense that for every pair of patches v and v' in $V(j)$, there exists a translation of $\mathbb{R}^2$ which maps v onto v', and every patch in the sub-tree of root v onto a patch in the sub-tree of root v'.

The ground property of the feature representations $\phi_w(f) : \mathrm{Im}(R) \to \mathcal{H}_w$ is their hierarchical organization, in fact $\phi_w(f)$ depends on the image f only through the feature representations $\phi_v(f)$ localized on the patches v in $\mathrm{Ch}(w)$. The construction of the feature mapping ϕ_w from the mappings ϕ_v is conveniently implemented in various steps.

First, we define the direct sum

$$\mathcal{H}_{\widehat{w}} = \bigoplus_{v \in \mathrm{Ch}(w)} \mathcal{H}_v$$

the Hilbert space of "normalized" inner product

$$\langle (h_1, h_2, \ldots), (h'_1, h'_2, \ldots) \rangle_{\widehat{w}} = \frac{1}{|\mathrm{Ch}(w)|} \sum_{i=1}^{|\mathrm{Ch}(w)|} \langle h_i, h'_i \rangle_{v_i}$$

where we have enumerated the patches $\{v_1, v_2, \dots\}$ in $\mathrm{Ch}(w)$, and $\langle \cdot, \cdot \rangle_v$, $\langle \cdot, \cdot \rangle_{\widehat{w}}$ denote the scalar products in $\mathcal{H}_v$ and $\mathcal{H}_{\widehat{w}}$ respectively. Hence the "child" mappings $\{\phi_v | v \in \mathrm{Ch}(w)\}$ are stacked in the "direct sum" mapping $\phi_{\widehat{w}} : \mathrm{Im}(R) \to \mathcal{H}_{\widehat{w}}$

$$\forall f \in \mathrm{Im}(R) \quad \phi_{\widehat{w}}(f) = (\phi_{v_1}(f), \phi_{v_2}(f), \dots). \tag{3}$$

Second, a linear operator $\Pi_w : \mathcal{H}_{\widehat{w}} \to \mathcal{H}_{\widehat{w}}$ is constructed using available sample data; we generally assume that Π_w is bounded, symmetric and positive semi-definite. In the following sections we will present two detailed examples of such a construction (see Definitions 3.1 and 5.1).

Finally, given a positive constant λ_{j+1} depending only on the depth $j+1$ of the w's layer, we postulate that for every pair of images f and f', the "parent" mapping ϕ_w satisfies

$$\langle \phi_w(f), \phi_w(f') \rangle_w \quad = \quad \exp\left(-\lambda_{j+1} \left\| \Pi_w(\phi_{\widehat{w}}(f) - \phi_{\widehat{w}}(f')) \right\|_{\widehat{w}}^2\right) \tag{4}$$

In order to prove that Equation (4) defines $\mathcal{H}_w$ and ϕ_w up to isometries, we need the following assumption on the space of images.

Hypothesis 2.1 *We assume that* $\mathrm{Im}(R)$ *is a compact subset of* $L^2(R, \{-1, 1\})$*, the space of square-integrable functions on* R *taking values in* $\{-1, 1\}$.

Using Hypothesis 2.1 we can prove the well-definiteness of ϕ_w.

Proposition 2.1 *For every* $v \in \mathrm{Ch}(w)$*, let the separable Hilbert space* $\mathcal{H}_v$ *and the continuous mapping* $\phi_v : \mathrm{Im}(R) \to \mathcal{H}_v$ *be given. Let* Π_w *be a bounded linear operator on* $\mathcal{H}_{\widehat{w}}$*. Then there exists a separable Hilbert space* $\mathcal{H}_w$ *and a continuous mapping* $\phi_w : \mathrm{Im}(R) \to \mathcal{H}_w$ *which fulfill Equation (4). Moreover for any other mapping* $\phi'_w : \mathrm{Im}(R) \to \mathcal{H}'_w$ *fulfilling Equation (4) there exists a unitary operator* $U : \mathcal{H}_w \to \mathcal{H}'_w$ *such that* $\phi'_w = U \circ \phi_w$.

Note that the proof of Proposition 2.1 gives an explicit construction of ϕ_w as the canonical embedding of $\mathrm{Im}(R)$ into the reproducing kernel Hilbert space of kernel $K_w(f, f') := \langle \phi_w(f), \phi_w(f') \rangle_w$ given by Equation (4). For sake of simplicity in the following we will often use the simplified notations $d_v(f, f') := \phi_v(f) - \phi_v(f')$ and $d_{\widehat{w}}(f, f') := \phi_{\widehat{w}}(f) - \phi_{\widehat{w}}(f')$.

We have seen that Equations (3) and (4) define the feature mappings on the $(j+1)$-st layer from feature mappings on the j-th layer, therefore in order to implement a recursive construction we have to define the lowest layer's feature mappings. These mappings ϕ_v, for all $v \in V(0)$, are naturally induced by the L^2 metrics on $\mathrm{Im}(R)$, we define $\mathcal{H}_v = L^2(R, \{-1, 1\})$ and ϕ_v the identity mapping

$$\forall f, f' \in \mathrm{Im}(R) \quad \langle \phi_v(f), \phi_v(f') \rangle_v = \frac{1}{\mathcal{A}(v)} \int_v f(x) f'(x) dx \tag{5}$$

where $\mathcal{A}(v)$ is the area of v. Note that since by Hypothesis 2.1 the images take values in $\{-1, 1\}$, by Equations (4) and (5), for every patch v and image f it holds

$$\|\phi_v(f)\|_v = 1.$$

So far we have briefly described the general recipe to construct the feature mappings on the hierarchy of patches, in the following we will show how to specialize this construction to two particularly interesting cases.

3 The "soft" model

In this section we show that the presented framework is suitable for the description of models similar to the one presented in the Introduction. At this aim we proceed to the definition of the operator Π_w for some $w \in V(j+1)$, with $0 \leq j \leq K-1$. First we need to introduce a formal notion of "templates": a (finite) collection T_j of images in $\mathrm{Im}(R)$. Each image $t \in T_j$ represents a basic shape or "template" involved in the construction of the "neural responses" to visual stimuli on the patch w. For example, templates relative to the layer $j = 0$ might be (as in [21]) simple oriented bars, while templates relative to deeper layers might be complex combinations of oriented bars forming contours or boundary conformations. We assume that the templates in T_j are normalized and centered on an arbitrary reference patch $v^\star \in V(j)$, and we also assume without loss of generality that they take value 0 off this reference patch. Therefore "templates" $H_v(t)$ centered on general patches $v \in V(j)$ are defined as follows

$$(H_v(t))(x) = \begin{cases} t(h_v(x)) & x \in v, \\ 0 & x \notin v, \end{cases}$$

where h_v is the translation in $\mathbb{R}^2$ which maps v onto $v^\star$. We need the weak technical assumption

Hypothesis 3.1 *For every $v \in V(j)$ and $t \in T_j$, the function $H_v(t)$ belongs to $\mathrm{Im}(R)$.*

Hypothesis 3.1 is required in the definition of the operator Π_w, which is expressed in terms of the vectors $\phi_v(H_v(t))$ with $v \in \mathrm{Ch}(w)$ and $t \in T_j$.

Definition 3.1 *For every $w \in V(j+1)$, with $0 \leq j \leq K-1$, let Π_w be the bounded symmetric positive semi-definite operator on $\mathcal{H}_{\widehat{w}}$, defined by*

$$\Pi_w^2 = \operatorname*{Av}_{t \in T_j} \mathrm{Q}[I_{\widehat{w}}(t)] \tag{6}$$

where the average Av is relative to the uniform probability measure on T_j, $\mathcal{H}_{\widehat{w}} \ni I_{\widehat{w}}(t) = ((\phi_{v_1}(H_{v_1}(t)), \phi_{v_2}(H_{v_2}(t)), \dots))$, and $\mathrm{Q}[h]$ is the projection operator $\mathrm{Q}[h]u = h \langle h, u \rangle_{\widehat{w}}$.

Note that Π_w are bounded since

$$\left\|\Pi_w^2\right\| \leq \underset{t\in T_j}{\mathrm{Av}} \|\mathrm{Q}[I_{\widehat{w}}(t)]\| = \underset{t\in T_j}{\mathrm{Av}} \|I_{\widehat{w}}(t)\|_{\widehat{w}}^2 = \underset{t\in T_j}{\mathrm{Av}} \underset{v\in \mathrm{Ch}(w)}{\mathrm{Av}} \|\phi_v(H_v(t))\|_v^2 = 1.$$

From this Definition and Equations (3) and (4) it follows that identifying a complex unit of layer j with a pair $(w,t) \in V_{j+1} \times T_j$, and denoting by $f \in \mathrm{Im}(R)$ the visual stimulus, the input to the unit can be represented by the vector of components

$$X_v(\mathbf{C}(w,t)) := K_v(f, H_v(t)) \qquad v \in \mathrm{Ch}(w)$$

the unit executes a "soft" version of the pooling operation in Equation (2), returning as output the average, rather than the maximum, of its inputs

$$Y(\mathbf{C}(w,t)) := \underset{v\in \mathrm{Ch}w}{\mathrm{Av}} X_v(\mathbf{C}(w,t)).$$

At the next stage the vector of components

$$X_t(\mathbf{S}(w,\tilde{t})) = Y(\mathbf{C}(w,t)) \qquad t \in T_j$$

is elaborated by the simple unit of layer $j+1$ represented by a pair $(w,\tilde{t}) \in V(j+1)\times T_{j+1}$. The unit executes some function and outputs the result $Y(\mathbf{S}(w,\tilde{t}))$.

Finally the outputs $Y(\mathbf{S}(w,\tilde{t}))$ will serve as input components to the complex unit of layer $j+1$ parameterized by the pair $(\tilde{w},\tilde{t}) \in V(j+2) \times T_{j+1}$ with $\tilde{w}$ the parent of w

$$X_w(\mathbf{C}(\tilde{w},\tilde{t})) = Y(\mathbf{S}(w,\tilde{t})) \qquad w \in \mathrm{Ch}(\tilde{w})$$

and so on.

The function executed by the simple unit $\mathbf{S}(w,\tilde{t})$ is simply obtained using Definition 3.1 and Equation (4), in fact we get

$$\begin{aligned}
Y(\mathbf{S}(w,\tilde{t})) &= X_w(\mathbf{C}(\tilde{w},\tilde{t})) = K_w(f, H_w(\tilde{t})) \\
&= \exp\left(-\lambda_{j+1} \left\langle \phi_{\widehat{w}}(f) - \phi_{\widehat{w}}(H_w(\tilde{t})), \Pi_w^2(\phi_{\widehat{w}}(f) - \phi_{\widehat{w}}(H_w(\tilde{t})))\right\rangle_{\widehat{w}}\right) \\
&= \exp\left(-\lambda_{j+1} \underset{t\in T_j}{\mathrm{Av}} (\left\langle \phi_{\widehat{w}}(H_w(\tilde{t})), I_{\widehat{w}}(t)\right\rangle_{\widehat{w}} - \left\langle \phi_{\widehat{w}}(f), I_{\widehat{w}}(t)\right\rangle_{\widehat{w}})^2\right) \\
&= \exp\left(-\lambda_{j+1} \underset{t\in T_j}{\mathrm{Av}} (\underset{v\in \mathrm{Ch}w}{\mathrm{Av}} (K_v(H_w(\tilde{t}), H_v(t)) - K_v(f, H_v(t))))^2\right) \\
&= \exp\left(-\lambda_{j+1} \underset{t\in T_j}{\mathrm{Av}} (W_t(\mathbf{S}(w,\tilde{t})) - X_t(\mathbf{S}(w,\tilde{t})))^2\right)
\end{aligned}$$

with $W_t(\mathbf{S}(w,\tilde{t}))$ the weight vector of the unit, defined by

$$W_t(\mathbf{S}(w,\tilde{t})) := \underset{v\in \mathrm{Ch}w}{\mathrm{Av}} K_v(H_w(\tilde{t}), H_v(t)) \qquad t \in T_j.$$

This last relations are analogous to Equation (1) in the Introduction

This connection with the usual formalism of the model of the ventral stream shows that the framework presented in Section 2 is general enough to encompass that type of algorithm. In Section 5 we will present a different choice for the operators Π_w, but first we give a quantitative error analysis for the general algorithm in presence of dynamics of the input visual stimulus.

4 Dynamic visual stimuli and error analysis

The algorithms referred to in the Introduction were to some extent designed on the basis of physiological and psychophysical evidence, trying to model the functional structure of the primary visual cortex. These algorithms have been proved competitive in terms of performance on a variety pattern recognition applications, however so far no solid mathematical theory accounting for their effectiveness is available. A tentative step in this direction has been proposed by Földiak [10] and developed by Wiskott [27]. Theses authors start their analysis from the general principle of "slowness", according to which the *sensory signals vary more quickly than their significance*. In this perspective, the local feature representations would be able to filter out the "fast" components of the input signals and retain the more significative "slow" components.

Recently, Maurer [16] by elaborating on this idea, developed a new dimensionality reduction technique based on *hyperbolic-PCA* [14]. In [16] the time sequence of sensory signals is modelled by a stationary stochastic process taking values over $\mathrm{Im}(R)$, and a projector on $\mathrm{Im}(R)$ is selected on the basis of a criterion which rewards data-variance and penalizes abrupt changes of the projected signal. Using a representation of finite dimensional projections as bounded linear functionals on the space of Hilbert-Schmidt operators on $\mathrm{Im}(R)$, [16] gives some *PAC-type performance guarantees* for the resulting feature maps.

In this section we develop an error analysis for the performance of our hierarchy of feature mappings based on the framework presented in [16].

The time evolution of the visual stimulus is modelled by a *discrete-time stationary process* taking values in $\mathrm{Im}(R)$

$$\mathbf{F} = \{F_\tau\}_{\tau\in\mathbb{Z}}$$

Here the integer τ represents time, and the stationarity assumption means that for any δ, the shifted process $\mathbf{F}_\delta = \{F_{\tau+\delta}\}_{\tau\in\mathbb{Z}}$ has the same distribution has $\mathbf{F}$. We will often need to introduce random variables (r.v.) independent and identically distributed (i.i.d.) with $\mathbf{F}$, we denote these r.v. by $\mathbf{F}' = \{F'_\tau\}_{\tau\in\mathbb{Z}}$.

We will also assume that the $\mathcal{H}_v$-valued r.v. $\phi_v(F_0)$ for all the patches v in the layer $V(j)$ are identically distributed up to isometry, in the sense that

Hypothesis 4.1 *Let $0 \le j \le K$, and $v, v' \in V(j)$. Then there exists an isometric isomorphism $U : \mathcal{H}_v \to \mathcal{H}_{v'}$ such that the random variables $(\phi_v(F_0), \phi_v(F_1), \dots)$ and $(U\phi_{v'}(F_0), U\phi_{v'}(F_1), \dots)$ are identically distributed.*

Since in Section 2 we have already assumed that the patches are evenly arranged, Hypothesis 4.1 essentially amounts to the assumption that the distribution of the visual stimuli restricted to some mask $M \subset R$ is not affected by translations of M. Under these hypothesis we can identify spaces $\mathcal{H}_v$ and operators Π_v relative to different patches v of the same layer.

We are now ready to introduce the pattern recognition tasks which will be used in the assessment of the performance of our algorithms. We associate to

every $v \in V(j)$ a denumerable partition of $\mathrm{Im}(R)$

$$\mathcal{C}_v = \{C_v(k)\}_k.$$

Different $C_v(k)$ represent classes of images whose restrictions on v share the same pattern or category. For example for v on a superficial layer (small j), $\mathcal{C}_v$ might be a partition of $\mathrm{Im}(R)$ according to local properties (such as texture, main directional orientation or color) within v. On the contrary on deeper layers (large j) $\mathcal{C}_v$ might represent some complex categorization of the image in v (*e.g.* separating cats from dogs).

Broadly speaking the partition $\mathcal{C}_v$ codes the range of *significance* of a visual stimulus at the characteristic spatial scale of the patch $v \in V(j)$ (i.e. the radius σ_j). According to the general principle that *sensory signals vary more quickly than their significance*, we expect that within intervals of time-length τ_j characteristic of the the layer j, the typical stimulus might be subject to considerable changes but should persist within some fixed $C_v(k)$. Following [16] we express this believe by the following Hypothesis.

Hypothesis 4.2 *For every $0 \leq j \leq K$ there exists a positive integer τ_j such that*

$$\forall k \quad \forall A, B \subseteq C_v(k) \quad \mathbb{P}\left[F_{\tau_j} \in B | F_0 \in A\right] \geq \mathbb{P}\left[F_{\tau_j} \in B\right].$$

It is intuitive that the larger is the spatial scale σ_j of a layer, the larger will be the characteristic time of persistency τ_j (texture and color of a tiny detail of an object may change rapidly while the object retain its overall identity). Therefore we may expect the chain of inequalities

$$\tau_0 < \tau_1 < \cdots < \tau_{K-1} < \tau_K,$$

for simplicity in Proposition 4.2 we will assume that τ_{j+1} is a multiple of τ_j.

The main results of this section, Propositions 4.1 and 4.2, give bounds on the error probability of a simple algorithm using the feature mapping ϕ_v to identify similarity or dissimilarity relations between couples of randomly drawn images. Given a threshold parameter $\sigma \in [0, 2]$, and two images f and f', the algorithm compares the distance $\|d_v(f, f')\|_v^2$ over σ. If the distance is less than σ the two images are identified as similar, otherwise they are classified as dissimilar. The error probability is computed assuming that f and f' are drawn independently from the distribution of F_0. Therefore we define

$$\mathrm{Err}_v(\sigma) = \mathbb{P}\left[\|d_v(F_0, F_0')\|_v^2 < \sigma \nLeftrightarrow (F_0, F_0') \in \bigcup_k C_v(k) \times C_v(k)\right]$$

The first result does not involve the hierarchical structure of the patches. It shows that for a suitable value $\sigma^\star$ of the threshold parameter σ, the error $\mathrm{Err}_v(\sigma^\star)$ is bound by a simple expression, $\check{\mathrm{Err}}_v$, involving "variance" and "persistency" on time-scale τ_j of the feature representation $\phi_v(f)$.

Proposition 4.1 *For every $v \in V(j)$, define the quantities*

$$\epsilon_v = \sum_k \mathbb{P}\left[F_0 \in C_v(k)\right]^2,$$

and

$$\sigma^* = 2\left(1 + \sqrt{\frac{\mathbb{E}\left[K_v(F_0, F_0')\right]}{\mathbb{E}\left[\|d_v(F_0, F_{\tau_j})\|_v^2\right]}}\right)^{-1}.$$

then it holds

$$\mathrm{Err}_v(\sigma^*) \leq \check{\mathrm{Err}}_v := \left(\sqrt{\mathbb{E}\left[K_v(F_0, F_0')\right]} + \sqrt{\mathbb{E}\left[\|d_v(F_0, F_{\tau_j})\|_v^2\right]}\right)^2 - \epsilon_v.$$

The term $\mathbb{E}\left[\|d_v(F_0, F_{\tau_j})\|_v^2\right]$ is related to the "persistency" of the feature representation, its contribution to the error gets small when the map ϕ_v filters out the features that typically vary on a interval of time-length τ_j or shorter. On the other hand $\phi_v(f)$ should retain as much "information" as possible about f in order to have large variance and small $\mathbb{E}\left[K_v(F_0, F_0')\right] = 1 - \mathrm{Var}(\phi_v(F_0))$. It is interesting to note that when ϕ_v is the ideal classifier

$$\phi_v(f) = e_k \Leftrightarrow f \in C_v(k)$$

where the vectors e_k form an orthonormal system (that is $\langle e_h, e_k \rangle_v = \delta_{hk}$), then $\mathbb{E}\left[K_v(F_0, F_0')\right] = \epsilon_v$; and if the categories $C_v(k)$ are "persistent" (in the sense that $\mathbb{E}\left[\|d_v(F_0, F_{\tau_j})\|_v^2\right] \to 0$), then $\mathrm{Err}_v \to 0$ and our bound is tight.

The second result relates the value $\check{\mathrm{Err}}_w$ relative to the patch $w \in V(j+1)$ and the value $\check{\mathrm{Err}}_v$ relative to the children patches $v \in V(j)$. Note that by Hypothesis 4.1 the quantities $\mathbb{E}\left[K_v(F_0, F_0')\right]$ and $\mathbb{E}\left[\|d_v(F_0, F_{\tau_j})\|_v^2\right]$ depend on v only through the depth of its layer, and therefore the same holds for the sum $\check{\mathrm{Err}}_v + \epsilon_v$.

The increase of error rate from one layer to the next is expressed in terms of the two parameters

$$a_w(\Pi_w) := \frac{\mathbb{E}\left[\|\Pi_w d_{\widehat{w}}(F_0, F_0')\|_{\widehat{w}}^2\right]}{\mathbb{E}\left[\|d_{\widehat{w}}(F_0, F_0')\|_{\widehat{w}}^2\right]} \qquad b_w(\Pi_w) := \frac{\mathbb{E}\left[\|\Pi_w d_{\widehat{w}}(F_0, F_{\tau_{j+1}})\|_{\widehat{w}}^2\right]}{\mathbb{E}\left[\|d_{\widehat{w}}(F_0, F_{\tau_{j+1}})\|_{\widehat{w}}^2\right]} \quad (7)$$

which quantify the relative reductions of "variance" and "persistency" of feature representations due to the action of an operator Π_w.

Proposition 4.2 *For every $0 \leq j < K$ and $w \in V(j+1)$, let τ_{j+1} be a multiple of τ_j, Π_w be a bounded symmetric positive semi-definite operator fulfilling*

$$\|\Pi_w\| \leq 1, \quad (8)$$

and

$$\lambda_{j+1} \leq \frac{1}{2b_w(\Pi_w)}\left(\frac{\tau_j}{\tau_{j+1}}\right)^2, \quad (9)$$

then for any $v \in \mathrm{Ch}(w)$

$$\sqrt{\check{\mathrm{Err}}_w + \epsilon_w} \quad \leq \quad \sqrt{\check{\mathrm{Err}}_v + \epsilon_v} + \sqrt{1 - a_w(\Pi_w) + \exp(-2\lambda_{j+1})}. \tag{10}$$

The previous Proposition gives a range of suitable values for the parameter λ_{j+1}, however in the light of the presented error bound, the best choice corresponds to the minimum value of the inter-layer performance degradation term

$$\delta_w(\Pi_w, \lambda_{j+1}) := 1 - a_w(\Pi_w) + \exp(-2\lambda_{j+1}), \tag{11}$$

that is

$$\lambda_{j+1} = \frac{1}{2b_w(\Pi_w)} \left(\frac{\tau_j}{\tau_{j+1}} \right)^2. \tag{12}$$

Note that, since $\|\Pi_w\| \leq 1$, by Equations (7), $a_w(\Pi_w)$ and $b_w(\Pi_w)$ are numbers in the interval $[0, 1]$. The ideal choice of the operator Π_w is the one which minimizes the degradation term $\delta_w(\Pi_w, \lambda_{j+1})$ in the error bound, that would correspond to $b_w(\Pi_w)$ close to 0 (thorough filtering out of "fast" features) and $a_w(\Pi_w)$ close to 1 (retaining as much variance of the representation as possible). In the next Section we will develop this criteria for the choice of the operators Π_w, giving an alternative to the option of Definition 3.1 in Section 3.

5 An algorithm for dynamic stimuli

Following the discussion at the end of the previous Section, we now proceed to the presentation of an alternative to Definition 3.1 for the operator Π_w. A by-product of the proposed approach is a choice for the "tuning sharpness" parameter λ_{j+1}. The Definition 5.1 for Π_w and λ_{j+1} below, follows naturally from Proposition 4.2. Then in Proposition 5.1, under a suitable technical condition, a spectral characterization of Π_w is presented. This characterization is expressed in terms of an unknown real parameter b and some averages w.r.t. the stochastic process $\mathbf{F}$. We conclude the Section with a discussion on how to use the spectral characterization given in Proposition 5.1 to actually estimate Π_w from a finite set of sample images suitably sampled from $\mathbf{F}$. We will not attempt a quantitative assessment of the error introduced by this estimation step from samples; some results in this direction can be found in [14] and [16].

The following Definition is directly motivated by the text of Proposition 4.2, and it is aimed at improving the bound (10).

Definition 5.1 *For every* $w \in V(j+1)$, *with* $0 \leq j \leq K-1$, *let* Π_w *be a bounded symmetric positive semi-definite operator on* $\mathcal{H}_{\widehat{w}}$, *and* λ_{j+1} *a positive number which minimize the inter-layer performance degradation* $\delta_w(\Pi_w, \lambda_{j+1})$ *defined by Equation (11), the minimization being subject to the constraints (8) and (9).*

We state below a result which gives a spectral characterization for the solution of the minimization problem stated in Definition 5.1.

Proposition 5.1 *Let us assume that for some operator Π_w defined according to Definition 5.1, it holds*

$$b := b_w(\Pi_w) \leq \frac{1}{2}\left(\frac{\tau_j}{\tau_{j+1}}\right)^2.$$

Moreover let us define the bounded symmetric positive semi-definite operators on $\mathcal{H}_{\widehat{w}}$

$$A_w = \frac{\mathbb{E}\left[\mathrm{Q}[d_{\widehat{w}}(F_0, F_0')]\right]}{\mathbb{E}\left[\|d_{\widehat{w}}(F_0, F_0')\|_{\widehat{w}}^2\right]} \qquad B_w = \frac{\mathbb{E}\left[\mathrm{Q}[d_{\widehat{w}}(F_0, F_{\tau_{j+1}})]\right]}{\mathbb{E}\left[\|d_{\widehat{w}}(F_0, F_{\tau_{j+1}})\|_{\widehat{w}}^2\right]}$$

where $\mathrm{Q}[h]$ is the projection operator $\mathrm{Q}[h]u = h\,\langle h, u\rangle_{\widehat{w}}$, and introduce the functions

$$\varphi(x) = \exp\left(-\frac{1}{x}\left(\frac{\tau_j}{\tau_{j+1}}\right)^2\right)$$

and

$$\theta(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0. \end{cases}$$

Then the following pair fulfills Definition 5.1

$$\bar{\Pi}_w = \theta(A_w - \varphi'(b)B_w), \qquad \bar{\lambda}_{j+1} = \frac{1}{2b}\left(\frac{\tau_j}{\tau_{j+1}}\right)^2, \tag{13}$$

where in the first expression, $\theta(\cdot)$ is intended as a spectral function.

The main difficulty with the solution $\bar{\Pi}_w$ given in Equation (13) above is that it is expressed in terms of the eigensystem of the operator $A_w - \varphi'(b)B_w$, with b an unknown parameter, and A_w and B_w defined as averages of functions of the random variables F_0, F_0' and $F_{\tau_{j+1}}$. However, in practice only a finite set of empirical samples from these r.v. is available, and it is natural to replace the expressions for A_w and B_w with suitable averages over the available empirical samples.

We assume that n independent samples $(f_1, f_1'), (f_2, f_2'), \ldots, (f_n, f_n')$ of the r.v. (F_0, F_0') are available; these images play the role of dissimilar example pairs. Moreover the n independent samples $(f_{n+1}, f_{n+1}'), (f_{n+2}, f_{n+2}'), \ldots, (f_{2n}, f_{2n}')$ of the r.v. $(F_0, F_{\tau_{j+1}})$ represent similar example pairs. Given these samples, the operator $\theta(A_w - \varphi'(b)B_w)$ can be estimated by the empirical operator

$$\theta\left(\sum_{i=1}^{n} \mathrm{Q}[d_{\widehat{w}}(f_i, f_i')] - \alpha \sum_{i=n+1}^{2n} \mathrm{Q}[d_{\widehat{w}}(f_i, f_i')]\right), \tag{14}$$

for some positive α.

Using this estimate for Π_w an algorithm for the computation of the kernel K_w from the kernels K_v on $v \in \mathrm{Ch}(w)$ is given by the following Proposition.

Proposition 5.2 *Let K_w be the kernel defined by Equations (3) and (4) with Π_w given by Equation (14).*

For all (f, f') and (g, g') in $\mathrm{Im}(R)^2$, define

$$\langle (f, f'), (g, g') \rangle = \underset{v \in \mathrm{Ch}(w)}{\mathrm{Av}} \left[K_v(f, g) + K_v(f', g') - K_v(f, g') - K_v(f', g) \right]$$

where the average is relative to the uniform probability measure on $\mathrm{Ch}(w)$.

Define the two $2n \times 2n$ matrices $\mathbf{G}$ and $\mathbf{P}$

$$\mathbf{G}_{lm} = \langle (f_l, f'_l), (f_m, f'_m) \rangle , \qquad \mathbf{P}_{lm} = \sum_{i=1}^{n} \mathbf{G}_{li} \mathbf{G}_{im} - \alpha \sum_{i=n+1}^{2n} \mathbf{G}_{li} \mathbf{G}_{im}.$$

Finally denote by $\mathbf{u}_1, \ldots, \mathbf{u}_N$ an orthonormal system of column eigenvectors of $\mathbf{G}^{-\frac{1}{2}} \mathbf{P} \mathbf{G}^{-\frac{1}{2}}$ associated with positive eigenvalues, then for all $(f, f') \in \mathrm{Im}(R)^2$ it holds

$$K_w(f, f') = \exp \left(-\lambda_{j+1} \sum_{h=1}^{N} \left(\sum_{i=1}^{2n} \left(\mathbf{G}^{-\frac{1}{2}} \mathbf{u}_h \right)_i \langle (f_i, f'_i), (f, f') \rangle \right)^2 \right).$$

References

[1] Y. Amit and M. Mascaro. An integrated network for invariant visual detection and recognition. *Vision Research*, 43:2073–2088(16), September 2003.

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[3] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a Mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6:937–965, 2005.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[5] Y. Bengio, J. F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouiment. Out-of-sample extensions for LLE Isomap, MDS, Eigenmaps, and Spectral Clustering. In *Advances in Neural Information Processing Systems*, 2003.

[6] G. Chechik and N. Tishby. Extracting relevant structures with side information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems* 15. MIT Press, 2003.

[7] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similiarity metric discriminatively, with application to face verification. *CVPR*, 2005.

[8] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[9] G. Donoho. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *PNAS*, 100(10), 2003.

[10] P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 1991.

[11] K. Fukushima. Neocognitron:a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[12] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.*, 195:215–243, 1968.

[13] Yann Le Cun and Yoshua Bengio. Convolutional networks for images, speech, and time-series. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 255–257. MIT Press, 1995.

[14] Andreas Maurer. Generalization bounds for subspace selection and hyperbolic pca. In *Subspace, Latent Structure and Feature Selection. LNCS*, number 3940, pages 185–197. Springer, 2006.

[15] Andreas Maurer. Learning to compare using operator-valued large-margin classifiers. In *Advances in Neural Information Processing Systems, LTCE Workshop*, 2006.

[16] Andreas Maurer. Unsupervised slow subspace-learning from stationary processes. In *The 17th international conference on Algorithmic Learning Theory*, pages 363–377, 2006.

[17] J. Mutch and G. Lowe. Multiclass object recognition with sparse, localized features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–18, New York, NY, USA, June 2006.

[18] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen. A multiscale routing circuit for forming size- and position-invariant object representations. *The Journal of Computational Neuroscience*, 2:45–62, 1995.

[19] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

[20] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems* 16. MIT Press, 2004.

[21] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Science*, 104(15):6424–6429, April 2007.

[22] T. Serre and M. Riesenhuber. Realistic modeling of simple and complex cell tuning in the hmax model, and implications for invariant object recognition in cortex. *CBCL Paper #239/AI Memo #2004‒004, Massachusetts Institute of Technology, Cambridge, MA*, July 2004.

[23] N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment learning and relevant component analysis. In *The Seventh European Conference on Computer Vision*, volume 4, pages 776–792, Copenhagen, Denmark, 2002.

[24] J. Tenenbaum, Vin de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2000.

[25] I. W. Tsang, P. M. Cheung, and J. T. Kwok. Kernel relevant component analysis for distance metric learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'05)*, pages 954–959, Montreal, Canada, July 2005.

[26] H. Wersing and E. Korner. Learning optimized features for hierarchical models of invariant object recognition. *Neural Comput.*, 7(15):1559–1588, July 2003.

[27] T. Wiskott and T. Sejnowski. Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, 14:715–770, 2003.

[28] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russel. Distance metric learning with application to clustering with side information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems* 14. MIT Press, 2002.

A Proofs

PROOF OF PROPOSITION 2.1:
For every $f, f' \in \mathrm{Im}(R)$ define

$$K_w(f, f') := \exp\left(-\lambda_{j+1} \left\|\Pi_w(\phi_{\widehat{w}}(f) - \phi_{\widehat{w}}(f'))\right\|_{\widehat{w}}^2\right).$$

Since by assumption Π_w is bounded and $\phi_{\widehat{w}}$ is continuous, then $K_w(f, f')$ is continuous on $\mathrm{Im}(R)^2$. Moreover, since the Gaussian kernel is positive definite, then $K_w(f, f')$ is a Mercer's kernel ([2], [8]). Let $\mathcal{H}_w$ be the reproducing kernel Hilbert space associated with the kernel K_w, and define ϕ_w the canonical embedding of $\mathrm{Im}(R)$ into $\mathcal{H}_w$, that is

$$\phi_w(f)(\cdot) := K_w(f, \cdot).$$

Since, by Hypothesis 2.1, $\mathrm{Im}(R)$ is compact then $\mathcal{H}_w$ is separable. Moreover $\phi_w(f)$ is continuous, since K_w is continuous and by the reproducing property

$$\|\phi_w(f) - \phi_w(f')\|_w^2 = K_w(f, f) + K_w(f', f') - 2K_w(f, f').$$

Finally $\phi_w(f)$ fulfills Equation (4), since by the reproducing property it holds

$$\langle \phi_w(f), \phi_w(f') \rangle_w = K_w(f, f').$$

This concludes the proof of existence of $\mathcal{H}_w$ and ϕ_w satisfying the conditions in the text of the Proposition.

Now let us assume that $\mathcal{H}'_w$ and ϕ'_w is different solution. Then we can define the unitary operator U on $\mathrm{Range}(\phi_w)$ by

$$\forall f \in \mathrm{Im}(R) \quad U\phi_w(f) = \phi'_w(f),$$

this is a good definition since by assumption

$$\langle \phi_w(f), \phi_w(f') \rangle_w = \langle U\phi_w(f), U\phi_w(f') \rangle_{w'}.$$

Finally, since $\mathrm{Range}(\phi_w)$ generates $\mathcal{H}_w$, we extend U to $\mathcal{H}_w$ by linearity.

PROOF OF PROPOSITION 4.1: For sake of simplicity we omit the pedex $\cdot_v$ from most of the notation which follows (ϕ means ϕ_v, $\|\cdot\|$ means $\|\cdot\|_v$, etc), we also use the abbreviated notation $\Delta(f, g) := \|d_v(f, g)\|_v^2$. The indicator function of a predicate p is denoted by $\mathbf{1}\{p\}$.

Let us begin by estimating the rate of false positives

$$\begin{aligned}
& \sum_k \mathbb{E}\left[\mathbf{1}\{\Delta(F_0, F_0') \geq \sigma^*\}\mathbf{1}\{(F_0, F_0') \in C_v(k) \times C_v(k)\}\right] \qquad (15)\\
\leq\ & \sum_k \mathbb{E}\left[\frac{\Delta(F_0, F_0')}{\sigma^*}\mathbf{1}\{(F_0, F_0') \in C_v(k) \times C_v(k)\}\right]\\
\leq\ & \frac{1}{\sigma^*}\sum_k \mathbb{E}\left[\Delta(F_0, F_{\tau_j})\mathbf{1}\left\{(F_0, F_{\tau_j}) \in C_v(k) \times C_v(k)\right\}\right]\\
=\ & \frac{1}{\sigma^*}\mathbb{E}\left[\Delta(F_0, F_{\tau_j})\mathbf{1}\left\{(F_0, F_{\tau_j}) \in \bigcup_k C_v(k) \times C_v(k)\right\}\right]\\
\leq\ & \frac{1}{\sigma^*}\mathbb{E}\left[\Delta(F_0, F_{\tau_j})\right]
\end{aligned}$$

where the first inequality follows from the inequality $\mathbf{1}\{\Delta(f,g) \geq \sigma^*\} \leq \Delta(f,g)/\sigma^*$, and the second inequality follows from Hypothesis 4.2, since for every nonnegative function $p = p(f,g)$ and every k it holds

$$\mathbb{E}\left[p(F_0, F_0')\mathbf{1}\{(F_0, F_0') \in C_v(k) \times C_v(k)\}\right] \leq \mathbb{E}\left[p(F_0, F_{\tau_j})\mathbf{1}\left\{(F_0, F_{\tau_j}) \in C_v(k) \times C_v(k)\right\}\right],$$

as it can be shown approximating p by simple functions.

Let us proceed estimating the rate of false negatives

$$\begin{aligned}
& \sum_{k,l:k\neq l} \mathbb{E}\left[\mathbf{1}\{\Delta(F_0, F_0') < \sigma^*\}\mathbf{1}\{(F_0, F_0') \in C_v(k) \times C_v(l)\}\right] \qquad (16)\\
=\ & \mathbb{E}\left[\mathbf{1}\{\Delta(F_0, F_0') < \sigma^*\}\right] + \sum_k \mathbb{E}\left[\mathbf{1}\{\Delta(F_0, F_0') \geq \sigma^*\}\mathbf{1}\{(F_0, F_0') \in C_v(k) \times C_v(k)\}\right] - \epsilon_v\\
\leq\ & \mathbb{E}\left[\mathbf{1}\{\Delta(F_0, F_0') < \sigma^*\}\right] + \frac{1}{\sigma^*}\mathbb{E}\left[\Delta(F_0, F_{\tau_j})\right] - \epsilon_v\\
\leq\ & \mathbb{E}\left[\frac{2 - \Delta(F_0, F_0')}{2 - \sigma^*}\right] + \frac{1}{\sigma^*}\mathbb{E}\left[\Delta(F_0, F_{\tau_j})\right] - \epsilon_v
\end{aligned}$$

where for the first inequality we used (15), and for the second inequality the bound $\mathbf{1}\{\Delta(f,g) \leq \sigma^*\} \leq (2 - \Delta(f,g))/(2 - \sigma^*)$, which holds because both $\Delta(f,g)$ and σ^* are no greater than 2.

Finally, recalling the definition of $\check{\mathrm{Err}}_v$, inequalities (15) and (16), and the definition of σ^*

$$\sigma^* = 2\left(1 + \sqrt{\frac{\mathbb{E}\left[\langle\phi(F_0), \phi(F_0')\rangle\right]}{\mathbb{E}\left[\left\|\phi(F_0) - \phi(F_{\tau_j})\right\|^2\right]}}\right)^{-1}$$

by substitution we get

$$
\begin{aligned}
\mathrm{Err}_v(\sigma^*) &\leq \frac{2-\mathbb{E}\left[\Delta(F_0,F_0')\right]}{2-\sigma^*}+\frac{2}{\sigma^*}\mathbb{E}\left[\Delta(F_0,F_{\tau_j})\right]-\epsilon_v \\
&= \frac{2\mathbb{E}\left[\langle\phi(F_0),\phi(F_0')\rangle\right]}{2-\sigma^*}+\frac{2\mathbb{E}\left[\left\|\phi(F_0)-\phi(F_{\tau_j})\right\|^2\right]}{\sigma^*}-\epsilon_v \\
&= \left(\sqrt{\mathbb{E}\left[\langle\phi(F_0),\phi(F_0')\rangle\right]}+\sqrt{\mathbb{E}\left[\left\|\phi(F_0)-\phi(F_{\tau_j})\right\|^2\right]}\right)^2-\epsilon_v \\
&= \check{\mathrm{Err}}_v
\end{aligned}
$$

which completes the proof.

PROOF OF PROPOSITION 4.2: We use the simplified notations $a_w := a_w(\Pi_w)$ and $b_w := b_w(\Pi_w)$. From the definition of b_w, Equation (9) and observing that, by convexity, $1-\exp(-z) \leq z$ for every nonnegative z, we get

$$
\begin{aligned}
\mathbb{E}\left[\left\|\phi_w(F_0)-\phi_w(F_{\tau_{j+1}})\right\|_w^2\right] &= 2\mathbb{E}\left[1-\exp\left(-\lambda_{j+1}\left\|\Pi_w(\phi_{\widehat{w}}(F_0)-\phi_{\widehat{w}}(F_{\tau_{j+1}}))\right\|_{\widehat{w}}^2\right)\right] \\
&\leq 2\lambda_{j+1}\mathbb{E}\left[\left\|\Pi_w(\phi_{\widehat{w}}(F_0)-\phi_{\widehat{w}}(F_{\tau_{j+1}}))\right\|_{\widehat{w}}^2\right] \\
&= 2\lambda_{j+1}b_w\mathbb{E}\left[\left\|\phi_{\widehat{w}}(F_0)-\phi_{\widehat{w}}(F_{\tau_{j+1}})\right\|_{\widehat{w}}^2\right] \\
&= 2\lambda_{j+1}b_w\mathbb{E}\left[\left\|\sum_{i=1}^{\tau_{j+1}/\tau_j}(\phi_{\widehat{w}}(F_{(i-1)\tau_j})-\phi_{\widehat{w}}(F_{i\tau_j}))\right\|_{\widehat{w}}^2\right] \\
&\leq 2\lambda_{j+1}b_w\left(\frac{\tau_{j+1}}{\tau_j}\right)^2\mathbb{E}\left[\left\|\phi_{\widehat{w}}(F_0)-\phi_{\widehat{w}}(F_{\tau_j})\right\|_{\widehat{w}}^2\right] \\
&= 2\lambda_{j+1}b_w\left(\frac{\tau_{j+1}}{\tau_j}\right)^2\operatorname*{Av}_{v\in\mathrm{Ch}(w)}\mathbb{E}\left[\left\|\phi_v(F_0)-\phi_v(F_{\tau_j})\right\|_v^2\right] \\
&\leq \operatorname*{Av}_{v\in\mathrm{Ch}(w)}\mathbb{E}\left[\left\|\phi_v(F_0)-\phi_v(F_{\tau_j})\right\|_v^2\right] \qquad (17)
\end{aligned}
$$

where the second inequality follows from the general property of Hilbert space norms[1]

$$
\left\|\sum_{i=1}^n \mathrm{v}_i\right\|^2 \leq n\sum_{i=1}^n\|\mathrm{v}_i\|^2,
$$

[1]This property can be derived by induction on n. In fact from the inductive hypothesis we get

$$
\left\|\sum_{i=1}^n \mathrm{v}_i\right\|^2 \leq (n-1)\sum_{i=1}^{n-1}\|\mathrm{v}_i\|^2+\|\mathrm{v}_n\|^2+2\sum_{i=1}^{n-1}\langle\mathrm{v}_n,\mathrm{v}_i\rangle \leq \sum_{i=1}^n\|\mathrm{v}_i\|^2,
$$

which holds since

$$
2\langle\mathrm{v}_n,\mathrm{v}_i\rangle \leq \|\mathrm{v}_n\|^2+\|\mathrm{v}_i\|^2.
$$

and the stationarity of the stochastic process $\mathbf{F}$.

Observe that since

$$\|\Pi_w\| \leq 1,$$

and by definition $\langle \phi_{\widehat{w}}(f), \phi_{\widehat{w}}(g) \rangle_{\widehat{w}} \in [0,1]$, it holds

$$\|\Pi_w(\phi_{\widehat{w}}(f) - \phi_{\widehat{w}}(g))\|_{\widehat{w}}^2 \leq \|\Pi_w\|^2 \|\phi_{\widehat{w}}(f) - \phi_{\widehat{w}}(g)\|_{\widehat{w}}^2 \leq 2.$$

From the previous inequality and noticing that, for every $z \in [0, Z]$ it holds $\exp(-z) \leq 1 + (\exp(-Z) - 1)z/Z$, we get

$$\begin{aligned}
\mathbb{E}\left[\langle \phi_w(F_0), \phi_w(F_0') \rangle_w\right] &= \mathbb{E}\left[\exp\left(-\lambda_{j+1} \|\Pi_w(\phi_{\widehat{w}}(F_0) - \phi_{\widehat{w}}(F_0'))\|_{\widehat{w}}^2\right)\right] \\
&\leq 1 + \frac{1}{2}(\exp(-2\lambda_{j+1}) - 1)\mathbb{E}\left[\|\Pi_w(\phi_{\widehat{w}}(F_0) - \phi_{\widehat{w}}(F_0'))\|_{\widehat{w}}^2\right] \\
&= 1 + \frac{1}{2}a_w(\exp(-2\lambda_{j+1}) - 1)\mathbb{E}\left[\|\phi_{\widehat{w}}(F_0) - \phi_{\widehat{w}}(F_0')\|_{\widehat{w}}^2\right] \\
&= 1 + a_w(\exp(-2\lambda_{j+1}) - 1)\left(1 - \operatorname*{Av}_{v \in \mathrm{Ch}(w)} \mathbb{E}\left[\langle \phi_v(F_0), \phi_v(F_0') \rangle_v\right]\right) \\
&\leq 1 - a_w + \exp(-2\lambda_{j+1}) + \operatorname*{Av}_{v \in \mathrm{Ch}(w)} \mathbb{E}\left[\langle \phi_v(F_0), \phi_v(F_0') \rangle_v\right] \qquad (18)
\end{aligned}$$

where in the last inequality we used the fact that $a_w \in [0,1]$. In the following, for sake of brevity, we use the notation $\delta_w := 1 - a_w + \exp(-2\lambda_{j+1})$.

Finally, by the definition of $\check{\mathrm{Err}}_w$, equations (17) and (18), Hypothesis 4.1, and the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ which hold for every nonnegative x and y, we get

$$\begin{aligned}
\sqrt{\check{\mathrm{Err}}_w + \epsilon_w} &= \sqrt{\mathbb{E}\left[\langle \phi_w(F_0), \phi_w(F_0') \rangle_w\right]} + \sqrt{\mathbb{E}\left[\left\|\phi_w(F_0) - \phi_w(F_{\tau_{j+1}})\right\|_w^2\right]} \\
&\leq \sqrt{\delta_w + \operatorname*{Av}_{v \in \mathrm{Ch}(w)} \mathbb{E}\left[\langle \phi_v(F_0), \phi_v(F_0') \rangle_v\right]} + \sqrt{\operatorname*{Av}_{v \in \mathrm{Ch}(w)} \mathbb{E}\left[\left\|\phi_v(F_0) - \phi_v(F_{\tau_j})\right\|_v^2\right]} \\
&\leq \sqrt{\delta_w} + \sqrt{\mathbb{E}\left[\langle \phi_v(F_0), \phi_v(F_0') \rangle_v\right]} + \sqrt{\mathbb{E}\left[\left\|\phi_v(F_0) - \phi_v(F_{\tau_j})\right\|_v^2\right]} \\
&= \sqrt{\delta_w} + \sqrt{(\check{\mathrm{Err}}_v + \epsilon_v)}
\end{aligned}$$

which completes the proof.

PROOF OF PROPOSITION 5.1:

Let us start with some preliminary observations. First, notice that

$$a_w(\Pi) = \mathrm{Tr}\left[\Pi^2 A_w\right] \qquad b_w(\Pi) = \mathrm{Tr}\left[\Pi^2 B_w\right].$$

Second, notice that $\Lambda := A_w - \varphi'(b)B_w$ is symmetric and trace class since it is the sum of two symmetric and trace class operators; let $(\lambda_i^+, \psi_i^+)_i$ and $(\lambda_i^-, \psi_i^-)_i$

be the parts of Λ's eigensystem with positive and negative eigenvalues respectively. Third, let us introduce $\mathcal{D}$, the set of symmetric positive semi-definite operators on $\mathcal{H}_{\widehat{w}}$ with operator norm bounded by 1.

Now, since $\delta_w(\Pi, \lambda)$ is a strictly decreasing function of λ, Definition 5.1 implies that Equation (12) holds and that Π_w is a solution of the problem

$$\max \left[\mathrm{Tr}\left[\Pi^2 A_w\right] - \varphi(\mathrm{Tr}\left[\Pi^2 B_w\right]) | \Pi^2 \in \mathcal{D} \right]. \tag{19}$$

Moreover, by assumption, for the solution of this problem Π_w in the text of the Proposition, it holds $\mathrm{Tr}\left[\Pi_w^2 B_w\right] = b$, therefore $\bar{\lambda}_{j+1}$ defined in Equation (13) and any solution of the problem

$$\max \left[\mathrm{Tr}\left[\Pi^2 A_w\right] | \mathrm{Tr}\left[\Pi^2 B_w\right] = b, \Pi^2 \in \mathcal{D} \right] \tag{20}$$

is a pair fulfilling Definition 5.1. We are left to prove that $\bar{\Pi}_w$ defined in Equation (13) is a solution of (20).

In order to prove that $\bar{\Pi}_w$ is a solution of (20), let us first observe that since Π_w in the text of the Proposition is a solution of (19) and $\mathcal{D}$ is convex, then for all $\Pi^2 \in \mathcal{D}$ and $\alpha \in [0, 1]$, for $\Pi_\alpha^2 := (1-\alpha)\Pi_w^2 + \alpha\Pi^2$, it holds

$$\mathrm{Tr}\left[\Pi_w^2 A_w\right] - \mathrm{Tr}\left[\Pi_\alpha^2 A_w\right] - \varphi(\mathrm{Tr}\left[\Pi_w^2 B_w\right]) + \varphi(\mathrm{Tr}\left[\Pi_\alpha^2 B_w\right]) \geq 0.$$

Dividing by α, and letting α go to 0, the previous relation becomes

$$\forall \Pi^2 \in \mathcal{D} \qquad \mathrm{Tr}\left[(\Pi_w^2 - \Pi^2)\Lambda\right] \geq 0,$$

that is, Π_w^2 is a solution of

$$\max \left[\mathrm{Tr}\left[\Pi^2 \Lambda\right] | \Pi^2 \in \mathcal{D} \right]. \tag{21}$$

Since for every $\Pi^2 \in \mathcal{D}$

$$\mathrm{Tr}\left[\Pi^2 \Lambda\right] = \sum_i \lambda_i^+ \left\langle \psi_i^+, \Pi^2 \psi_i^+ \right\rangle_{\widehat{w}} + \sum_i \lambda_i^- \left\langle \psi_i^-, \Pi^2 \psi_i^- \right\rangle_{\widehat{w}} \leq \sum_i \lambda_i^+ = \mathrm{Tr}\left[\bar{\Pi}^2 \Lambda\right]$$

then $\bar{\Pi}^2$ is also a solution of problem (21), and $\left\langle \psi_i^+, \Pi_w^2 \psi_i^+ \right\rangle_{\widehat{w}} = 1$, $\left\langle \psi_i^-, \Pi_w^2 \psi_i^- \right\rangle_{\widehat{w}} = 0$. From this conditions, the fact that $\|\Pi_w\| \leq 1$ and Cauchy−Schwarz inequality, it follows that $\Pi_w \psi_i^+ = \psi_i^+$ and that if $\psi^0 \in \mathrm{Ker}(\Lambda)$ then $|\left\langle \psi^0, \Pi_w^2 \psi_i^- \right\rangle_{\widehat{w}}|^2 \leq \left\|\Pi_w \psi^0\right\|_{\widehat{w}} \left\|\Pi_w \psi_i^-\right\|_{\widehat{w}} = 0$. Therefore for some $R \in \mathcal{D}$ with $\mathrm{Range}(R) \subseteq \mathrm{Ker}(\Lambda)$, it holds

$$\Pi_w^2 = \bar{\Pi}^2 + R.$$

Now, since both Π_w^2 and $\bar{\Pi}^2$ are solutions of (21), then

$$\mathrm{Tr}[A_w R] = \varphi'(b)\mathrm{Tr}[B_w R]$$

and, since Π_w^2 is a solution of (19), then

$$\mathrm{Tr}[A_w R] \geq \varphi(b) - \varphi(b - \mathrm{Tr}[B_w R]),$$

which implies

$$\varphi(b) \leq \varphi(b - \mathrm{Tr}[B_w R]) + \varphi'(b)\mathrm{Tr}[B_w R]\,.$$

However, since φ is strictly convex in $\left[0, \frac{1}{2}\left(\frac{\tau_j}{\tau_{j+1}}\right)^2\right]$ and by assumption $b \leq \frac{1}{2}\left(\frac{\tau_j}{\tau_{j+1}}\right)^2$, from the previous inequality it follows that $\mathrm{Tr}[B_w R] = 0$. This means that $\mathrm{Tr}\left[\bar{\Pi}^2 B_w\right] = b$ and $\bar{\Pi}_w^2$ is a solution of (20) as claimed.

PROOF OF PROPOSITION 5.2:

Let $(\psi_1, \psi_2, \ldots, \psi_N)$ be the system of eigenvectors of the operator $\sum_{i=1}^{n} \mathrm{Q}[d_{\widehat{w}}(f_i, f_i')] - \alpha \sum_{i=n+1}^{2n} \mathrm{Q}[d_{\widehat{w}}(f_i, f_i')]$ associated to positive eigenvectors.

The projector Π_w defined by Equation (14) is equal to $\sum_{h=1}^{N} \mathrm{Q}[\psi_i]$; since it is idempotent, Equation (4) becomes

$$\begin{aligned} K_w(f, f') &= \exp\left(-\lambda_{j+1} \left\langle d_{\widehat{w}}(f, f'), \Pi_w d_{\widehat{w}}(f, f')\right\rangle_{\widehat{w}}\right) \\ &= \exp\left(-\lambda_{j+1} \sum_h \left\langle \psi_h, d_{\widehat{w}}(f, f')\right\rangle_{\widehat{w}}^2\right). \end{aligned} \tag{22}$$

Observe that by the definition of $d_{\widehat{w}}$ it holds

$$\forall l, m \qquad \mathbf{G}_{lm} = \langle (f_l, f_l'), (f_m, f_m')\rangle = \left\langle d_{\widehat{w}}(f_l, f_l'), d_{\widehat{w}}(f_m, f_m')\right\rangle_{\widehat{w}}\,.$$

Since the range of Π_w is included in $\mathrm{span}\{d_{\widehat{w}}(f_l, f_l') | 1 \leq l \leq 2n\}$ from the equations

$$\forall l, m \qquad \mathbf{P}_{lm} = \left\langle d_{\widehat{w}}(f_l, f_l'), \Pi_w d_{\widehat{w}}(f_m, f_m')\right\rangle_{\widehat{w}}$$

we conclude that $\psi_h = \sum_{i=1}^{2n} \left(\mathbf{G}^{-\frac{1}{2}} \mathbf{u}_h\right)_i d_{\widehat{w}}(f_i, f_i')$.

The proposition follows by substituting this expression for ψ_h in Equation (22).