On Invariance and Selectivity in Representation Learning

Fabio Anselmi[1,2],Lorenzo Rosasco[1,2,3] and Tomaso Poggio[1,2]

March 23, 2015

Abstract

We discuss data representation which can be learned automatically from data, are invariant to transformations, and at the same time selective, in the sense that two points have the same representation only if they are one the transformation of the other. The mathematical results here sharpen some of the key claims of *i-theory* – a recent theory of feedforward processing in sensory cortex. [3, 4, 5].

Keywords:*Invariance, Machine Learning*
The paper is submitted to Information and Inference Journal.

1 Introduction

This paper considers the problem of learning *"good"* data representation which can lower the need of labeled data (sample complexity) in machine learning (ML). Indeed, while current ML systems have achieved impressive results in a variety of tasks, an obvious bottleneck appears to be the huge amount of labeled data needed. This paper builds on the idea that data representation, which are learned in an unsupervised manner, can be key to solve the problem. Classical statistical learning theory focuses on supervised learning and postulates that a suitable hypothesis space is given. In turn, under very general conditions, the latter can be seen to be equivalent to a data representation. In other words, data representation and how to select and learn it, is classically not considered to be part of the learning problem, but rather as a prior information. In practice ad hoc solutions are often empirically found for each problem.

The study in this paper is a step towards developing a theory of learning data representation. Our starting point is the intuition that, since many learning

[1] Center for Brains Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
[2] Laboratory for Computational Learning, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology
[3] DIBRIS, Universitá degli studi di Genova, Italy, 16146

tasks are invariant to transformations of the data, learning invariant representation from “unsupervised” experiences can significantly lower the ”size” of the problem, effectively decreasing the need of labeled data. In the following, we formalize the above idea and discuss how such invariant representations can be learned. Crucial to our reasoning is the requirement for invariant representations to satisfy a form of selectivity, broadly referred to as the property of distinguishing images which are not one the transformation of the other. Indeed, it is this latter requirement that informs the design of non trivial invariant representations. Our work is motivated by a theory of cortex and in particular visual cortex [5].

Data representation is a classical concept in harmonic analysis and signal processing. Here representations are typically designed on the basis of prior information assumed to be available. More recently, there has been an effort to automatically learn adaptive representation on the basis of data samples. Examples in this class of methods include so called dictionary learning [30], autoencoders [6] and metric learning techniques (see e.g. [33]). The idea of deriving invariant data representation has been considered before. For example in the analysis of shapes [19] and more generally in computational topology [10], or in the design of positive definite functions associated to reproducing kernel Hilbert spaces [12]. However, in these lines of study the selectivity properties of the representations have hardly been considered. The ideas in [22, 28] are close in spirit to the study in this paper. In particular, the results in [22] develop a different invariant and stable representation within a signal processing framework. In [28] an information theoretic perspective is considered to formalize the problem of learning invariant/selective representations.

In this work we develop a machine learning perspective closely following computational neuroscience models of the information processing in the visual cortex [15, 16, 26]. Our first and main result shows that, for compact groups, representation defined by nonlinear group averages can be shown to be invariant, as well as selective, to the action of the group. While invariance follows from the properties of the Haar measure associated to the group, selectivity is shown using probabilistic results that characterize a probability measure in terms of one dimensional projections. This set of ideas, which form the core of the paper, is then extended to local transformations, and multilayer architectures. These results bear some understanding to the nature of certain deep architecture, in particular neural networks of the convolution type.

The rest of the paper is organized as follows. We describe the concept of invariance and selective representation in Section 2 and their role for learning in Section 3. We discuss a family of invariant/selective representation for transformations which belong to compact groups in Section 4 that we further develop in Sections 5 and 6. Finally we conclude in Section 7 with some final comments.

2 Invariant and Selective Data Representations

We next formalize and discuss the notion of *invariant and selective* data representation, which is the main focus of the rest of the paper.

We model the data space as a (real separable) Hilbert space $\mathcal{I}$ and denote by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ the inner product and norm, respectively. Example of data spaces are one dimensional signals (as in audio data), where we could let $\mathcal{I} \subset L^2(\mathbb{R})$, or two dimensional signals (such as images), where we could let $\mathcal{I} \subset L^2(\mathbb{R}^2)$. After discretization, data can often be seen as vectors in high-dimensional Euclidean spaces, e.g. $\mathcal{I} = \mathbb{R}^d$. The case of (digital) images serves as a main example throughout the paper.

A data representation is a map from the data space in a suitable representation space, that is

$$\mu : \mathcal{I} \to \mathcal{F}.$$

Indeed, the above concept appears under different names in various branch of pure and applied sciences, e.g. it is called an encoding (information theory), a feature map (learning theory), a transform (harmonic analysis/signal processing) or an embedding (computational geometry).

In this paper, we are interested in representations which are invariant (see below) to suitable sets of transformations. The latter can be seen as a set of maps

$$\mathcal{G} \subset \{g \mid g : \mathcal{I} \to \mathcal{I}\}.$$

Many interesting examples of transformations have a group structure. Recall that a group is a set endowed with a well defined *composition/multiplication* operation satisfying four basic properties,

- closure: $gg' \in \mathcal{G}$, for all $g, g' \in \mathcal{G}$
- associativity: $(gg')g'' = g(g'g'')$, for all $g, g', g'' \in \mathcal{G}$
- identity: there exists $\mathrm{Id} \in \mathcal{G}$ such that $\mathrm{Id}g = g\mathrm{Id} = g$, for all $g \in \mathcal{G}$.
- invertibility: for all $g \in \mathcal{G}$ there exists $g^{-1} \in \mathcal{G}$ such that $(gg^{-1}) = \mathrm{Id}$.

There are different kind of groups. In particular, "small" groups such as compact (or locally compact, i.e. a group that admits a locally compact Hausdorff topology such that the group operations of composition and inversion are continuous.) groups, or "large" groups which are not locally compact. In the case of images, examples of locally compact groups include affine transformations (e.g. scaling, translations, rotations and their combinations) which can be thought of as suitable *viewpoint* changes. Examples of non locally compact groups are diffeomorphisms, which can be thought of as various kind of local or global *deformations.*

Example 1. *Let $I \in L^2(\mathbb{R})$. A basic example of group transformation is given by the translation group, which can be represented as a family of linear operators*

$$T_\tau : L^2(\mathbb{R}) \to L^2(\mathbb{R}), \quad T_\tau I(p) = I(p - \tau), \quad \forall p \in \mathbb{R}, I \in \mathcal{I},$$

for $\tau \in \mathbb{R}$. Other basic examples of locally compact groups include scaling (the multiplication group) and affine transformations (affine group). Given a smooth map $d : \mathbb{R} \to \mathbb{R}$ a diffeomorphism can also be seen as a linear operator given by

$$D_d : L^2(\mathbb{R}) \to L^2(\mathbb{R}), \quad D_d I(p) = I(d(p)), \quad \forall p \in \mathbb{R}, I \in \mathcal{I}.$$

Note that also in this case the representation is linear.

Clearly, not all transformations have a group structure– think for example of images obtained from three dimensional rotations of an object.
Given the above premise, we next discuss, properties of data representation with respect to transformations. We first add one remark about the notation.

Remark 1 (Notation: Group Action and Representation)**.** *If $\mathcal{G}$ is a group and $\mathcal{I}$ a set, the group action is the map $(g, x) \mapsto g.x \in \mathcal{I}$. In the following, with an abuse of notation we will denote by gx the group action. Indeed, when $\mathcal{I}$ is a linear space, we also often denote by g both a group element and its representation, so that g can be identified with a linear operator. Throughout the article we assume the group representation to be unitary [25].*

To introduce the notion of invariant representation, we recall that an orbit associated to an element $I \in \mathcal{I}$ is the set $O_I \subset \mathcal{I}$ given by $O_I = \{I' \in \mathcal{I} \mid I' = gI, \;\; g \in \mathcal{G}\}$. Orbits form a partition of $\mathcal{I}$ in equivalence classes, with respect to the equivalence relation,

$$I \sim I' \quad \Leftrightarrow \quad \exists\, g \in \mathcal{G} \text{ such that } gI = I',$$

for all $I, I' \in \mathcal{I}$. We have the following definition.

Definition 1 (Invariant Representation)**.** *We say that a representation μ is invariant with respect to $\mathcal{G}$ if*

$$I \sim I' \Rightarrow \mu(I) = \mu(I'),$$

for all $I, I' \in \mathcal{I}$.

In words, the above definition states that if two data points are one the transformation of the other, than they will have the same representation. Indeed, if a representation μ is invariant

$$\mu(I) = \mu(gI)$$

for all $I \in \mathcal{I}$, $g \in \mathcal{G}$. Clearly, trivial invariant representations can be defined, e.g. the constant function. This motivates a second requirement, namely selectivity.

Definition 2 (Selective Representation)**.** *We say that a representation μ is selective with respect to $\mathcal{G}$ if*

$$\mu(I) = \mu(I') \Rightarrow I \sim I',$$

for all $I, I' \in \mathcal{I}$.

Together with invariance, selectivity asserts that two points have the same representation *if and only* if they are one a transformation of the other. Several comments are in order. First, the requirement of exact invariance as in Definition 1, seems desirable for (locally) compact groups, but not for non locally compact group such as diffeomorphisms. In this case, requiring a form of stability to *small* transformations seems to be natural, as it is more generally to require stability to small perturbations, e.g. noise (see [22]). Second, the concept of selectivity is natural and requires that no two orbits are mapped in the same representation. It corresponds to an injectivity property of a representation on the quotient space $\mathcal{I}/\sim$. Assuming $\mathcal{F}$ to be endowed with a metric $d_{\mathcal{F}}$, a stronger requirement would be to characterize the metric embedding induced by μ, that is to control the ratio (or the deviation) of the distance of two representation and the distance of two orbits. Indeed, the problem of finding invariant and selective representation, is tightly related to the problem of finding an injective embedding of the quotient space $\mathcal{I}/\sim$.

We next provide a discussion of the potential impact of invariant representations on the solution of subsequent learning tasks.

3 From Invariance to Low Sample Complexity

In this section we first recall how the concepts of data representation and hypothesis space are closely related, and how the sample complexity of a supervised problem can be characterized by the covering numbers of the hypothesis space. Then, we discuss how invariant representations can lower the sample complexity of a supervised learning problem.

Supervised learning amounts to finding an input-output relationship on the basis of a *training* set of input-output pairs. Outputs can be scalar or vector valued, as in regression, or categorical, as in multi-category or multi-label classification, binary classification being a basic example. The bulk of statistical learning theory is devoted to study conditions under which learning problems can be *solved*, approximately and up to a certain confidence, provided a suitable hypothesis space is given. A hypotheses space is a subset

$$\mathcal{H} \subset \{f \mid f : \mathcal{I} \to \mathcal{Y}\},$$

of the set of all possible input output relations. As we comment below, under very general assumptions *hypothesis spaces and data representations are equivalent concepts.*

3.1 Data Representation and Hypothesis Space

Indeed, practically useful hypothesis spaces are typically endowed with a Hilbert space structure, since it is in this setting that most computational solutions can be developed. A further natural requirement is for evaluation functions to be well defined and continuous. This latter property allows to give a well defined meaning of the evaluation of a function at every points, a property which is arguably natural since we are interested in making predictions. The requirements of 1) being a Hilbert space of of functions and 2) have continuous evaluation functionals, define so called reproducing kernel Hilbert spaces [24]. Among other properties, these spaces of functions are characterized by the existence of a feature map $\mu : \mathcal{I} \to \mathcal{F}$, which is a map from the data space into a feature space which is itself a Hilbert space. Roughly speaking, functions in a RKHS $\mathcal{H}$ with an associated feature map μ can be seen as *hyperplanes* in the feature space, in the sense that $\forall f \in \mathcal{H}$, there exists $w \in \mathcal{F}$ such that

$$f(I) = \langle w, \mu(I) \rangle_{\mathcal{F}}, \quad \forall I \in \mathcal{I}.$$

The above discussion illustrates how, under mild assumptions, the choice of a hypothesis space is equivalent to the choice of a data representation (a feature map). In the next section, we recall how hypothesis spaces, hence data representation, are usually assumed to be given in statistical learning theory and are characterized in terms of sample complexity.

3.2 Sample Complexity in Supervised Learning

Supervised statistical learning theory characterizes the difficulty of a learning problem in terms of the "size" of the considered hypothesis space, as measured by suitable capacity measures. More precisely, given a measurable loss function $V : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$, for any measurable function $f : \mathcal{I} \to \mathcal{Y}$ the expected error is defined as

$$\mathcal{E}(f) = \int V(f(I), y) d\rho(I, y)$$

where ρ is a probability measure on $\mathcal{I} \times \mathcal{Y}$. Given a training set $S_n = \{(I_1, y_1), \ldots, (I_n, y_n)\}$ of input-output pairs sampled identically and independently with respect to ρ, and a hypothesis space $\mathcal{H}$, the goal of learning is to find an approximate solution $f_n = f_{S_n} \in \mathcal{H}$ to the problem

$$\inf_{f \in \mathcal{H}} \mathcal{E}(f)$$

The difficulty of a learning problem is captured by the following definition.

Definition 3 (Learnability and Sample Complexity)**.** *A hypothesis space $\mathcal{H}$ is said to be learnable if, for all $\epsilon \in [0, \infty)$, $\delta \in [0, 1]$, there exists $n(\epsilon, \delta, \mathcal{H}) \in \mathbb{N}$ such that*

$$\inf_{f_n} \sup_{\rho} \mathbb{P}\left(\mathcal{E}(f_n) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) \geq \epsilon \right) \leq \delta. \tag{1}$$

The quantity $n(\epsilon, \delta, \mathcal{H})$ is called the sample complexity of the problem.

The above definition characterizes the complexity of the learning problem associated to a hypothesis space $\mathcal{H}$, in terms of the existence of an algorithm that, provided with at least $n(\epsilon, \delta, \mathcal{H})$ training set points, can *approximately* solve the learning problem on $\mathcal{H}$ with *accuracy* ϵ and *confidence* δ.

The sample complexity associated to a hypothesis space $\mathcal{H}$ can be derived from suitable notions of covering numbers, and related quantities, that characterize the size of $\mathcal{H}$. Recall that, roughly speaking, the covering number N_ϵ associated to a (metric) space is defined as the minimal number of ϵ balls needed to cover the space. The sample complexity can be shown [31, 9] to be proportional to the logarithm of the covering number, i.e.

$$n(\epsilon, \delta, \mathcal{H}) \propto \frac{1}{\epsilon^2} \log \frac{N_\epsilon}{\delta}.$$

As a basic example, consider $\mathcal{I}$ to be d-dimensional and a hypothesis space of linear functions

$$f(I) = \langle w, I \rangle \,, \quad \forall I \in \mathcal{I}, w \in \mathcal{I},$$

so that the data representation is simply the identity. Then the ϵ-covering number of the set of linear functions with $\|w\| \leq 1$ is given by

$$N_\epsilon \sim \epsilon^{-d}.$$

If the input data lie in a subspace of dimension $s \leq d$ then the covering number of the space of linear functions becomes $N_\epsilon \sim \epsilon^{-s}$. In the next section, we further comment on the above example and provide an argument to illustrate the potential benefits of invariant representations.

3.3 Sample Complexity of the Invariance Oracle

Consider the simple example of a set of images of $p \times p$ pixels each containing an object within a (square) window of $k \times k$ pixels and surrounded by a uniform background. Imagine the object positions to be possibly anywhere in the image. Then it is easy to see that as soon as objects are translated so that they not overlap we get an orthogonal subspace. Then, we see that there are $r^2 = (p/k)^2$ possible subspaces of dimension k^2, that is the set of translated images can be seen as a distribution of vectors supported within a ball in $d = p^2$ dimensions. Following the discussion in the previous section the best algorithm based on a linear hypothesis space will incur in a sample complexity proportional to d. Assume now to have access to an *oracle* that can "register" each image so that each object occupies the centered position. In this case, the distribution of images is effectively supported within a ball in $s = k^2$ dimensions and the sample complexity is proportional to s rather than d. In other words a linear learning algorithm would need

$$r^2 = d/s$$

less examples to achieve the same accuracy. The idea is that invariant representations can act as an invariance oracle, and have the same impact on the sample complexity. We add a few comments. First, while the above reasoning is developed for linear hypothesis space, a similar conclusion holds if non linear hypothesis spaces are considered. Second, one can see that the set of images obtained by translation is a low dimensional manifold, embedded in a very high dimensional space. Other transformations, such as small deformation, while being more complex, would have a much milder effect on the dimensionality of the embedded space. Finally, the natural question is how invariant representations can be learned, a topic we address next.

4 Compact Group Invariant Representations

Consider a set of transformations $\mathcal{G}$ which is a locally compact group. Recall that each locally compact groups has a finite measure naturally associated to it, the so called Haar measure. The key feature of the Haar measure is its invariance to the group action, and in particular for all measurable functions $f : \mathcal{G} \to \mathbb{R}$, and $g' \in \mathcal{G}$, it holds

$$\int dg f(g) = \int dg f(g'g).$$

The above equation is reminding of the invariance to translation of Lebesgue integrals and indeed, the Lebesgue measure can be shown to be the Haar measure associated to the translation group. The invariance property of the Haar measure associated to a locally compact group, is key to our development of invariant representation, as we describe next.

4.1 Invariance via Group Averaging

The starting point for deriving invariant representations is the following direct application of the invariance property of the Haar measure.

Theorem 1. *Let $\psi : \mathcal{I} \to \mathbb{R}$ be a, possibly non linear, functional on $\mathcal{I}$. Then, the functional defined by*

$$\mu : \mathcal{I} \to \mathbb{R}, \qquad \mu(I) = \int dg \psi(gI), \quad I \in \mathcal{I} \tag{2}$$

is invariant in the sense of Definition 1.

The functionals ψ, μ can be thought to be measurements, or features, of the data. In the following we are interested in measurements of the form

$$\psi : \mathcal{I} \to \mathbb{R}, \qquad \psi(I) = \eta(\langle gI, t \rangle), \quad I \in \mathcal{I}, g \in \mathcal{G} \tag{3}$$

where $t \in \mathcal{T} \subseteq \mathcal{I}$ the set of unit vectors in $\mathcal{I}$ and $\eta : \mathbb{R} \to \mathbb{R}$ is a possibly non linear function. As discussed in [4], the main motivation for considering measurements of the above form is their interpretation in terms of biological or artificial neural networks, see the following remarks.

Remark 2 (Hubel and Wiesel Simple and Complex Cells [14])**.** *A measurement as in* (3) *can be interpreted as the output of a* neuron *which computes a possibly high-dimensional inner product with a template $t \in \mathcal{T}$. In this interpretation,*

η can be seen as a, so called, activation function, for which natural choices are sigmoidal functions, such as the hyperbolic tangent or rectifying functions such as the hinge. The functional μ, obtained plugging (3) *in* (2) *can be seen as the output of a second neuron which aggregates the output of other neurons by a simple averaging operation. Neurons of the former kind are similar to simple cells, whereas neurons of the second kind are similar to complex cells in the visual cortex.*

Remark 3 (Convolutional Neural Networks [20])**.** *The computation of a measurement obtained plugging* (3) *in* (2) *can also be seen as the output of a so called convolutional neural network where each neuron, ψ is performing the inner product operation between the input, I, and its synaptic weights, t, followed by a pointwise nonlinearity η and a pooling layer.*

A second, reason to consider measurements of the form (3) is computational and, as shown later, have direct implications for learning. Indeed, to compute an invariant feature, according to (2) it is necessary to be able to compute the action of any element $I \in \mathcal{I}$ for which we wish to compute the invariant measurement. However, a simple observation suggests an alternative strategy. Indeed, since the group representation is unitary, then

$$\langle gI, I' \rangle = \left\langle I, g^{-1}I' \right\rangle, \quad \forall I, I' \in \mathcal{I}$$

so that in particular we can compute ψ by considering

$$\psi(I) = \int dg \eta(\langle I, gt \rangle), \quad \forall I \in \mathcal{I}, \tag{4}$$

where we used the invariance of the Haar measure. The above reasoning implies that an invariant feature can be computed for any point provided that for $t \in \mathcal{T}$, the sequence gt, $g \in \mathcal{G}$ is available. This observation has the following interpretation: if we view a sequence gt, $g \in \mathcal{G}$, as a "movie" of an object undergoing a family of transformations, then the idea is that invariant features can be computed for any new image provided that a movie of the template is available.

While group averaging provides a natural way to tackle the problem of invariant representation, it is not clear how a family of invariant measurements can be ensured to be selective. Indeed, in the case of compact groups selectivity can be provably characterized using a probabilistic argument summarized in the following three steps:

1. A unique probability distribution can be naturally associated to each orbit.

2. Each such probability distributions can be characterized in terms of one-dimensional projections.

3. One dimensional probability distributions are easy to characterize, e.g. in terms of their cumulative distribution or their moments.

We note in passing that the above development, which we describe in detail next, naturally provides as a byproduct indications on how the non linearity in (3) needs to be chosen and thus gives insights on the nature of the *pooling* operation.

4.2 A Probabilistic Approach to Selectivity

Let $\mathcal{I} = \mathbb{R}^d$, and $\mathcal{P}(\mathcal{I})$ the space of probability measures on $\mathcal{I}$. Recall that for any compact group, the Haar measure is finite, so that, if appropriately normalized, it correspond to a probability measure.

Assumption 1. *In the following we assume $\mathcal{G}$ to be Abelian and compact and the corresponding Haar measure to be normalized.*

The first step in our reasoning is the following definition.

Definition 4 (Representation via Orbit Probability)**.** *For all $I \in \mathcal{I}$, define the random variable*

$$Z_I : (\mathcal{G}, dg) \to \mathcal{I}, \quad Z_I(g) = gI, \quad \forall g \in \mathcal{G},$$

with law

$$\rho_I(A) = \int_{Z_I^{-1}(A)} dg,$$

for all measurable sets $A \subset \mathcal{I}$. Let

$$P : \mathcal{I} \to \mathcal{P}(\mathcal{I}), \quad P(I) = \rho_I, \quad \forall I \in \mathcal{I}.$$

The map P associates to each point a corresponding probability distribution. From the above definition we see that we are essentially viewing an orbit as a distribution of points, and mapping each point in one such distribution. Then we have the following result.

Theorem 2. *For all $I, I' \in \mathcal{I}$*

$$I \sim I' \quad \Leftrightarrow \quad P(I) = P(I'). \tag{5}$$

Proof. We first prove that $I \sim I' \;\Rightarrow\; \rho_I = \rho_{I'}$. Recalling that if $\mathcal{C}_c(\mathcal{I})$ is the set of continuous functions on $\mathcal{I}$ with compact support, ρ_I can be alternatively defined as the unique probability distribution such that

$$\int f(z) d\rho_I(z) = \int f(Z_I(g)) dg, \quad \forall f \in \mathcal{C}_c(\mathcal{I}). \tag{6}$$

Therefore $\rho_I = \rho_{I'}$ if and only if for any $f \in \mathcal{C}_c(\mathcal{I})$, we have $\int_{\mathcal{G}} f(Z_I(g))dg = \int_{\mathcal{G}} f(Z_{I'}(g))dg$ which follows immediately by a change of variable and invariance of the Haar measure:

$$\int_{\mathcal{G}} f(Z_I(g))dg = \int_{\mathcal{G}} f(gI)dg = \int_{\mathcal{G}} f(gI')dg = \int_{\mathcal{G}} f(g\tilde{g}I)dg = \int_{\mathcal{G}} f(\hat{g}I)d\hat{g}$$

To prove that $\rho_I = \rho_{I'} \quad \Rightarrow \quad I \sim I'$, note that $\rho_I(A) - \rho_{I'}(A) = 0$ for all measurable sets $A \subseteq \mathcal{I}$ implies in particular that the support of the probability distributions of I has non null intersection on a set of non zero measure. Since the support of the distributions $\rho_I, \rho_{I'}$ are exactly the orbits associated to I, I' respectively, then the orbits coincide, that is $I \sim I'$. □

The above result shows that an invariant representation can be defined considering the probability distribution naturally associated to each orbit, however its computational realization would require dealing with high-dimensional distributions. Indeed, we next show that the above representation can be further developed to consider only probability distributions on the real line.

4.2.1 Tomographic Probabilistic Representations

We need to introduce some notation and definitions. Let $\mathcal{T} = \mathcal{S}$, the unit sphere in $\mathcal{I}$, and let $\mathcal{P}(\mathbb{R})$ denote the set of probability measures on the real line. For each $t \in \mathcal{T}$, let

$$\pi_t : \mathcal{I} \to \mathbb{R}, \quad \pi_t(I) = \langle I, t \rangle, \quad \forall I \in \mathcal{I}.$$

If $\rho \in \mathcal{P}(\mathcal{I})$, for all $t \in \mathcal{T}$ we denote by $\rho^t \in \mathcal{P}(\mathbb{R})$ the random variable with law given by

$$\rho^t(B) = \int_{\pi_t^{-1}(B)} d\rho,$$

for all measurable sets $B \subset \mathbb{R}$.

Definition 5 (Radon Embedding)**.** *Let $\mathcal{P}(\mathbb{R})^{\mathcal{T}} = \{h \mid h : \mathcal{T} \to \mathcal{P}(\mathbb{R})\}$ and define*

$$R : \mathcal{P}(\mathcal{I}) \to \mathcal{P}(\mathbb{R})^{\mathcal{T}}, \quad R(\rho)(t) = \rho^t, \quad \forall I \in \mathcal{I}.$$

The above map associates to each probability distribution a (continuous) *family* of probability distributions on the real line defined by one dimensional projections (*tomographies*). Interestingly, R can be shown to be a generalization of the Radon Transform to probability distributions [17]. We are going to use it to define the following data representation.

Definition 6 (TP Representation)**.** *We define the Tomographic Probabilistic (TP) representation as*

$$\Psi : \mathcal{I} \to \mathcal{P}(\mathbb{R})^{\mathcal{T}}, \quad \Psi = R \circ P,$$

with P and R as in Definitions 4, 5, respectively.

The TP representation is obtained by first mapping each point in the distribution supported on its orbit and then in a (continuous) family of corresponding one dimensional distributions. The following result characterizes the invariance/selectivity property of the TP representation.

Theorem 3. *Let Ψ be the TP representation in Definition 6, then for all $I, I' \in \mathcal{I}$*

$$I \sim I' \quad \Leftrightarrow \quad \Psi(I) = \Psi(I'). \tag{7}$$

The proof of the above result is obtained combining Theorem 2 with the following well known result, characterizing probability distributions in terms of their one dimensional projections.

Theorem 4 (Cramer-Wold [8]). *For any $\rho, \gamma \in \mathcal{P}(\mathcal{I})$, it holds*

$$\rho = \gamma \quad \Leftrightarrow \quad \rho^t = \gamma^t, \quad \forall t \ \in \mathcal{S}. \tag{8}$$

Through the TP representation, the problem of finding invariant/selective representations reduces to the study of one dimensional distributions, as we discuss next.

4.2.2 CDF Representation

A natural way to describe a one-dimensional probability distribution is to consider the associated cumulative distribution function (CDF). Recall that if $\xi : (\Omega, p) \to \mathbb{R}$ is a random variable with law $q \in \mathcal{P}(\mathbb{R})$, then the associated CDF is given by

$$f_q(b) = q((\infty, b]) = \int dp(a) H(b - \xi(a)), \quad b \in \mathbb{R}, \tag{9}$$

where where H is the Heaviside step function. Also recall that the CDF uniquely defines a probability distribution since, by the Fundamental Theorem of Calculus, we have

$$\frac{d}{db} f_q(b) = \frac{d}{db} \int dp(a) H(b - \xi(a)) = \frac{d}{db} \int_{-\infty}^{b} dp(a) = p(b).$$

We consider the following map.

Definition 7 (CDF Vector Map). *Let $\mathcal{F}(\mathbb{R}) = \{h \mid h : \mathbb{R} \to [0,1]\}$, and*

$$\mathcal{F}(\mathbb{R})^{\mathcal{T}} = \{h \mid h : \mathcal{T} \to \mathcal{F}(\mathbb{R})\}.$$

Define

$$F : \mathcal{P}(\mathbb{R})^{\mathcal{T}} \to \mathcal{F}(\mathbb{R})^{\mathcal{T}}, \quad F(\overline{\gamma})(t) = f_{\overline{\gamma}^t}$$

for $\overline{\gamma} \in \mathcal{P}(\mathbb{R})^{\mathcal{T}}$ and where we let $\overline{\gamma}^t = \overline{\gamma}(t)$ for all $t \in \mathcal{T}$.

The above map associates to a family of probability distributions on the real line their corresponding CDFs. We can then define the following representation.

Definition 8 (CDF Representation). *Let*

$$\mu : \mathcal{I} \to \mathcal{F}(\mathbb{R})^{\mathcal{T}}, \quad \mu = F \circ R \circ P,$$

with F,P and R as in Definitions 7, 4, 5, respectively.

Then, the following result holds.

Theorem 5. *For all $I \in \mathcal{I}$ and $t \in \mathcal{T}$*

$$\mu^t(I)(b) = \int dg \eta_b(\langle I, gt \rangle), \quad b \in \mathbb{R}, \tag{10}$$

where we let $\mu^t(I) = \mu(I)(t)$ and, for all $b \in \mathbb{R}$, $\eta_b : \mathbb{R} \to \mathbb{R}$, is given by $\eta_b(a) = H(b - a), \quad a \in \mathbb{R}$. Moreover, for all $I, I' \in \mathcal{I}$

$$I \sim I' \quad \Leftrightarrow \quad \mu(I) = \mu(I').$$

Proof. The proof follows noting that μ is the composition of the one to one maps F, R and a map P that is one to one w.r.t. the equivalence classes induced by the group of transformations G. Therefore μ is one to one w.r.t. the equivalence classes i.e. $I \sim I' \quad \Leftrightarrow \quad \mu(I) = \mu(I')$. □

We note that, from a direct comparison, one can see that (10) is of the form (4). Different measurements correspond to different choices of the threshold b.

Remark 4. *[Pooling Functions: from CDF to Moments and Beyond] The above reasoning suggests that a principled choice for the non linearity in* (4) *is a step function, which in practice could be replaced by a smooth approximation such a sigmoidal function. Interestingly, other choices of non linearities could be considered. For example, considering different powers would yield information on the moments of the distributions (more general non linear function than powers would yield generalized moments). This latter point of view is discussed in some detail in Appendix A.*

4.3 Templates Sampling and Metric Embedings

We next discuss what happens if only a finite number of (possibly random) templates are available. In this case, while invariance can be ensured, in general we cannot expect selectivity to be preserved. However, it is possible to show that the representation is *almost* selective (see below) if a sufficiently large number number of templates is available.

Towards this end we introduce a metric structure on the representation space. Recall that if $\rho, \rho' \in \mathcal{P}(\mathbb{R})$ are two probability distributions on the real line and $f_\rho, f_{\rho'}$ their cumulative distributions functions, then the uniform Kolmogorov-Smirnov (KS) metric is induced by the uniform norm of the cumulative distributions that is

$$d_\infty(f_\rho, f_{\rho'}) = \sup_{s \in \mathbb{R}} |f_\rho(s) - f_{\rho'}(s)|,$$

and takes values in $[0, 1]$. Then, if μ is the representation in (10) we can consider the metric

$$d(I, I') = \int du(t) d_\infty(\mu^t(I), \mu^t(I')) \tag{11}$$

where u is the (normalized) uniform measure on the sphere $\mathcal{S}$. We note that, theorems 4 and 5 ensure that (11) is a well defined metric on the quotient space induced by the group transformations, in particular

$$d(I, I') = 0 \Leftrightarrow I \sim I'.$$

If we consider the case in which only a finite set $\mathcal{T}_k = \{t_1, \ldots, t_k\} \subset \mathcal{S}$ of k templates is available, each point is mapped in a finite sequence of probability distributions or CDFs and (11) is replaced by

$$\widehat{d}(I, I') = \frac{1}{k} \sum_{i=1}^{k} d_\infty(\mu^{t_i}(I), \mu^{t_i}(I')) \tag{12}$$

Clearly, in this case we cannot expect to be able to discriminate every pair of points, however we have the following result.

Theorem 6. *Consider n images $\mathcal{I}_n$ in $\mathcal{I}$. Let $k \geq \frac{2}{c\epsilon^2} \log \frac{n}{\delta}$, where c is a constant. Then with probability $1 - \delta^2$,*

$$|d(I, I') - \widehat{d}(I, I')| \leq \epsilon. \tag{13}$$

for all $I, I' \in \mathcal{I}_n$.

Proof. The proof follows from a direct application of Höeffding's inequality and a union bound. Fix $I, I' \in \mathcal{I}_n$. Define the real random variable $Z : \mathcal{S} \to [0, 1]$,

$$Z(t_i) = d_\infty(\mu^{t_i}(I), \mu^{t_i}(I')), \quad i = 1, \ldots, k.$$

From the definitions it follows that $\|Z\| \leq 1$ and $\mathbb{E}(Z) = d(I, I')$. Then, Höeffding inequality implies

$$|d(I, I') - \widehat{d}(I, I')| = |\frac{1}{k} \sum_{i=1}^{k} \mathbb{E}(Z) - Z(t_i)| \geq \epsilon,$$

with probability at most $2e^{-\epsilon^2 k}$. A union bound implies that the result holds uniformly on $\mathcal{I}_n$ with probability at least $n^2 2e^{-\epsilon^2 k}$. The proof is concluded setting this probability to δ^2 and taking $k \geq \frac{2}{c\epsilon^2} \log \frac{n}{\delta}$. □

We note that, while we considered the KS metric for convenience, other metrics over probability distributions can be considered. Also, we note that a natural further question is how discretization/sampling of the group affects the representation. The above reasoning could be extended to yield results in this latter case. Finally, we note that, when compared to classical results on distance preserving embedding, such as Johnson Linderstrauss Lemma [18], Theorem 12 only ensures distance preservation up to a given accuracy which increases with a larger number of projections. This is hardly surprising, since the problem of finding suitable embedding for probability spaces is known to be considerably harder than the analogue problem for vector spaces [2]. The question of how devise strategies to define distance preserving embedding is an interesting open problem.

5 Locally Invariant and Covariant Representations

We consider the case where a representation is given by collection of *"local"* group averages, and refer to this situation as the partially observable group (POG) case. Roughly speaking, the idea is that this kind of measurements can be invariant to sufficiently small transformations, i.e. be locally invariant. Moreover, representations given by collections of POG averages can be shown to be *covariant* (see section 5.2 for a definition).

5.1 Partially Observable Group Averages

For a subset $\mathcal{G}_0 \subset \mathcal{G}$ consider a POG measurement of the form

$$\psi(I) = \int_{\mathcal{G}_0} dg \eta(\langle I, gt \rangle). \tag{14}$$

The above quantity can be interpreted as the "response" of a cell that can perceive visual stimuli within a "window" (receptive field) of size $\mathcal{G}_0$. A POG measurement corresponds to a local group average restricted to a subset of transformations $\mathcal{G}_0$. Clearly, such a measurement will not in general be invariant. Consider a POG measurement on a transformed point

$$\int_{\mathcal{G}_0} dg \eta(\langle \tilde{g} I, gt \rangle) = \int_{\mathcal{G}_0} dg \eta(\langle I, \tilde{g}^{-1} gt \rangle) = \int_{\tilde{g}\mathcal{G}_0} dg \eta(\langle I, gt \rangle).$$

If we compare the POG measurements on the same point with and without a transformation, we have

$$| \int_{\mathcal{G}_0} dg \eta(\langle I, gt \rangle) - \int_{\tilde{g}\mathcal{G}_0} dg \eta(\langle I, gt \rangle)|. \tag{15}$$

While there are several situations in which the above difference can be zero, the intuition from the vision interpretation is that the same response should be obtained if a sufficiently small object does not move (transform) too much with respect to the receptive field size. This latter situation can be described by the assumption that the function

$$h : \mathcal{G} \to \mathbb{R}, \quad h(g) = \eta(\langle I, gt \rangle)$$

is zero outside of the intersection of $\tilde{g}\mathcal{G}_0 \cap \mathcal{G}_0$. Indeed, for all $\tilde{g} \in \mathcal{G}$ satisfying this latter assumption, the difference in (15) would clearly be zero. The above reasoning results in the following theorem.

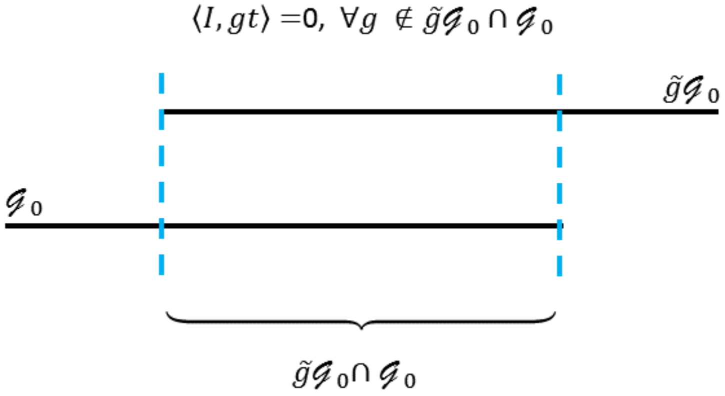
Theorem 7. *Given $I \in \mathcal{I}$ and $t \in \mathcal{T}$, assume that there exists a set $\tilde{\mathcal{G}} \subset \mathcal{G}$ such that, for all $\tilde{g} \in \tilde{\mathcal{G}}$,*

$$\eta(\langle I, gt \rangle) = 0 \quad \forall g \notin \tilde{g}\mathcal{G}_0 \cap \mathcal{G}_0. \tag{16}$$

Then for $\tilde{g} \in \tilde{\mathcal{G}}$

$$\psi(I) = \psi(\tilde{g} I),$$

with ψ as in (14).



$\langle I, gt \rangle = 0, \ \forall g \notin \tilde{g}\mathcal{G}_0 \cap \mathcal{G}_0$
$\tilde{g}\mathcal{G}_0$
$\mathcal{G}_0$
$\tilde{g}\mathcal{G}_0 \cap \mathcal{G}_0$

Figure 1: A sufficient condition for invariance for locally compact groups: if $\langle gI, t \rangle = 0$ for all $g \in \tilde{g}\mathcal{G}_0 \Delta \mathcal{G}_0$, the integral of $\eta_b \langle I, gt \rangle$ over $\mathcal{G}_0$ or $\tilde{g}\mathcal{G}_0$ will be equal.

We add a few comments. First, we note that condition (16) can be weakened requiring only $\eta(\langle I, gt \rangle) = 0$ for all $g \in \tilde{g}\mathcal{G}_0 \Delta \mathcal{G}_0$, where we denote by Δ the symmetric difference of two sets ($A \Delta B = (A \cup B)/(A \cap B)$ with A, B sets). Second, we note that if the non linearity η is zero only in zero, then we can rewrite condition (16) as

$$\langle I, gt \rangle = 0, \quad \forall g \in \tilde{g}\mathcal{G}_0 \Delta \mathcal{G}_0.$$

Finally, we note that the latter expression has a simple interpretation in the case of the translation group. In fact, we can interpret (16) as a spatial localization condition on the image I and the template t (assumed to be positive valued functions), see Figure 1. We conclude with the following remark.

Remark 5 (Localization Condition and V1)**.** *Regarding the localization condition discussed above, as we comment elsewhere [3], the fact that a template needs to be localized could have implications from a biological modeling standpoint. More precisely, it could provides a theoretical foundation of the Gabor like shape of the responses observed in V1 cells in the visual cortex [23, 3, 5].*

Remark 6 (More on the Localization Condition)**.** *From a more mathematical point of view, an interesting question is about conditions under which whether the localization condition* (16) *is also necessary rather than only sufficient.*

5.2 POG Representation

For all $\bar{g} \in G$, let $\bar{g}\mathcal{G}_0 = \{g \in \mathcal{G} \mid g = \bar{g}g', \quad g' \in \mathcal{G}_0\}$, the collection of "local" subsets of the group obtained from the subset $\mathcal{G}_0$. Moreover, let

$$V = \int_{\mathcal{G}_0} dg.$$

Clearly, by the invariance of the measure, we have $\int_{\bar{g}\mathcal{G}_0} dg = V$, for all $\bar{g} \in \mathcal{G}$. Then, for all $I \in \mathcal{I}$, $\bar{g} \in \mathcal{G}$, define the random variables

$$Z_{I,\bar{g}} : \bar{g}\mathcal{G}_0 \to \mathcal{I}, \quad Z_{I,\bar{g}}(g) = gI, \quad g \in \bar{g}\mathcal{G}_0, \tag{17}$$

with laws

$$\rho_{I,\bar{g}}(A) = \frac{1}{V} \int_{Z^{-1}_{I,\bar{g}}(A)} dg, t$$

for all measurable sets $A \subset \mathcal{I}$. For each $I \in \mathcal{I}$, $\bar{g} \in \mathcal{G}$, the measure $\rho_{I,\bar{g}}$ corresponds to the distribution on the fraction of the orbit corresponding to the observable group subset $\bar{g}\mathcal{G}_0$. Then we can represent each point with a collection of POG distributions.

Definition 9 (Representation via POG Probabilities)**.** *Let $\mathcal{P}(\mathcal{I})^{\mathcal{G}} = \{h \mid h : \mathcal{G} \to \mathcal{P}(\mathcal{I})\}$ and define*

$$\bar{P} : \mathcal{I} \to \mathcal{P}(\mathcal{I})^{\mathcal{G}}, \quad \bar{P}(I)(g) = \rho_{I,g} \quad \forall I \in \mathcal{I}, g \in \mathcal{G}$$

Each point is mapped in the collection of distributions obtained considering all possible fractions of the orbit corresponding to $\bar{g}\mathcal{G}_0$, $\bar{g} \in \mathcal{G}$. Note that, the action of an element $\tilde{g} \in \mathcal{G}$ of the group on the POG probability representation is given by

$$\tilde{g}\bar{P}(I)(g) = \bar{P}(I)(\tilde{g}g)$$

for all $g \in \mathcal{G}$. The following result holds.

Theorem 8. *Let $\bar{P}$ as in Definition* (9)*. Then for all $I, I' \in \mathcal{I}$ if*

$$I \sim I' \quad \Rightarrow \quad \exists \tilde{g} \in \mathcal{G} \text{ such that } \bar{P}(I') = \tilde{g}\bar{P}(I). \tag{18}$$

Equivalently, for all $I, I' \in \mathcal{I}$ if

$$I' = \tilde{g}I$$

then

$$\bar{P}(I')(g) = \bar{P}(I)(g\tilde{g}), \quad \forall g \in \mathcal{G}. \tag{19}$$

i.e. $\bar{P}$ is covariant*.*

Proof. The proof follows noting that $\rho_{I',\bar{g}} = \rho_{I,\bar{g}\tilde{g}}$ holds since, using the same characterization of ρ as in (6),we have that for any $f \in \mathcal{C}_c(\mathcal{I})$

$$\int_{\bar{g}\mathcal{G}_0} f(Z_{I',\bar{g}}(g))dg = \int_{\bar{g}\mathcal{G}_0} f(gI')dg = \int_{\bar{g}\mathcal{G}_0} f(g\tilde{g}I)dg = \int_{\bar{g}\mathcal{G}_0\tilde{g}} f(gI)dg$$

where we used the invariance of the measure. □

Following the reasoning in the previous sections and recalling Definition 5, we consider the mapping given by one dimensional projections (tomographies) and corresponding representations.

Definition 10 (TP-POG Representation)**.** *Let $\mathcal{P}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}} = \{h \mid h : \mathcal{G}\times\mathcal{T} \to \mathcal{P}(\mathbb{R})\}$ and define*

$$\bar{R} : \mathcal{P}(\mathcal{I})^{\mathcal{G}} \to \mathcal{P}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}}, \quad \bar{R}(h)(g,t) = R(h(g))(t) = h^t(g),$$

for all $h \in \mathcal{P}(\mathcal{I})^{\mathcal{G}}, g \in \mathcal{G}, t \in \mathcal{T}$. Moreover, we define the Tomographic Probabilistic POG representation as

$$\bar{\Psi} : \mathcal{I} \to \mathcal{P}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}}, \quad \bar{\Psi} = \bar{R}\circ\bar{P},$$

with $\bar{P}$ as in Definition 9.

We have the following result:

Theorem 9. *The representation $\bar{\Psi}$ defined in 10 is covariant, i.e. $\bar{\Psi}(\tilde{g}I)(g) = \bar{\Psi}(I)(\tilde{g}g)$.*

Proof. The map $\bar{\Psi} = \bar{R}\circ\bar{P}$ is covariant if both $\bar{R}$ and $\bar{P}$ are covariant. The map $\bar{P}$ was proven to be covariant in Theorem 8. We then need to prove the covariance of $\bar{R}$ i.e. $\tilde{g}\bar{R}(h)(g,t) = \bar{R}(h)(\tilde{g}g,t)$ for all $h \in \mathcal{P}(\mathcal{I})^{\mathcal{G}}$. This follows from

$$\bar{R}(\tilde{g}h)(g,t) = R(\tilde{g}h(g))(t) = R(h(\tilde{g}g))(t) = R(h)(\tilde{g}g,t).$$

□

The TP-POG representation is obtained by first mapping each point I in the family of distributions $\rho_{I,g}$, $g \in \mathcal{G}$ supported on the orbit fragments corresponding to POG and then in a (continuous) family of corresponding one dimensional distributions $\rho^t_{I,g}$, $g \in \mathcal{G}$, $t \in \mathcal{T}$. Finally, we can consider the representation obtained representing each distribution via the corresponding CDF.

Definition 11 (CDF-POG Representation)**.** *Let $\mathcal{F}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}} = \{h \mid h : \mathcal{G}\times\mathcal{T} \to \mathcal{F}(\mathbb{R})\}$ and define*

$$\bar{F} : \mathcal{P}(\mathcal{I})^{\mathcal{G}\times\mathcal{T}} \to \mathcal{P}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}}, \quad \bar{F}(h)(g,t) = F(h(g,t)) = f_{h(g,t)},$$

for all $h \in \mathcal{P}(\mathcal{I})^{\mathcal{G}\times\mathcal{T}}$ and $g \in \mathcal{G}, t \in \mathcal{T}$. Moreover, define the CDF-POG representation as

$$\bar{\mu} : \mathcal{I} \to \mathcal{F}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}}, \quad \bar{\mu} = \bar{F}\circ\bar{R}\circ\bar{P},$$

with $\bar{P}$,$\bar{F}$ as in Definition 9, 10, respectively.

It is easy to show that

$$\mu_{\bar{g},t}(I)(b) = \int_{\bar{g}\mathcal{G}_0} \eta_b(\langle I, gt\rangle)dg. \tag{20}$$

where we let $\mu_{\bar{g},t}(I) = \mu(I)(\bar{g},t)$.

6 Further Developments: Hierarchical Representation

In this section we discuss some further developments of the framework presented in the previous section. In particular, we sketch how multi-layer (deep) representations can be obtained abstracting and iterating the basic ideas introduced before.

Hierarchical representations, based on multiple layers of computations, have naturally arisen from models of information processing in the brain [11, 26]. They have also been critically important in recent machine learning successes in a variety of engineering applications, see e.g. [27]. In this section we address the question of how to generalize the framework previously introduced to consider multi-layer representations.

Recall that the basic idea for building invariant/selective representation is to consider local (or global) measurements of the form

$$\int_{\mathcal{G}_0} \eta(\langle I, gt \rangle) dg, \tag{21}$$

with $\mathcal{G}_0 \subseteq \mathcal{G}$. A main difficulty to iterate this idea is that, following the development in previous sections, the representation (11)-(20), induced by collection of (local) group averages, maps the data space $\mathcal{I}$ in the space $\mathcal{P}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}}$. The latter space lacks an inner product as well as natural linear structure needed to define the measurements in (21). One possibility to overcome this problem is to consider an embedding in a suitable Hilbert space. The first step in this direction is to consider an embedding of the probability space $\mathcal{P}(\mathbb{R})$ in a (real separable) Hilbert space $\mathcal{H}$. Interestingly, this can be achieved considering a variety of reproducing kernels over probability distributions, as we describe in Appendix B. Here we note that if $\Phi : \mathcal{P}(\mathbb{R}) \to \mathcal{H}$ is one such embeddings, then we could consider a corresponding embedding of $\mathcal{P}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}}$ in the space

$$L^2(\mathcal{G}\times\mathcal{T}, \mathcal{H}) = \{h : \mathcal{G}\times\mathcal{T} \to \mathcal{H} \mid \int \|h(g,t)\|^2 \, dgdu(t)\}$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm induced by the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ in $\mathcal{H}$ and u is the uniform measure on the sphere $\mathcal{S} \subset \mathcal{I}$. The space $L^2(\mathcal{G}\times\mathcal{T}, \mathcal{H})$ is endowed with the inner product

$$\langle h, h' \rangle_{\mathcal{H}} = \int \langle h(g,t), h'(g,t) \rangle_{\mathcal{H}}^2 \, dgdu(t),$$

for all $h, h' \in L^2(\mathcal{G}\times\mathcal{T}, \mathcal{H})$, so that the corresponding norm is exactly

$$\|h\|_{\mathcal{H}}^2 = \int \|h(g,t)\|^2 \, dgdu(t).$$

The embedding of $\mathcal{P}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}}$ in $L^2(\mathcal{G}\times\mathcal{T}, \mathcal{H})$ is simply given by

$$J_\Phi : \mathcal{P}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}} \to L^2(\mathcal{G}\times\mathcal{T}, \mathcal{H}), \quad J_\Phi(\rho)(g,t) = \Phi(\rho(g,t)) \quad \text{i.e.}$$

for all $\rho \in \mathcal{P}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}}$. Provided with above notation we have the following result.

Theorem 10. *The representation defined by*

$$\bar{Q} : \mathcal{I} \to L^2(\mathcal{G} \times \mathcal{T}, \mathcal{H}), \quad \bar{Q} = J_\Phi \circ \bar{\Psi}. \tag{22}$$

with $\bar{\Psi}$ as in Definition 10, is covariant, in the sense that,

$$\bar{Q}(gI) = g\bar{Q}(I)$$

for all $I \in \mathcal{I}$, $g \in \mathcal{G}$.

Proof. The proof follows checking that by definition both $\bar{R}$ and J_Φ are covariant and using Theorem 8. The fact that $\bar{R}$ is covariant was proven in Th. 9. The covariance of J_Φ, i.e. $\tilde{g}J_\Phi(h)(g,t) = J_\Phi(h)(\tilde{g}g,t)$ for all $h \in \mathcal{P}(\mathbb{R})^{\mathcal{G}\times\mathcal{T}}$, follows from

$$J_\Phi(\tilde{g}h)(g,t) = \Phi(\tilde{g}h(g,t)) = \Phi(h(\tilde{g}g,t)) = J_\Phi(h)(\tilde{g}g,t).$$

Now since $\bar{P}$ was already proven covariant in Th. 8 we have that, being $\bar{Q} = J_\Phi \circ \bar{R} \circ \bar{P}$ composition of covariant representations, $\bar{Q}$ is covariant i.e. $\tilde{g}\bar{Q}(I) = \bar{Q}(\tilde{g}I)$. □

Using the above definitions a *second layer* invariant measurement can be defined considering,

$$v : \mathcal{I} \to \mathbb{R}, \quad v(I) = \int_{\mathcal{G}_0} \eta(\langle \bar{Q}(x), g\tau \rangle_2) dg \tag{23}$$

where $\tau \in L^2(\mathcal{G} \times \mathcal{T}, \mathcal{H})$ has unit norm.

We add several comments. First, following the analysis in the previous sections Equation (23) can be used to define invariant (or locally invariant) measurements and hence representations defined by collections of measurements. Second, the construction can be further iterated to consider multi-layer representations, where at each layer an intermediate representation is obtained considering "distributions of distributions". Third, considering multiple layers naturally begs the question of how the number and properties of each layer affect the properties of the representation. Preliminary answers to these questions are described in [3, 4, 21, 23]. A full mathematical treatment is beyond the scope of the current paper which however provides a formal framework to tackle them in future work.

7 Discussion

Motivated by the goal of characterizing good data representation that can be learned, this paper studies the mathematics of an approach to learn data representation that are invariant and selective to suitable transformations. While invariance can be proved rather directly from the invariance of the Haar measure associated with the group, characterizing selectivity requires a novel probabilistic argument developed in the previous sections.

Several extensions of the theory are natural and have been sketched with preliminary results in [3, 4, 21, 23]. The main directions that need a rigorous theory extending the results of this paper are:

- Hierarchical architectures. We described how the theory can be used to analyze local invariance properties, in particular for locally compact groups. We described covariance properties. Covariant layers can integrate representations that are locally invariant into representations that are more globally invariant.

- Approximate invariance for transformations that are not groups. The same basic algorithm analyzed in this paper is used to yield approximate invariance, provided the templates transforms as the image, which requires the templates to be tuned to specific object classes.

We conclude with a few general remarks connecting our paper with this special issue on deep learning and especially with an eventual theory of such networks. *Hierarchical architectures of simple and complex units.* Feedforward architecture with n layers, consisting of dot products and nonlinear pooling functions, are quite general computing devices, basically equivalent to Turing machines running for n time points (for example the layers of the HMAX architecture in [26] can be described as AND operations (dot products) followed by OR operations (pooling), i.e. as disjunctions of conjunctions.). Given a very large set of labeled examples it is not too surprising that greedy algorithms such as stochastic gradient descent can find satisfactory parameters in such an architecture, as shown by the recent successes of Deep Convolutional Networks. Supervised learning with millions of examples, however, is not, in general, biologically plausible. Our theory can be seen as proving that a form of unsupervised learning in convolutional architectures is possible and effective, because it provides invariant representations with small sample complexity.
Two stages: group and non-group transformations. The core of the theory applies to compact groups such as rotations of the image in the image plane. Exact invariance for each module is equivalent to a localization condition which could be interpreted as a form of sparsity [3]. If the condition is relaxed to hold approximately it becomes a *sparsity condition for the class of images w.r.t. the dictionary t^k under the group G* when restricted to a subclass of similar images. This property, which is similar to compressive sensing "incoherence" (but in a group context), requires that I and t^k have a representation with rather sharply peaked autocorrelation (and correlation) and guarantees approximate invariance for transformations which do not have group structure, see [21].
Robustness of pooling. It is interesting that the theory is robust with respect to the pooling nonlinearity. Indeed, as discussed, very general class of nonlinearities will work, see Appendix A. Any nonlinearity will provide invariance, if the nonlinearity does not change with time and is the same for all the simple cells pooled by the same complex cells. A sufficient number of different nonlinearities, each corresponding to a complex cell, can provide selectivity [3].
Biological predictions and biophysics, including dimensionality reduction and

PCAs. There are at least two possible biophysical models for the theory. The first is the original Hubel and Wiesel model of simple cells feeding into a complex cell. The theory proposes the "ideal" computation of a CDF, in which case the nonlinearity at the output of the simple cells is a threshold. A complex cell, summating the outputs of a set of simple cells, would then represent a bin of the histogram; a different complex cell in the same position pooling a set of similar simple cells with a different threshold would represent another bin of the histogram.

The second biophysical model for the HW module that implements the computation required by i-theory consists of a single cell where dendritic branches play the role of simple cells (each branch containing a set of synapses with weights providing, for instance, Gabor-like tuning of the dendritic branch) with inputs from the LGN; active properties of the dendritic membrane distal to the soma provide separate threshold-like nonlinearities for each branch separately, while the soma summates the contributions for all the branches. This model would solve the puzzle that so far there seems to be no morphological difference between pyramidal cells classified as simple vs complex by physiologists. Further if the synapses are Hebbian it can be proved that Hebb's rule, appropriately modified with a normalization factor, is an online algorithm to compute the eigenvectors of the input covariance matrix, therefore tuning the dendritic branches weights to principal components and thus providing an efficient dimensionality reduction.

*($n \rightarrow 1$).*The present phase of Machine Learning is characterized by supervised learning algorithms relying on large sets of labeled examples ($n \rightarrow \infty$). The next phase is likely to focus on algorithms capable of learning from very few labeled examples ($n \rightarrow 1$), like humans seem able to do. We propose and analyze a possible approach to this problem based on the unsupervised, automatic learning of a good representation for supervised learning, characterized by small sample complexity (n). In this view we take a step towards a major challenge in learning theory beyond the supervised learning, that is the problem of *representation learning*, formulated here as the unsupervised learning of invariant representations that significantly reduce the sample complexity of the supervised learning stage.

Acknowledgment

We would like to thank the McGovern Institute for Brain Research for their support. This research was sponsored by grants from the National Science Foundation, AFSOR-THRL (FA8650-05-C-7262). Additional support was provided by the Eugene McDermott Foundation.

References

[1] Akhiezer, N. (1965) *The classical moment problem: and some related questions in analysis*, University mathematical monographs. Oliver & Boyd.

[2] Andoni, A., Ba, K. D., Indyk, P. & Woodruff, D. P. (2009) Efficient Sketches for Earth-Mover Distance, with Applications.. in *FOCS*, pp. 324–330. IEEE Computer Society.

[3] Anselmi, F., Leibo, J. Z., Rosasco, L., Mutch, J., Tacchetti, A. & Poggio, T. (2013) Unsupervised Learning of Invariant Representations in Hierarchical Architectures. *arXiv preprint 1311.4158*.

[4] Anselmi, F. & Poggio, T. (2010) Representation Learning in Sensory Cortex: a theory. *CBMM memo n 26*.

[5] Anselmi F. Leibo J.Z. Rosasco L. Mutch J., T. A. P. T. (2013) Magic Materials: a theory of deep hierarchical architectures for learning sensory representations. *CBCL paper*.

[6] Bengio, Y. (2009) Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning 2*.

[7] Berlinet, A. & Thomas-Agnan, C. (2004) *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, Boston.

[8] Cramer, H. & Wold, H. (1936) Some theorems on distribution functions. *J. London Math. Soc.*, **4**, 290–294.

[9] Cucker, F. & Smale, S. (2002) On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, (39), 1–49.

[10] Edelsbrunner, H. & Harer, J. L. (2010) *Computational Topology, An Introduction*. American Mathematical Society.

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

[12] Haasdonk, B. & Burkhardt, H. (2007) Invariant Kernel Functions for Pattern Analysis and Machine Learning. *Mach. Learn.*, **68**(1), 35–61.

[13] Hein, M. & Bousquet, O. (2005) Hilbertian Metrics and Positive Definite Kernels on Probability Measures. in *AISTATS 2005*, ed. by Z. G. Cowell, R., pp. 136–143. Max-Planck-Gesellschaft.

[14] Hubel, D. & Wiesel, T. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, **160**(1), 106.

[15] Hubel, D. & Wiesel, T. (1965) Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, **28**(2), 229.

[16] Hubel, D. & Wiesel, T. (1968) Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, **195**(1), 215.

[17] Jan Boman, F. L. (2009) Support Theorems for the Radon Transform and Cramr-Wold Theorems. *Journal of Theoretical Probability,*.

[18] Johnson, W.B. Lindenstrauss, J. (1984) Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, **26**.

[19] Kazhdan, M., Funkhouser, T. & Rusinkiewicz, S. (2003) Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors. in *Proceedings of the 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, SGP '03, pp. 156–164.

[20] LeCun, Y. & Bengio, Y. (1995) Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, pp. 255–258.

[21] Leibo, J. Z., Liao, Q., Anselmi, F. & Poggio, T. (2014) The invariance hypothesis implies domain-specific regions in visual cortex. *http://dx.doi.org/10.1101/004473.*

[22] Mallat, S. (2012) Group Invariant Scattering. *Communications on Pure and Applied Mathematics*, **65**(10), 1331–1398.

[23] Poggio, T., Mutch, J., Anselmi, F., Tacchetti, A., Rosasco, L. & Leibo, J. Z. (2013) Does invariant recognition predict tuning of neurons in sensory cortex?. *MIT-CSAIL-TR-2013-019, CBCL-313.*

[24] Ramm, A. G. (1998) On the theory of reproducing kernel hilbert spaces. .

[25] Reed, M. & Simon, B. (1978) *Methods of modern mathematical physics. II. , Fourier Analysis, Self-Adjointness.* Academic Press, London.

[26] Riesenhuber, M. & Poggio, T. (1999) Hierarchical models of object recognition in cortex. *Nature Neuroscience*, **2**(11), 1019–1025.

[27] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. & LeCun, Y. (2014) OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. in *International Conference on Learning Representations (ICLR2014).* CBLS.

[28] Soatto, S. (2009) Actionable information in vision. in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2138–2145. IEEE.

[29] Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B. & Lanckriet, G. R. G. (2010) Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, **11**, 1517–1561.

[30] Tosic, I. & Frossard, P. (2011) Dictionary learning: What is the right representation for my signal?. *IEEE Signal Processing Magazine*, **28**(2), 27–38.

[31] Vapnik, V. (1982) *Estimation of dependencies based on empirical data.* Springer Verlag.

[32] Vedaldi, A. & Zisserman, A. (2011) Efficient Additive Kernels via Explicit Feature Maps. *Pattern Analysis and Machine Intellingence*, **34**(3).

[33] Xing, E. P., Ng, A. Y., Jordan, M. I. & Russell, S. (2003) Distance Metric Learning, With Application To Clustering With Side-Information. in *Advances in neural information processing systems*, pp. 505–512. MIT Press.

A Representation Via Moments

In Section 4.2.2 we have discussed the derivation of invariant selective representation considering the CDFs of suitable one dimensional probability distributions. As we commented in Remark 4 alternative representations are possible, for example by considering moments. Here we discuss this point of view in some more detail.

Recall that if $\xi : (\Omega, p) \to \mathbb{R}$ is a random variables with law $q \in \mathcal{P}(\mathbb{R})$, then the associated moment vector is given is given by

$$m_q^r = \mathbb{E}|\xi|^r = \int dq|\xi|^r, \quad r \in N. \tag{24}$$

In this case we have the following definitions and results.

Definition 12 (Moments Vector Map)**.** *Let $\mathcal{M}(\mathbb{R}) = \{h \mid h : \mathbb{N} \to \mathbb{R}\}$, and*

$$\mathcal{M}(\mathbb{R})^{\mathcal{T}} = \{h \mid h : \mathcal{T} \to \mathcal{M}(\mathbb{R})\}.$$

Define

$$M : \mathcal{P}(\mathbb{R})^{\mathcal{T}} \to \mathcal{M}(\mathbb{R})^{\mathcal{T}}, \quad M(\overline{\mu})(t) = m_{\mu^t}$$

for $\overline{\mu} \in \mathcal{P}(\mathbb{R})$ and where we let $\overline{\mu}(t) = \overline{\mu}^t$, for all $t \in \mathcal{T}$.

The above mapping associates to each one dimensional distribution the corresponding vector of moments. Recall that this association uniquely determines the probability distribution if the so called Carleman's condition is satisfied:

$$\sum_{r=1}^{\infty} m_{2r}^{-\frac{1}{2r}} = +\infty$$

where m_r is the set of moments of the distribution.
We can then define the following representation.

Definition 13 (Moments Representation)**.** *Let*

$$\mu : \mathcal{I} \to \mathcal{M}(\mathbb{R})^{\mathcal{T}}, \quad \mu = M \circ R \circ P,$$

with M,P and R as in Definitions 12, 4, 5, respectively.

Then, the following result holds.

Theorem 11. *For all $I \in \mathcal{I}$ and $t \in \mathcal{T}$*

$$\mu^t(I)(r) = \int dg| \langle I, gt \rangle |^r, \quad r \in \mathbb{N},$$

where we let $\mu(I)(t) = \mu^t(I)$. Moreover, for all $I, I' \in \mathcal{I}$

$$I \sim I' \quad \Leftrightarrow \quad \mu(I) = \mu(I').$$

Proof. $\mu = M \circ R \circ P$ is a composition of a one to one map R, a map P that is one to one w.r.t. the equivalence classes induced by the group of transformations $\mathcal{G}$ and a map M that is one to one since Carleman's condition is satisfied. Indeed, we have,

$$\sum_{r=1}^{\infty} \Big(\int \; dg \, \langle I, gt \rangle^{2r} \Big)^{-\frac{1}{2r}} \leq \sum_{r=1}^{\infty} \Big(\int \; dg| \langle I, gt \rangle | \Big)^{-\frac{1}{2r}2r} = \sum_{r=1}^{\infty} \frac{1}{C} = +\infty$$

where $C = \int \; dg| \langle I, gt \rangle |$. Therefore μ is one to one w.r.t. the equivalence classes i.e. $I \sim I' \quad \Leftrightarrow \quad \mu(I) = \mu(I')$. □

We add one remark regarding possible developments of the above result.

Remark 7. *Note that the above result essentially depends on the characterization of the moment problem of probability distributions on the real line. In this view, it could be further developed to consider for example the* truncated *case when only a finite number of moments is considered or the generalized moments problem, where families of (nonlinear) continuous functions, more general than powers, are considered (see e.g. [1]).*

B Kernels on probability distributions

To consider multi-layers within the framework proposed in the paper we need to embed probability spaces in Hilbert spaces. A natural way to do so is by considering appropriate positive definite (PD) kernels, that is symmetric functions $K : X \times X \to \mathbb{R}$ such that

$$\sum_{i,j=1}^{n} K(\rho_i, \rho_j)\alpha_i\alpha_j \geq 0$$

for all $\forall \rho_1, \dots, \rho_n \in X, \alpha_1, \dots, \alpha_n \in \mathbb{R}$ and where X is any set, e.g. $X = \mathbb{R}$ or $X = \mathcal{P}(\mathbb{R})$. Indeed, PD kernels are known to define a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ for which they correspond to reproducing kernels, in the sense that if $\mathcal{H}_K$ is the RKHS defined by K, then $K_x = K(x, \cdot) \in \mathcal{H}_K$ for all $x \in X$ and

$$\langle f, K_x \rangle_K = f(x), \quad \forall f \in \mathcal{H}_K, x \in X, \tag{25}$$

where $\langle \cdot, \cdot \rangle_K$ is the inner product in $\mathcal{H}_K$ (see for example [7] for an introduction to RKHS).

Many examples of kernels on distributions are known and have been studied. For example [13, 32] discuss a variety of kernels of the form

$$K(\rho, \rho') = \int \int d\gamma(x) \kappa(p_\rho(x), p_{\rho'}(x))$$

where $p_\rho, p_{\rho'}$ are the densities of the measures ρ, ρ' with respect to a dominating measure γ (which is assumed to exist) and $\kappa : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}$ is a PD kernel. Recalling that a PD kernel defines a pseudo-metric via the equation

$$d_K(\rho, \rho')^2 = K(\rho, \rho) + K(\rho', \rho) - 2K(\rho, \rho').$$

it is shown in [13, 32] how different classic metric on probability distributions can be recovered by suitable choices of the kernel κ. For example,

$$\kappa(x, x') = \sqrt{xx'},$$

corresponds to the Hellinger's distance, see [13, 32] for other examples.

A different approach is based on defining kernels of the form

$$K(\rho, \rho') = \int \int d\rho(x) d\rho'(x') k(x, x'), \tag{26}$$

where $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a PD kernel. Using the reproducing property of k we can write

$$K(\rho, \rho') = \left\langle \int d\rho(x') k_x, \int d\rho(x) k_{x'} \right\rangle_k = \langle \Phi(\rho), \Phi(\rho') \rangle$$

where $\Phi : \mathcal{P}(\mathbb{R}) \to \mathcal{H}$ is the embedding $\Phi(x) = \int d\rho(x') k_x$ mapping each distribution in a corresponding kernel *mean*, see e.g. [7]. Condition on the kernel k, hence on K, ensuring that the corresponding function d_K is a metric have been studied in detail, see e.g. [29].