Categorical and geometric methods in statistical, manifold, and machine learning

Hông Vân Lê,[0000−0003−4822−6466]
Hà Quang Minh,[0000−0003−3926−8875]
Frederic Protin,[000−0003−1702−9451]
Wilderich Tuschmann[0000−002−1149−2800]

Abstract We present and discuss applications of the category of probabilistic morphisms, initially developed in [37], as well as some geometric methods to several classes of problems in statistical, machine and manifold learning which shall be, along with many other topics, considered in depth in the forthcoming book [39].

1 Introduction

Let us start with a general concept of learning and machine learning. *Learning* is a process of gaining new knowledge in the sense of obtaining new correlations between features of observables by the examination of empirical data associated with a given *finite set* of observables. When these characterizations can be tested and validated through the examination of new associated data and their accuracy, and expressive and predictive power does improve by feeding such data, we speak of *successful learning*. Nowadays, *machine learning* refers to any learning process in this sense which can be implemented into and performed by a computing device. Finally, the

Hông Vân Lê
Institute of Mathematics, Czech Academy of Sciences, Zitna 25, Praha, 11567, Czechia e-mail: hvle@math.cas.cz

Hà Quang Minh
RIKEN Center for Advanced Intelligence Project, Tokyo, Japan e-mail: minh.haquang@riken.jp

Frederic Protin
Torus Action SAS, 3 Avenue Didier Daurat, Toulouse, 31400, France & Institut de Mathematiques de Toulouse, Université Toulouse 3, 18 Route de Narbonne,Toulouse, 31400 France e-mail: fredprotin@yahoo.fr

Wilderich Tuschmann
Faculty of Mathematics, Karlsruhe Institute of Technology, Englerstr. 2, Karlsruhe, D-76131, Germany e-mail: tuschmann@kit.edu

theory or science of *Machine Learning* as a whole denotes the study, creation and application of machine learning processes and techniques.

The mathematical theory for learning from data, using probability theory and mathematical statistics, is called statistical learning theory. In statistical learning theory, to learn from data, we need to perform the following steps.

Step 1. Construct a mathematical model of learning, using probability theory and mathematical statistics to model incomplete information of the observed data. More precisely, given a sample space $\mathcal{X}$, we need to construct a hypothesis space $\mathcal{H}$ of possible decisions and model our incomplete information of data $S_n \in \mathcal{X}^n$, our uncertainty of a correct choice of a decision $h \in \mathcal{H}$, given data S_n, by using probability theory and mathematical statistics.

Step 2. Find a learning algorithm A, i.e. a map $A : \cup_{n=1}^{\infty} \mathcal{X}^n \to \mathcal{H}$.

Step 3. Estimate the error of a learning algorithm $A(S_n) \in \mathcal{H}$, to make sure that the learning algorithm is successful.

In Step 1 we consider different types of learning problems, in particular types of training data, i.e., the data available to the learner before making a decision/prediction. A main type of statistical learning problems is supervised learning where training data are labeled, see Definition 3, Example 2, which we shall consider in greater detail in Section 2, using the categorical concept of Markov kernels.

There are two main (non-Bayesian and Bayesian) approaches to Step 1, which yield different probabilistic modelings of training data and their stochastic relations and hence to Step 2, see [63], [28], [29], [60], [46]. In both approaches of probabilistic modeling, the concept of a Markov kernel is of central importance. In our paper we discuss applications of probabilistic morphisms, which are categorical realizations of Markov kernels, to statistical learning theory. The category of probabilistic morphisms was proposed independently by Lawvere in 1962 [36], motivated by discrete stochastic processes, and by Chentsov in 1972 [16] as *the category of statistical decisions* [16, §5, p. 65], which is a natural continuation of Chentsov' work in 1965 [15]. Methods of probabilistic morphisms were developed for generative models of supervised learning by Lê in [37] and applied by us to problems of statistical learning and machine learning in [39]. We also present geometric methods to several problems in statistical learning theory, which are complementary to the categorical approach in our book [39].

Our paper is organized as follows. In Section 2 we first recall the notion of the category of probabilistic morphisms. Basic examples of probabilistic morphisms are measurable mappings, Markov kernels and regular conditional probability measures that are of central importance in mathematical statistics and statistical learning theory. Using the concept of probabilistic morphisms we present the notion of a generative model of supervised learning. Using Lê's characterization of regular conditional probability measures, based on the concept of a graph of a probabilistic morphism (Theorem 1), we give an example of a correct loss function of a generative model of supervised learning (Examples 2, 3). In the last part of this section we outline the proof of Lê's Theorem on the existence of over-parameterized supervised learnable models (Proposition 3), which is essentially based on the concept $\mathbf{Meas}(\mathcal{X}, \mathcal{Y})$ of a correct loss function. The learnability of this model is proved using a version of

Vapnik-Stephanyuk's method of solving stochastic ill-posed problems (Proposition 4). In Section 3 we discuss the problem of generalizing the Gaussian kernel defined on Euclidean spaces to metric spaces, in particular to Riemannian manifolds. After discussing the difficulties of such a generalization to general Riemannian manifolds, we present positive definite kernels defined using the Log-Euclidean metric on the manifold of symmetric positive definite (SPD) matrices and its infinite-dimensional generalization, the Log-Hilbert-Schmidt metric on the set of positive definite unitized Hilbert-Schmidt operators on a Hilbert space. We then discuss a special setting for the Log-Hilbert-Schmidt metric, namely, that of reproducing kernel Hilbert space (RKHS) covariance operators, where all quantities of interest admit closed form expressions via finite-dimensional Gram matrices and thus can be applied in practical applications. In Section 4 we discuss several important topics of manifold learning, which lies at the intersection of geometry and statistical learning. In the final section we summarize the main results of this paper and discuss some related problems.

2 Probabilistic morphisms and generative models of supervised learning

2.1 Notation and conventions

- For a measurable space $\mathcal{X}$ we denote by $\Sigma_{\mathcal{X}}$ the σ-algebra of $\mathcal{X}$, and by $\mathcal{S}(\mathcal{Y}), \mathcal{M}(\mathcal{Y}), \mathcal{P}(\mathcal{Y})$ the spaces of all signed finite measures, (nonnegative) finite measures, probability measures, respectively, on $\mathcal{X}$. If $\mathcal{X}$ is a topological space, we always consider the Borel σ-algebra $\mathcal{B}(\mathcal{X})$, unless otherwise stated.

- For a nonnegative measure μ on a measurable space $\mathcal{X}$ we denote by $\mathcal{L}^p(\mathcal{X}, \mu)$, or just $\mathcal{L}^p(\mu)$, the class of all μ-measurable functions f such that $|f^p|$ is a μ-integrable function. Denote by $L^p(\mu)$ the quotient space of $\mathcal{L}^p(\mu)$ with respect to the equivalence relation $f \overset{\mu}{\sim} g$ if $f = g$ μ-a.e.. Following tradition, we use the expression "a function f in $L^p(\mu)$", where one should say "a function f in $\mathcal{L}^p(\mu)$" when this does not lead to misunderstanding.

- For any measurable space $\mathcal{X}$ we denote by Σ_w the smallest σ-algebra on $\mathcal{P}(\mathcal{X})$ such that for any $A \in \Sigma_{\mathcal{X}}$ the function $I_{1_A} : \mathcal{P}(\mathcal{X}) \to \mathbb{R}, \mu \mapsto \int 1_A d\mu$ is measurable. Here 1_A is the characteristic function of A. In our paper we always consider $\mathcal{P}(\mathcal{X})$ as a measurable space with the σ-algebra Σ_w, unless otherwise stated.

- A Markov kernel $T : \mathcal{X} \times \Sigma_{\mathcal{Y}} \to [0, 1]$ is uniquely defined by the map $\overline{T} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ such that $\overline{T}(x)(A) = T(x, A)$ for all $x \in \mathcal{X}, A \in \Sigma_{\mathcal{Y}}$. We shall also use notations $T(A|x) := T(x, A)$ and $\overline{T}(A|x) := \overline{T}(x)(A)$.

- A *probabilistic morphism* $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ is an arrow assigned to a measurable mapping, denoted by $\overline{T}$ from $\mathcal{X}$ to $\mathcal{P}(\mathcal{Y})$. We say that T is generated by $\overline{T}$. For a measurable mapping $T : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ we denote by $\underline{T} : \mathcal{X} \rightsquigarrow \mathcal{Y}$ the generated probabilistic morphism.

- For probabilistic morphisms $T_{\mathcal{Y}|\mathcal{X}} : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $T_{\mathcal{Z}|\mathcal{Y}} : \mathcal{Y} \rightsquigarrow \mathcal{Z}$ their composition is the probabilistic morphism

$$T_{\mathcal{Z}|\mathcal{X}} := T_{\mathcal{Z}|\mathcal{Y}} \circ T_{\mathcal{Y}|\mathcal{X}} : \mathcal{X} \rightsquigarrow \mathcal{Z}$$

$$(T_{\mathcal{Z}|\mathcal{Y}} \circ T_{\mathcal{Y}|\mathcal{X}})(x, C) := \int_{\mathcal{Y}} T_{\mathcal{Z}|\mathcal{Y}}(y, C) T_{\mathcal{Y}|\mathcal{X}}(dy|x)$$

for $x \in \mathcal{X}$ and $C \in \Sigma_{\mathcal{Z}}$. This corresponds to the composition of the associated Markov kernels, and hence the composition is associative.

- We denote by $\mathbf{Meas}(\mathcal{X}, \mathcal{Y})$ the set of all measurable mappings from a measurable space $\mathcal{X}$ to a measurable space $\mathcal{Y}$, and by $\mathbf{Probm}(\mathcal{X}, \mathcal{Y})$ the set of all probabilistic morphisms from $\mathcal{X}$ to $\mathcal{Y}$. We denote by $\mathcal{Y}^{\mathcal{X}}$ the set of all mappings from $\mathcal{X}$ to $\mathcal{Y}$. For any $\mathcal{X}$ we denote by $\mathrm{Id}_{\mathcal{X}}$ the identity map on $\mathcal{X}$. For a product space $\mathcal{X} \times \mathcal{Y}$ we denote by $\Pi_{\mathcal{X}}$ the canonical projection to the factor $\mathcal{X}$. For $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denote by $\mu_{\mathcal{X}}$ the marginal probability measure $(\Pi_{\mathcal{X}})_* \mu \in \mathcal{P}(\mathcal{X})$.

Important examples of probabilistic morphisms are measurable mappings $\kappa : \mathcal{X} \to \mathcal{Y}$, since the Dirac map $\delta : \mathcal{X} \to \mathcal{P}(\mathcal{X}), x \mapsto \delta_x$, is measurable [27], and hence we regard κ as a probabilistic morphism which is generated by the measurable mapping $\overline{\kappa} := \delta \circ \kappa : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$. We shall denote measurable mappings by straight arrows and probabilistic morphisms by curved arrows. Hence $\mathbf{Meas}(\mathcal{X}, \mathcal{Y})$ can be regarded as a subset of $\mathbf{Probm}(\mathcal{X}, \mathcal{Y})$. Other important examples of probabilistic morphisms in statistical learning are regular conditional probability measures, whose definition we recall now. Let μ be a finite measure on $(\mathcal{X} \times \mathcal{Y}, \Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{Y}})$. A *product regular conditional probability measure* for μ with respect to the projection $\Pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ is a Markov kernel $\mu_{\mathcal{Y}|\mathcal{X}} : \mathcal{X} \times \Sigma_{\mathcal{Y}} \to [0, 1]$ such that

$$\mu(A \times B) = \int_A \mu_{\mathcal{Y}|\mathcal{X}}(x, B)\, d(\Pi_{\mathcal{X}})_* \mu(x) \tag{1}$$

for any $A \in \Sigma_{\mathcal{X}}$ and $B \in \Sigma_{\mathcal{Y}}$. We shall simply call $\mu_{\mathcal{Y}|\mathcal{X}}$ a *regular conditional probability measure* for μ and identify $\mu_{\mathcal{Y}|\mathcal{X}}$ with the generating measurable map $\overline{\mu_{\mathcal{Y}|\mathcal{X}}} : \mathcal{X} \to \mathcal{P}(\mathcal{Y}), x \mapsto \mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$.

- We denote by **Meas** the category of measurable spaces, whose objects are measurable spaces and morphisms are measurable mappings.

2.2 The category Probm and characterizations of regular conditional probability measures

2.2.1 Category of probabilistic morphisms

We denote by **Probm** the category whose objects are measurable spaces and morphisms are of Markov kernels $T : \mathcal{X} \times \Sigma_{\mathcal{Y}} \to [0, 1]$ as morphisms from $\mathcal{X}$ to $\mathcal{Y}$. The identity Markov kernel $\underline{\delta \circ \mathrm{Id}_{\mathcal{X}}} : \mathcal{X} \times \Sigma_{\mathcal{X}} \to [0, 1]$is defined as follows:

$(x, A) \mapsto \delta_x(A)$. For $\kappa \in \mathbf{Meas}(\mathcal{X}, \mathcal{Y})$ we also use the shorthand notation

$$\overline{\kappa} := \delta \circ \kappa. \tag{2}$$

Identifying $\mathbf{Meas}(\mathcal{X}, \mathcal{Y})$ with a subset in $\mathbf{Probm}(\mathcal{X}, \mathcal{Y}) = \mathbf{Meas}(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$ via the composition with the Dirac map δ: $\mathbf{Meas}(\mathcal{X}, \mathcal{Y}) \ni \kappa \mapsto \delta \circ \kappa \in \mathbf{Probm}(\mathcal{X}, \mathcal{Y})$, we regard **Meas** as a subcategory of **Probm**.

The category **Probm** admits a faithful functor S to the category **Ban** of Banach spaces whose morphisms are bounded linear mappings of operator norm less than or equal to one as follows. For any $T \in \mathbf{Probm}(\mathcal{X}, \mathcal{Y})$, $\mu \in \mathcal{S}(\mathcal{X})$ and $B \in \Sigma_{\mathcal{Y}}$ we set

$$S(T)_*\mu(B) = \int_{\mathcal{X}} \overline{T}(B|x)\, d\mu(x). \tag{3}$$

Proposition 1 *[16, Lemmas 5.9, 5.10] Let S assign to each measurable space $\mathcal{X}$ the Banach space $\mathcal{S}(\mathcal{X})$ of finite signed measures on $\mathcal{X}$ endowed with the total variation norm $\|\cdot\|_{TV}$ and to every Markov kernel $T_{\mathcal{Y}|\mathcal{X}}$ the Markov homomorphism $(T_{\mathcal{Y}|\mathcal{X}})_*$. Then S is a faithful functor from the category* **Probm** *to the category* **Ban**.

Remark 1 (1) It is known that the restriction $M_*(T)$ of $S_*(T)$ to $\mathcal{M}(\mathcal{X})$ and the restriction $P_*(T)$ of $S_*(T)$ to $\mathcal{P}(\mathcal{X})$ maps $\mathcal{M}(\mathcal{X})$ to $\mathcal{M}(\mathcal{Y})$ and $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}(\mathcal{Y})$, respectively [16, Lemma 5.9, p. 72]. We shall use the shorthand notation T_* for $S_*(T)$, $M_*(T)$ and $P_*(T)$, if no misunderstanding occurs.

(2) Let $T \in \mathbf{Probm}(\mathcal{X}, \mathcal{Y})$. Then $T_* : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$ is a measurable mapping [27, Theorem 1].

(3) Let $T \in \mathbf{Probm}(\mathcal{X}, \mathcal{Y})$ and $\nu \ll \mu \in \mathcal{S}(\mathcal{X})$. Then $T_*(\nu) \ll T_*(\mu)$ by a result due to Ay-Jost-Lê-Schwachhöfer [1, Remark 5.4, p. 255], which generalizes Morse-Sacksteder's result [47, Proposition 5.1], see also [32, Theorem 2 (2)] for an alternative proof.

Using the functor S, we shall characterize probabilistic regular conditional probability measures. Given two measurable mappings $\overline{T}_i : \mathcal{X} \to \mathcal{P}(\mathcal{Y}_i)$, where i = 1,2, let us consider the map

$$\overline{T_1 \cdot T_2} : \mathcal{X} \to \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_2),\ x \mapsto \mathfrak{m}(\overline{T}_1(x), \overline{T}_2(x)),$$

where the multiplication $\mathfrak{m}$ is defined as follows:

$$\mathfrak{m}(\mu_1, \mu_2) = \mu_1 \otimes \mu_2.$$

It is easy to see that $\mathfrak{m}$ is a measurable mapping. The map $\overline{T_1 \cdot T_2} : \mathcal{X} \to \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_2)$ is a measurable mapping since it is the composition of two measurable mappings $(\overline{T}_1, \overline{T}_2) : \mathcal{X} \to \mathcal{P}(\mathcal{Y}_1) \times \mathcal{P}(\mathcal{Y}_2)$ and $\mathfrak{m} : \mathcal{P}(\mathcal{Y}_1) \times \mathcal{P}(\mathcal{Y}_2) \to \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_2)$.

Definition 1 (1) Given two probabilistic morphisms $T_i : \mathcal{X} \rightsquigarrow \mathcal{Y}_i$ for i = 1,2, *the join of T_1 and T_2* is the probabilistic morphism $T_1 \cdot T_2 : \mathcal{X} \rightsquigarrow \mathcal{Y}_1 \times \mathcal{Y}_2$ whose generating mapping is $\overline{T_1 \cdot T_2} : \mathcal{X} \to \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_2)$ given by

$$\overline{T_1 \cdot T_2}(x) := \mathfrak{m}(\overline{T}_1(x), \overline{T}_2(x)).$$

(2) Given a probabilistic morphism $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ we denote the join of $\mathrm{Id}_{\mathcal{X}}$ with T by $\Gamma_T : \mathcal{X} \rightsquigarrow \mathcal{X} \times \mathcal{Y}$ and call it the *graph of* T.

Remark 2 (1) If $\kappa : \mathcal{X} \to \mathcal{Y}$ is a measurable mapping, regarded as a probabilistic morphism, then the graph $\Gamma_\kappa : \mathcal{X} \rightsquigarrow \mathcal{X} \times \mathcal{Y}$ is a measurable mapping defined by $x \mapsto (x, \kappa(x))$ for $x \in \mathcal{X}$, since $\overline{\mathrm{Id}}_{\mathcal{X}} = \delta \circ \mathrm{Id}_{\mathcal{X}}$, and therefore $\mathfrak{m}(\overline{\mathrm{Id}}_{\mathcal{X}}(x), \overline{\kappa}(x)) = \delta_x \otimes \delta_{\kappa(x)} = \delta \circ \Gamma_\kappa(x)$.

(2) For historical notes on the concept of the join of two probabilistic morphisms, we refer to [37, Remark 2.17 (1)].

Definition 2 (Almost surely equality) [37], [26] Let $\mu \in \mathcal{P}(\mathcal{X})$. Two measurable mappings $T, T' : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ will be called *equal μ-a.e.* (with the shorthand notation $T = T'$ μ-a.e.), if for any $B \in \Sigma_{\mathcal{Y}}$

$$\mu\{x \in \mathcal{X} : T(x)(B) \neq T'(x)(B)\} = 0.$$

2.2.2 Characterizations of regular conditional probability measures

Theorem 1 (Characterization of regular conditional probability measures)

(i) A measurable mapping $\overline{T} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ is a regular conditional probability measure for $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with respect to the projection $\Pi_{\mathcal{X}}$ if and only if

$$(\Gamma_T)_* \mu_{\mathcal{X}} = \mu. \tag{4}$$

(ii) If $\overline{T}, \overline{T}' : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ are regular conditional probability measures for $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with respect to the projection $\Pi_{\mathcal{X}}$, then $\overline{T} = \overline{T}'$ $\mu_{\mathcal{X}}$-a.e. Conversely, if T is a regular conditional probability measure for $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with respect to the projection $\Pi_{\mathcal{X}}$ and $\overline{T} = \overline{T}'$ $\mu_{\mathcal{X}}$-a.e., then $\overline{T}' : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ is a regular conditional probability measure for $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with respect to the projection $\Pi_{\mathcal{X}}$.

Theorem 1 follows easily from the definition of a regular conditional probability measure (Equation 1), see also [37, Theorem 2.22] for a slight generalization.

Example 1 (Measurement error) Assume that $x \in (\mathcal{X}, \mu_{\mathcal{X}})$, $\mu_{\mathcal{X}} \in \mathcal{P}(\mathcal{X})$, and $y \in \mathbb{R}^n$ are related by the following equation

$$y = f(x) + \varepsilon \tag{5}$$

where $f \in \mathbb{R}^{\mathcal{X}}$ and $\varepsilon \in (\mathbb{R}^n, \mu_\varepsilon)$. One regards ε being generated by a measurable mapping $g_\varepsilon : (\Omega, \mu_\Omega) \to \mathbb{R}^n$, where (Ω, μ_Ω) is a latent source space, $\mu_\Omega \in \mathbf{p}(\Omega)$ and $\mu_\varepsilon = (g_\varepsilon)_* \mu_\Omega$. One interprets the assumption that the noise ε is independent of $x \in \mathcal{X}$, and the measurement error equation (5) as the equation for the conditional probability $\mu_{\mathbb{R}^n|\mathcal{X}}(x)$ of y given x, which satisfies

$$\overline{\mu_{\mathbb{R}^n|\mathcal{X}}}(x) = \delta_{f(x)} * \mu_\varepsilon \in \mathcal{P}(\mathbb{R}^n). \tag{6}$$

Here $\mu_1 * \mu_2 = A_*(\mu_1 \otimes \mu_2)$ is the convolution for $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^n)$, noting that $A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n, (x, y) \mapsto (x+y)$ is a measurable map. Since for any $\mu_0 \in \mathcal{P}(\mathbb{R}^n)$ one easily shows that the embedding $I_{\mu_0} : \mathcal{P}(\mathbb{R}^n) \to \mathcal{P}(\mathbb{R}^n \otimes \mathbb{R}^n), \mu \mapsto \mu \otimes \mu_0$, is a measurable mapping, we conclude that the map $\overline{\mu_{\mathbb{R}^n|\mathcal{X}}} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ defined in (6) is a measurable map, if $f \in \mathbf{Meas}(\mathcal{X}, \mathbb{R}^n)$. In this case (x, y) is distributed jointly by the probability measure $\mu \in \mathcal{P}(\mathcal{X} \times \mathbb{R}^n)$ that is defined uniquely by its marginal measure $\mu_\mathcal{X} = (\Pi_\mathcal{X})_*\mu$ and its regular conditional probability measure that is generated by the map $\overline{\mu_{\mathbb{R}^n|\mathcal{X}}} : \mathcal{X} \to \mathcal{P}(\mathbb{R}^n)$ defined in (6), i.e. $\mu = (\Gamma_{\mu_{\mathbb{R}^n|\mathcal{X}}})_*\mu_\mathcal{X}$. If the mean $m(\mu_\varepsilon) := \int_{\mathbb{R}^n} y \, d\mu_\varepsilon(y)$ of μ_ε is zero, then one has

$$f(x) = r_\mu(x) := \int_{\mathbb{R}^n} y \, d\mu_{\mathbb{R}^n|\mathcal{X}}(y|x). \tag{7}$$

Formula (7) explains the importance and popularity of the problem of estimating the regression function r_μ, where $\mu \in \mathcal{P}(\mathcal{X} \times \mathbb{R}^n)$ is (unknown) probability measure, in supervised learning, see e.g. [62, §1.4, p. 26], [17].

2.3 Generative models of supervised learning

2.3.1 Generative models of supervised learning and correct loss functions

In supervised learning, given a data set of labeled items $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, the aim of a learner is to find a best approximation f_{S_n} of the stochastic relation between $x \in \mathcal{X}$ and its labels $y \in \mathcal{Y}$, formalized as a version of the conditional probability measure $\mu_{\mathcal{Y}|\mathcal{X}} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ expressing the probability of the label $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$. In this subsection we consider the setting of non-Bayesian supervised learning, where (x_i, y_i) are assumed to be i.i.d., i.e., for any n, $S_n \in ((\mathcal{X} \times \mathcal{Y})^n, \mu^n)$ for some unknown $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Furthermore, we assume that μ admits a regular conditional probability measure $\mu_{\mathcal{Y}|\mathcal{X}}$ with respect to the projection $\Pi_\mathcal{X} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$. It is known that if $\mathcal{Y}$ is a Polish space, and $\mathcal{X}$ is a measurable space, then for any $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ there exists a regular conditional measure $\mu_{\mathcal{Y}|\mathcal{X}}$ with respect to $\Pi_\mathcal{X} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$, see e.g. [10, Corollary 10.4.15, p. 366, vol. 2], or [38, Theorem 3.1] for a more general result.

The concept of a best approximation requires a specification of a hypothesis $\mathcal{H}$ of possible predictors as well a the notion of a *correct* loss function that measures the deviation of a possible predictor from a regular conditional probability measure $\mu_{\mathcal{Y}|\mathcal{X}}$. The loss function concept in statistical analysis was introduced by Wald in his work on statistical decision theory [67] and has been intensively discussed in statistical learning theory [62], [58]. Now we perform Step 1 of modeling a supervised learning problem under the above assumptions in the following definition.

Definition 3 (cf. [37, Definitions 3.1, 3.3]) *A generative model of supervised learning* is given by a quintuple $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, R, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$, where $\mathcal{X}$ and $\mathcal{Y}$ are measurable spaces, $\mathcal{H}$ is a family of measurable mappings $h : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \subset \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ contains all possible probability measures governing the distributions of labeled pairs (x, y), and $R : \mathcal{H} \times \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R} \cup \{+\infty\}$ is a risk/loss function such that $\inf_{h\in\mathcal{H}} R_\mu(h) \neq \pm\infty$ for any $\mu \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. If $R(h, \mu) = \mathbb{E}_\mu(L(h))$ where $L : \mathcal{X}\times\mathcal{Y}\times\mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ is an instantaneous loss measurable function then we shall write the model as $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$. Note that $\mathbb{E}_\mu f$ is well-defined, iff $h \in L^1(\mu)$, otherwise we let $\mathbb{E}_\mu(h) := +\infty$. If $\mathcal{H} \subset \mathbf{Meas}\,(\mathcal{X}, \mathcal{Y}) \subset \mathbf{Meas}\,(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$ we shall say that $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, R, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$ is a *discriminative* model of supervised learning. If R can be extended to a loss function, also denoted by $R : \mathcal{H}\times(\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\cup\mathcal{P}_{emp}(\mathcal{X}\times\mathcal{Y})) \to \mathbb{R} \cup \{+\infty\}$, then R is called *empirically definable*.

A loss function $R : \mathcal{H}\times\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}\cup\{+\infty\}$ will be called $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$*-correct*, if there exists a set $\widetilde{\mathcal{H}} \subset \mathbf{Meas}(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$ such that the following three conditions hold:

(1) $\mathcal{H} \subset \widetilde{\mathcal{H}}$.

(2) For any $\mu \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ there exists $h \in \widetilde{\mathcal{H}}$ such that h is a regular conditional measure for μ relative to the projection $\Pi_{\mathcal{X}}$.

(3) R is the restriction of a loss function $\widetilde{R} : \widetilde{\mathcal{H}} \times \left(\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \cup \mathcal{P}_{emp}(\mathcal{X}\times\mathcal{Y})\right) \to \mathbb{R} \cup \{+\infty\}$ such that for any $\mu \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$

$$\arg\min_{h\in\widetilde{\mathcal{H}}} \widetilde{R}(h, \mu) = \{h \in \widetilde{\mathcal{H}} |\ h \text{ is a regular conditional probability measure for } \mu\}.$$

A loss function $R : \mathcal{H} \times \left(\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\right) \to \mathbb{R} \cup \{+\infty\}$ will be called *correct*, if R is the restriction of a $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$-correct loss function $\widetilde{R} : \mathcal{H} \times \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$.

Example 2 Let us consider the following generative model of classification problems $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$, a particular case of supervised learning where $\mathcal{Y} = \{0, 1\}$, and L shall be defined below in (8). Since $\mathcal{Y} = \{0, 1\}$, one can identify $\mathcal{P}(\mathcal{Y})$ with the interval $[0, 1]$, namely, we associate a probability measure $\mathbf{p} \in \mathcal{P}(\mathcal{Y})$ with the value $\mathbf{p}(\{1\}) \in [0, 1]$. Since $\mathcal{Y}$ is finite, the strong topology on $\mathcal{P}(\mathcal{Y})$ coincides with the weak*-topology τ_w, and therefore $\mathcal{P}(\mathcal{Y})$ is a measurable space isomorphic to $([0, 1], \mathcal{B}([0, 1])$. Thus

$$\mathbf{Meas}\,(\mathcal{X}, \mathcal{P}(\mathcal{Y})) = \mathbf{Meas}\,(\mathcal{X}, [0, 1]).$$

Recall that $\mu_{\mathcal{Y}|\mathcal{X}} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})(\cdot|x), x \in \mathcal{X}$, is a regular conditional probability measure for μ and

$$r_\mu(x) = \int_{\mathcal{Y}} y\, d\mu_{\mathcal{Y}|\mathcal{X}}(y|x) \in [0, 1].$$

Since $r_\mu(x) = \mu_{\mathcal{Y}|\mathcal{X}}(\{1\}|x)$, $\mu_{\mathcal{Y}|\mathcal{X}}$ and r_μ define each other. Now assume that

$$\mathcal{H} \subset \mathbf{Meas}\,(\mathcal{X}, [0, 1]) = \mathbf{Probm}\,(\mathcal{X}, \mathcal{Y}),$$

where, recall that, in the last equality, for $h \in \mathbf{Meas}\,(\mathcal{X}, [0, 1])$, we identify $h(x)$ with the measure $\bar{h}(x) \in \mathcal{P}(\mathcal{Y})$ defined by $\bar{h}(x)(\{1\}) = h(x)$ and $\bar{h}(x)(\{0\}) = 1 - h(x)$.

Clearly $\mathbf{Meas}(\mathcal{X}, [0,1]) \subset \mathcal{L}^2(\mathcal{X}, (\Pi_{\mathcal{X}})_*\mu)$ for any $\mu \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$. Next it is known that r_μ is the minimizer of the expected loss function

$$R^L_\mu : \mathcal{L}^2(\mathcal{X}, (\Pi_{\mathcal{X}})_*\mu) \to \mathbb{R}_{\geq 0},\ h \mapsto \mathbb{E}_\mu L(\cdot,\cdot,h)$$

where L is the instantaneous quadratic function:

$$L(x,y,h) = (y - h(x))^2, \tag{8}$$

since r_μ is a regular version of the conditional expectation $\mathbb{E}_\mu(\mathrm{Id}_{\mathbb{R}}|\Pi_{\mathcal{X}})$, see e.g. [17, Proposition 1] for a slight generalization of this fact. Hence r_μ is a minimizer of $R^L_\mu : \mathbf{Meas}(\mathcal{X}, [0,1]) \to [0,1]$. Moreover, any minimizer R^L_μ to $\mathbf{Meas}(\mathcal{X}, [0,1])$ coincides with r_μ $\mu_{\mathcal{X}}$-a.e. Thus the risk function R^L_μ in the generative model $(\mathcal{X}, \mathcal{Y} = \{0,1\}, \mathbf{Meas}(\mathcal{X}, \mathcal{P}(\mathcal{Y})), L, \mathcal{P}(\mathcal{X}\times\mathcal{Y}))$ is correct. Hence for any generative model $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$, where, abusing notation, the restriction of L to $\mathcal{H}$ is also denoted by L, the risk function R^L_μ is also correct. In particular the instantaneous 0-1 loss function $L^{0,1} : \mathcal{X}\times\mathcal{Y}\times\mathbf{Meas}(\mathcal{X},\mathcal{Y}), (x,y,h) \mapsto d^{0-1}(y,h(x))$, where d^{0-1} is the 0-1 distance, generates a correct loss function.

Remark 3 (1) The class of generative models of supervised learning of the type $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, R, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$ encompasses al models for density estimation on $\mathcal{Y}$ by letting $\mathcal{X}$ consist of a single point.

(2) In classical statistical learning theory one usually considers loss functions $R^L : \mathcal{H}\times\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}_{\geq 0}\cup\{+\infty\}$ which are generated by instantaneous loss functions $L : \mathcal{X}\times\mathcal{Y}\times\mathcal{H}$, as well perturbations of R^L, see, e.g., [62], [58]. Note that any loss function which is generated by an instantaneous loss function is empirically definable.

2.3.2 RKHSs and correct loss functions

In what follows we shall provide a natural example of empirically definable correct loss functions which may not be generated by instantaneous loss functions and their perturbations, using our characterization of regular conditional probability measures (Theorem 1) and kernel mean embeddings.

Let $K : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ be a measurable positive definite symmetric (PDS) kernel on a measurable space $\mathcal{Y}$. For $y \in \mathcal{Y}$ let K_y be the function on $\mathcal{Y}$ defined by

$$K_y(y') = K(y,y') \text{ for } y' \in \mathcal{Y}.$$

We denote by $\mathcal{H}(K)$ the associated RKHS [2], see also [11] i.e.

$$\mathcal{H}(K) = \overline{\mathrm{span}}\{K_y, y \in \mathcal{Y}\},$$

where the closure is taken with respect to the $\mathcal{H}(K)$-norm defined by

$$\langle K_y, K_{y'}\rangle_{\mathcal{H}(K)} = K(y,y').$$

Then for any $f \in \mathcal{H}(K)$ we have

$$f(y) = \langle f, K_y \rangle_{\mathcal{H}(K)}. \tag{9}$$

By Bochner's theorem, $\int_{\mathcal{Y}} \sqrt{K(y,y)}\, d\mu(y) < \infty$ for $\mu \in \mathcal{P}(\mathcal{Y})$ if and only if the *kernel mean embedding* $\mathfrak{M}_K(\mu)$ *of* μ via the Bochner integral is well-defined [11], where

$$\mathfrak{M}_K(\mu) = \int_{\mathcal{Y}} K_y d\mu(y) \in \mathcal{H}(K). \tag{10}$$

If $\mathfrak{M}_K(\mu)$ is well-defined for all $\mu \in \mathcal{P}(\mathcal{X})$, the kernel mean embedding $\mathfrak{M}_K : \mathcal{P}(\mathcal{X}) \to \mathcal{H}(K)$ extends to a linear map, also denoted by $\mathfrak{M}_K$ from $\mathcal{S}(\mathcal{Y})$ to $\mathcal{H}(K)$.

Example 3 (cf. [37, Example 3.4 (2)]) Assume K be a measurable positive definite symmetric (PDS) kernel on a measurable space $\mathcal{X} \times \mathcal{Y}$ such that $\mathfrak{M}_K : \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}(K)$ is an embedding. Then the loss function

$$R^K : \mathbf{Probm}(\mathcal{X}, \mathcal{Y}) \times \mathcal{P}(\mathcal{X}, \mathcal{Y}) \to \mathbb{R}_{\geq 0}, (h, \mu) \mapsto \|\mathfrak{M}_K(\Gamma_h)_* \mu_{\mathcal{X}} - \mathfrak{M}_k(\mu)\|_{\mathcal{H}(K)} \tag{11}$$

is a correct loss function by Theorem 1.

It is known that if $\mathcal{X} \times \mathcal{Y}$ is a Polish subspace in $\mathbb{R}^n$ then there are many PDS kernels on $\mathcal{X} \times \mathcal{Y}$ such that $\mathfrak{M}_K : \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}(K)$ is an embedding [56, Theorem 3.2], e.g. K is the restriction of the Gaussian kernels $K_\sigma : \mathbb{R}^{n+m} \times \mathbb{R}^{n+m} \to \mathbb{R}_{\geq 0}$ for $\sigma > 0$: $K_\sigma(z, z') = \exp(-\sigma \|z - z'\|^2)$. Here $\|\cdot\|$ denotes the norm in $\mathbb{R}^{n+m}$ generated by the Euclidean metric.

Remark 4 (1) If $\mathcal{X}$ consists of a single point, the loss function R^K was used by Lopez-Paz, Muandet, Schölkopf, and Tolstikhin for probability measure estimation problem. They proved the following beautiful result on estimating probab.

Proposition 2 *[41, Theorem 1] Let $K : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a measurable kernel such that $\mathfrak{M}_K : \mathcal{S}(\mathcal{Y}) \to \mathcal{H}(K)$ is well-defined. Assume that $\|f\|_\infty \leq 1$ for all $f \in \mathcal{H}(K)$ with $\|f\|_{\mathcal{H}(K)} \leq 1$. Then for any $\varepsilon \in (0,1)$ we have*

$$\mu\Big\{S_n \in \mathcal{Y}^n : \|\mathfrak{M}_K(\mu_{S_n}) - \mathfrak{M}_K(\mu)\|_{\mathcal{H}(K)} \leq 2\sqrt{\frac{\int_{\mathcal{Y}} K(y,y) d\mu(y)}{n}} + \sqrt{\frac{2\log\frac{1}{\varepsilon}}{n}}\Big\} \geq 1 - \varepsilon. \tag{12}$$

(2) For more examples of correct loss functions, see [37, Example 3.4, Theorem 4.6].

2.4 Leanability of overparameterized supervised learning models

In this subsection we shall specify the notion of successful learning of a procedure of statistical learning from data as the concept of learnability of a generative model supervised (Definition 4) learning and present examples of learning algorithms (Definition 5). Then we demonstrate an application of categorical and geometrical methods in proving the learnability of overparameterized supervised learning models.

2.4.1 Learnability of a supervised learning model and regularized ERM algorithms

Given a supervised learning model $(\mathcal{X} \times \mathcal{Y}, \mathcal{H}, R, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$ and $\mu \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, we set

$$R_{\mu,\mathcal{H}} := \inf_{h\in\mathcal{H}} R_\mu(h). \tag{13}$$

For $h \in \mathcal{H}$, recall that its *estimation error* is defined as follows (cf. [17, Chapter I, §3]):

$$\mathcal{E}_{\mathcal{H},R,\mu}(h) := R_\mu(h) - R_{\mu,\mathcal{H}}. \tag{14}$$

If $R = R^L$ we shall write $\mathcal{E}_{\mathcal{H},L,\mu}$ instead of $\mathcal{E}_{\mathcal{H},R^L,\mu}$. Let μ_* denote the inner measure defined by $\mu \in \mathcal{M}(\mathcal{X})$.

Definition 4 cf. [37, Definition 3.8] A supervised learning model $(\mathcal{X}\times\mathcal{Y}, \mathcal{H}, R, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$ will be said to have a *generalization ability* or will be called *learnable*, if there exists a uniformly consistent learning algorithm

$$A : \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H},$$

i.e. for any $(\varepsilon, \delta) \in (0,1)^2$ there exists a number $m_A(\varepsilon, \delta)$ such that for any $m \geq m_A(\varepsilon, \delta)$ and any $\mu \in \mathcal{P}_{\mathcal{Z}}$ we have

$$(\mu^n)_*\{S \in (\mathcal{X} \times \mathcal{Y})^n, \mathcal{E}_{\mathcal{H},R,\mu}(A(S)) \leq \varepsilon\} \geq 1 - \delta. \tag{15}$$

In this case A will be called *uniformly consistent*.

Remark 5 (1) In [37, Definition 3.8] Lê considered a general statistical learning model $(\mathcal{Z}, \mathcal{H}, R, \mathcal{P}_{\mathcal{Z}})$ where $\mathcal{Z}$ is a measurable (sample) space, $\mathcal{H}$ is a class of mappings, $\mathcal{P}_{\mathcal{Z}} \subset \mathcal{P}(\mathcal{Z})$ is the set of all possible probability measures governing distribution of observable data $z_i \in \mathcal{Z}$ and $R : \mathcal{H}\times\mathcal{P}_{\mathcal{Z}} \to \mathbb{R}\cup+\infty$ is a loss function.

(2)The current definition of uniform consistency, also called generalizability, of a learning algorithm A in the standard literature is almost identical to our definition, but we relax the convergence in probability in the classical requirement to the convergence in outer probability for the sequence of functions

$\mathcal{E}_{\mathcal{H},R,\mu} \circ A : (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}$, $n \in \mathbb{N}$, whose μ^n-measurability needed in the definition of the convergence in probability represents a quite strong assumption.

From now on we assume that the loss function $R : \mathcal{H} \times \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R} \cup \{+\infty\}$ is empirically definable. For a data $\{S_n \in (\mathcal{X} \times \mathcal{Y})^n\}$ we define the empirical risk

$$\widehat{R}_{S_n} : \mathcal{H} \to \mathbb{R},\ h \mapsto R_{\mu_{S_n}}(h).$$

It is well-known in analysis that to solve an equation $A(f) = F$ it is often useful to look at its perturbed equation $A_\varepsilon(f) = F_\varepsilon$. A successful perturbation method has been used in approximation theory under the name of Tikhonov's regularization method, which has been developed further by Vapnik-Stephaniuk as a method of solving stochastic ill-posed problem [66], see also [62, Chapter 7], [63, Chapter 7]. The Vapnik-Stephaniuk method of solving stochastic ill-posed problem fits particularly well to our equation for regular conditional probability measures (Theorem 1) which leads to the correct loss function in Example 3.

Definition 5 Assume that $(\mathcal{X}\times\mathcal{Y}, \mathcal{H}, R, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$ is a generative model ofs supervised learning. Let $W : \mathcal{H} \to \mathbb{R}_{\geq 0}$ be a function. Given a sequence of numbers $\Gamma := (\gamma_1 > \gamma_2 > \ldots > 0 : \lim_{n\to\infty} \gamma_n = 0)$, we shall call a sequence of the following regularized loss functions R_{γ_n} for R

$$R_{\gamma_n} : \mathcal{H} \times \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R} \cup \{+\infty\}, (h, \mu) \mapsto R(h, \mu) + \gamma_n W(h) \tag{16}$$

a sequence of (W, Γ)-*regularized loss functions for* R. If $\gamma_i = 0$ for all i we write $\Gamma = 0$. Let $C = (c_1 \geq \ldots \geq c_n \geq \ldots : c_i \geq 0)$. A learning algorithm

$$A : \bigcup_{n\in\mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$$

will be called a (C, Γ)- *regularized empirical risk minimizing* algorithm, abbreviated as (C, Γ)-regularized ERM algorithm, if there exists a function $W : \mathcal{H} \to \mathbb{R}_{\geq 0}$ such that for any $n \in \mathbb{N}$ and any $S \in (\mathcal{X} \times \mathcal{Y})^n$, we have

$$R_{\gamma_n}(A(S), \mu_S) - \inf_{h\in\mathcal{H}} R_{\gamma_n}(h, \mu_S) \leq c_n.$$

If $c_i = 0$ for all i we write $C = 0$,

Remark 6 (1) (C, Γ)-regularized ERM algorithms are the most frequently used algorithms in statistical learning theory [62]. If $\Gamma = 0$, we shall write C-ERM algorithm instead of $(C, 0)$-regularized ERM algorithm. In this case a specification of C is often suppressed [62, p. 80]. In general cases, we need to specify C to ensure that a (C, Γ)-regularized ERM algorithm is uniformly consistent.

(2) The uniform consistency of C-ERM algorithms, which is also called ERM-algorithms by suppressing C, in the case that a loss function $R = R^L$ is generated by an instantaneous loss function $L : \mathcal{X}\times\mathcal{Y}\times\mathcal{P}(\mathcal{X}\times\mathcal{Y}) \to \mathbb{R}_{\geq 0}$ and $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} = \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ has been discussed in depth by Vapnik [62], [63]. Vapnik emphasized the relation

between the consistency of ERM algorithms and the convergence in probability of associated empirical processes [62, §3.3, p. 86], which could be regarded as a uniform law of large numbers in a functional space [62, §3.3.1, p. 87].

Berner-Groh-Kutyniok-Peterson note that Vapnik's theory for ERM algorithms with respect to distribution free models, i.e., where $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} = \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, is not sufficient to explain the learnability of neural networks, and they asked for a new statistical learning theory, which must take into account also geometry of underlying statistical model $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ [5, §1.1.3].

Our examples of overparametrized supervised learning models (Proposition 3) in the next subsection satisfy these requirements, since our theorem takes into account the interplay between the geometry of a statistical model $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ and the geometry of a hypothesis space $\mathcal{H} \subset C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2})$.

2.4.2 Examples of overparameterized supervised learning models

• Let $\mathcal{X}$ be a compact subset in $\mathbb{R}^n\times\{0\} \subset \mathbb{R}^{n+m}$ and $\mathcal{Y}$ a compact subset in $\{0\}\times\mathbb{R}^m$.
• Let $K_2 : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ be the restriction of a Gaussian kernel $K = K_\sigma : \mathbb{R}^{m+n} \times \mathbb{R}^{m+n} \to \mathbb{R}_{\geq 0}$ to $\mathcal{Y}\times\mathcal{Y}$. Let $K_1 : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ be the restriction of the kernel K to $(\mathcal{X}\times\mathcal{Y})$.
• It is known that the kernel mean embeddings $\mathfrak{M}_{K_1} : \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \to \mathcal{H}(K_1)$ and $\mathfrak{M}_{K_2} : \mathcal{P}(\mathcal{Y}) \to \mathcal{H}(K_2)$ extend linearly to embeddings, denoted by $\mathfrak{M}_{K_1} : \mathcal{S}(\mathcal{X}\times\mathcal{Y}) \to \mathcal{H}(K_1)$ and , $\mathfrak{M}_{K_2} : \mathcal{S}(\mathcal{Y}) \to \mathcal{H}(K_2)$ respectively. We denote by $\|\cdot\|_{\tilde{K}_1}$ and $\|\cdot\|_{\tilde{K}_2}$ the pull-back of the norm $\|\cdot\|_{\mathcal{H}(K_1)}$ on $\mathcal{S}(\mathcal{X}\times\mathcal{Y})$ and on $\mathcal{S}(\mathcal{Y})$ respectively. It is known that the restriction of $\|\cdot\|_{\tilde{K}_1}$ to $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ and the restriction $\|\cdot\|_{\tilde{K}_2}$ to $\mathcal{P}(\mathcal{Y})$ generate metrics that induce the weak*-topology on the spaces, [56, Example 1, Theorem 3.2]. We denote by $\mathcal{P}(\mathcal{X}\times\mathcal{Y})_{\tilde{K}_1}$ and $\mathcal{P}(\mathcal{Y})_{\tilde{K}_2}$ the corresponding metric spaces.
• For metric spaces (F_1, d_1), (F_2, d_2), we denote by $C_{Lip}(F_1, F_2)$ the space of all Lipschitz continuous mappings from F_1 to F_2. For $h \in C_{Lip}(F_1, F_2)$ we denote by $L(h)$ the *Lipschitz constant* of h namely the nonnegative number

$$L(h) := \sup_{x\neq y} \frac{d_2(h(x), h(y))}{d_1(x, y)}.$$

• Denote by $\mathcal{P}_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}, vol_{\mathcal{X}})$ the set of all probability measures $\mu \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ such that:
(i) sppt $\mu_{\mathcal{X}} = \mathcal{X}$, where $\mu_{\mathcal{X}} = (\Pi_{\mathcal{X}})_*\mu$;
(ii) there exists a regular conditional measure $\mu_{\mathcal{Y}|\mathcal{X}} \in C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2})$ for μ with respect to the projection $\Pi_{\mathcal{X}} : \mathcal{X}\times\mathcal{Y} \to \mathcal{X}$.

We define a loss function (cf. Example 3)

$$R^{K_1} : C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}) \times \mathcal{P}_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}, vol_{\mathcal{X}}) \to \mathbb{R}_{\geq 0},$$
$$(h, \mu) \mapsto \|(\Gamma_h)_*\mu_{\mathcal{X}} - \mu\|_{\tilde{K}_1}. \tag{17}$$

Proposition 3 *([37, Corollary 6.3]) Assume that $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \subset \mathcal{P}_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}, vol_{\mathcal{X}})$ satisfies the following condition (L).*

(L) The function $(\mathcal{P}_{\mathcal{X}\times\mathcal{Y}})_{\tilde{K}_1} \to \mathbb{R}, \mu \mapsto L(\mu_{\mathcal{Y}|\mathcal{X}})$, where $\mu_{\mathcal{Y}|\mathcal{X}} \in C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2})$, takes values in a finite interval $[a, b] \subset \mathbb{R}$.

(1) Then there exists a (C, Γ)-regularized ERM algorithm A for the loss function R^{K_1} which is uniformly consistent for the supervised learning model $(\mathcal{X}, \mathcal{Y}, C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}), \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$.

(2) Assume further that $\mathcal{H}$ is a subspace of the space $C_{Lip}(\mathcal{X}, \mathcal{Y})$. Then the supervised learning model $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, R^{K_1}, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$ has a generalization ability.

Remark 7 Here is a concrete example of $\mathcal{X}, \mathcal{Y}$ and $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \subset \mathcal{P}_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}, vol_{\mathcal{X}})$ satisfying the condition (L) in Proposition 3. Let $\mathcal{X} = [0, 1]^n \subset \mathbb{R}^n \times \{0\} \subset \mathbb{R}^{n+m}$ and $\mathcal{Y} = [0, 1]^m \subset \{0\} \times \mathbb{R}^m \subset \mathbb{R}^{n+m}$. Denote by dx the restriction of the Lebesgue measure on $\mathbb{R}^n$ to $\mathcal{X}$ and by dy the restriction of the Lebesgue measure on $\mathbb{R}^m$ to $\mathcal{Y}$. Let $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ consist of all probability measures $\mu_f := f dxdy$ such that there exists $c_0 > 0$ with the following property: $f \in C^1(\mathcal{X} \times \mathcal{Y})$, $L(f) \leq c_0$, and $f(x, y) \geq c_0$ for all $x, y \in \mathcal{X} \times \mathcal{Y}$.

Outline of the proof of Proposition 3

The proof of Proposition 3 uses Vapnik-Stefanyuk's method of solving stochastic ill-posed problem [66], [57], [62, Theorem 7.3, p. 299] which has been slightly improved by Lê in [37, Theorem 6.1]. The main idea of Proposition 3 uses our characterization of regular conditional probability measure in Theorem 1 in a variational form, namely, the loss function $R_\mu^{K_1} : \mathcal{H} \to \mathbb{R}$, where $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is the (unknown) joint distribution of x, y. Since μ is unknown, we wish to use the empirical measures μ_{S_n} where $S_n \in (\mathcal{X}\times\mathcal{Y})^n$ are observed data to approximate μ in the definition of the loss function $R_\mu : \mathcal{H} \to \mathbb{R}_{\geq 0}$. But such a direct approximation method may not work and we have to perturb the loss function $R_\mu^{K_1} : \mathcal{H} \to \mathbb{R}$ to a sequence of (W, Γ)-regularized loss functions. Vapnik-Stefanyuk's method and its variant [37, Theorem 6.1] provide conditions when a (C, Γ)- ERM algorithm is consistent.

So let us first give a general setting of Vapnik-Stefanyuk's methods of solving stochastic ill-posed problems.

We consider the following operator equation

$$Af = F \tag{18}$$

defined by a continuous operator A which maps in a one-to-one manner the elements f of a metric space (E_1, ρ_{E_1}) into the elements of a metric space (E_2, ρ_{E_2}), assuming that a solution $f \in E_1$ of (18) exists and is unique.

Assume that A belongs to a space $\mathcal{A}$ and instead of Equation (18) we are given a sequence $\{F_{S_l} \in E_2, l \in \mathbb{N}^+\}$, a sequence $\{A_{S_l} \in \mathcal{A}, l \in \mathbb{N}^+\}$, where S_l belongs to a probability space $(\mathcal{X}_l, \mu_l)$ and A_{S_l}, F_{S_l} are defined by a family of maps $\mathcal{X}_l \to E_2$, $S_l \mapsto F_{S_l}$, and $\mathcal{X}_l \to \mathcal{A}$, $S_l \mapsto A_{S_l}$.

Let $W : E_1 \to \mathbb{R}_{\geq 0}$ be a lower semi-continuous function that satisfies the following property (W).

(W) The sets $\mathcal{M}_c = W^{-1}([0,c])$ for $c \geq 0$ are all compact.

Given A_{S_l}, F_{S_l}, and $\gamma_l > 0$, let us define a regularized risk function $R^*_{\gamma_l}(\cdot, F_{S_l}, A_{S_l}) : E_1 \to \mathbb{R}$ by

$$R^*_{\gamma_l}(\hat{f}, F_{S_l}, A_{S_l}) = \rho^2_{E_2}(A_{S_l}\hat{f}, F_{S_l}) + \gamma_l W(\hat{f}). \tag{19}$$

We shall say that $f_{S_l} \in E_1$ is an ε_l-*minimizer* of $R^*_{\gamma_l}$ if

$$R^*_{\gamma_l}(f_{S_l}, F_{S_l}, A_{S_l}) \leq R_{\gamma_l}(\hat{f}, F_{S_l}, A_{S_l}) + \varepsilon_l \text{ for all } \hat{f} \in E_1. \tag{20}$$

We shall also use the shorthand notation A_l for A_{S_l}, F_l for F_{S_l}, f_l for f_{S_l}, ρ_2 for ρ_{E_2}, ρ_1 for ρ_{E_1}. For any $\varepsilon_l > 0$, an ε_l-minimizer of $R^*_{\gamma_l}$ exists. We will measure the closeness of an operator A and an operator A_l by the distance

$$\|A_l - A\| := \sup_{\hat{f} \in E_1} \frac{\|A_l\hat{f} - A\hat{f}\|_{E_2}}{W^{1/2}(\hat{f})}. \tag{21}$$

The following theorem proved by Lê in [37, Theorem 6.1] is a slight improvement of Stefanyuk's theorem [57], [62, Theorem 7.3, p. 299].

Proposition 4 *cf. [62, Theorem 7.3, p. 299] Let f_{S_l} be a γ_l^2-minimizer of $R^*_{\gamma_l}$ in* (19) *and f the solution of* (18)*. For any $\varepsilon > 0$ and any constant $C_1, C_2 > 0$ there exists a value $\gamma(\varepsilon, C_1, C_2) > 0$ such that for any $\gamma_l \leq \gamma(\varepsilon, C_1, C_2)$*

$$\begin{aligned}(\mu_l)^*\{S_l \in \mathcal{X}_l : \rho_1(f_{S_l}, f) > \varepsilon\} \leq (\mu_l)^*\{S_l \in \mathcal{X}_l : \rho_2(F_{S_l}, F) > C_1\sqrt{\gamma_l}\} \\ +(\mu_l)^*\{S_l \in \mathcal{X}_l : \|A_{S_l} - A\| > C_2\sqrt{\gamma_l}\}\end{aligned} \tag{22}$$

holds true.

In the first step of proof of Proposition 3, we shall find a condition for a generative model of supervised learning together with the perturbation term W such that the conditions of Proposition 4 are met. This first step of defining W has been done using a new distance d_M on $C(\mathcal{X}, \mathcal{M}(\mathcal{Y})_{\tilde{K}_2})$ as follows:

$$d_M(f, f') := \sup_{x \in \mathcal{X}}(\|(f - f')(x)\|_{\tilde{K}_2} + \|\Gamma_f(x) - \Gamma_{f'}(x)\|_{\tilde{K}_2} \tag{23}$$

In other worlds, the metric d_M is induced by the norm $\|\cdot\|_M$ on the space $C(\mathcal{X}, S(\mathcal{Y})_{\tilde{K}_2})$ given as follows

$$\|f\|_M = \sup_{x \in \mathcal{X}}(\|f(x)\|_{\tilde{K}_2} + \|\Gamma_f(x)\|_{\tilde{K}_1}).$$

Lemma 1 *Let $W : C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2})_M \to \mathbb{R}_{\geq 0}$ be defined as follows*

$$W(f) := \|f\|_M + L(f) + \|\Gamma_{\underline{f}}\|_{\tilde{K}_3, \tilde{K}_1} \tag{24}$$

Then $W : C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2})_M \to \mathbb{R}_{\geq 0}$ is a lower semi-continuous function. Furthermore, for any $c \geq 0$ the set $W^{-1}[0,c]$ is a compact set in $C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2})_M$.

For the proof of Lemma 1 we refer to [37, Proposition 6.1].

Using Lemma 1 and Proposition 4, we obtain the following

Proposition 5 *Let $\mathcal{X}$ be a compact subset in $\mathbb{R}^n \times \{0\} \subset \mathbb{R}^{n+m}$ and $\mathcal{Y}$ a compact subset in $\{0\} \times \mathbb{R}^m$. Let $K_2 : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be the restriction of the Gaussian kernel $K : \mathbb{R}^{m+n} \times \mathbb{R}^{m+n} \to \mathbb{R}$. Denote by $\mathcal{P}_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}, vol_{\mathcal{X}})$ the set of all probability measures $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that:*
(i) $\mathrm{sppt}\, \mu_{\mathcal{X}} = \mathcal{X}$, *where $\mu_{\mathcal{X}} = (\Pi_{\mathcal{X}})_* \mu$;*
(ii) there exists a regular conditional measure $\mu_{\mathcal{Y}|\mathcal{X}} \in C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2})$ for μ with respect to the projection $\Pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$.

Let $K_1 : \mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ be the restriction of the kernel K to $(\mathcal{X} \times \mathcal{Y})$. We define a loss function

$$R^{K_1} : C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}) \times \mathcal{P}_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}, vol_{\mathcal{X}}) \to \mathbb{R}_{\geq 0}, (h, \mu) \mapsto \|(\Gamma_h)_* \mu_{\mathcal{X}} - \mu\|_{\tilde{K}_1}$$

as in (17)*. Then for any $\mu \in \mathcal{P}_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}, vol_{\mathcal{X}})$, there exists a consistent (C, Γ)-regularized ERM algorithm A for the supervised learning model $(\mathcal{X}, \mathcal{Y}, C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{K_2}), R^{K_1}, \mathcal{P}_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}, vol_{\mathcal{X}}))$. Moreover, for any $\varepsilon, \delta > 0$ there exists $N(\varepsilon, \delta)$ such that for any $n \geq N(\varepsilon, \delta)$ we have*

$$(\mu^n)^* \{S_n \in (\mathcal{X} \times \mathcal{Y})^n : d_M(A(S_n), \mu_{\mathcal{Y}|\mathcal{X}}) > \varepsilon\} \leq \delta, \tag{25}$$

where $\mu_{\mathcal{Y}|\mathcal{X}} \in C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}$ is the unique regular conditional probability measure for μ with respect to the projection $\mathcal{P}_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$.

For a detailed proof of Proposition 5, we refer to [37, Theorem 6.5].

Proposition 3 is derived from Proposition 5. Under the condition (L) of Proposition 3, $W(f)$ is bounded. Hence we can bound the RHS of (22) in the condition of Proposition 5 by using the estimation in (12). It is not hard to see that a (C, Γ) regularized algorithm using the regularized risk function $R^*_{\gamma_n}$ in (20) in the condition of Proposition 5 is uniformly consistent, if $\lim_{n\to\infty} \gamma_n = 0$ and $\lim_{n\to\infty} n\gamma_n = \infty$. Note that the difference between Proposition 3 and Proposition 5 is that in Proposition 5 we choose a subset $\mathcal{H} \subset C_{Lip}(\mathcal{X}, \mathcal{Y}) \subset C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2})$; moreover, we require the condition (L). Since $\mathcal{H} \subset C_{Lip}(\mathcal{X}, \mathcal{Y}) \subset C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2})$, using the explicit form of the loss function R^{K_1} we can find a learning algorithm for the supervised learning model $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, R^{K_1}, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$ from the one for the larger model $(\mathcal{X}, \mathcal{Y}, C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}_2}), R^{K_1}, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$, which is ensured by Proposition 5 and the condition (L).

Our examples in Proposition 3 are of a different nature than the consistent non-parametric regression estimators described in [61]. In [63, §7.9.1, Theorem 5.7, p. 2.5.2] Vapnik also suggested a sufficient condition on the function W for the uniform consistency of the (C, Γ)-algorithm in Stefanyuk's theorem [57], [62, Theorem 7.3, p. 299]. His condition resembles the sufficient condition for the learnability in Proposition 3.

3 Geometric kernels and applications

Kernel methods form a very important paradigm in modern machine learning. In the literature, most positive definite kernels are defined over inner product spaces. However, in many applications, including brain computer interfaces, radar signal processing, and computer vision, the actual data can have much richer geometrical structures beyond the Euclidean space setting, including in particular those of a smooth manifold. There are thus both mathematical and practical interests to investigate positive definite kernels over non-Euclidean structures. In this section, we present several results on positive definite kernels defined over the set of SPD (symmetric positive definite) matrices along with corresponding generalizations to the set of positive Hilbert-Schmidt operators on a Hilbert space.

Consider first the general setting of a metric space (M, d). It is natural to investigate whether the generalization of the Gaussian kernel $K_\gamma : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $K(x, y) = \exp(-\gamma||x - y||^2)$, $\gamma > 0$, to this setting, namely $K_\gamma : M \times M \to \mathbb{R}$, $K(x, y) = \exp(-\gamma d^2(x, y))$, is positive definite.

We first briefly review the concepts of positive definite and negative definite kernels, see e.g. [4], Chapter 3. Let X be a nonempty set. A function $\varphi : X \times X \to \mathbb{R}$ is said to be a *positive definite kernel* if and only if it is symmetric and

$$\sum_{j,k=1}^{N} c_j c_k \varphi(x_j, x_k) \geq 0 \tag{26}$$

where $N \in \mathbb{N}$, $\forall \{x_j\}_{j=1}^N \subset X$, $\{c_j\}_{j=1}^N \subset \mathbb{R}$. A function $\varphi : X \times X \to \mathbb{R}$ is said to be a *negative definite kernel* if and only if it is symmetric and

$$\sum_{j,k=1}^{N} c_j c_k \varphi(x_j, x_k) \leq 0 \tag{27}$$

where $N \geq 2$, $\forall \{x_j\}_{j=1}^N \subset X$, $\{c_j\}_{j=1}^N \subset \mathbb{R}$, with $\sum_{j=1}^N c_j = 0$.

Positive definite and negative definite kernels are closely related. The following is Theorem 3.2.2 in [4], a generalization of Theorem 1 in [52], which is stated for $\varphi(x, y) = d^2(x, y)$ on a separable semi-metric space (M, d) (see Theorem 3 below):

Theorem 2 *Let X be a nonempty set. Then $\varphi : X \times X \to \mathbb{R}$ is negative definite if and only if $\exp(-\gamma\varphi)$ is positive definite $\forall \gamma > 0$.*

In particular, on a Hilbert space $(\mathcal{H}, \langle, \rangle, || \, ||)$, it can be readily verified that the kernel $\varphi(x, y) = ||x-y||^2$ is negative definite, since the condition $\sum_{j=1}^N c_j = 0$ implies that $\sum_{j,k=1}^N c_j c_k ||x_j - x_k||^2 = -|| \sum_{j=1}^N c_j x_j||^2 \leq 0$ always. Thus the Gaussian kernel $K_\gamma(x, y) = \exp(-\gamma||x-y||^2)$ is positive definite $\forall \gamma > 0$. Furthermore, it can be shown that $K(x, y) = \exp(-\gamma||x - y||^p)$ is positive definite $\forall \gamma > 0$ if and only if $0 < p \leq 2$ ([52], Corollary 3).

More generally, let X be a nonempty set and suppose that there exists a map $\psi : X \to \mathcal{H}$. Then the kernel $\varphi : X \times X \to \mathbb{R}$ defined by $\varphi(x, y) = ||\psi(x) - \psi(y)||^2$ is negative definite, hence $K : X \times X \to \mathbb{R}$, $K(x, y) = \exp(-\gamma||\psi(x) - \psi(y)||^2)$ is positive definite $\forall \gamma > 0$.

In general, on a metric space (M, d), by Theorem 2, the kernel $K_\gamma(x, y) = \exp(-\gamma d^2(x, y))$ is positive definite $\forall \gamma > 0$ if and only if $\varphi(x, y) = d^2(x, y)$ is negative definite. Schoenberg [52] proved that this happens if and only if (M, d) is isometrically embeddable into a Hilbert space $\mathcal{H}$. The following is essentially Theorem 1 in [52], see also Proposition 3.3.2 in [4] for a more general version.

Theorem 3 *Let (M, d) be a metric space. The kernel $K_\gamma : M \times M \to \mathbb{R}$, $K_\gamma(x, y) = \exp(-\gamma d^2(x, y))$, is positive definite $\forall \gamma > 0$, or equivalently, the kernel $\varphi(x, y) = d^2(x, y)$ is negative definite, if and only if there exists a Hilbert space $\mathcal{H}$ and a map $\psi : M \to \mathcal{H}$ such that $d(x, y) = ||\psi(x) - \psi(y)||$.*

Consider now the setting where (M, g) is a geodesically complete finite-dimensional Riemannian manifold. Let d_g denote the corresponding metric on M induced by g. Then (M, d_g) is a complete metric space. Building upon Theorem 3, we have the following result (see [30], Theorem 6.2, and [25], Theorem 2)

Theorem 4 *Let (M, g) be a geodesically complete finite-dimensional Riemannian manifold. The kernel $K_\gamma : M \times M \to \mathbb{R}$, $K_\gamma(x, y) = \exp(-\gamma d^2(x, y))$, is positive definite $\forall \gamma > 0$ if and only if M is isometric, in the Riemannian sense, to Euclidean space $\mathbb{R}^n$, for some $n \in \mathbb{N}$.*

Thus on a complete Riemannian manifold (M, g), the Gaussian kernel $K_\gamma(x, y) = \exp(-\gamma d^2(x, y))$ is positive definite $\forall \gamma > 0$ if and only if M is isometric to Euclidean space. This means that if M has nonzero curvature, then the Gaussian kernel on M *cannot* be positive definite $\forall \gamma > 0$. Subsequently, we apply the above general results to construct positive definite kernels on the sets of SPD matrices and positive Hilbert-Schmidt operators on a Hilbert space.

3.1 Kernels defined via Riemannian metrics

Throughout the following, let $\mathrm{M}(n)$ denote the set of real $n \times n$ matrices. Then $(\mathrm{M}(n), +, \cdot, \langle , \rangle_F)$ is an inner product space, where $+$ and $\cdot$ denote standard matrix addition and scalar multiplication, respectively, and $\langle , \rangle_F$ denotes the Frobenius inner product, with $\langle A, B\rangle_F = \mathrm{trace}(A^T B)$. Subsequently, we refer to $\mathrm{M}(n)$ with this inner product space structure, which can be identified with the Euclidean space $(\mathbb{R}^{n^2}, \langle , \rangle)$ under the canonical inner product. Let $\mathrm{Sym}(n)$ denote the set of real, $n \times n$ symmetric matrices. Then $(\mathrm{Sym}(n), +, \cdot, \langle , \rangle_F)$ is a subspace of $\mathrm{M}(n)$.

Let $\mathrm{Sym}^{++}(n)$ denote the set of real $n \times n$ SPD matrices. Then it is an open convex cone in $\mathrm{Sym}(n)$, being closed under scalar multiplication by positive numbers. Thus $\mathrm{Sym}^{++}(n)$ can be viewed as a smooth manifold, with tangent space $T_P(\mathrm{Sym}^{++}(n)) \cong \mathrm{Sym}(n)$ $\forall P \in \mathrm{Sym}^{++}(n)$, and can be equipped with a Riemannian metric.

In the following, we investigate the Gaussian kernel defined using the induced Riemannian distance corresponding to three commonly used Riemannian metrics on $\mathrm{Sym}^{++}(n)$, namely the affine-invariant, Bures-Wasserstein, and Log-Euclidean metrics.

Affine-invariant Riemannian metric. The most well-known Riemannian metric on $\mathrm{Sym}^{++}(n)$ is the affine-invariant Riemannian metric, the study of which goes as far back as [53] and [48]. This Riemannian metric g_{ai} is defined by, $\forall P \in \mathrm{Sym}^{++}(n)$ and $U, V \in T_P(\mathrm{Sym}^{++}(n)) \cong \mathrm{Sym}(n)$,

$$g_{\mathrm{ai}}(P)(U,V) = \langle P^{-1/2}UP^{-1/2}, P^{-1/2}VP^{-1/2}\rangle_F = \mathrm{trace}(P^{-1}UP^{-1}V), \tag{28}$$

where $\langle , \rangle_F$ denotes the Frobenius inner product. It corresponds to the Fisher-Rao metric on the set of zero-mean Gaussian densities on $\mathbb{R}^n$. The Riemannian manifold $(\mathrm{Sym}^{++}(n), g_{\mathrm{ai}})$ is a Cartan-Hadamard manifold, that is it is geodesically complete, simply connected, and with nonpositive sectional curvature (see e.g. [34], chapter XII). There is a unique geodesic $\gamma_{\mathrm{ai}}^{AB}$ connecting any pair $A, B \in \mathrm{Sym}^{++}(n)$, with closed form expression

$$\gamma_{\mathrm{ai}}^{AB}(t) = A^{1/2}\exp[t\log(A^{-1/2}BA^{-1/2})]A^{1/2}, \quad t \in [0,1]. \tag{29}$$

The Riemannian distance between A and B is the length of this geodesic,

$$d_{\mathrm{ai}}(A,B) = ||\log(A^{-1/2}BA^{-1/2})||_F, \tag{30}$$

where $||\ ||_F$ denotes the Frobenius norm. Since $(\mathrm{Sym}^{++}(n), g_{\mathrm{ai}})$ has *nonpositive sectional curvature*, for $n \geq 2$, it cannot be isometric to Euclidean space, and thus by Theorem 4, the kernel $K_\gamma : \mathrm{Sym}^{++}(n) \times \mathrm{Sym}^{++}(n) \to \mathbb{R}$ defined by

$$K_\gamma(A,B) = \exp(-\gamma d_{\mathrm{ai}}^2(A,B)) = \exp(-\gamma||\log(A^{-1/2}BA^{-1/2})||_F^2), \ \gamma > 0, \tag{31}$$

cannot be positive definite $\forall \gamma > 0$ (but may be positive definite for some $\gamma > 0$).

Bures-Wasserstein metric. Another commonly used Riemannian metric on $\mathrm{Sym}^{++}(n)$ is the Bures-Wasserstein metric g_{bw}, see, e.g., [59, 6, 42]. It corresponds to the 2-Wasserstein distance between zero-mean Gaussian measures on $\mathbb{R}^n$. The Bures-Wasserstein metric is given by, $\forall P \in \mathrm{Sym}^{++}(n)$,

$$g_{\mathrm{bw}}(P)(U,V) = \mathrm{trace}(L_P(U)PL_P(V)), \quad U, V \in \mathrm{Sym}(n), \tag{32}$$

where $L_P(V) \in \mathrm{Sym}(n)$ is the unique solution of the Lyapunov equation $XP + PX = V$. $(\mathrm{Sym}^{++}(n), g_{\mathrm{bw}})$ is a Riemannian manifold with *nonnegative sectional curvature*. The Riemannian distance between $A, B \in \mathrm{Sym}^{++}(n)$ is the Bures-Wasserstein distance, which admits the following closed form expression:

$$d_{\mathrm{bw}}(A,B) = \sqrt{\mathrm{trace}(A) + \mathrm{trace}(B) - 2\mathrm{trace}(A^{1/2}BA^{1/2})^{1/2}}. \tag{33}$$

It is the length of the geodesic curve

$$\gamma_{\text{bw}}^{AB}(t) = (1-t)^2 A + t^2 B + t(1-t)[(AB)^{1/2} + (BA)^{1/2}]. \tag{34}$$

Similar to the case of the affine-invariant metric, $(\text{Sym}^{++}(n), g_{\text{bw}})$ *cannot* be isometric to Euclidean space for $n \geq 2$. Thus, by Theorem 4, the kernel $K_\gamma : \text{Sym}^{++}(n) \times \text{Sym}^{++}(n) \to \mathbb{R}$ defined by

$$K_\gamma(A, B) = \exp(-\gamma d_{\text{bw}}^2(A, B)), \quad \gamma > 0, \tag{35}$$

cannot be positive definite $\forall \gamma > 0$ (but may be positive definite for some $\gamma > 0$).

Log-Euclidean metric. We consider next the Log-Euclidean metric, which is widely used in many applications. The Log-Euclidean Riemannian metric on $\text{Sym}^{++}(n)$ was formulated by the authors of [3]. It is the Riemannian metric arising from the following commutative Lie group multiplication on $\text{Sym}^{++}(n)$,

$$\begin{aligned} \odot : \text{Sym}^{++}(n) \times \text{Sym}^{++}(n) &\to \text{Sym}^{++}(n), \\ A \odot B &= \exp(\log(A) + \log(B)), \end{aligned} \tag{36}$$

where log denotes the principal matrix logarithm. The Log-Euclidean metric is a *bi-invariant* Riemannian metric on $(\text{Sym}^{++}(n), \odot)$. Let us fix the inner product on $T_I(\text{Sym}^{++}(n)) \cong \text{Sym}(n)$ to be the Frobenius inner product. Then the Log-Euclidean metric is given by, $\forall P \in \text{Sym}^{++}(n)$,

$$g_{\text{logE}}(P)(U, V) = \langle D\log(P)(U), D\log(P)(V)\rangle_F, \quad U, V \in \text{Sym}(n), \tag{37}$$

where $D\log(P)$ denotes the Fréchet derivative of the principal logarithm log at P. $(\text{Sym}^{++}(n), g_{\text{logE}})$ is a Riemannian manifold with *zero* sectional curvature. There is a unique geodesic joining any pair $A, B \in \text{Sym}^{++}(n)$, with closed form expression

$$\gamma_{\text{logE}}^{AB}(t) = \exp((1-t)\log(A) + t\log(B)). \tag{38}$$

The Riemannian distance between A and B is the length of this geodesic:

$$d_{\text{logE}}(A, B) = ||\log(A) - \log(B)||_F. \tag{39}$$

Vector space structure of $\text{Sym}^{++}(n)$. While $\text{Sym}^{++}(n)$ is not a vector subspace of the Euclidean space $(\text{Sym}(n), +, \cdot, \langle , \rangle_F)$, it admits the following special vector space structure. Together with the abelian group operation $\odot$, we define the following scalar multiplication on $\text{Sym}^{++}(n)$ [3]:

$$\begin{aligned} \circledast : \mathbb{R} \times \text{Sym}^{++}(n) &\to \text{Sym}^{++}(n), \\ \lambda \circledast A &= \exp(\lambda \log(A)) = A^\lambda, \quad \lambda \in \mathbb{R}. \end{aligned} \tag{40}$$

Endowed with the abelian group operation $\odot$ and the scalar multiplication $\circledast$, the vector space axioms can be readily verified to show that $(\text{Sym}^{++}(n), \odot, \circledast)$ is a vector space [3].

Inner product space structure on $\mathrm{Sym}^{++}(n)$. On top of the vector space structure $(\mathrm{Sym}^{++}(n), \odot, \circledast)$, we define the following *Log-Euclidean inner product*:

$$\langle A, B\rangle_{\mathrm{logE}} = \langle \log(A), \log(B)\rangle_F = \mathrm{trace}[\log(A)\log(B)]. \tag{41}$$

along with the corresponding *Log-Euclidean norm*

$$||A||^2_{\mathrm{logE}} = \langle \log(A), \log(A)\rangle_F = \mathrm{trace}[\log^2(A)]. \tag{42}$$

The axioms of inner product, namely symmetry, positivity, and linearity with respect to the operations $(\odot, \circledast)$ can be readily verified to show that

$$(\mathrm{Sym}^{++}(n), \odot, \circledast, \langle\ ,\ \rangle_{\mathrm{logE}}) \tag{43}$$

is an inner product space, as first discussed in [40]. Furthermore, the following map is an isometrical isomorphism of inner product spaces:

$$\log : (\mathrm{Sym}^{++}(n), \odot, \circledast, \langle\ ,\ \rangle_{\mathrm{logE}}) \to (\mathrm{Sym}(n), +, \cdot, \langle\ ,\ \rangle_F),\ A \to \log(A). \tag{44}$$

In terms of the Log-Euclidean inner product and Log-Euclidean norm, the Log-Euclidean distance d_{logE} in Eq.(39) is expressed as

$$d_{\mathrm{logE}}(A, B) = ||\log(A) - \log(B)||_F = ||A \odot B^{-1}||_{\mathrm{logE}}. \tag{45}$$

The inner product space structure of $(\mathrm{Sym}^{++}(n), \odot, \circledast, \langle\ ,\ \rangle_{\mathrm{logE}})$ allows us to define positive definite kernels on $\mathrm{Sym}^{++}(n)$ using the inner product $\langle\ ,\ \rangle_{\mathrm{logE}}$ and the norm $||\ ||_{\mathrm{logE}}$. We have the following result.

Theorem 5 *The following kernels* $K : \mathrm{Sym}^{++}(n) \times \mathrm{Sym}^{++}(n) \to \mathbb{R}$ *are positive definite:*

$$K(A, B) = (\langle A, B\rangle_{\mathrm{logE}} + c)^d = (\langle \log(A), \log(B)\rangle_F + c)^d, \quad c \geq 0, d \in \mathbb{N}. \tag{46}$$

$$\begin{aligned} K(A, B) &= \exp\left(-\frac{1}{\sigma^2}||A \odot B^{-1}||^p_{\mathrm{logE}})\right) \\ &= \exp\left(-\frac{1}{\sigma^2}||\log(A) - \log(B)||^p_F\right), \quad \sigma \neq 0, 0 < p \leq 2. \end{aligned} \tag{47}$$

3.2 Kernels defined with Bregman divergences

In the previous section, we discussed $\mathrm{Sym}^{++}(n)$ from the viewpoint of a Riemannian manifold, along with the corresponding induced geodesic distance. In this section, we consider the Bregman divergences, which are distance-like functions arising from the open convex cone structure of $\mathrm{Sym}^{++}(n)$. In particular, we discuss the Alpha Log-Determinant (Log-Det) divergences, which are obtained based on the strictly convex function $\phi(X) = -\log\det(X)$, $X \in \mathrm{Sym}^{++}(n)$.

Let us first recall the concept of Bregman divergence [12]. Let $\Omega \subset \mathbb{R}^n$ be a convex set and $\phi : \Omega \to \mathbb{R}$ be a differentiable and strictly convex function. Then it defines the following *divergence* function on Ω:

$$B_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle. \tag{48}$$

For example, with $\Omega = \mathbb{R}^n$ and $\phi(x) = ||x||^2$, we obtain the squared Euclidean distance $B_\phi(x, y) = ||x-y||^2$. More generally, ϕ defines following family of divergences [68], parametrized by a parameter $\alpha \in \mathbb{R}$,

$$d^\alpha_\phi(x, y) = \frac{4}{1-\alpha^2}\left[\frac{1-\alpha}{2}\phi(x) + \frac{1+\alpha}{2}\phi(y) - \phi\left(\frac{1-\alpha}{2}x + \frac{1+\alpha}{2}y\right)\right], \tag{49}$$

with $d^{\pm 1}_\phi$ defined as the limits of d^α_ϕ as $\alpha \to \pm 1$. In fact, we have

$$d^1_\phi(x, y) = \lim_{\alpha\to 1} d^\alpha_\phi(x, y) = B_\phi(x, y), \tag{50}$$

$$d^{-1}_\phi(x, y) = \lim_{\alpha\to -1} d^\alpha_\phi(x, y) = B_\phi(y, x), \tag{51}$$

In general, it can be readily verified that d^α_ϕ can be expressed in terms of the Bregman divergence B_ϕ $\forall \alpha \in \mathbb{R}$, as follows:

$$d^\alpha_\phi(x, y) = \frac{4}{1-\alpha^2}\left[\frac{1-\alpha}{2}B_\phi\left(x, \frac{1-\alpha}{2}x + \frac{1+\alpha}{2}y\right) + \frac{1+\alpha}{2}B_\phi\left(y, \frac{1-\alpha}{2}x + \frac{1+\alpha}{2}y\right)\right]. \tag{52}$$

Alpha Log-Det divergences. Consider $\Omega = \mathrm{Sym}^{++}(n)$ together with the function $\phi(X) = -\log\det(X)$, $X \in \mathrm{Sym}^{++}(n)$. Fan's inequality [21] on the log-concavity of the matrix determinant function on $\mathrm{Sym}^{++}(n)$ states that

$$\det[\alpha A + (1-\alpha)B] \geq \det(A)^\alpha \det(B)^{1-\alpha},\ \forall A, B \in \mathrm{Sym}^{++}(n),\ 0 \leq \alpha \leq 1. \tag{53}$$

For $0 < \alpha < 1$, equality occurs if and only if $A = B$. Thus the function $\phi(X) = -\log\det(X)$ is strictly convex on $\mathrm{Sym}^{++}(n)$. Hence, based on Eq. (49), we obtain the parametrized family of Alpha Log-Det divergences, as defined in [14]

$$d^\alpha_{\text{logdet}}(A, B) = \frac{4}{1-\alpha^2}\log\left[\frac{\det(\frac{1-\alpha}{2}A + \frac{1+\alpha}{2}B)}{\det(A)^{\frac{1-\alpha}{2}}\det(B)^{\frac{1+\alpha}{2}}}\right],\ -1 < \alpha < 1, \tag{54}$$

with the limiting cases $\alpha = \pm 1$, obtained via L'Hopital's rule, given by

$$d^1_{\text{logdet}}(A, B) = \lim_{\alpha\to 1} d^\alpha_{\text{logdet}}(A, B) = \mathrm{trace}(B^{-1}A - I) - \log\det(B^{-1}A), \tag{55}$$

$$d^{-1}_{\text{logdet}}(A, B) = \lim_{\alpha\to -1} d^\alpha_{\text{logdet}}(A, B) = \mathrm{trace}(A^{-1}B - I) - \log\det(A^{-1}B). \tag{56}$$

The following properties are immediate from Fan's inequality

$$d^{\alpha}_{\text{logdet}}(A, B) \geq 0, \tag{57}$$

$$d^{\alpha}_{\text{logdet}}(A, B) = 0 \Longleftrightarrow A = B. \tag{58}$$

Instead of symmetry, $d^{\alpha}_{\text{logdet}}$ satisfies the *dual symmetry* property

$$d^{\alpha}_{\text{logdet}}(A, B) = d^{-\alpha}_{\text{logdet}}(B, A). \tag{59}$$

In particular, $d^{\alpha}_{\text{logdet}}$ is symmetric if and only if $\alpha = 0$, that is,

$$d^{0}_{\text{logdet}}(A, B) = d^{0}_{\text{logdet}}(B, A). \tag{60}$$

Having defined the Alpha Log-Det divergences $d^{\alpha}_{\text{logdet}}$ on $\text{Sym}^{++}(n)$, as in the previous section, a natural question that arises is whether Gaussian-like kernels can be defined using $d^{\alpha}_{\text{logdet}}$. By the symmetry property of kernels, the only case that could be considered is $\alpha = 0$. We have the following result from [54, 55].

Theorem 6 ([54, 55]) *The kernel $K : \text{Sym}^{++}(n) \times \text{Sym}^{++}(n) \to \mathbb{R}$, defined by*

$$K(A, B) = \exp\left(-\frac{\sigma}{4} d^{0}_{\text{logdet}}(A, B)\right), \tag{61}$$

is positive definite if and only if σ satisfies

$$\sigma \in \left\{\frac{1}{2}, 1, \ldots, \frac{n-1}{2}\right\} \cup \left\{\sigma \in \mathbb{R}, \sigma > \frac{n-1}{2}\right\}. \tag{62}$$

Proof. This theorem is a special case of Theorem VII.3.1 in [22] in the general setting of Euclidean Jordan algebras and symmetric cones, with $\text{Sym}(n)$ being the Euclidean Jordan algebra under the Jordan product $A \circ B = \frac{1}{2}(AB + BA)$ and $\text{Sym}^{++}(n)$ being the associated symmetric cone.

The following elementary proof for the *if* part is based on that given in [54, 55]. Assume that σ satisfies (62). By definition of d^{0}_{logdet},

$$K(A, B) = \frac{\det(A)^{\frac{\sigma}{2}} \det(B)^{\frac{\sigma}{2}}}{\det(\frac{A+B}{2})^{\sigma}}.$$

It thus suffices to show that the kernel function $H_{\sigma} : \text{Sym}^{++}(n) \times \text{Sym}^{++}(n) \to \mathbb{R}$ defined by $H_{\sigma}(X_1, X_2) = \det(X_1 + X_2)^{-\sigma}$ is positive definite with σ as given in (62). For $X \in \text{Sym}^{++}(n)$, we have the Gaussian integral

$$\int_{\mathbb{R}^n} e^{-y^T X y} dy = \pi^{n/2} \det(X)^{-1/2}.$$

Define the feature map $\varphi : \text{Sym}^{++}(n) \to L^2(\mathbb{R}^n)$ by $\varphi(X)(y) = \frac{1}{\pi^{n/4}} e^{-y^T X y}$. Then for any pair $X_1, X_2 \in \text{Sym}^{++}(n)$,

$$\langle \varphi(X_1), \varphi(X_2) \rangle_{L^2(\mathbb{R}^n)} = \frac{1}{\pi^{n/2}} \int_{\mathbb{R}^n} e^{-y^T (X_1+X_2)y} dy = \det(X_1 + X_2)^{-1/2}.$$

It follows that $H_{1/2}$ is positive definite. Consequently, H_σ is positive definite whenever $\sigma = \frac{k}{2}\ \forall k \in \mathbb{N}$.

For $\sigma > \frac{n-1}{2}$, $\sigma \in \mathbb{R}$, consider the matrix-variate Gamma function (see e.g. [43], Chapter 5, Eq.(5.1.2))

$$\Gamma_n(\sigma) = \int_{\mathrm{Sym}^{++}(n)} e^{-\mathrm{trace}(X)} \det(X)^{\sigma - \frac{n+1}{2}} dX, \quad \sigma > \frac{n-1}{2}, \tag{63}$$

along with the following identity ([43], Eq.(5.2.2)), where $\forall B \in \mathrm{Sym}^{++}(n)$,

$$\det(B)^{-\sigma} = \frac{1}{\Gamma_n(\sigma)} \int_{\mathrm{Sym}^{++}(n)} e^{-\mathrm{trace}(BX)} \det(X)^{\sigma - \frac{n+1}{2}} dX, \quad \sigma > \frac{n-1}{2}. \tag{64}$$

Define the feature map $\psi : \mathrm{Sym}^{++}(n) \to L^2(\mathrm{Sym}^{++}(n), \nu)$, where $d\nu(X) = \frac{1}{\Gamma_n(\sigma)} \det(X)^{\sigma - \frac{n+1}{2}} dX$, by $\psi(X)(Y) = e^{-\mathrm{trace}(XY)}$. Then for any pair $X_1, X_2 \in \mathrm{Sym}^{++}(n)$,

$$\langle \psi(X_1), \psi(X_2) \rangle_{L^2(\mathrm{Sym}^{++}(n), \nu)} = \det(X_1 + X_2)^{-\sigma}.$$

It thus follows that H_σ is positive definite $\forall \sigma > \frac{n-1}{2}$. □

3.3 Kernels defined with the Log-Hilbert-Schmidt metric

In this section, we describe the generalization of the Log-Euclidean metric on $\mathrm{Sym}^{++}(n)$ and its corresponding kernels, as described in Section 3.1, to the infinite-dimensional setting of positive definite Hilbert-Schmidt operators on a Hilbert space. This generalization was first carried out in [44].

Throughout the following, let $(\mathcal{H}, \langle , \rangle)$ be a real, separable Hilbert space, with $\dim(\mathcal{H}) = \infty$ unless explicitly stated otherwise. For two separable Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$, let $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ denote the Banach space of bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$, with operator norm $||A|| = \sup_{||x||_1 \leq 1} ||Ax||_2$. For $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$, we use the notation $\mathcal{L}(\mathcal{H})$. Let $\mathrm{Sym}(\mathcal{H}) \subset \mathcal{L}(\mathcal{H})$ be the set of bounded, self-adjoint linear operators on $\mathcal{H}$. Let $\mathrm{Sym}^+(\mathcal{H}) \subset \mathrm{Sym}(\mathcal{H})$ be the set of self-adjoint, *positive* operators on $\mathcal{H}$, i.e. $A \in \mathrm{Sym}^+(\mathcal{H}) \iff A^* = A, \langle Ax, x \rangle \geq 0 \forall x \in \mathcal{H}$. Let $\mathrm{Sym}^{++}(\mathcal{H}) \subset \mathrm{Sym}^+(\mathcal{H})$ be the set of self-adjoint, *strictly positive* operators on $\mathcal{H}$, i.e $A \in \mathrm{Sym}^{++}(\mathcal{H}) \iff A^* = A, \langle x, Ax \rangle > 0\ \forall x \in \mathcal{H}, x \neq 0$. We write $A \geq 0$ for $A \in \mathrm{Sym}^+(\mathcal{H})$ and $A > 0$ for $A \in \mathrm{Sym}^{++}(\mathcal{H})$. If $\gamma I + A > 0$, where I is the identity operator,$\gamma \in \mathbb{R}, \gamma > 0$, then $\gamma I + A$ is also invertible, in which case it is called *positive definite*. In general, $A \in \mathrm{Sym}(\mathcal{H})$ is said to be positive definite if $\exists M_A > 0$ such that $\langle x, Ax \rangle \geq M_A ||x||^2\ \forall x \in \mathcal{H}$ - this condition is equivalent to A being both strictly positive and invertible, see, e.g., [51].The Hilbert space $\mathrm{HS}(\mathcal{H}_1, \mathcal{H}_2)$ of Hilbert-

Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$ is defined by (see, e.g., [33]) $\mathrm{HS}(\mathcal{H}_1, \mathcal{H}_2) = \{A \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2) : ||A||^2_{\mathrm{HS}} = \mathrm{trace}(A^*A) = \sum_{k=1}^{\infty} ||Ae_k||_2^2 < \infty\}$, for any orthonormal basis $\{e_k\}_{k\in\mathbb{N}}$ in $\mathcal{H}_1$, with inner product $\langle A, B\rangle_{\mathrm{HS}} = \mathrm{trace}(A^*B)$. For $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$, we write $\mathrm{HS}(\mathcal{H})$.

We now seek to generalize the expression $||\log(A) - \log(B)||_F, A, B \in \mathrm{Sym}^{++}(n)$, to the setting where A, B are self-adjoint positive Hilbert-Schmidt operators on a separable Hilbert space. We first note the following *two crucial differences between the finite and infinite-dimensional settings*.

(i) Assume that $1 \leq \dim(\mathcal{H}) \leq \infty$. Assume that $A \in \mathrm{Sym}(\mathcal{H})$ is compact and strictly positive. Then A has a countable spectrum of positive eigenvalues $\{\lambda_k(A)\}_{k=1}^{\dim(\mathcal{H})}$, counting multiplicities, with $\lim_{k\to\infty} \lambda_k(A) = 0$ if $\dim(\mathcal{H}) = \infty$. If $\{\phi_k(A)\}_{k=1}^{\dim(\mathcal{H})}$ denote the corresponding normalized eigenvectors, then A admits the spectral decomposition

$$A = \sum_{k=1}^{\dim(\mathcal{H})} \lambda_k(A)\phi_k(A) \otimes \phi_k(A), \tag{65}$$

where $\phi_k(A)\otimes\phi_k(A) : \mathcal{H} \to \mathcal{H}$ is defined by $(\phi_k(A)\otimes\phi_k(A))w = \langle w, \phi_k(A)\rangle\phi_k(A)$, $w \in \mathcal{H}$. The principal logarithm of A is then given by

$$\log(A) = \sum_{k=1}^{\dim(\mathcal{H})} \log(\lambda_k(A))\phi_k(A) \otimes \phi_k(A). \tag{66}$$

Clearly, $\log(A)$ is bounded if and only if $\dim(\mathcal{H}) < \infty$, since for $\dim(\mathcal{H}) = \infty$, we have $\lim_{k\to\infty} \log(\lambda_k(A)) = -\infty$. Thus, when $\dim(\mathcal{H}) = \infty$, the condition that A be strictly positive is *not* sufficient for $\log(A)$ to be bounded. This problem is resolved by considering the *regularized* or *unitized* operator $A + \gamma I$, $\gamma \in \mathbb{R}$, $\gamma > 0$, which is *positive definite*, so that the following operator is always bounded:

$$\log(A + \gamma I) = \sum_{k=1}^{\infty} \log(\lambda_k(A) + \gamma)\phi_k(A) \otimes \phi_k(A). \tag{67}$$

(ii) The generalization of the Frobenius norm $||\ ||_F$ to the Hilbert space setting is the Hilbert-Schmidt norm $||\ ||_{\mathrm{HS}}$. Consider now the operators $A+\gamma I > 0, B+\mu I > 0$, with $A, B \in \mathrm{HS}(\mathcal{H})$. The expression

$$||\log(A + \gamma I) - \log(B + \mu I)||_{\mathrm{HS}} \tag{68}$$

is generally infinite, however, since the identity operator I is not Hilbert-Schmidt when $\dim(\mathcal{H}) = \infty$, with $||I||_{\mathrm{HS}} = \infty$. Thus, for $A = B = 0$, with $\gamma \neq \mu > 0$, we have $||\log(\gamma I) - \log(\mu I)||_{\mathrm{HS}} = |\log(\gamma) - \log(\mu)|\ ||I||_{\mathrm{HS}} = \infty$. This problem is fully resolved by enlarging the set of Hilbert-Schmidt operators to include the identity operator, via the *extended Hilbert-Schmidt norm* and *inner product*, as follows.

Extended Hilbert-Schmidt operators. In [35], the author considered the following set of *extended*, or *unitized*, Hilbert-Schmidt operators

$$\mathrm{HS}_X(\mathcal{H}) = \{A + \gamma I : A \in \mathrm{HS}(\mathcal{H}), \gamma \in \mathbb{R}\}. \tag{69}$$

This set is a Hilbert space under the *extended Hilbert-Schmidt inner product*, under which the Hilbert-Schmidt and scalar operators are orthogonal,

$$\langle A + \gamma I, B + \mu I\rangle_{\mathrm{HS}_X} = \langle A, B\rangle_{\mathrm{HS}} + \gamma\mu. \tag{70}$$

The corresponding *extended Hilbert-Schmidt norm* is given by

$$||A + \gamma I||^2_{\mathrm{HS}_X} = ||A||^2_{\mathrm{HS}} + \gamma^2. \tag{71}$$

In particular, $||I||_{\mathrm{HS}_X} = 1$, in contrast to $||I||_{\mathrm{HS}} = \infty$.

The Hilbert manifold of positive definite Hilbert-Schmidt operators. Consider the following subset of *(unitized) positive definite Hilbert-Schmidt operators*

$$\mathscr{P}\mathscr{C}_2(\mathcal{H}) = \{A + \gamma I > 0 : A \in \mathrm{Sym}(\mathcal{H}) \cap \mathrm{HS}(\mathcal{H}), \gamma \in \mathbb{R}\} \subset \mathrm{HS}_X(\mathcal{H}). \tag{72}$$

The set $\mathscr{P}\mathscr{C}_2(\mathcal{H})$ is an open subset of the Hilbert space

$$\mathrm{Sym}(\mathcal{H}) \cap \mathrm{HS}_X(\mathcal{H}) = \{A + \gamma I : A = A^*, A \in \mathrm{HS}(\mathcal{H}), \gamma \in \mathbb{R}\}. \tag{73}$$

Thus $\mathscr{P}\mathscr{C}_2(\mathcal{H})$ is a Hilbert manifold modeled on $\mathrm{Sym}(\mathcal{H}) \cap \mathrm{HS}_X(\mathcal{H})$. If $(A+\gamma I) \in \mathscr{P}\mathscr{C}_2(\mathcal{H})$, then it has a countable spectrum $\{\lambda_k(A) + \gamma\}_{k=1}^{\infty}$ satisfying $\lambda_k + \gamma \geq M_A$ for some constant $M_A > 0$, and $\log(A + \gamma I)$ as defined by (66) is well-defined and bounded, with $\log(A + \gamma I) \in \mathrm{Sym}(\mathcal{H}) \cap \mathrm{HS}_X(\mathcal{H})$.

The operations $(\odot, \circledast)$ on $\mathrm{Sym}^{++}(n)$ as defined in Section 3.1, can be readily generalized to $\mathscr{P}\mathscr{C}_2(\mathcal{H})$ as follows [44]. First, the commutative Lie group multiplication operation $\odot$ on $\mathscr{P}\mathscr{C}_2(\mathcal{H})$ is defined by

$$\begin{aligned} \odot : \mathscr{P}\mathscr{C}_2(\mathcal{H}) \times \mathscr{P}\mathscr{C}_2(\mathcal{H}) &\to \mathscr{P}\mathscr{C}_2(\mathcal{H}) \\ (A + \gamma I) \odot (B + \mu I) &= \exp(\log(A + \gamma I) + \log(B + \mu I)). \end{aligned} \tag{74}$$

The Log-Hilbert-Schmidt metric is then a bi-invariant metric on $(\mathscr{P}\mathscr{C}_2(\mathcal{H}), \odot)$. Let us choose the inner product on $T_I(\mathscr{P}\mathscr{C}_2(\mathcal{H})) \cong \mathrm{Sym}(\mathcal{H}) \cap \mathrm{HS}_X(\mathcal{H})$ to be the extended Hilbert-Schmidt inner product. Then we have the following generalization of the Log-Euclidean metric in Section 3.1

Theorem 7 *The Log-Hilbert-Schmidt metric is given by,* $\forall P \in \mathscr{P}\mathscr{C}_2(\mathcal{H})$, $\forall U, V \in \mathrm{Sym}(\mathcal{H}) \cap \mathrm{HS}_X(\mathcal{H})$,

$$g_{\mathrm{logHS}}(P)(U, V) = \langle D\log(P)(U), D\log(P)(V)\rangle_{\mathrm{HS}_X}, \tag{75}$$

where $D\log(P)$ *denotes the Fréchet derivative of the principal logarithm* log *at* P. $(\mathscr{P}\mathscr{C}_2, g_{\mathrm{logHS}})$ *is an infinite-dimensional Riemannian manifold with zero sectional curvature. There is a unique geodesic joining any pair* $(A+\gamma I), (B+\mu I) \in \mathscr{P}\mathscr{C}_2(\mathcal{H})$,

with closed form expression

$$\gamma_{\text{logHS}}^{(A+\gamma I),(B+\mu I)}(t) = \exp((1-t)\log(A+\gamma I) + t\log(B+\mu I)). \tag{76}$$

The induced Riemannian distance between $(A+\gamma I)$ *and* $(B+\mu I)$ *is the length of this geodesic:*

$$d_{\text{logHS}}[(A+\gamma I),(B+\mu I)] = ||\log(A+\gamma I) - \log(B+\mu I)||_{\text{HS}_X}. \tag{77}$$

For any pair of operators $(A+\gamma I),(B+\mu I) \in \mathscr{PC}_2(\mathcal{H})$, the distance $d_{\text{logHS}}[(A+\gamma I),(B+\mu I)] = ||\log(A+\gamma I)-\log(B+\mu I)||_{\text{HS}_X}$ is always finite. Furthermore, when $\dim(\mathcal{H}) = \infty$, by the orthogonality of the scalar and Hilbert-Schmidt operators, we have the decomposition

$$||\log(A+\gamma I) - \log(B+\mu I)||^2_{\text{HS}_X} = \left\|\log\left(\frac{A}{\gamma}+I\right) - \log\left(\frac{B}{\mu}+I\right)\right\|^2_{\text{HS}} + \left(\log\frac{\gamma}{\mu}\right)^2. \tag{78}$$

In particular, for $A = B = 0$, Eq. (78) gives

$$d_{\text{logHS}}[\gamma I, \mu I] = ||\log(\gamma I) - \log(\mu I)||_{\text{HS}_X} = |\log(\gamma/\mu)|. \tag{79}$$

Thus the second term on the right hand side of Eq. (78) is precisely the squared Log-Hilbert-Schmidt distance between the scalar operators γI and μI (this distance would be infinite if measured in the Hilbert-Schmidt norm $||\ ||_{\text{HS}}$).

Vector space structure of $\mathscr{PC}_2(\mathcal{H})$. Together with the group operation $\odot$, we define the following scalar multiplication on $\mathscr{PC}_2(\mathcal{H})$

$$\begin{aligned} &\circledast : \mathbb{R} \times \mathscr{PC}_2(\mathcal{H}) \to \mathscr{PC}_2(\mathcal{H}), \\ &\quad \lambda \circledast (A+\gamma I) = \exp(\lambda \log(A+\gamma I)) = (A+\gamma I)^{\lambda}, \quad \lambda \in \mathbb{R}. \end{aligned} \tag{80}$$

Endowed with the commutative group multiplication $\odot$ and the scalar multiplication $\circledast$, the vector space axioms can be readily verified to show that $(\mathscr{PC}_2(\mathcal{H}), \odot, \circledast)$ is a vector space.

Hilbert space structure on $\mathscr{PC}_2(\mathcal{H})$. On top of the vector space structure $(\mathscr{PC}_2(\mathcal{H}), \odot, \circledast)$, we define the following *Log-Hilbert-Schmidt inner product* on $(\mathscr{PC}_2(\mathcal{H}), \odot, \circledast)$ by

$$\langle (A+\gamma I),(B+\mu I)\rangle_{\text{logHS}} = \langle \log(A+\gamma I), \log(B+\mu I)\rangle_{\text{HS}_X}, \tag{81}$$

along with the corresponding *Log-Hilbert-Schmidt norm*

$$||A+\gamma I||^2_{\text{logHS}} = \langle \log(A+\gamma I), \log(A+\gamma I)\rangle_{\text{HS}_X}. \tag{82}$$

The axioms of inner product, namely symmetry, positivity, and linearity with respect to the operations $(\odot, \circledast)$ can be verified (see [44]) to show that

$$(\mathscr{PC}_2(\mathcal{H}), \odot, \circledast, \langle\ ,\ \rangle_{\text{logHS}}) \tag{83}$$

is a complete inner product space, that is, a Hilbert space. This Hilbert space structure was first discussed in [44] and generalizes the finite-dimensional inner product space

$$(\text{Sym}^{++}(n), \odot, \circledast, \langle\ ,\ \rangle_{\text{logE}})$$

in Section 3.1. Also generalizing from Section 3.1, the map

$$\log : (\mathscr{PC}_2(\mathcal{H}), \odot, \circledast, \langle\ ,\ \rangle_{\text{logHS}}) \to (\text{Sym}(\mathcal{H}) \cap \text{HS}_X(\mathcal{H}), +, \cdot, \langle\ ,\ \rangle_{\text{HS}_X}) \tag{84}$$

is an isometrical isomorphism of Hilbert spaces, where the operations $(+, \cdot)$ are the standard addition and scalar multiplication operations, respectively.

In terms of the Log-Hilbert-Schmidt inner product and Log-Hilbert-Schmidt norm, the Log-Hilbert-Schmidt distance d_{logHS} in Eq. (77) is expressed as

$$\begin{aligned} d_{\text{logHS}}(A+\gamma I), (B+\mu I)] &= ||\log(A+\gamma I) - \log(B+\mu I)||_{\text{HS}_X} \\ &= ||(A+\gamma I) \odot (B+\mu I)^{-1}||_{\text{logHS}} \\ &= \sqrt{\langle (A+\gamma I) \odot (B+\mu I)^{-1}, (A+\gamma I) \odot (B+\mu I)^{-1} \rangle_{\text{logHS}}}. \end{aligned} \tag{85}$$

The Hilbert space structure of $(\mathscr{PC}_2(\mathcal{H}), \odot, \circledast, \langle\ ,\ \rangle_{\text{logHS}})$ allows us to define positive definite kernels on $\mathscr{PC}_2(\mathcal{H})$ using the inner product $\langle\ ,\ \rangle_{\text{logHS}}$ and the norm $||\ ||_{\text{logHS}}$. The following are the infinite-dimensional generalizations of the Log-Euclidean kernels defined in Section 3.1, see [44]:

Theorem 8 *The following kernels $K : \mathscr{PC}_2(\mathcal{H}) \times \mathscr{PC}_2(\mathcal{H}) \to \mathbb{R}$ are positive definite:*

$$\begin{aligned} K[(A+\gamma I), (B+\mu I)] &= (\langle (A+\gamma I), (B+\mu I) \rangle_{\text{logHS}} + c)^d, \quad c \geq 0,\ d \in \mathbb{N} \\ &= (c + \langle \log(A+\gamma I), \log(B+\mu I) \rangle_{\text{HS}_X})^d, \end{aligned} \tag{86}$$

$$\begin{aligned} K[(A+\gamma I), (B+\mu I)] &= \exp\left(-\frac{||(A+\gamma I) \odot (B+\mu I)^{-1}||^p_{\text{logHS}}}{\sigma^2} \right) \\ &= \exp\left(-\frac{||\log(A+\gamma I) - \log(B+\mu I)||^p_{\text{HS}_X}}{\sigma^2} \right), \end{aligned} \tag{87}$$

for $\sigma \neq 0$, $0 < p \leq 2$.

Theorem 8 thus generalizes Theorem 5 to the infinite-dimensional setting. In particular, for $\mathcal{H} = \mathbb{R}^n$, $\mathscr{PC}_2(\mathcal{H}) = \text{Sym}^{++}(n)$, $\gamma = \mu = 0$, and $A, B \in \text{Sym}^{++}(n)$, we recover Theorem 5.

The RKHS setting We now present a concrete setting, namely that of RKHS covariance operators, which is of particular interest computationally and practically, since many quantities admit closed forms via kernel Gram matrices which can be

efficiently computed. This setting has been applied in problems in machine learning and computer vision, see e.g. [69, 44, 45].

In the following, we assume that $\mathcal{X}$ is a complete separable metric space, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous positive definite kernel, ρ is a Borel probability measure on $\mathcal{X}$, such that $\int_{\mathcal{X}} K(x,x)d\rho(x) < \infty$. Let $\mathcal{H}(K)$ be the reproducing kernel Hilbert space (RKHS) induced by K, then $\mathcal{H}(K)$ is separable ([58], Lemma 4.33). Let $\Phi : \mathcal{X} \to \mathcal{H}(K)$ be the corresponding canonical feature map $\Phi(x) = K_x$, where $K_x : \mathcal{X} \to \mathbb{R}$ is defined by $K_x(y) = K(x,y) \forall y \in \mathcal{X}$. Then $K(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}(K)}$ $\forall (x,y) \in \mathcal{X} \times \mathcal{X}$ and the probability measure ρ satisfies

$$\int_{\mathcal{X}} ||\Phi(x)||^2_{\mathcal{H}(K)} d\rho(x) = \int_{\mathcal{X}} K(x,x)d\rho(x) < \infty. \tag{88}$$

The following RKHS mean vector $\mu_\Phi \in \mathcal{H}(K)$ and RKHS covariance operator $C_\Phi : \mathcal{H}(K) \to \mathcal{H}(K)$ induced by the feature map Φ are then well-defined:

$$\mu_\Phi = \mu_{\Phi,\rho} = \int_{\mathcal{X}} \Phi(x)d\rho(x) \in \mathcal{H}(K), \tag{89}$$

$$C_\Phi = C_{\Phi,\rho} = \int_{\mathcal{X}} (\Phi(x) - \mu_\Phi) \otimes (\Phi(x) - \mu_\Phi)d\rho(x). \tag{90}$$

Here C_Φ is a positive, trace class operator on $\mathcal{H}(K)$. Let $\mathbf{X} = [x_1, \ldots, x_m]$,$m \in \mathbb{N}$, be a data matrix randomly sampled from $\mathcal{X}$ according to ρ, where $m \in \mathbb{N}$ is the number of observations. The feature map Φ on $\mathbf{X}$ defines the bounded linear operator $\Phi(\mathbf{X}) : \mathbb{R}^m \to \mathcal{H}(K), \Phi(\mathbf{X})\mathfrak{b} = \sum_{j=1}^m b_j \Phi(x_j), \mathfrak{b} \in \mathbb{R}^m$. The corresponding RKHS empirical mean vector and RKHS covariance operator for $\Phi(\mathbf{X})$ are defined to be

$$\mu_{\Phi(\mathbf{X})} = \frac{1}{m} \sum_{j=1}^m \Phi(x_j) = \frac{1}{m} \Phi(\mathbf{X})\mathbf{1}_m, \tag{91}$$

$$C_{\Phi(\mathbf{X})} = \frac{1}{m} \sum_{j=1}^m (\Phi(x_j) - \mu_{\Phi(\mathbf{X})}) \otimes (\Phi(x_j) - \mu_{\Phi(\mathbf{X})}) \tag{92}$$

$$= \frac{1}{m} \Phi(\mathbf{X}) J_m \Phi(\mathbf{X})^* : \mathcal{H}(K) \to \mathcal{H}(K), \tag{93}$$

where $J_m = I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T, \mathbf{1}_m = (1, \ldots, 1)^T \in \mathbb{R}^m$, is the centering matrix.

For concreteness, consider the task of image classification in computer vision, see e.g. [69, 44, 45]. In this setting, $\mathbf{X}$ is a data matrix of features obtained from an image and $C_{\Phi(\mathbf{X})}$ is a representation of that image, the so-called *covariance operator representation*. This representation gives rise to powerful nonlinear algorithms with substantial improvements over finite-dimensional covariance matrices, which are a special case of the RKHS formulation when K is the linear kernel, i.e. $K(x,y) = \langle x, y \rangle$ on $\mathbb{R}^n \times \mathbb{R}^n$. With each image being represented by an RKHS covariance operator, for the task of image classification it is necessary to have a measure of similarity/dissimilarity between them.

Let ρ_1, ρ_2 be two Borel probability measures on $\mathcal{X}$ satisfying $\int_{\mathcal{X}} K(x,x)d\rho_i(x) < \infty$, $i = 1, 2$. Let $\mathbf{X}^i = (x^i_j)_{j=1}^{m_i}$, $i = 1, 2$, be randomly sampled from $\mathcal{X}$ according to ρ_i. Let $\mu_{\Phi(\mathbf{X}^1)}, \mu_{\Phi(\mathbf{X}^2)}$ and $C_{\Phi(\mathbf{X}^1)}, C_{\Phi(\mathbf{X}^2)}$ be the corresponding mean vectors and covariance operators induced by the kernel K, respectively. Define the following finite Gram matrices:

$$K[\mathbf{X}^1] = \Phi(\mathbf{X}^1)^*\Phi(\mathbf{X}^1) \in \mathbb{R}^{m_1\times m_1},\ K[\mathbf{X}^2] = \Phi(\mathbf{X}^2)^*\Phi(\mathbf{X}^2) \in \mathbb{R}^{m_2\times m_2}, \tag{94}$$

$$K[\mathbf{X}^1, \mathbf{X}^2] = \Phi(\mathbf{X}^1)^*\Phi(\mathbf{X}^2) \in \mathbb{R}^{m_1\times m_2}, \tag{95}$$

$$(K[\mathbf{X}^1])_{jk} = K(x^1_j, x^1_k), (K[\mathbf{X}^2])_{jk} = K(x^2_j, x^2_k), (K[\mathbf{X}^1, \mathbf{X}^2])_{jk} = K(x^1_j, x^2_k). \tag{96}$$

Let $\gamma_i > 0$, $i = 1, 2$, be fixed. Then the Log-Hilbert-Schmidt distance between $(C_{\Phi(\mathbf{X}^1)} + \gamma_1 I_{\mathcal{H}(K)})$ and $(C_{\Phi(\mathbf{X}^2)} + \gamma_2 I_{\mathcal{H}(K)})$ admits a closed form formula via the Gram matrices, as follows.

$$\begin{aligned} &||\log(C_{\Phi(\mathbf{X}^1)} + \gamma_1 I_{\mathcal{H}(K)}) - \log(C_{\Phi(\mathbf{X}^2)} + \gamma_2 I_{\mathcal{H}(K)})||^2_{\mathrm{HS}_X} \\ &= ||\log(I_{m_1} + A^*A)||^2_F + ||\log(I_{m_2} + B^*B)||^2_F - 2\mathrm{trace}[B^*Ah(A^*A)A^*Bh(B^*B)] \\ &\quad + \left(\log\frac{\gamma_2}{\gamma_1}\right)^2. \end{aligned} \tag{97}$$

Here $A^*A = \frac{1}{m_1\gamma_1}J_{m_1}K[\mathbf{X}^1]J_{m_1}$, $B^*B = \frac{1}{m_2\gamma_2}J_{m_2}K[\mathbf{X}^2]J_{m_2}$, $A^*B = \frac{1}{\sqrt{m_1m_2\gamma_1\gamma_2}}J_{m_1}K[\mathbf{X}^1, \mathbf{X}^2]J_{m_2}$, $B^*A = \frac{1}{\sqrt{m_1m_2\gamma_1\gamma_2}}J_{m_2}K[\mathbf{X}^2, \mathbf{X}^1]J_{m_1}$, and $h(A) = A^{-1}\log(I + A)$, $A \in \mathrm{Sym}^+(\mathcal{H})$, with $h(0) = I$.

A two-layer kernel machine. The Log-Hilbert-Schmidt inner product and distance between RKHS covariance operators gives rise to the following two-layer kernel machine, used in, e.g., image classification, as follows. In the first layer, a kernel K_1 is applied to the extracted image features, so that each image is represented via an RKHS covariance operator. The Log-Hilbert-Schmidt distances/inner products between the RKHS covariance operators are then computed and another kernel K_2 is defined using these distances/inner products. A kernel algorithm, e.g. for classification, is then readily applied using the kernel K_2. All computations are carried out via the corresponding kernel Gram matrices. It has been demonstrated that the extra nonlinearity, via the addition of the first kernel layer, generally results in substantially better practical performances than kernel methods with the Log-Euclidean metric, where only the second kernel layer is present. We refer to [44, 45] for further details of the actual practical experiments.

4 Geometric manifold learning techniques

Data-driven sciences are widely regarded as the next paradigm that can fundamentally change sciences and pave the way for a new industrial revolution. In passing

now from (merely) topological to *geometric* data analysis we see now that differential, computational and discrete geometry have achieved first and great successes in data characterization and modelling. In particular, geometric deep learning has significantly advanced the capability of learning models for data with complicated topological and geometric structures. The combination of geometric methods with learning models has thus great potential to fundamentally change the data sciences, and the involved disciplines, methods, and techniques nowadays include, but are not confined to, discrete exterior calculus and Laplace Operators, discrete optimal transport, and geometric flow, discrete Ricci (like Olivier and Forman type) curvatures, conformal geometry, combinatorial Hodge theory, dimension reduction via manifold learning, Isomap, Laplacian eigenmaps, diffusion maps, hyperbolic geometry, Poincaré embeddings, etc., geometric signal processing and deep learning, graph, simplex and hypergraph neural networks, index theory, information geometry and (Gromov-) Hausdorff distance.

Given this huge and diverse amount of topics, it is clearly impossible to cover all of them here in desirable detail. In this section, we will therefore concentrate on just two, but paradigmatic ones, about 'learning' Riemannian manifolds, namely, the result of Belkin-Niyogi justifying mathematically, at least to a certain extent, the wide-spread use of Laplacian eigenmap techniques, and the foundational theoretical work of Fefferman et al. on (Riemannian) manifold reconstruction.

4.1 Some Generalities on Manifold Learning

Broadly speaking, manifold learning stands for a variety of techniques used in machine learning and data analysis to understand the underlying structure of high-dimensional data. Originally, "manifold learning techniques" refer to non-linear techniques of dimension reduction, i.e., techniques of transforming an initial representation of items in some high dimensional space into a representation of these items in a space of lower dimension while preserving certain relevant properties of the initial representation. We shall consider a more general concept of "manifold learning" by interpreting "relevant properties of the initial representation of items" as characteristics of the underlying geometric structure of a Riemannian (sub)manifold M underlying the items and residing in a metric space X. Thus, in this broader sense, "manifold learning" signifies learning these characteristics from data points $x_i \in M \subset X$, which have been distributed by the Riemannian volume form on M.

4.2 Manifold Learning using spectral properties à la Belkin-Niyogi

We consider the typical scenario in which the given data reside on a low-dimensional manifold embedded in a higher-dimensional Euclidean space. One popular approach for dealing with this situation is to construct a graph that approximately represents the

manifold and embed this graph into a low dimensional Euclidean space. Examples of methods utilizing this approach include Isomap, Locally Linear Embedding (LLE), and Laplacian eigenmaps, among many others. As an example, we now discuss a theorem by Belkin and Niyogi (Theorem 9) motivating their Laplacian eigenmap algorithm [7], which we interpret as an instance of Manifold Learning (Remark 8).

Let (M, g) be an n-dimensional compact connected Riemannian submanifold of Euclidean d-dimensional space $(\mathbb{R}^d, \langle , \rangle)$ for some $d > n$. [1]

We shall use the notation $\langle , \rangle$ and $\| \cdot \|$ for the Euclidean scalar product and the Euclidean norm, respectively, and also for the induced metric g and the associated norm on $TM \subset \mathbb{R}^n$, respectively. Given a set of data points $S_m = \{x_i\}_{i=1}^m$ in M, we now seek to find a map $f_{S_m} : S_m \to \mathbb{R}^d$ such that the Riemannian distances between each pair (x_i, x_j) are preserved as best as possible. The map f_{S_m} will thus preserve certain features of the Riemannian geometry of (M, g). Following Belkin and Niyogi [7, 8], we shall associate to S_m a one-parameter family of weighted full graphs $S_m(t)$, where $t \in \mathbb{R}^+$, whose vertices are elements of S_m and whose edge e_{ij} connecting x_i with x_j has weight

$$W_{ij}^t = e^{-\frac{||x_i - x_j||^2_{\mathbb{R}^d}}{4t}}.$$

Each of these graphs then encodes full information of the extrinsic distance between each pair (x_i, x_j) in S_m, when regarded as points in $\mathbb{R}^d$. Note that we can compute the extrinsic distance between x_i and x_j but not their intrinsic distance, i.e., the Riemannian distance, since the Riemannian geometry of M is unknown. However, a theorem due to Belkin and Niyogi [7, 9] states that we can learn the spectral geometry of M from the spectral geometry of graphs $S_m(t)$ in probability, as m goes to infinity.

To each graph $S_m(t)$ we assign a symmetric positive semi-definite bilinear form $\Delta^t_{S_m}$ on $\mathbb{R}^m$ by setting for any $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$

$$\Delta^t_{S_m}(z, z) := \frac{1}{2} \sum_{i,j=1}^m W_{ij}^t (z_i - z_j)^2. \tag{98}$$

The bilinear form $\Delta^t_{S_m}$ is called the *unnormalized graph Laplacian* associated with the graph $S_m(t)$. We also identify $\Delta^t_{S_m}$ with a self-adjoint operator on $\mathbb{R}^m$. For a function $f : M \to \mathbb{R}$, we have the vector $\mathbf{f} = (f(x_i))_{i=1}^m \in \mathbb{R}^m$ and we shall regard $\Delta^t_{S_m}$ as an operator acting on f at the data points x_i as follows:

$$(\Delta^t_{S_m} f)(x_i) := (\Delta^t_{S_m} \mathbf{f})(x_i) = f(x_i) \sum_{j=1}^m e^{-\frac{||x_i - x_j||^2_{\mathbb{R}^d}}{4t}} - \sum_{j=1}^m f(x_j) e^{-\frac{||x_i - x_j||^2_{\mathbb{R}^d}}{4t}}.$$

[1] By the famous Nash embedding theorem, any Riemannian manifold (M, g) can be isometrically embedded into a Euclidean space [49]). However, this result also comes with its own 'curse of dimensionality' - namely, if M has dimension n, then d is in general of order n^3.More precisely: if M is compact, $d = \frac{1}{2}n(3n + 1)$, otherwise $d = \frac{1}{2}n(3n + 1)(n + 1)$.

This motivates the definition of the following *point cloud Laplace operator* acting on any $f : M \to \mathbb{R}$, where $\forall x \in M$, one sets

$$(\Delta^t_{S_m} f)(x) = f(x)\frac{1}{m}\sum_{j=1}^{m} e^{-\frac{\|x-x_j\|^2_{\mathbb{R}^d}}{4t}} - \frac{1}{m}\sum_{j=1}^{m} f(x_j)e^{-\frac{\|x-x_j\|^2_{\mathbb{R}^d}}{4t}}.$$

Here $(\mathbf{\Delta}^t_{S_m} f)(x_i) = \frac{1}{m}(\Delta^t_{S_m} f)(x_i)$. Denote by vol_g the Riemannian volume form on (M, g). For a compact manifold M, its volume is finite and the Riemannian volume form gives rise to the uniform probability distribution $\mu \in \mathcal{P}(M)$ via $d\mu(p) = \frac{\mathrm{vol}_g(p)}{\mathrm{vol}_g(M)}$. The following result then shows that $\mathbf{\Delta}^t_{S_m}$ is an empirical version of Δ_g.

Theorem 9 ([9], Theorem 3.1) *Let M be a compact n-dimensional Riemannian submanifold in $\mathbb{R}^d$. Let $S_m = \{x_i\}_{i=1}^m$ be sampled according to the uniform probability distribution μ on M. Let $f \in C^\infty(M)$. Put $t_m = m^{-\frac{1}{n+2+\alpha}}$, where $\alpha > 0$. Then $\forall p \in M$, and for any fixed $\epsilon > 0$,*

$$\lim_{m\to\infty} \mu^m \left\{ S_m \in M^m : \left| \frac{1}{t_m(4\pi t_m)^{n/2}} \mathbf{\Delta}^t_{S_m} f(p) - \frac{1}{\mathrm{vol}_g(M)} (\Delta_g f)(p) \right| > \epsilon \right\} = 0.$$

Remark 8 (1) Theorem 9 can be interpreted as a statement on the successful learning of the Laplace operator of a Riemanian submanifold $M \overset{i}{\hookrightarrow} \mathbb{R}^N$ by empirical data $S_m = \{x_i\}_{i=1}^m$ in a setting of unsupervised learning. Here the hypothesis class of possible approximators of the Riemannian Laplace operator consists of point cloud Laplace operators. To measure the deviation of the point cloud Laplace operators from the Laplacian operator Δ_g of M we use the family of loss functions, parameterized by $p \in M$ and $f \in C^\infty(M)$ defined in Theorem 9.

(2) The investigation of the spectral geometry of a Riemannian manifold (M, g) from its discretizations has been initiated by Dodziuk [18] and Dodziuk-Patodi [19]. Since then there have been many papers in Riemannian geometry devoted to this problem, see e.g. Burago-Ivanov-Kurylev [13] and the references therein. A principal difference between Belkin-Niyogi's and the aforementioned results in Riemannian geometry is that the former considers the extrinsic distance between data points $x_i \in M \subset \mathbb{R}^d$, which is indeed very natural from a data science point of view. Another principal difference between Belkin-Niyogi's result and the ones just mentioned is that the convergence in Theorem 9 is convergence in probability, whereas the convergence in the others takes place under different assumptions. Notice, however, that spectral invariants are not complete invariants of a Riemannian manifold, i.e., isospectrality does not imply isometry, see, e.g., Gordon-Webb-Wolpert [20].

(3) One may also ask whether there are analogues or extensions of Belkin-Niyogi's result to other natural differential operators on Riemannian manifolds. For example, in the case where M is spin, is there a corresponding theorem for the Dirac operator on M?

4.3 Riemannian Manifold Reconstruction à la Fefferman et al.

Most results in manifold learning, as the one by Belkin-Niyogi discussed above, assume *a priori* the existence of a (Riemannian) manifold fitting the given data, though the manifold itself will actually essentially always remain unknown. However, let us now consider a sort of converse problem, namely: Suppose that we are given a set of data and distances between its respective elements, so that we may think of it as a metric space X (with, at least in data science applications, a usually finite, but nevertheless large number of elements). Will there then be an algorithm to construct a Riemannian manifold (M, g) from X so that the further study of X, might, in particular, be amenable to techniques from global analysis and differential geometry?

Fundamental, and at least in our opinion, not merely mathematically sound, but indeed seminal work in this direction has recently been done by Fefferman et al. in a series of papers which we would now like to draw attention upon. Compare here, in particular, [23] as well as [24] and the further references therein.

Indeed, in [23] the authors thoroughly investigate how a Riemannian manifold (M, g) may best interpolate a given metric space (X, d). For such an approximation, a smooth n-dimensional manifold with Riemannian metric has to be constructed, and determining such a Riemannian manifold does involve the construction of its topology, differentiable structure, and Riemannian metric. As their main result, the authors provide sufficient as well as necessary conditions to ensure that a metric space X can be approximated, in the Gromov–Hausdorff or quasi-isometric sense, by a Riemannian manifold (M, g) of a fixed dimension and with bounded diameter, sectional curvature, and injectivity radius, see ([23], Theorem 1). For this to hold, X should locally be metrically close to some Euclidean space and globally be endowed with an almost intrinsic metric, and M is constructed as a submanifold of a separable Hilbert space, that is either $\mathbb{R}^d$ or ℓ^2 (though the Riemannian metric g is, in general, not equal to the induced submanifold one).

Moreover, in [24] the authors take an important step further, in particular to challenges arising in machine learning, by considering the task of approximating a Riemannian manifold (M, g) from the geodesic distances of points in a discrete subset of M, where the Riemannian manifold (M, g) is considered as an abstract metric space with intrinsic distances (and not just as an embedded submanifold of an ambient Euclidean space), and, in addition, some 'noise' is present. Indeed, if (M, g) is an (unknown) Riemannian manifold, $X_1, X_2, \ldots, X_N$ a set of N sample points randomly sampled from M and, in terms of the induced geodesic distance d_M on M, $D_{jk} := d_M(X_j, X_k) + \eta_{jk}$, where $j, k = 1, 2, \ldots, N$ and η_{jk} are independent, identically distributed random variables (subject to certain natural conditions) are given, the authors prove that for N getting larger and larger, one can construct a Riemannian manifold (M^*, g^*) approximating (M, g) w.r.t. the Lipschitz distance in higher and higher probability.

5 Final remarks

We have demonstrated that learning, as most basic human activity, enjoys functoriality description and geometric expressibility which allows us to study learning processes using categorical languages and geometric methods. It remains an open problem to find new geometric methods for the construction of learning models and learning processes and new necessary and sufficient conditions under which a learning process is successful.

In this paper we demonstrated the usefulness of the languague of the Markov category of probabilistic morphisms in statistical learning theory. We showed that the learnability of learning algorithms in supervised learning can be derived in particular from geometric properties of the graph operator of probabilistic morphisms that formalize regular conditional probability operators in supervised learning theory. Our categorical and geometrical framework is a natural generalization of Vapnik-Stephanyuk's representation of regular conditional probabilities as a solution of multidimensional Fredholm integral equations [62] [63], [64].

Acknowledgments

The contributions by HVL and WT were supported by the Research Institute for Mathematical Sciences, an International Joint Usage/Research Center located in Kyoto University during their visit to RIMS in September 2023. All the authors would like to thank the ananymous referee for several helpful comments.

Further funding: The research of HVL was additionally supported by the Institute of Mathematics, Czech Academy of Sciences (RVO 67985840) and GAČR-project GA22-00091S, and the research of WT was additionally supported from the German-Japanese university consortium HeKKSaGOn.

References

1. N. Ay, J. Jost, H. V. Lê and L. Schwachhöfer, Information Geometry. Springer, 2017.
2. N. Aronszajn, Theory of reproducing kernels. Trans. Amer. Math. Soc. 68 (1950),337-404.
3. V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J. on Matrix An. and App., 29(1):328-347, 2007.
4. C. Berg, J. P. R. Christensen, and P. Ressel, Harmonic Analysis on Semigroups. Springer, 1984

5. J. Berner, P. Grohs, G. Kutyniok, P. Petersen, The Modern Mathematics of Deep Learning, in: Mathematical Aspects of Deep Learning", Edited by P. Grohs, G. Kutyniok, Cambridge University Press 2022, 1-111, arXiv:2105.04026.
6. R. Bhatia, T. Jain, and Y. Lim. On the Bures–Wasserstein distance between positive definite matrices. Expositiones Mathematicae, 37(2), 165-191, 2019.
7. M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Advances in Neural Information Processing Systems 14(6):585-591, 2001.
8. M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 15(6), pp.1373-1396, 2003.
9. M. Belkin and P. Niyogi, Towards a theoretical foundation for Laplacian-based manifold methods. Journal of Computer and System Sciences, 74(8), 1289-1308, 2008.
10. V. I. Bogachev, Measure Theory I, II. Springer, 2007.
11. A. Berlinet and C. Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer Academic Publishers, 2004.
12. L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR computational mathematics and mathematical physics, 7(3):200-217, 1967
13. D. Burago, S. Ivanov and Y. Kurylev, A graph discretization of the Laplace-Beltrami operator, J. Spectr. Theory 4 (2014), 675-714.
14. Z. Chebbi and M. Moakher. Means of Hermitian positive-definite matrices based on the log-determinant α-divergence function. Linear Algebra and its Applications, 436 (7):1872-1889, 2012.
15. N. N. Chentsov, The categories of mathematical statistics. (Russian) Dokl. Akad. Nauk SSSR 164 (1965), 511-514.
16. N. N. Chentsov, Statistical Decision Rules and Optimal Inference. Translation of mathematical monographs, AMS, Providence, Rhode Island, 1982, translation from Russian original, Nauka, Moscow, 1972.
17. F. Cucker and S. Smale, On mathematical foundations of learning. Bulletin of AMS, 39 (2002), 1-49.
18. J. Dodziuk, Finite-Difference Approach to the Hodge Theory of Harmonic Forms, Amer. J. of Math. 98, No. 1, 79-104.
19. J. Dodziuk, V. Patodi, Riemannian structures and triangulations of manifolds, J. Ind. Math. Soc. 40 (1976), p. 1-52.
20. C. Gordon, D. L. Webb, and S. Wolpert, One cannot hear the shape of a drum, Bulletin (New Series) of the AMS, Volume 27 1992), Number 1, p. 134-138.
21. K. Fan. On a theorem of Weyl concerning eigenvalues of linear transformations: II. Proceedings of the National Academy of Sciences of the United States of America, 36 (1):31, 1950.
22. J. Faraut, and A. Korányi, Analysis on symmetric cones. Oxford University Press, 1994.
23. Fefferman, Charles and Ivanov, Sergei and Kurylev, Yaroslav and Lassas, Matti and Narayanan, Hariharan, Reconstruction and interpolation of manifolds. I: The geometric Whitney problem, Foundations of Computational Mathematics, 20 (5) (2020), p. 1035–1133.
24. Fefferman, Charles; Ivanov, Sergei; Lassas, Matti; Narayanan, Hariharan. Reconstruction of a Riemannian manifold from noisy intrinsic distances. SIAM J. Math. Data Sci. 2 (3) (2020) p. 770-808.
25. A. Feragen, F. Lauze, F., and S. Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3032-3042, 2015.
26. T. Fritz, A synthetic approach to Markov kernel, conditional independence and theorem of sufficient statistics. Adv. Math. 370, 107239 (2020), arXiv:1908.07021.
27. M. Giry, A categorical approach to probability theory, In: B. Banaschewski, (editor) Categorical Aspects of Topology and Analysis, Lecture Notes in Mathematics 915, 68- 85, Springer, 1982.
28. U. Grenander and M. I. Miller, Pattern theory: From Representation to Inference, Oxford University Press, 2007.

29. T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer 2008.
30. S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel methods on Riemannian manifolds with Gaussian RBF kernels. IEEE transactions on pattern analysis and machine intelligence, 37(12), pp.2464-2477, 2015.
31. P. Joharinad and J. Jost, Mathematical Principles of Topological and Geometric Data Analysis, Springer 2023.
32. J. Jost, H. V. Lê, and T. D. Tran, Probabilistic morphisms and Bayesian nonparametrics. Eur. Phys. J. Plus 136, 441 (2021), arXiv:1905.11448.
33. R.V. Kadison and J.R. Ringrose, Fundamentals of the theory of operator algebras. Volume I: Elementary Theory. Academic Press, 1983.
34. S. Lang, Fundamentals of Differential Geometry. Springer, 1999.
35. G. Larotonda. Nonpositive curvature: A geometrical approach to Hilbert-Schmidt operators. Differential Geometry and its Applications, 25:679-700, 2007.
36. W. F. Lawvere, The category of probabilistic mappings (1962). Unpublished, Available at https://ncatlab.org/nlab/files/lawvereprobability1962.pdf.
37. H. V. Lê, Supervised learning with probabilistic morphisms and kernel mean embeddings, arXiv:2305.06348.
38. D Leao Jr., M. Fragoso and P. Ruffino, Regular conditional probability, disintegration of probability and Radon spaces. Proyecciones vol. 23 (2004)Nr. 1, Universidad Catolica Norte, Antofagasta, Chile, 15-29.
39. H. V. Lê, H. Q. Minh, F. Protin, W. Tuschmann, Mathematical Foundations of Machine Learning (book in preparation, to be published by Springer in 2025).
40. P. Li, Q. Wang, W. Zuo, and L. Zhang. Log-Euclidean kernels for sparse representation and dictionary learning. In International Conference on Computer Vision (ICCV), pages 1601-1608, 2013.
41. D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. In Proceedings of the 32nd International Conference on Machine Learning (ICML2015), 2015.
42. L. Malagò, L., L. Montrucchio, and G. Pistone, Wasserstein Riemannian geometry of Gaussian densities. Information Geometry, 1, 137-179, 2018.
43. A.M. Mathai, S.B. Provost, and H.J. Haubold, Multivariate statistical analysis in the real and complex domains. Springer Nature, 2022.
44. H.Q. Minh, M. San Biagio, and V. Murino, Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces. Advances in neural information processing systems 27 (2014).
45. H.Q. Minh and V. Murino. Covariances in computer vision and machine learning. Springer Synthesis Lectures on Computer Vision, 2018.
46. M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of Machine Learning. MIT Press, 2nd Edition, 2018.
47. N. Morse, and R. Sacksteder, Statistical isomorphism. Ann. Math. Statist. 37, 1 (1966), 203–214.
48. G. Mostow. Some new decomposition theorems for semi-simple groups. Memoirs of the American Mathematical Society, 14:31-54, 1955.
49. J. Nash, The imbedding problem for Riemannian manifolds, Ann. Math. 63(1956), 383-396.
50. V. I. Paulsen, M. Raghupathi, An introduction to the theory of reproducing kernel Hilbert spaces. Cambridge Studies in Advanced Mathematics, 152. Cambridge University Press, Cambridge, 2016.
51. W.V. Petryshyn, Direct and iterative methods for the solution of linear operator equations in Hilbert spaces. Transactions of the American Mathematical Society, 105:136-175, 1962.
52. I.J. Schoenberg. Metric spaces and positive definite functions. Transactions of the American Mathematical Society, 44(3), 522-536, 1938.
53. C. L. Siegel. Symplectic geometry. American Journal of Mathematics, 65(1):1-86, 1943.
54. S. Sra. A new metric on the manifold of kernel matrices with application to matrix geometric means. In Advances in Neural Information Processing Systems (NIPS), pages 144-152, 2012.

55. S. Sra. Positive definite matrices and the S-divergence. Proceedings of the American Mathematical Society, 144(7):2787-2797, 2016.
56. B. Sriperumbudur, On the optimal estimation of probability measures in weak and strong topologies. Bernoulli, 22(3):1839-1893, 08, 2016.
57. A. R. Stefanyuk, Estimation of the likelihood ratio function in the “disorder" problem of random process, Autom. Remote. Control 9 (1986), 53-59.
58. I. Steinwart and A. Christmann. Support vector machines. Springer, 2008.
59. A. Takatsu, Wasserstein geometry of Gaussian measures. Osaka Journal of Mathematics, 48:1005-1026, 2011.
60. S. Theodoridis, Machine Learning, a Bayesian and Optimization Perspective. Academic Press, 2015.
61. A. B. Tsybakov, Introduction to Nonparametric Estimation. Springer, 2009.
62. V. Vapnik, Statistical Learning Theory. John Willey & Sons, 1998.
63. V. Vapnik, The Nature of Statistical Learning Theory, Springer, 2nd Edition, 2000.
64. V. Vapnik and R. Izmailov, Synergy of Monotonic Rules,Journal of Machine Learning Research 17 (2016) 1-33.
65. V. Vapnik and R. Izmailov, Rethinking statistical learning theory: learning using statistical invariants, Machine Learning (2019) 108:381-423.
66. V. Vapnik and A. Stefanyuk, Nonparametric methods for estimating probability densities, Automation and Remote Control, 8 (1978), 38-52.
67. A. Wald, Statistical Decision Functions. Wiley, New York; Chapman & Hall, London, 1950.
68. J. Zhang. Divergence function, duality, and convex analysis. Neural Computation, 16 (1):159-195, 2004
69. S. K. Zhou and R. Chellappa. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(6):917-929, 2006.